

Data Science

Questions and Answers

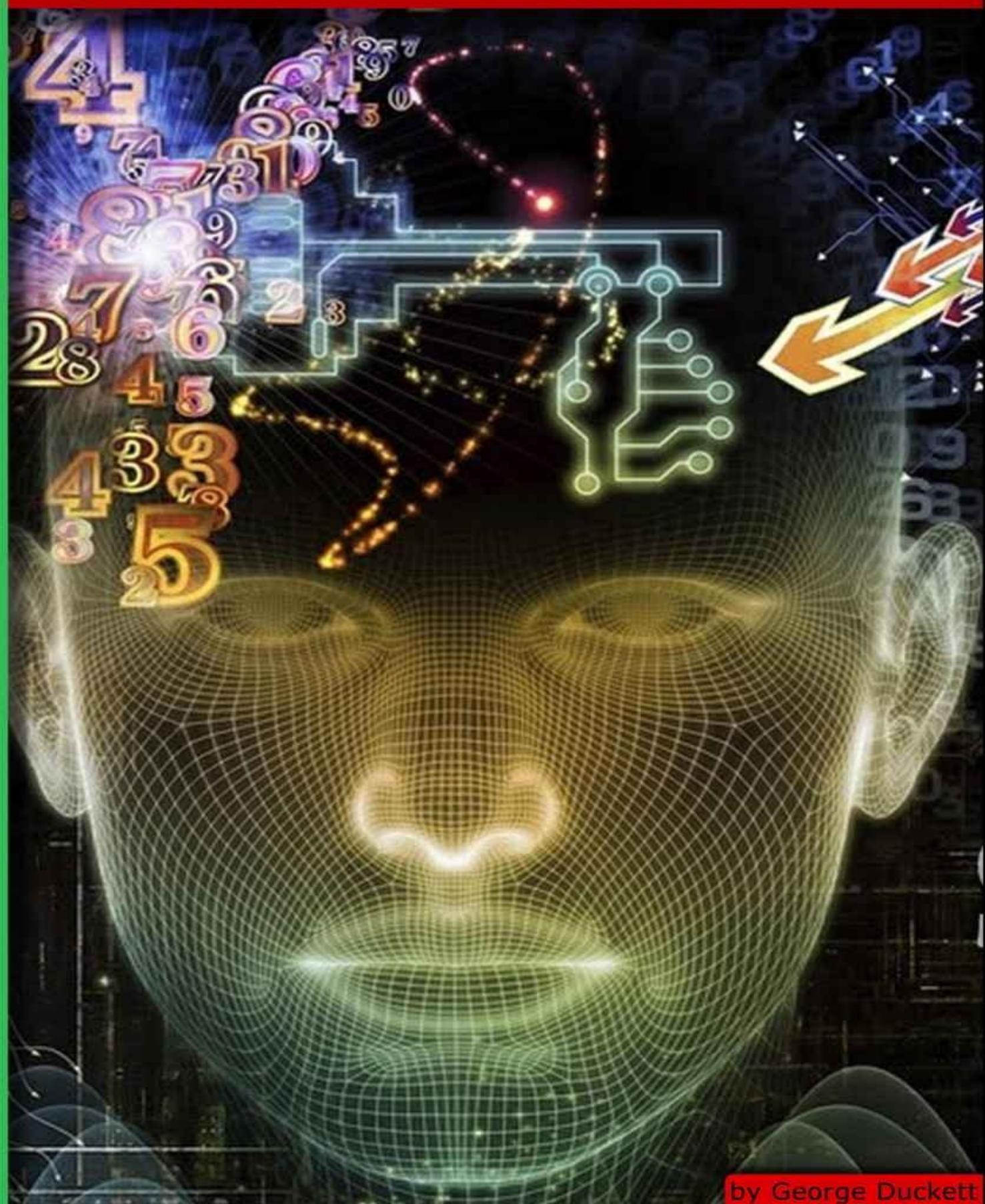


Table of Contents

[About this book](#)

[Machine Learning](#) (95 questions)

[Bigdata](#) (30 questions)

[Data Mining](#) (28 questions)

[Classification](#) (28 questions)

[Neuralnetwork](#) (23 questions)

[Statistics](#) (19 questions)

[Python](#) (19 questions)

[Clustering](#) (15 questions)

[R](#) (14 questions)

[Text Mining](#) (14 questions)

[NLP](#) (13 questions)

[Dataset](#) (12 questions)

[Efficiency](#) (11 questions)

[Algorithms](#) (11 questions)

[Hadoop](#) (11 questions)

[SVM](#) (11 questions)

[Tools](#) (9 questions)

[Recommendation](#) (9 questions)

[Visualization](#) (9 questions)

[Databases](#) (8 questions)

[Feature Selection](#) (8 questions)

[NoSQL](#) (7 questions)

[Predictive Modeling](#) (7 questions)

[Definitions](#) (6 questions)

[Education](#) (6 questions)

[Search](#) (5 questions)

[Similarity](#) (5 questions)

[Social Network Analysis](#) (5 questions)

[Time Series](#) (5 questions)

[Scalability](#) (4 questions)

[Beginner](#) (4 questions)

[Data Cleaning](#) (3 questions)

[Aws](#) (3 questions)

[Graphs](#) (3 questions)

[Cross Validation](#) (3 questions)

[Apache Spark](#) (3 questions)

[Categorical Data](#) (2 questions)

[Hierarchical Data Format](#) (2 questions)

[Xgboost](#) (2 questions)

[Sequence](#) (1 question)

[Copyright](#)

About this book

This book has been divided into categories where each question belongs to one or more categories. The categories are listed based on how many questions they have; the question appears in the most popular category. Everything is linked internally, so when browsing a category you can easily flip through the questions contained within it. Where possible links within questions and answers link to appropriate places within in the book. If a link doesn't link to within the book, then it gets a special icon, like [this](#).

Machine Learning

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Overview

From The Discipline of Machine Learning by Tom Mitchell:

The field of Machine Learning seeks to answer the question “How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?” This question covers a broad range of learning tasks, such as how to design autonomous mobile robots that learn to navigate from their own experience, how to data mine historical medical records to learn which future patients will respond best to which treatments, and how to build search engines that automatically customize to their user’s interests. To be more precise, we say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E. Depending on how we specify T, P, and E, the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

The following journals are dedicated to research in Machine Learning:

- [Journal of Machine Learning Research](#)  (Open Access)
 - [Machine Learning](#)  (Open Access)
 - [International Journal of Machine Learning and Cybernetics](#)  (Open Access)
 - [International Journal of Machine Learning and Computing](#)  (Open Access)
 - [International Journal of Machine Learning and Computing](#)  (Open Access)
-

Questions

[Q: Use liblinear on big data for semantic analysis](#)

Tags: [machine-learning](#) ([Next Q](#)), [bigdata](#) ([Next Q](#))

I use [Libsvm](#)  to train data and predict classification on **semantic analysis** problem. But it has a **performance** issue on large-scale data, because semantic analysis concerns ***n*-dimension** problem.

Last year, [Liblinear](#)  was released, and it can solve performance bottleneck. But it costs too much **memory**. Is **MapReduce** the only way to solve semantic analysis problem on big data? Or are there any other methods that can improve memory bottleneck on **Liblinear**?

Tags: [machine-learning](#) ([Next Q](#)), [bigdata](#) ([Next Q](#))

User: [puffin-gdi](#) 

[Answer](#)  by [sean-owen](#) 

Note that there is an early version of LIBLINEAR ported to [Apache Spark](#) . See [mailing list comments](#)  for some early details, and the [project site](#) .

[Answer](#)  by [marc-claesens](#) 

You can check out [vowpal wabbit](#) . It is quite popular for large-scale learning and includes parallel provisions.

From their website:

VW is the essence of speed in machine learning, able to learn from terafeature datasets with ease. Via parallel learning, it can exceed the throughput of any single machine network interface when doing linear learning, a first amongst learning algorithms.

Tags: [machine-learning](#) ([Next Q](#)), [bigdata](#) ([Next Q](#))

[Q: Why Is Overfitting Bad in Machine Learning?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Next Q](#))

Logic often states that by overfitting a model, its capacity to generalize is limited, though this might only mean that overfitting stops a model from improving after a certain complexity. Does overfitting cause models to become worse regardless of the complexity of data, and if so, why is this the case?

Related: Followup to the question above, “[When is a Model Underfitted?](#)” 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Next Q](#))

User: [blunders](#)

[Answer](#) by [alex-i](#)

Overfitting is *empirically* bad. Suppose you have a data set which you split in two, test and training. An overfitted model is one that performs much worse on the test dataset than on training dataset. It is often observed that models like that also in general perform worse on additional (new) test datasets than models which are not overfitted.

One way to understand that intuitively is that a model may use some relevant parts of the data (signal) and some irrelevant parts (noise). An overfitted model uses more of the noise, which increases its performance in the case of known noise (training data) and decreases its performance in the case of novel noise (test data). The difference in performance between training and test data indicates how much noise the model picks up; and picking up noise directly translates into worse performance on test data (including future data).

Summary: overfitting is bad by definition, this has not much to do with either complexity or ability to generalize, but rather has to do with mistaking noise for signal.

P.S. On the “ability to generalize” part of the question, it is very possible to have a model which has inherently limited ability to generalize due to the structure of the model (for example linear SVM, ...) but is still prone to overfitting. In a sense overfitting is just one way that generalization may fail.

[Answer](#) by [rubens](#)

Overfitting, in a nutshell, means take into account **too much** information from your data and/or prior knowledge, and use it in a model. To make it more straightforward, consider the following example: you’re hired by some scientists to provide them with a model to predict the growth of some kind of plants. The scientists have given you information collected from their work with such plants throughout a whole year, and they shall continuously give you information on the future development of their plantation.

So, you run through the data received, and build up a model out of it. Now suppose that, in your model, you considered just as many characteristics as possible to always find out the exact behavior of the plants you saw in the initial dataset. Now, as the production continues, you’ll always take into account those characteristics, and will produce very *fine-grained* results. However, if the plantation eventually suffer from some seasonal change, the results you will receive may fit your model in such a way that your predictions will begin to fail (either saying that the growth will slow down, while it shall actually speed up, or the opposite).

Apart from being unable to detect such small variations, and to usually classify your entries incorrectly, the *fine-grain* on the model, i.e., the great amount of variables, may cause the processing to be too costly. Now, imagine that your data is already complex. Overfitting your model to the data not only will make the classification/evaluation very complex, but will most probably make you error the prediction over the slightest variation you may have on the input.

Edit: [This](#) might as well be of some use, perhaps adding dynamicity to the above

explanation :D

[Answer](#) by [kim](#)

That's because something called [bias-variance dilemma](#). The overfitted model means that we will have more complex decision boundary if we give more variance on model. The thing is, not only too simple models but also complex models are likely to have dis-classified result on unseen data. Consequently, over-fitted model is not good as under-fitted model. That's why overfitting is bad and we need to fit the model somewhere in the middle.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Next Q](#))

[Q: Is there any APIs for crawling abstract of paper?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Next Q](#))

If I have a very long list of paper names, how could I get abstract of these papers from internet or any database?

The paper names are like "Assessment of Utility in Web Mining for the Domain of Public Health".

Does any one know any API that can give me a solution? I tried to crawl google scholar, however, google blocked my crawler.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Next Q](#))

User: [alex-gao](#)

[Answer](#) by [alex-i](#)

Look it up on:

- Google Scholar [link](#)
- Citeseer [link](#)

If you get a single exact title match then you have probably found the right article, and can fill in the rest of the info from there. Both give you download links and bibtex-style output. What you would likely want to do though to get perfect metadata is download and parse the pdf (if any) and look for DOI-style identifier.

Please be nice and rate-limit your requests if you do this.

[Answer](#) by [cwharland](#)

arXiv has an [API and bulk download](#) but if you want something for paid journals it will be hard to come by without paying an indexer like pubmed or elsevier or the like.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Next Q](#))

[Q: Machine learning techniques for estimating users' age based on Facebook sites they like](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Next Q](#))

I have a database from my Facebook application and I am trying to use machine learning to estimate users' age based on what Facebook sites they like.

There are three crucial characteristics of my database:

- the age distribution in my training set (12k of users in sum) is skewed towards younger users (i.e. I have 1157 users aged 27, and 23 users aged 65);
- many sites have no more than 5 likers (I filtered out the FB sites with less than 5 likers).
- there's many more features than samples.

So, my questions are: what strategy would you suggest to prepare the data for further analysis? Should I perform some sort of dimensionality reduction? Which ML method would be most appropriate to use in this case?

I mainly use Python, so Python-specific hints would be greatly appreciated.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Next Q](#))

User: [wojciech-walczak](#)

[Answer](#) by [cwharland](#)

One thing to start off with would be k-NN. The idea here is that you have a user/item matrix and for some of the users you have a reported age. The age for a person in the user item matrix might be well determined by something like the mean or median age of some nearest neighbors in the item space.

So you have each user expressed as a vector in item space, find the k nearest neighbors and assign the vector in question some summary stat of the nearest neighbor ages. You can choose k on a distance cutoff or more realistically by iteratively assigning ages to a train hold out and choosing the k that minimizes the error in that assignment.

If the dimensionality is a problem you can easily perform reduction in this setup by single value decomposition choosing the m vectors that capture the most variance across the group.

In all cases since each feature is binary it seems that cosine similarity would be your go to distance metric.

I need to think a bit more about other approaches (regression, rf, etc...) given the narrow focus of your feature space (all variants of the same action, liking) I think the user/item approach might be the best.

One note of caution, if the ages you have for train are self reported you might need to correct some of them. People on facebook tend to report ages in the decade they were

born. Plot a histogram of the birth dates (derived from ages) and see if you have spikes at decades like 70s, 80s, 90s.

[Answer](#) by [therriault](#)

I recently did a similar project in Python (predicting opinions using FB like data), and had good results with the following basic process:

1. Read in the training set ($n = N$) by iterating over comma-delimited like records line-by-line and use a counter to identify the most popular pages
2. For each of the K most popular pages (I used about 5000, but you can play around with different values), use `pandas.DataFrame.isin` to test whether each individual in the training set likes each page, then make a $N \times K$ dataframe of the results (I'll call it `xdata_train`)
3. Create a series (I'll call it `ydata_train`) containing all of the outcome variables (in my case opinions, in yours age) with the same index as `xdata_train`
4. Set up a random forest classifier through scikit-learn to predict `ydata_train` based on `xdata_train`
5. Use scikit-learn's cross-validation testing to tweak parameters and refine accuracy (tweaking number of popular pages, number of trees, min leaf size, etc.)
6. Output random forest classifier and list of most popular pages with pickle (or keep in memory if you are doing everything at once)
7. Load in the rest of your data, load the list of popular pages (if necessary), and repeat step 2 to produce `xdata_new`
8. Load the random forest classifier (if necessary) and use it to predict values for the `xdata_new` data
9. Output the predicted scores to a new CSV or other output format of your choosing

In your case, you'd need to swap out the classifier for a regressor (so see here:

<http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>) but otherwise the same process should work without much trouble.

Also, you should be aware of the most amazing feature of random forests in Python: instant parallelization! Those of us who started out doing this in R and then moved over are always amazed, especially when you get to work on a machine with a few dozen cores (see here: <http://blog.yhat.com/posts/comparing-random-forests-in-python-and-r.html>).

Finally, note that this would be a perfect application for network analysis if you have the data on friends as well as the individuals themselves. If you can analyze the ages of a user's friends, the age of the user will almost certainly be within a year or two of the median among his or her friends, particularly if the users are young enough to have built their friend networks while still in school (since most will be classmates). That prediction would likely trump any you would get from modeling—this is a textbook example of a problem where the right data > the right model every time.

Good luck!

[Answer](#) by [damienfrancois](#)

Another suggestion is to test the [logistic regression](#). As an added bonus, the weights (coefficients) of the model will give you an idea of which sites are age-discriminант.

Sklearn offers the [sklearn.linear_model.LogisticRegression](#) package that is designed to handle sparse data as well.

As mentionned in the comments, in the present case, with more input variables than samples, you need to regularize the model (with [sklearn.linear_model.LogisticRegression](#) use the `penalty='l1'` argument).

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Next Q](#))

[Q: Is Data Science just a trend or is a long term concept?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Next Q](#)), [databases](#) ([Next Q](#)), [education](#) ([Next Q](#))

I see a lot of courses in Data Science emerging in the last 2 years. Even big universities like Stanford and Columbia offers MS specifically in Data Science. But as long as I see, it looks like data science is just a mix of computer science and statistics techniques. So I always think about this. If it is just a trend and if in 10 years from now, someone will still mention Data Science as an entire field or just a subject/topic inside CS or stats. What do you think?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Next Q](#)), [databases](#) ([Next Q](#)), [education](#) ([Next Q](#))

User: [filipe-ferminiano](#)

[Answer](#) by [marco13](#)

The one thing that you can say for sure is: Nobody can say this for sure. And it might indeed be opinion-based to some extent. The introduction of terms like “Big Data” that some people consider as “hypes” or “buzzwords” don’t make it easier to flesh out an appropriate answer here. But I’ll try.

In general, interdisciplinary fields often seem to have the problem of not being taken serious by either of the fields they are spanning. However, the more research is invested into a particular field, the greater is the urge to split this field into several sub-topics. And these sub-topics sooner or later have to be re-combined in new ways, in order to prevent an overspecialization, and to increase and broaden the applicability of techniques that are developed by the (over?)specialized experts in the different fields.

And I consider “Data Science” as such an approach to combine the expertise and findings from different fields. You described it as

...a mix of computer science and statistics techniques

And indeed, several questions here aim at the differentiation between data science and

statistics. But a pure statistician will most likely not be able to set up a Hadoop cluster and show the results of his analysis in an interactive HTML5 dashboard. And someone who can implement a nice HTML5 dashboard might not be so familiar with the mathematical background of a Chi-Squared-Test.

It is reasonable to assume that giving students enough knowledge to *apply* the most important techniques from the different fields that are covered by data science will lead to new applications of these techniques, and be beneficial - also for the “purists” in these fields. The combination of these techniques is not straightforward in many cases, and can justify an own branch of research.

You also asked whether in 10 years, data science will be considered as “just a topic inside computer science”. Again: Nobody can say for sure. But I wonder at which point people stopped asking the question whether “Computer Science” will one day only be considered only as a mix of (or a subject of) Electrical Engineering and Mathematics...

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Next Q](#)), [databases](#) ([Next Q](#)), [education](#) ([Next Q](#))

[Q: How to specify important attributes?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Next Q](#))

Assume a set of loosely structured data (e.g. Web tables/Linked Open Data), composed of many data sources. There is no common schema followed by the data and each source can use synonym attributes to describe the values (e.g. “nationality” vs “bornIn”).

My goal is to find some “important” attributes that somehow “define” the entities that they describe. So, when I find the same value for such an attribute, I will know that the two descriptions are most likely about the same entity (e.g. the same person).

For example, the attribute “lastName” is more discriminative than the attribute “nationality”.

How could I (statistically) find such attributes that are more important than others?

A naive solution would be to take the average IDF of the values of each attribute and make this the “importance” factor of the attribute. A similar approach would be to count how many distinct values appear for each attribute.

I have seen the term feature, or attribute selection in machine learning, but I don’t want to discard the remaining attributes, I just want to put higher weights to the most important ones.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Next Q](#))

User: [yefthym](#) 

[Answer](#)  by [darklordofsoftware](#) 

Actually there are more than one question to answer here:

1. How to work on schemaless/loose/missing data
2. How to label a person (from what I understand unsupervised) and create an identifier
3. How to train your system so that it can tell you which attributes you should use in order to identify the person

As Rubens mentioned, you can use **decision tree** methods, specifically [Random Forests](#) for calculating the most important attributes based on information gain if you have already found a way to identify how to label a person.

However, if you do not have any label information maybe you can use some expert view for preliminary attribute selection. After that you make **unsupervised classification** in order to retrieve your labels. Lastly, you can select the most important fields using **Random Forest** or other methods like **Bayesian Belief Networks**.

In order to achieve all that, you also need complete data set. If your data set is loose you have to manually or heuristically find a way to couple attributes indicating same thing with different names. What is more, you can use *imputation* techniques such as [Expectation Maximization](#) method and complete your data set. Or you can also work with Bayesian Networks and can leave missing fields as they are.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Next Q](#))

[Q: Algorithm for generating classification rules](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Next Q](#))

So we have potential for a machine learning application that fits fairly neatly into the traditional problem domain solved by classifiers, i.e., we have a set of attributes describing an item and a “bucket” that they end up in. However, rather than create models of probabilities like in Naive Bayes or similar classifiers, we want our output to be a set of roughly human-readable rules that can be reviewed and modified by an end user.

Association rule learning looks like the family of algorithms that solves this type of problem, but these algorithms seem to focus on identifying common combinations of features and don't include the concept of a final bucket that those features might point to. For example, our data set looks something like this:

```
Item A { 4-door, small, steel } => { sedan }
Item B { 2-door, big, steel } => { truck }
Item C { 2-door, small, steel } => { coupe }
```

I just want the rules that say “if it's big and a 2-door, it's a truck,” not the rules that say “if it's a 4-door it's also small.”

One workaround I can think of is to simply use association rule learning algorithms and ignore the rules that don't involve an end bucket, but that seems a bit hacky. Have I missed some family of algorithms out there? Or perhaps I'm approaching the problem

incorrectly to begin with?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Next Q](#))

User: [super_seabass](#) 

[Answer](#)  by [rapaio](#) 

C45 made by Quinlan is able to produce rule for prediction. Check this [Wikipedia](#)  page. I know that in [Weka](#)  its name is J48. I have no idea which are implementations in R or Python. Anyway, from this kind of decision tree you should be able to infer rules for prediction.

Later edit

Also you might be interested in algorithms for directly inferring rules for classification. RIPPER is one, which again in Weka it received a different name JRip. See the original paper for RIPPER: [Fast Effective Rule Induction, W.W. Cohen 1995](#) 

[Answer](#)  by [therriault](#) 

It's actually even simpler than that, from what you describe—you're just looking for a basic classification tree algorithm (so no need for slightly more complex variants like C4.5 which are optimized for prediction accuracy). The canonical text is:

<http://www.amazon.com/Classification-Regression-Wadsworth-Statistics-Probability/dp/0412048418> 

This is readily implemented in R:

<http://cran.r-project.org/web/packages/tree/tree.pdf> 

and Python:

<http://scikit-learn.org/stable/modules/tree.html> 

[Answer](#)  by [ger](#) 

You could take a look at CN2 rule learner in Orange 2 <http://orange.biolab.si/orange2/> 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Next Q](#))

[Q: Human activity recognition using smartphone data set problem](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

I'm new to this community and hopefully my question will well fit in here. As part of my undergraduate data analytics course I have chosen to do the project on human activity recognition using smartphone data sets. As far as I'm concerned this topic relates to Machine Learning and Support Vector Machines. I'm not well familiar with these technologies yet so I will need some help.

I have decided to follow this project idea

<http://www.inf.ed.ac.uk/teaching/courses/dme/2014/datasets.html> (first project on the top) The project goal is determine what activity a person is engaging in (e.g., WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) from data recorded by a smartphone (Samsung Galaxy S II) on the subject's waist. Using its embedded accelerometer and gyroscope, the data includes 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz.

All the data set is given in one folder with some description and feature labels. The data is divided for 'test' and 'train' files in which data is represented in this format:

```
2.5717778e-001 -2.3285230e-002 -1.4653762e-002 -9.3840400e-001 -9.2009078e-001 -6.6768331e-001 -9.5
```

And that's only a very small sample of what the file contain.

I don't really know what this data represents and how can be interpreted. Also for analyzing, classification and clustering of the data, what tools will I need to use? Is there any way I can put this data into excel with labels included and for example use R or python to extract sample data and work on this?

Any hints/tips would be much appreciated.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

User: [jakubee](#)

[Answer](#) by [mcp_infiltrator](#)

The data set definitions are on the page here:

[Attribute Information at the bottom](#)

or you can see inside the ZIP folder the file named activity_labels, that has your column headings inside of it, make sure you read the README carefully, it has some good info in it. You can easily bring in a .csv file in R using the read.csv command.

For example if you name you file samsungdata you can open R and run this command:

```
data <- read.csv("directory/where/file/is/located/samsungdata.csv", header = TRUE)
```

Or if you are already inside of the working directory in R you can just run the following

```
data <- read.csv("samsungdata.csv", header = TRUE)
```

Where the name data can be changed to whatever you want to call your data set.

[Answer](#) by [damian-melniczuk](#)

It looks like this (or very similar data set) is used for Coursera courses. Cleaning this dataset is task for [Getting and Cleaning Data](#), but it is also used for case study for [Exploratory Data analysis](#). Video from this case study is available in videos for week 4 of EDA course-ware. It might help you with starting with this data.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev](#)

[Q: Can machine learning algorithms predict sports scores or plays?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I have a variety of NFL datasets that I think might make a good side-project, but I haven't done anything with them just yet.

Coming to this site made me think of machine learning algorithms and I was wondering how good they might be at either predicting the outcome of football games or even the next play.

It seems to me that there would be some trends that could be identified - on 3rd down and 1, a team with a strong running back *theoretically should* have a tendency to run the ball in that situation.

Scoring might be more difficult to predict, but the winning team might be.

My question is whether these are good questions to throw at a machine learning algorithm. It could be that a thousand people have tried it before, but the nature of sports makes it an unreliable topic.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [steve-kallestad](#) 

[Answer](#)  by [lsdr](#) 

There are a lot of good questions about Football (and sports, in general) that would be awesome to throw to an algorithm and see what comes out. The tricky part is to know *what* to throw to the algorithm.

A team with a good RB could just pass on 3rd-and-short just because the opponents would probably expect run, for instance. So, in order to actually produce some worthy results, I'd break the problem in smaller pieces and analyse them statistically while throwing them to the machines!

There are a few (good) websites that try to do the same, you should check'em out and use whatever they found to help you out:

- [Football Outsiders](#) 
- [Advanced Football Analytics](#) 

And if you truly want to explore Sports Data Analysis, you should definitely check the [Sloan Sports Conference](#)  videos. There's a lot of them spread on Youtube.

[Answer](#)  by [binga](#) 

Yes. Why not?! With so much of data being recorded in each sport in each game, smart use of data could lead us in obtaining important insights regarding player performance.

Some examples:

- **Baseball:** In the movie Moneyball (which is an adaptation of the MoneyBall book), Brad Pitt plays a character who analyses player statistics to come up with a team that performs tremendously well! It was a depiction of the real life story of Oakland Athletics baseball team. For more info, <http://www.theatlantic.com/entertainment/archive/2013/09/forget-2002-this-years-oakland-as-are-the-real-em-moneyball-em-team/279927/>
- **Cricket:** SAP Labs has come up with an auction analytics tool that has given insights about impact players to buy in the 2014 Indian Premier League auction for the Kolkata Knight Riders team, which eventually went on to win the 2014 IPL **Championship.** For more info, <http://scn.sap.com/community/hana-in-memory/blog/2014/06/10/sap-hana-academy-cricket-demo—how-sap-hana-powered-the-kolkata-knight-riders-to-ipl-championship>

So, yes, statistical analysis of the player records can give us insights about **which players are more likely to perform but not which players will perform.** So, machine learning, a close cousin of statistical analysis will be proving to be a game changer.

[Answer](#) by [ihars](#)

Definitely they can. I can target you to a [nice paper](#). Once I used it for soccer league results prediction algorithm implementation, primarily aiming at having some value against bookmakers.

From paper's abstract:

a Bayesian dynamic generalized model to estimate the time dependent skills of all teams in a league, and to predict next weekend's soccer matches.

Keywords:

Dynamic Models, Generalized Linear Models, Graphical Models, Markov Chain Monte Carlo Methods, Prediction of Soccer Matches

Citation:

Rue, Havard, and Oyvind Salvesen. "Prediction and retrospective analysis of soccer matches in a league." Journal of the Royal Statistical Society: Series D (The Statistician) 49.3 (2000): 399-418.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Q: What are some easy to learn machine-learning applications?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Being new to machine-learning in general, I'd like to start playing around and see what the possibilities are.

I'm curious as to what applications you might recommend that would offer the fastest time from installation to producing a meaningful result.

Also, any recommendations for good getting-started materials on the subject of machine-learning in general would be appreciated.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [steve-kallestad](#) 

[Answer](#)  by [stanpol](#) 

I would recommend to start with some MOOC on machine learning. For example Andrew Ng's [course](#)  at coursera.

You should also take a look at [Orange](#)  application. It has a graphical interface and probably it is easier to understand some ML techniques using it.

[Answer](#)  by [justin](#) 

To be honest, I think that doing some projects will teach you much more than doing a full course. One reason is that doing a project is more motivating and open-ended than doing assignments.

A course, if you have the time AND motivation (real motivation), is better than doing a project. The other commentators have made good platform recommendations on tech.

I think, from a fun project standpoint, you should ask a question and get a computer to learn to answer it.

Some good classic questions that have good examples are:

- Neural Networks for recognizing hand written digits
- Spam email classification using logistic regression
- Classification of objects using Gaussian Mixture models
- Some use of linear regression, perhaps forecasting of grocery prices given neighborhoods

These projects have the math done, code done, and can be found with Google readily.

Other cool subjects can be done by you!

Lastly, I research robotics, so for me the most FUN applications are behavioral. Examples can include (if you can play with an arduino)

Create a application, that uses logistic regression perhaps, that learns when to turn the fan off and on given the inner temperature, and the status of the light in the room.

Create an application that teaches a robot to move an actuator, perhaps a wheel, based on sensor input (perhaps a button press), using Gaussian Mixture Models (learning from demonstration).

Anyway, those are pretty advanced. The point I'm making is that if you pick a project that you (really really) like, and spend a few week on it, you will learn a massive amount, and understand so much more than you will get doing a few assignments.

[Answer](#) by [iliasfl](#)

I think [Weka](#) is a good starting point. You can do a bunch of stuff like supervised learning or clustering and easily compare a large set of algorithms na methodologies.

Weka's manual is actually a book on machine learning and data mining that can be used as introductory material.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Difference between using RMSE and nDCG to evaluate Recommender Systems](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Next Q](#))

What kind of error measures do RMSE and nDCG give while evaluating a recommender system, and how do I know when to use one over the other? If you could give an example of when to use each, that would be great as well!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Next Q](#))

User: [user3004041](#)

[Answer](#) by [debasis](#)

nDCG is used to evaluate a golden ranked list (typically human judged) against your output ranked list. The more is the correlation between the two ranked lists, i.e. the more similar are the ranks of the relevant items in the two lists, the closer is the value of nDCG to 1.

RMSE (Root Mean Squared Error) is typically used to evaluate regression problems where the output (a predicted scalar value) is compared with the true scalar value output for a given data point.

So, if you are simply recommending a score (such as recommending a movie rating), then use RMSE. Whereas, if you are recommending a list of items (such as a list of related movies), then use nDCG.

[Answer](#) by [emre](#)

nDCG is a ranking metric and RMSE is not. In the context of recommender systems, you would use a ranking metric when your ratings are implicit (e.g., item skipped vs. item consumed) rather than explicit (the user provides an actual number, a la Netflix).

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Next Q](#))

[Q: Choosing a learning rate](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Next Q](#)), [algorithms](#) ([Next Q](#))

I'm currently working on implementing Stochastic Gradient Descent (SGD) for neural nets using backpropagation, and while I understand its purpose I have some questions about how to choose values for the learning rate.

- Is the learning rate related to the shape of the error gradient, as it dictates the rate of descent?
- If so, how do you use this information to inform your decision about a value?
- If it's not what sort of values should I choose, and how should I choose them?
- It seems like you would want small values to avoid overshooting, but how do you

- choose one such that you don't get stuck in local minima or take too long to descend?
- Does it make sense to have a constant learning rate, or should I use some metric to alter its value as I get nearer a minimum in the gradient?

In short: How do I choose the learning rate for SGD?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Next Q](#)), [algorithms](#) ([Next Q](#))

User: [ragingsloth](#) 

[Answer](#)  by [brnguyen](#) 

Below is a very good note (page 12) on learning rate in Neural Nets (Back Propagation) by Andrew Ng. You will find details relating to learning rate.

http://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf 

For your 4th point, you're right that normally one has to choose a "balanced" learning rate, that should neither overshoot nor converge too slowly. One can plot the learning rate w.r.t. the descent of the cost function to diagnose/fine tune. In practice, Andrew normally uses the L-BFGS algorithm (mentioned in page 12) to get a "good enough" learning rate.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Next Q](#)), [algorithms](#) ([Next Q](#))

Q: Looking for example infrastructure stacks/workflows/pipelines

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Next Q](#)), [scalability](#) ([Next Q](#))

I'm trying to understand how all the "big data" components play together in a real world use case, e.g. hadoop, monogodb/nosql, storm, kafka, ... I know that this is quite a wide range of tools used for different types, but I'd like to get to know more about their interaction in applications, e.g. thinking machine learning for an app, webapp, online shop.

I have visitors/session, transaction data etc and store that; but if I want to make recommendations on the fly, I can't run slow map/reduce jobs for that on some big database of logs I have. Where can I learn more about the infrastructure aspects? I think I can use most of the tools on their own, but plugging them into each other seems to be an art of its own.

Are there any public examples/use cases etc available? I understand that the individual pipelines strongly depend on the use case and the user, but just examples will probably be very useful to me.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Next Q](#)), [scalability](#) ([Next Q](#))

User: [chrshmmmr](#) 

[Answer](#)  by [j_houg](#) 

In order to understand the variety of ways machine learning can be integrated into production applications, I think it is useful to look at open source projects and papers/blog posts from companies describing their infrastructure.

The common theme that these systems have is the separation of model training from model application. In production systems, model application needs to be fast, on the order of 100s of ms, but there is more freedom in how frequently fitted model parameters (or equivalent) need to be updated.

People use a wide range of solutions for model training and deployment:

- Build a model, then export and deploy it with PMML
 - [AirBnB describes their model training](#) in R/Python and deployment of PMML models via OpenScoring.
 - [Pattern](#) is project related to [Cascading](#) that can consume PMML and deploy predictive models.
- Build a model in MapReduce and access values in a custom system
 - [Conjecture is an open source project from Etsy](#) that allows for model training with [Scalding](#), an easier to use scala wrapper around MapReduce, and deployment via Php.
 - [Kiji is an open source project from WibiData](#) that allows for real-time model scoring (application) as well as functionality for persisting user data and training models on that data via [Scalding](#).
- Use an online system that allows for continuously updating model parameters.
 - [Google released a great paper about an online collaborative filtering](#) they implemented to deal with recommendations in Google News.

[Answer](#) by [tchakravarty](#)

One of the most detailed and clear explanations of setting up a complex analytics pipeline is from the folks over at [Twitch](#).

They give detailed motivations of each of the architecture choices for collection, transportation, coordination, processing, storage, and querying their data. Compelling reading! Find it [here](#) and [here](#).

[Answer](#) by [trey](#)

[Airbnb](#) and [Etsy](#) both recently posted detailed information about their workflows.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Next Q](#)), [scalability](#) ([Next Q](#))

[Q: What are the implications for training a Tree Ensemble with highly biased datasets?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

I have a highly biased binary dataset - I have 1000x more examples of the negative class than the positive class. I would like to train a Tree Ensemble (like Extra Random Trees or a Random Forest) on this data but it's difficult to create training datasets that contain enough examples of the positive class.

What would be the implications of doing a stratified sampling approach to normalize the number of positive and negative examples? In other words, is it a bad idea to, for instance, artificially inflate (by resampling) the number of positive class examples in the training set?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

User: [gallamine](#)

[Answer](#) by [mattbagg](#)

Yes, it's problematic. If you oversample the minority, you risk overfitting. If you undersample the majority, you risk missing aspects of the majority class. Stratified sampling, btw, is the equivalent to assigning non-uniform misclassification costs.

Alternatives:

(1) Independently sampling several subsets from the majority class and making multiple classifiers by combining each subset with all the minority class data, as suggested in the answer from @Debasis and described in this [EasyEnsemble paper](#),

(2) [SMOTE \(Synthetic Minority Oversampling Technique\)](#) or [SMOTEB](#), [\(combining SMOTE with boosting\)](#) to create synthetic instances of the minority class by making nearest neighbors in the feature space. SMOTE is implemented in R in [the DMwR package](#).

[Answer](#) by [indico](#)

I would recommend training on more balanced subsets of your data. Training random forest on sets of randomly selected positive example with a similar number of negative samples. In particular if the discriminative features exhibit a lot of variance this will be fairly effective and avoid over-fitting. However in stratification it is important to find balance as over-fitting can become a problem regardless. I would suggest seeing how the model does with the whole data set then progressively increasing the ratio of positive to negative samples approaching an even ratio, and selecting for the one that maximizes your performance metric on some representative hold out data.

This paper seems fairly relevant <http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf> it talks about a weighted Random Forest which more heavily penalizes misclassification of the minority class.

[Answer](#) by [debasis](#)

A fast, easy and often effective way to approach this imbalance would be to randomly subsample the bigger class (which in your case is the negative class), run the classification N number of times with members from the two classes (one full and the other subsampled) and report the average metric values, the average being computed over N (say 1000) iterations.

A more methodical approach would be to execute the Mapping Convergence (MC) algorithm, which involves identifying a subset of strong negative samples with the help of a one-class classifier, such as OSVM or SVDD, and then iteratively execute binary classification on the set of strong negative and positive samples. More details of the MC algorithm can be found in this [paper](#).

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

[Q: Is GLM a statistical or machine learning model?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

I thought that generalized linear model (GLM) would be considered a statistical model, but a friend told me that some papers classify it as a machine learning technique. Which one is true (or more precise)? Any explanation would be appreciated.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [user77571](#)

[Answer](#) by [ben](#)

A GLM is absolutely a statistical model, but statistical models and machine learning techniques are not mutually exclusive. In general, statistics is more concerned with inferring parameters, whereas in machine learning, prediction is the ultimate goal.

[Answer](#) by [rapaio](#)

Regarding prediction, statistics and machine learning sciences started to solve mostly the same problem from different perspectives.

Basically statistics assumes that the data were produced by a given stochastic model. So, from a statistical perspective, a model is assumed and given various assumptions the errors are treated and the model parameters and other questions are inferred.

Machine learning comes from a computer science perspective. The models are algorithmic and usually very few assumptions are required regarding the data. We work with hypothesis space and learning bias. The best exposition of machine learning I found is contained in Tom Mitchell's book called [Machine Learning](#).

For a more exhaustive and complete idea regarding the two cultures you can read the Leo Brozman paper called [Statistical Modeling: The Two Cultures](#)

However what must be added is that even if the two sciences started with different perspectives, both of them now share a fair amount of common knowledge and techniques. Why, because the problems were the same, but the tools were different. So now machine learning is mostly treated from a statistical perspective (check the Hastie, Tibshirani, Friedman book [The Elements of Statistical Learning](#) from a machine learning point of view with a statistical treatment, and perhaps Kevin P. Murphy 's book [Machine Learning: A probabilistic perspective](#), to name just a few of the best books available today).

Even the history of the development of this field show the benefits of this merge of perspectives. I will describe two events.

The first is the creation of CART trees, which was created by Breiman with a solid statistical background. At approximately the same time, Quinlan developed ID3,C45,See5, and so on, decision tree suite with a more computer science background. Now both this families of trees and the ensemble methods like bagging and forests become quite similar.

The second story is about boosting. Initially they were developed by Freund and Shapire when they discovered AdaBoost. The choices for designing AdaBoost were done mostly from a computational perspective. Even the authors did not understand well why it works. Only 5 years later Breiman (again!) described the adaboost model from a statistical perspective and gave an explanation for why that works. Since then, various eminent scientists, with both type of backgrounds, developed further those ideas leading to a Pleiads of boosting algorithms, like logistic boosting, gradient boosting, gentle boosting ans so on. It is hard now to think at boosting without a solid statistical background.

GLM is a statistical development. However new Bayesian treatments puts this algorithm also in machine learning playground. So I believe both claims could be right, since the interpretation and treatment of how it works could be different.

[Answer](#) by [binga](#)

In addition to Ben's answer, the subtle distinction between statistical models and machine learning models is that, in statistical models, you explicitly decide the output equation structure prior to building the model. The model is built to compute the parameters/coefficients.

Take linear model or GLM for example,

```
y = a1x1 + a2x2 + a3x3
```

Your independent variables are x1, x2, x3 and the coefficients to be determined are a1,a2,a3. You define your equation structure this way prior to building the model and compute a1,a2,a3. If you believe that y is somehow correlated to x2 in a non-linear way, you could try something like this.

```
y = a1x1 + a2(x2)^2 + a3x3.
```

Thus, you put a restriction in terms of the output structure. Inherently statistical models are linear models unless you explicitly apply transformations like sigmoid or kernel to make them nonlinear (GLM and SVM).

In case of machine learning models, you rarely specify output structure and algorithms like decision trees are inherently non-linear and work efficiently.

Contrary to what Ben pointed out, machine learning models aren't just about prediction, they do classification, regression etc which can be used to make predictions which are also done by various statistical models.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

[Q: Word2Vec for Named Entity Recognition](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

I'm looking to use google's word2vec implementation to build a named entity recognition system. I've heard that recursive neural nets with back propagation through structure are well suited for named entity recognition tasks, but I've been unable to find a decent implementation or a decent tutorial for that type of model. Because I'm working with an atypical corpus, standard NER tools in NLTK and similar have performed very poorly, and it looks like I'll have to train my own system.

In short, what resources are available for this kind of problem? Is there a standard recursive neural net implementation available?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

User: [madison-may](#) 

[Answer](#)  by [mrmeritology](#) 

Instead of "recursive neural nets with back propagation" you might consider the approach used by Frantzi, et. al. at National Centre for Text Mining (NaCTeM) at University of Manchester for *Termine* (see: <http://www.nactem.ac.uk/index.php>  and <http://personalpages.manchester.ac.uk/staff/sophia.ananiadou/IJODL2000.pdf> 

[Answer](#)  by [shark8me](#) 

Two recent papers use a Deep learning architecture called CharWNN to address this problem. CharWNN was first used to [get state of the art results](#)  (without handcrafted features) on Part of Speech (POS) tagging on an English corpus.

The [second paper](#)  by the same author uses the same (or similar) architecture for predicting whether a word belongs to 10 Named Entity classes, with apparent state of the art results.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

Q: What statistical model should I use to analyze the likelihood that a single event influenced longitudinal data

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

I am trying to find a formula, method, or model to use to analyze the likelihood that a specific event influenced some longitudinal data. I am having difficulty figuring out what to search for on Google.

Here is an example scenario:

Image you own a business that has an average of 100 walk-in customers every day. One day, you decide you want to increase the number of walk-in customers arriving at your store each day, so you pull a crazy stunt outside your store to get attention. Over the next week, you see on average 125 customers a day.

Over the next few months, you again decide that you want to get some more business, and perhaps sustain it a bit longer, so you try some other random things to get more customers in your store. Unfortunately, you are not the best marketer, and some of your tactics have little or no effect, and others even have a negative impact.

What methodology could I use to determine the probability that any one individual event positively or negatively impacted the number of walk-in customers? I am fully aware that correlation does not necessarily equal causation, but what methods could I use to determine the likely increase or decrease in your business's daily walk in client's following a specific event?

I am not interested in analyzing whether or not there is a correlation between your attempts to increase the number of walk-in customers, but rather whether or not any one single event, independent of all others, was impactful.

I realize that this example is rather contrived and simplistic, so I will also give you a brief description of the actual data that I am using:

I am attempting to determine the impact that a particular marketing agency has on their client's website when they publish new content, perform social media campaigns, etc. For any one specific agency, they may have anywhere from 1 to 500 clients. Each client has websites ranging in size from 5 pages to well over 1 million. Over the course of the past 5 years, each agency has annotated all of their work for each client, including the type of work that was done, the number of webpages on a website that were influenced, the number of hours spent, etc.

Using the above data, which I have assembled into a data warehouse (placed into a bunch of star/snowflake schemas), I need to determine how likely it was that any one piece of work (any one event in time) had an impact on the traffic hitting any/all pages influenced by a specific piece of work. I have created models for 40 different types of content that are found on a website that describes the typical traffic pattern a page with said content type might experience from launch date until present. Normalized relative to the appropriate model, I need to determine the highest and lowest number of increased or decreased visitors a specific page received as the result of a specific piece of work.

While I have experience with basic data analysis (linear and multiple regression, correlation, etc), I am at a loss for how to approach solving this problem. Whereas in the past I have typically analyzed data with multiple measurements for a given axis (for example temperature vs thirst vs animal and determined the impact on thirst that increased temperate has across animals), I feel that above, I am attempting to analyze the impact of a single event at some point in time for a non-linear, but predictable (or at least model-able), longitudinal dataset. I am stumped :(

Any help, tips, pointers, recommendations, or directions would be extremely helpful and I would be eternally grateful!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [peter-kirby](#) 

[Answer](#)  by [neone4373](#) 

Back in my data analyst days this type of problem was pretty typical. Basically, everyone in marketing would come up with a crazy idea that they sold to higher ups as the single event that would boost KPI's by 2000%. The higher ups would approve them and then they would begin their "test". Results would come back, and management would dump it on the data analysts to determine what worked and who did it.

The short answer is you can't really know if it wasn't run as a random A/B style test on like time periods. But I am very aware of how deficient that answer is, especially if the fact that a pure answer doesn't exist is irrelevant to the urgency of future business decisions. Here are some of the techniques I would use to salvage the analysis in this situation, bear in mind this is more of an art than a science.

Handles

A handle is something that exists in the data that you can hold onto. From what you are telling me in your situation you have a lot of info on who the marketing agency is, when they tried a tactic, and to which site they applied it to. These are your starting point and information like this going to be the corner stone of your analysis.

Methodology

The methodology is going to probably hold the strongest impact on which agencies are given credit for any and all gains so you are going to need to make sure that it is clearly outlined and all stakeholders agree that it makes sense. If you can't do that it is going to be difficult for people to trust your analysis.

An example of this are conversions. Say the marketing department purchases some leads and they arrive at our landing page, we would track them for 3 days, if they made a purchase within that time we would count them as having been converted. Why 3 days, why not 5 or 1? That's not important as long as everyone agrees, you now have a definition you can build off of.

Comparisons

In an ideal world you would have a nice A/B test to prove a definitive relationship, I am

going to assume that you are running short on those, still, you can learn something from a simple comparison of like data. When companies are trying to determine the efficacy of radio advertising they will often run ads on offset months in the same market, or for several months in one market and compare that with the results in a separate but similar market. It's doesn't pass for science, but even with all that noise a strong results will almost always be noticeable.

I would combine these in your case to determine how long an event is given to register an effect. Once you have the data from that time period run it against your modeled out traffic prediction, week over week growth, month over month etc. Which, can then allow a meaningful comparison between agencies, and across time periods.

Pragmatism

The aspiration is to be able to provide a deep understanding of cause and effect, but it is probably not realistic. Because of how messy outside factors make your analysis, you are constantly going to run up against the question over and over again: Did this event raise volume/sales/click throughs, or would doing anything at all have had the same effect? The best advise I can give for this is set very realistic goals for what you are looking to measure. A good starting point is, within the methodology you have, which event had the largest impact. Once you have those open your aperture from there.

Summary

Once you have reasoned out all of these aspects you can go about building a general solution which can then be automated. The advantage to designing your solution in this manner is that the business logic is already built in. This will make your results much more approachable and intuitive to non-technical business leaders.

[Answer](#) by [mlespiau](#)

Edit: Warning, i leave my message but my answer seems wrong, please check out the comment below!

I'm not an expert but I guess the main problem is to answer this question:

Has an/any event affected the number of hits on a certain day?

But I don't know how to treat multiple events, so I would try to answer this question:

- Does event X affected the number of hits on a certain day?

Which can be answered using hypothesis testing with p-values (what scientist do to evaluate for instance if a medicine affects a disease or not).

By using p-values, you could determinate if the number of hits in a certain day were mere random and acceptable under normal circumstances or that they must correspond to a change in your model.

You can read more about p-values in [Open Intro to Statistics Book](#), I've actually learned about them from there.

Then, the other parts of the problem are how to identify your events and calculate the

necessary parameters to answer your question (average/median, variance, etc.) and also how to keep that up-to-date and working.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

Q: How to select algorithms for ensemble methods?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

There is a general recommendation that algorithms in ensemble learning combinations should be different in nature. Is there a classification table, a scale or some rules that allow to evaluate how far away are the algorithms from each other? What are the best combinations?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [tomaskazemekas](#) 

[Answer](#)  by [iliasfl](#) 

In general in an ensemble you try to combine the opinions of multiple classifiers. The idea is like asking a bunch of experts on the same thing. You get multiple opinions and you later have to combine their answers (e.g. by a voting scheme). For this trick to work you want the classifiers to be different from each other, that is you don't want to ask the same "expert" twice for the same thing.

In practice, the classifiers do not have to be different in the sense of a different algorithm. What you can do is train the same algorithm with different subset of the data or a different subset of features (or both). If you use different training sets you end up with different models and different "independent" classifiers.

There is no golden rule on what works best in general. You have to try to see if there is an improvement for your specific problem.

[Answer](#)  by [adesantos](#) 

As a rule of thumb I always propose three different options:

- Use a bagging learning technique, similar to that one followed by Random Forest. This technique allows the training of 'small' classifiers which see a small portion of the whole data. Afterwards, a simple voting scheme (as in Random Forest) will lead you to a very interesting and robust classification.
- Use any technique related to fusioning information or probabilistic fusion. This is a very suitable solution in order to combine different likelihoods from different classifiers.
- My last suggestion is the use of fuzzy logic, a very adequate tool in order to combine information properly from a probabilistic (belonging) perspective.

The selection of specific methods or strategies will depend enormously on the data.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Binary classification model for sparse / biased data](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

I have a dataset with following specifications:

- Training dataset with 193176 samples with 2821 positives
- Test Dataset with 82887 samples with 673 positives
- There are 10 features.

I want to perform a binary classification (say, 0/1). The issue I am facing is that the data is very biased or rather sparse. After normalization and scaling the data along with some feature engineering and using a couple of different algorithms, these are the best results I could achieve:

```
mean square error : 0.00804710026904
Confusion matrix : [[82214  667]
                     [    0   6]]
```

i.e only 6 correct positive hits. This is using logistic regression. Here are the various things I tried with this:

- Different algorithms like RandomForest, DecisionTree, SVM
- Changing parameters value to call the function
- Some intuition based feature engineering to include compounded features

Now, my questions are:

1. What can I do to improve the number of positive hits ?
2. How can one determine if there is an overfit in such a case ? (I have tried plotting etc.)
3. At what point could one conclude if maybe this is the best possible fit I could have? (which seems sad considering only 6 hits out of 673)
4. Is there a way I could make the positive sample instances weigh more so the pattern recognition improves leading to more hits ?
5. Which graphical plots could help detect outliers or some intuition about which pattern would fit the best?

I am using the scikit-learn library with Python and all implementations are library functions.

edit:

Here are the results with a few other algorithms:

Random Forest Classifier(n_estimators=100)

[[82211	667]
3	6]]

Decision Trees:

[[78611	635]
3603	38]]

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

User: [tejas](#) 

[Answer](#)  by [insys](#) 

1. Since you are doing binary classification, have you tried adjusting the classification threshold? Since your algorithm seems rather insensitive, I would try lowering it and check if there is an improvement.
2. You can always use [Learning Curves](#) , or a plot of one model parameter vs. Training and Validation error to determine whether your model is overfitting. It seems it is under fitting in your case, but that's just intuition.
3. Well, ultimately it depends on your dataset, and the different models you have tried. At this point, and without further testing, there can not be a definite answer.
4. Without claiming to be an expert on the topic, there are a number of different techniques you may follow (hint: [first link on google](#) ) , but in my opinion you should first make sure you choose your cost function carefully, so that it represents what you are actually looking for.
5. Not sure what you mean by pattern intuition, can you elaborate?

By the way, what were your results with the different algorithms you tried? Were they any different?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: Handling a regularly increasing feature set](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#))

I'm working on a fraud detection system. In this field, new frauds appear regularly, so that new features have to be added to the model on ongoing basis.

I wonder what is the best way to handle it (from the development process perspective)? Just adding a new feature into the feature vector and re-training the classifier seems to be a naive approach, because too much time will be spent for re-learning of the old features.

I'm thinking along the way of training a classifier for each feature (or a couple of related

features), and then combining the results of those classifiers with an overall classifier. Are there any drawbacks of this approach? How can I choose an algorithm for the overall classifier?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#))

User: [maxim-fridental](#) 

[Answer](#)  by [sean-owen](#) 

In an ideal world, you retain all of your historical data, and do indeed run a new model with the new feature extracted retroactively from historical data. I'd argue that the computing resource spent on this is quite useful actually. Is it really a problem?

Yes, it's a widely accepted technique to build an ensemble of classifiers and combine their results. You can build a new model in parallel just on new features and average in its prediction. This should add value, but, you will never capture interaction between the new and old features this way, since they will never appear together in a classifier.

[Answer](#)  by [insys](#) 

Here's an idea that just popped out of the blue – what if you make use of [Random Subspace Sampling](#)  (as in fact Sean Owen already suggested) to train a bunch of new classifiers every time a new feature appears (using a random feature subset, including the new set of features). You could train those models on a subset of samples as well to save some training time.

This way you can have new classifiers possibly taking on both new and old features, and at the same time keeping your old classifiers. You might even, perhaps using a cross validation technique to measure each classifier's performance, be able to kill-off the worst performing ones after a while, to avoid a bloated model.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#))

[**Q: What are some standard ways of computing the distance between documents?**](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Next Q](#))

When I say "document", I have in mind web pages like Wikipedia articles and news stories. I prefer answers giving either vanilla lexical distance metrics or state-of-the-art semantic distance metrics, with stronger preference for the latter.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Next Q](#))

User: [matt](#) 

[Answer](#)  by [indico](#) 

There's a number of different ways of going about this depending on exactly how much semantic information you want to retain and how easy your documents are to tokenize (html documents would probably be pretty difficult to tokenize, but you could conceivably do something with tags and context.)

Some of them have been mentioned by ffriend, and the paragraph vectors by user1133029 is a really solid one, but I just figured I would go into some more depth about plusses and minuses of different approaches.

- [Cosine Distance](#) - Tried a true, cosine distance is probably the most common distance metric used generically across multiple domains. With that said, there's very little information in cosine distance that can actually be mapped back to anything semantic, which seems to be non-ideal for this situation.
- [Levenshtein Distance](#) - Also known as edit distance, this is usually just used on the individual token level (words, bigrams, etc...). In general I wouldn't recommend this metric as it not only discards any semantic information, but also tends to treat very different word alterations very similarly, but it is an extremely common metric for this kind of thing
- [LSA](#) - Is a part of a large arsenal of techniques when it comes to evaluating document similarity called topic modeling. LSA has gone out of fashion pretty recently, and in my experience, it's not quite the strongest topic modeling approach, but it is relatively straightforward to implement and has a few open source implementations
- [LDA](#) - Is also a technique used for topic modeling, but it's different from LSA in that it actually learns internal representations that tend to be more smooth and intuitive. In general, the results you get from LDA are better for modeling document similarity than LSA, but not quite as good for learning how to discriminate strongly between topics.
- [Pachinko Allocation](#) - Is a really neat extension on top of LDA. In general, this is just a significantly improved version of LDA, with the only downside being that it takes a bit longer to train and open-source implementations are a little harder to come by
- [word2vec](#) - Google has been working on a series of techniques for intelligently reducing words and documents to more reasonable vectors than the sparse vectors yielded by techniques such as Count Vectorizers and TF-IDF. Word2vec is great because it has a number of open source implementations. Once you have the vector, any other similarity metric (like cosine distance) can be used on top of it with significantly more efficacy.
- [doc2vec](#) - Also known as paragraph vectors, this is the latest and greatest in a series of papers by Google, looking into dense vector representations of documents. The gensim library in python has an implementation of word2vec that is straightforward enough that it can pretty reasonably be leveraged to build doc2vec, but make sure to keep the license in mind if you want to go down this route

Hope that helps, let me know if you've got any questions.

[Answer](#) by [ffriend](#)

There's a number of semantic distance measures, each with its pros and cons. Here are just a few of them:

- [cosine distance](#), inner product between document feature vectors;
- [LSA](#), another vector-based model, but utilizing SVD for de-noising original term-document matrix;
- [WordNet](#)-based, human verified, though hardly extensible.

Start with a simplest approach and then move further based on issues for your specific case.

[Answer](#) by [user1133029](#)

State of the art appears to be “paragraph vectors” introduced in a recent paper:
http://cs.stanford.edu/~quocle/paragraph_vector.pdf. Cosine/Euclidean distance between paragraph vectors would likely work better than any other approach. This probably isn't feasible yet due to lack of open source implementations.

Next best thing is cosine distance between LSA vectors or cosine distance between raw BOW vectors. Sometimes it works better to choose different weighting schemes, like TF-IDF.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Next Q](#))

[Q: What are some standard ways of computing the distance between individual search queries?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Next Q](#))

I made a similar question asking about distance between “documents” (Wikipedia articles, news stories, etc.). I made this a separate question because search queries are considerably smaller than documents and are considerably noisier. I hence don’t know (and doubt) if the same distance metrics would be used here.

Either vanilla lexical distance metrics or state-of-the-art semantic distance metrics are preferred, with stronger preference for the latter.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Next Q](#))

User: [matt](#)

[Answer](#) by [alx49](#)

From my experience only some classes of queries can be classified on lexical features (due to ambiguity of natural language). Instead you can try to use boolean search results (sites or segments of sites, not documents, without ranking) as features for classification (instead on words). This approach works well in classes where there is a big lexical ambiguity in a query but exists a lot of good sites relevant to the query (e.g. movies, music, commercial queries and so on).

Also, for offline classification you can do LSI on query-site matrix. See “Introduction to Information Retrieval” book for details.

[Answer](#) by [simon](#)

The cosine similarity metric does a good (if not perfect) job of controlling for the document length, so comparing the similarity of 2 documents or 2 queries using the cosine metric and tf idf weights for the words should work well in either case. I would also recommend doing LSA first on tf idf weights, and then computing the cosine distance\similarities.

If you are trying to build a search engine, I would recommend using a free open source search engine like solr or elastic search, or just the raw lucene libraries, as they do most of the work for you, and have good built in methods for handling the query to document similarity problem.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Next Q](#))

[Q: Best python library for neural networks](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

I'm using Neural Networks to solve different Machine learning problems. I'm using Python and [pybrain](#) but this library is almost discontinued. Are there other good alternatives in Python?

Thanks

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

User: [marcodena](#)

[Answer](#) by [madison-may](#)

UPDATE: the landscape has changed quite a bit since I answered this question in July '14, and some new players have entered the space. In particular I would recommend checking out:

- Lasagne: <https://github.com/Lasagne/Lasagne>
- Keras: <https://github.com/fchollet/keras>
- Deepy: <https://github.com/uaca/deepy>
- Nolearn: <https://github.com/dnouri/nolearn>
- Blocks: <https://github.com/mila-udem/blocks>
- TensorFlow: <https://github.com/tensorflow/tensorflow>

They each have their strengths and weaknesses, so give them all a go and see which best suits your use case. Although I would have recommended using pylearn2 a year ago, the community is no longer active so I would recommend looking elsewhere. My original response to the answer is included below, but is largely irrelevant at this point.

[Pylearn2](#) is generally considered the library of choice for neural networks and deep learning in python. It's designed for easy scientific experimentation rather than ease of use, so the learning curve is rather steep, but if you take your time and follow the tutorials I think you'll be happy with the functionality it provides. Everything from standard Multilayer Perceptrons to Restricted Boltzmann Machines to Convolutional Nets to Autoencoders is provided. There's great GPU support and everything is built on top of Theano, so performance is typically quite good. The source for Pylearn2 is available [on github](#).

Be aware that Pylearn2 has the opposite problem of pybrain at the moment — rather than being abandoned, Pylearn2 is under active development and is subject to frequent changes.

[Answer](#) by [martin-thoma](#)

[Tensor Flow](#) ([docs](#)) by Google is another nice framework which has automatic differentiation. I've written down some [quick thoughts about Google Tensor Flow](#) on my blog, together with the MNIST example which they have in their tutorial.

[Lasagne](#) ([docs](#)) is very nice, as it uses theano (→ you can use the GPU) and makes it simpler to use. The author of lasagne won the Kaggle Galaxy challenge, as far as I know.

It is nice with [nolearn](#). Here is an MNIST example network:

[Skip code block](#)

```
#!/usr/bin/env python

import lasagne
from lasagne import layers
from lasagne.updates import nesterov_momentum
from nolearn.lasagne import NeuralNet

import sys
import os
import gzip
import pickle
import numpy

PY2 = sys.version_info[0] == 2

if PY2:
    from urllib import urlretrieve

    def pickle_load(f, encoding):
        return pickle.load(f)
else:
    from urllib.request import urlretrieve

    def pickle_load(f, encoding):
        return pickle.load(f, encoding=encoding)

DATA_URL = 'http://deeplearning.net/data/mnist/mnist.pkl.gz'
DATA_FILENAME = 'mnist.pkl.gz'

def _load_data(url=DATA_URL, filename=DATA_FILENAME):
    """Load data from `url` and store the result in `filename`."""
    if not os.path.exists(filename):
        print("Downloading MNIST dataset")
        urlretrieve(url, filename)

    with gzip.open(filename, 'rb') as f:
        return pickle_load(f, encoding='latin-1')

def load_data():
    """Get data with labels, split into training, validation and test set."""
    data = _load_data()
    X_train, y_train = data[0]
    X_valid, y_valid = data[1]
    X_test, y_test = data[2]
    y_train = numpy.asarray(y_train, dtype=numpy.int32)
    y_valid = numpy.asarray(y_valid, dtype=numpy.int32)
    y_test = numpy.asarray(y_test, dtype=numpy.int32)

    return dict(
        X_train=X_train,
        y_train=y_train,
        X_valid=X_valid,
        y_valid=y_valid,
        X_test=X_test,
        y_test=y_test,
        num_examples_train=X_train.shape[0],
        num_examples_valid=X_valid.shape[0],
        num_examples_test=X_test.shape[0],
        input_dim=X_train.shape[1],
        output_dim=10,
    )

def nn_example(data):
    net1 = NeuralNet(
        layers=[('input', layers.InputLayer),
                ('hidden', layers.DenseLayer),
                ('output', layers.DenseLayer),
                ],
        )
```

```

# layer parameters:
input_shape=(None, 28*28),
hidden_num_units=100, # number of units in 'hidden' layer
output_nonlinearity=lasagne.nonlinearities.softmax,
output_num_units=10, # 10 target values for the digits 0, 1, 2, ..., 9

# optimization method:
update=nesterov_momentum,
update_learning_rate=0.01,
update_momentum=0.9,

max_epochs=10,
verbose=1,
)

# Train the network
net1.fit(data['X_train'], data['y_train'])

# Try the network on new data
print("Feature vector (100-110): %s" % data['X_test'][0][100:110])
print("Label: %s" % str(data['y_test'][0]))
print("Predicted: %s" % str(net1.predict([data['X_test'][0]])))

def main():
    data = load_data()
    print("Got %i testing datasets." % len(data['X_train']))
    nn_example(data)

if __name__ == '__main__':
    main()

```

[Caffe](#) is a C++ library, but has Python bindings. You can do most stuff by configuration files (prototxt). It has a lot of options and can also make use of the GPU.

[Answer](#) by [jnovacho](#)

Pylearn relies on Theano and as mentioned in the other answer to use the library is quite complicated, until you get the hold of it.

In the meantime I would suggest using [Theanets](#). It also built on top of Theano, but is much more easier to work with. It might be true, that it doesn't have all the features of Pylearn, but for the basic work it's sufficient.

Also it's open source, so you can add custom networks on the fly, if you dare. :)

EDIT: Dec 2015. Recently I have started using [Keras](#). It is a bit lower level than Theanets, but much more powerful. For basic tests the Theanets is appropriate. But if you want to do some research in field of ANN Keras is much more flexible. Plus the Keras can use **Tensorflow** as a backend.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

Q: Efficient dynamic clustering

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [hierarchical-data-format](#) ([Next Q](#))

I have a set of datapoints from the unit interval (i.e. 1-dimensional dataset with numerical values). I receive some additional datapoints online, and moreover the value of some

datapoints might change dynamically. I'm looking for an ideal clustering algorithm which can handle these issues efficiently.

I know [sequential k-means clustering](#) copes with the addition of new instances, and I suppose with minor modification it can work with dynamic instance values (i.e. first taking the modified instance from the respective cluster, then updating the mean of the cluster and finally giving the modified instance as an input to the algorithm just as the addition of an unseen instance).

My concern with using the k-means algorithm is the requirement of supplying the number of clusters as an input. I know that they beat other clustering algorithms (GAs, MSTs, Hierarchical Methods etc.) in time&space complexity. Honestly I'm not sure, but maybe I can get away with using one of the aforementioned algorithms. Even that my datasets are relatively large, the existence of a single dimension makes me wonder.

More specifically a typical test case of mine would contain about 10K-200K 1-dimensional datapoints. I would like to complete the clustering preferably under a second. The dynamic changes in the value points are assumed to be smooth, i.e. relatively small. Thus being able to use existing solutions (i.e. being able to continue clustering on the existing one when a value is changed or new one is added) is highly preferred.

So all in all:

Can you think of an algorithm which will provide a sweet spot between computational efficiency and the accuracy of clusters wrt. the problem defined above?

Are there some nice heuristics for the k-means algorithm to automatically compute the value of K beforehand?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [hierarchical-data-format](#) ([Next Q](#))

User: [alex](#)

[Answer](#) by [denpost](#)

I think hierarchical clustering would be more time efficient in your case (with a single dimension). Depending on your task, you may implement something like this:

Having N datapoints d_i with their 1-dimension value x_i :

1. Sort datapoints based on their x_i value.
2. Calculate distances between adjacent datapoints ($N-1$ distances). Each distance must be assigned a pair of original datapoints (d_i, d_j).
3. Sort distances in descending order to generate list of datapoint pairs (d_i, d_j), starting from the closest one.
4. Iteratively unite datapoints (d_i, d_j) into clusters, starting from beginning of the list (the closest pair). (Depending on current state of d_i and d_j , uniting them means: (a) creating new cluster for two unclustered datapoints, (b) adding a datapoint to existing

cluster and (c) uniting two clusters.)

5. Stop uniting, if the distance is over some threshold.
6. Create singleton clusters for datapoints which did not get into clusters.

This algorithm implements [single linkage](#) clustering. It can be tuned easily to implement average linkage. [Complete linkage](#) will be less efficient, but maybe easier ones will give good results depending on your data and task.

I believe for 200K datapoints it must take under second, if you use proper data structures for above operations.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [hierarchical-data-format](#) ([Next Q](#))

Q: How to choose the features for a neural network?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

I know that there is no a clear answer for this question, but let's suppose that I have a huge neural network, with a lot of data and I want to add a new feature in input. The "best" way would be to test the network with the new feature and see the results, but is there a method to test if the feature IS UNLIKELY helpful? Like correlation measures (<http://www3.nd.edu/~mclark19/learn/CorrelationComparison.pdf>) etc?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

User: [marcodena](#)

[Answer](#) by [madison-may](#)

A very strong correlation between the new feature and an existing feature is a fairly good sign that the new feature provides little new information. A low correlation between the new feature and existing features is likely preferable.

A strong linear correlation between the new feature and the predicted variable is a good sign that a new feature will be valuable, but the absence of a high correlation is not necessarily a sign of a poor feature, because neural networks are not restricted to linear combinations of variables.

If the new feature was manually constructed from a combination of existing features, consider leaving it out. The beauty of neural networks is that little feature engineering and preprocessing is required — features are instead learned by intermediate layers. Whenever possible, prefer learning features to engineering them.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

Q: How to increase accuracy of classifiers?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Next Q](#))

I am using OpenCV letter_recog.cpp example to experiment on random trees and other classifiers. This example has implementations of six classifiers - random trees, boosting, MLP, kNN, naive Bayes and SVM. UCI letter recognition dataset with 20000 instances and 16 features is used, which I split in half for training and testing. I have experience with SVM so I quickly set its recognition error to 3.3%. After some experimentation what I got was:

UCI letter recognition:

- RTrees - 5.3%
- Boost - 13%
- MLP - 7.9%
- kNN(k=3) - 6.5%
- Bayes - 11.5%
- SVM - 3.3%

Parameters used:

- RTrees - max_num_of_trees_in_the_forrest=200, max_depth=20, min_sample_count=1
- Boost - boost_type=REAL, weak_count=200, weight_trim_rate=0.95, max_depth=7
- MLP - method=BACKPROP, param=0.001, max_iter=300 (default values - too slow to experiment)
- kNN(k=3) - k=3
- Bayes - none
- SVM - RBF kernel, C=10, gamma=0.01

After that I used same parameters and tested on Digits and MNIST datasets by extracting gradient features first (vector size 200 elements):

Digits:

- RTrees - 5.1%
- Boost - 23.4%
- MLP - 4.3%
- kNN(k=3) - 7.3%
- Bayes - 17.7%
- SVM - 4.2%

MNIST:

- RTrees - 1.4%
- Boost - out of memory
- MLP - 1.0%
- kNN($k=3$) - 1.2%
- Bayes - 34.33%
- SVM - 0.6%

I am new to all classifiers except SVM and kNN, for these two I can say the results seem fine. What about others? I expected more from random trees, on MNIST kNN gives better accuracy, any ideas how to get it higher? Boost and Bayes give very low accuracy. In the end I'd like to use these classifiers to make a multiple classifier system. Any advice?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Next Q](#))

User: [mika](#) 

[Answer](#)  by [sebastian-raschka](#) 

Dimensionality Reduction

Another important procedure is to compare the error rates on training and test dataset to see if you are overfitting (due to the “curse of dimensionality”). E.g., if your error rate on the test dataset is much larger than the error on the training data set, this would be one indicator.

In this case, you could try dimensionality reduction techniques, such as PCA or LDA.

If you are interested, I have written about PCA, LDA and some other techniques here: http://sebastianraschka.com/index.html#machine_learning  and in my GitHub repo here: https://github.com/rasbt/pattern_classification 

Cross validation

Also you may want to take a look at cross-validation techniques in order to evaluate the performance of your classifiers in a more objective manner

[Answer](#)  by [debasis](#) 

I expected more from random trees:

- With random forests, typically for N features, \sqrt{N} features are used for each decision tree construction. Since in your case $N=20$, you could try setting *max_depth* (the number of sub-features to construct each decision tree) to 5.
- Instead of decision trees, linear models have been proposed and evaluated as base estimators in random forests, in particular multinomial logistic regression and naive Bayes. This might improve your accuracy.

On MNIST kNN gives better accuracy, any ideas how to get it higher?

- Try with a higher value of K (say 5 or 7). A higher value of K would give you more supportive evidence about the class label of a point.

- You could run PCA or Fisher's Linear Discriminant Analysis before running k-nearest neighbour. By this you could potentially get rid of correlated features while computing distances between the points, and hence your k neighbours would be more robust.
 - Try different K values for different points based on the variance in the distances between the K neighbours.
-

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Next Q](#))

Q: Clustering geo location coordinates (lat,long pairs)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

What is the right approach and clustering algorithm for geo location clustering?

I'm using the following code to cluster geolocation coordinates :

[Skip code block](#)

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.vq import kmeans2, whiten

coordinates= np.array([
    [lat, long],
    [lat, long],
    ...
    [lat, long]
])
x, y = kmeans2(whiten(coordinates), 3, iter = 20)
plt.scatter(coordinates[:,0], coordinates[:,1], c=y);
plt.show()
```

Is it right to use Kmeans for location clustering, as it uses Euclidean distance and not Haversine formula as a distance function?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [user1264304](#) 

[Answer](#)  by [mike1886](#) 

K-means should be right in this case. Since k-means tries to group based solely on euclidean distance between objects you will get back clusters of locations that are close to each other.

To find the optimal number of clusters you can try making an 'elbow' plot of the within group sum of square distance. This may be helpful

(<http://nbviewer.ipython.org/github/nborwankar/LearnDataScience/blob/master/notebooks/Means%20Clustering%20Analysis.ipynb>)

[Answer](#)  by [anony-mousse](#) 

K-means is not the most appropriate algorithm here.

The reason is that k-means is designed to **minimize variance**. This is, of course, appearing from a statistical and signal processing point of view, but your data is not “linear”.

Since your data is in latitude, longitude format, you should use an algorithm that can handle *arbitrary* distance functions, in particular geodetic distance functions. Hierarchical clustering, PAM, CLARA, and DBSCAN are popular examples of this.

The problems of k-means are easy to see when you consider points close to the +180 degrees wrap-around. Even if you hacked k-means to use Haversine distance, in the update step when it recomputes the *mean* the result will be badly screwed. **Worst case is, k-means will never converge!**

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

[Q: t-SNE Python implementation: Kullback-Leibler divergence](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

t-SNE, as in [1], works by progressively reducing the Kullback-Leibler (KL) divergence, until a certain condition is met. The creators of t-SNE suggests to use KL divergence as a performance criterion for the visualizations:

you can compare the Kullback-Leibler divergences that t-SNE reports. It is perfectly fine to run t-SNE ten times, and select the solution with the lowest KL divergence [2]

I tried two implementations of t-SNE:

- **python:** `sklearn.manifold.TSNE()`.
- **R:** `tsne`, from library(`tsne`).

Both these implementations, when verbosity is set, print the error (Kullback-Leibler divergence) for each iteration. However, they don't allow the user to get this information, which looks a bit strange to me.

For example, the code:

```
import numpy as np
from sklearn.manifold import TSNE
X = np.array([[0, 0, 0], [0, 1, 1], [1, 0, 1], [1, 1, 1]])
model = TSNE(n_components=2, verbose=2, n_iter=200)
t = model.fit_transform(X)
```

produces:

```
[t-SNE] Computing pairwise distances...
[t-SNE] Computed conditional probabilities for sample 4 / 4
[t-SNE] Mean sigma: 1125899906842624.000000
[t-SNE] Iteration 10: error = 6.7213750, gradient norm = 0.0012028
[t-SNE] Iteration 20: error = 6.7192064, gradient norm = 0.0012062
[t-SNE] Iteration 30: error = 6.7178683, gradient norm = 0.0012114...
[t-SNE] Error after 200 iterations: 0.270186
```

Now, as far as I understand, **0.270186** should be the KL divergence. However i cannot get this information, neither from **model** nor from **t** (which is a simple numpy.ndarray).

To solve this problem I could: i) Calculate KL divergence by my self, ii) Do something nasty in python for capturing and parsing TSNE() function's output [3]. However: i) would be quite stupid to re-calculate KL divergence, when TSNE() has already computed it, ii) would be a bit unusual in terms of code.

Do you have any other suggestion? Is there a standard way to get this information using this library?

I mentioned I tried *R*'s tsne library, but I'd prefer the answers to focus on the *python* sklearn implementation.

References

- [1] <http://nbviewer.ipython.org/urls/gist.githubusercontent.com/AlexanderFabisch/1a0c648de2SNE.ipynb> 
- [2] <http://homepage.tudelft.nl/19j49/t-SNE.html> 
- [3] <http://stackoverflow.com/questions/16571150/how-to-capture-stdout-output-from-a-python-function-call> 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

User: [joker](#) 

[Answer](#)  by [trey](#) 

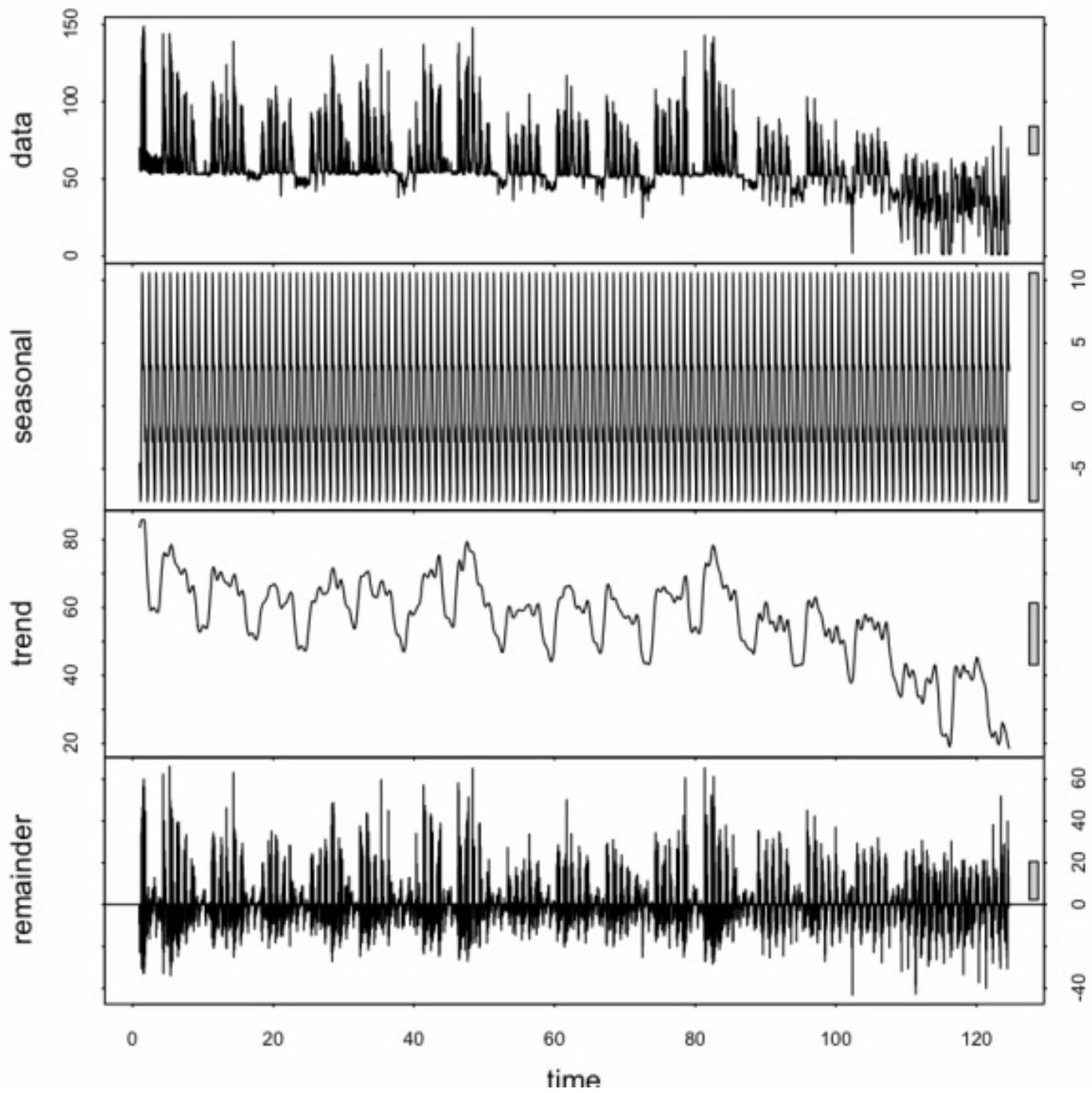
The TSNE source in scikit-learn is in pure Python. Fit `fit_transform()` method is actually calling a private `_fit()` function which then calls a private `_tsne()` function. That `_tsne()` function has a local variable `error` which is printed out at the end of the fit. Seems like you could pretty easily change one or two lines of source code to have that value returned to `fit_transform()`.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: Why should I care about seasonal data when I forecast?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Next Q](#))

I have a timeseries with hourly gas consumption. I want to use [ARMA](#) /[ARIMA](#)  to forecast the consumption on the next hour, basing on the previous. Why should I analyze/find the seasonality (with [Seasonal and Trend decomposition using Loess](#)  (STL))?



Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Next Q](#))

User: [marcodena](#)

[Answer](#) by [spacedman](#)

“Because its there”.

The data has a seasonal pattern. So you model it. The data has a trend. So you model it. Maybe the data is correlated with the number of sunspots. So you model that. Eventually you hope to get nothing left to model than uncorrelated random noise.

But I think you've screwed up your STL computation here. Your residuals are clearly not serially uncorrelated. I rather suspect you've not told the function that your “seasonality” is a 24-hour cycle rather than an annual one. But hey you haven't given us any code or data so we don't really have a clue what you've done, do we? What do you think

“seasonality” even means here? Do you have any idea?

Your data seems to have three peaks every 24 hours. Really? Is this ‘gas’=‘gasoline’=‘petrol’ or gas in some heating/electric generating system? Either way if you know a priori there’s an 8 hour cycle, or an 8 hour cycle on top of a 24 hour cycle on top of what looks like a very high frequency one or two hour cycle you **put that in your model.**

Actually you don’t even say what your x-axis is so maybe its days and then I’d fit a daily cycle, a weekly cycle, and then an annual cycle. But given how it all changes at time=85 or so I’d not expect a model to do well on both sides of that.

With statistics (which is what this is, sorry to disappoint you but you’re not a data scientist yet) you don’t just robotically go “And.. Now.. I.. Fit.. An... S TL model....”. You look at your data, try and get some understanding, then propose a model, fit it, test it, and use the parameters it make inferences about the data. Fitting cyclic seasonal patterns is part of that.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Next Q](#))

[Q: Advantages of AUC vs standard accuracy](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I was starting to look into area under curve(AUC) and am a little confused about its usefulness. When first explained to me, AUC seemed to be a great measure of performance but in my research I’ve found that some claim its advantage is mostly marginal in that it is best for catching ‘lucky’ models with high standard accuracy measurements and low AUC.

So should I avoid relying on AUC for validating models or would a combination be best? Thanks for all your help.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [aidankmcl](#) 

[Answer](#)  by [sean-owen](#) 

AUC and accuracy are fairly different things. AUC applies to binary classifiers that have some notion of a decision threshold internally. For example logistic regression returns positive/negative depending on whether the logistic function is greater/smaller than a threshold, usually 0.5 by default. When you choose your threshold, you have a classifier. You have to choose one.

For a given choice of threshold, you can compute accuracy, which is the proportion of true positives and negatives in the whole data set.

AUC measures how true positive rate (recall) and false positive rate trade off, so in that sense it is already measuring something else. More importantly, AUC is not a function of threshold. It is an evaluation of the classifier as threshold varies over all possible values. It

is in a sense a broader metric, testing the quality of the internal value that the classifier generates and then compares to a threshold. It is not testing the quality of a particular choice of threshold.

AUC has a different interpretation, and that is that it's also the probability that a randomly chosen positive example is ranked above a randomly chosen negative example, according to the classifier's internal value for the examples.

AUC is computable even if you have an algorithm that only produces a ranking on examples. AUC is not computable if you truly only have a black-box classifier, and not one with an internal threshold. These would usually dictate which of the two is even available to a problem at hand.

AUC is, I think, a more comprehensive measure, although applicable in fewer situations. It's not strictly better than accuracy; it's different. It depends in part on whether you care more about true positives, false negatives, etc.

F-measure is more like accuracy in the sense that it's a function of a classifier and its threshold setting. But it measures precision vs recall (true positive rate), which is not the same as either above.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[**Q: Statistics + Computer Science = Data Science?**](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

i want to become a **data scientist**. I studied applied **statistics** (actuarial science), so i have a great statistical background (regression, stochastic process, time series, just for mention a few). But now, I am going to do a master degree in **Computer Science** focus in Intelligent Systems.

Here is my study plan:

- Machine learning
- Advanced machine learning
- Data mining
- Fuzzy logic
- Recommendation Systems
- Distributed Data Systems
- Cloud Computing
- Knowledge discovery
- Business Intelligence
- Information retrieval
- Text mining

At the end, with all my statistical and computer science knowledge, can i call myself a data scientist? , or am i wrong?

Thanks for the answers.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [user3643160](#) 

[Answer](#)  by [samthebest](#) 

Well it depends on what kind of “Data Science” you wish to get in to. For basic analytics and reporting statistics will certainly help, but for Machine Learning and Artificial Intelligence then you’ll want a few more skills

- **Probability theory** - you must have a solid background in pure probability so that you can decompose any problem, whether seen before or not, into probabilistic principles. Statistics helps a lot for already solved problems, but new and unsolved problems require a deep understanding of probability so that you can design appropriate techniques.
 - **Information Theory** - this (relative to statistics) is quite a new field (though still decades old), the most important work was by Shannon, but even more important and often neglected note in literature is work by Hobson that proved that Kullback-Leibler Divergence is the only mathematical definition that truly captures the notion of a “*measure of information*”. Now fundamental to artificial intelligence is being able to quantify information. Suggest reading “Concepts in Statistical Mechanics” - Arthur Hobson (very expensive book, only available in academic libraries).
 - **Complexity Theory** - A big problem many Data Scientists face that do not have a solid complexity theory background is that their algorithms do not scale, or just take an extremely long time to run on large data. Take PCA for example, many peoples favourite answer to the interview question “how do you reduce the number of features in our dataset”, but even if you tell the candidate “the data set is really really really large” they still propose various forms of PCA that are $O(n^3)$. If you want to stand out, you want to be able to solve each problem on it’s own, NOT throw some text book solution at it designed a long time ago before Big Data was such a hip thing. For that you need to understand how long things take to run, not only theoretically, but practically - so how to use a cluster of computers to distribute an algorithm, or which data structures take up less memory.
 - **Communication Skills** - A huge part of Data Science is understanding business. Whether it’s inventing a product driven by data science, or giving business insight driven by data science, being able to communicate well with both the Project and Product Managers, the tech teams, and your fellow data scientists is very important. You can have an amazing idea, say an awesome AI solution, but if you cannot effectively (a) communicate WHY that will make the business money, (b) convince your colleagues it will work and (c) explain to tech people how you need their help to build it, then it wont get done.
-

[Answer](#)  by [sebastian-raschka](#) 

Data scientist (to me) a big umbrella term. I would see a data scientist as a person who can proficiently use techniques from the fields of data mining, machine learning, pattern

classification, and statistics.

However, those terms are intertwined to: machine learning is tied together with pattern classification, and also data mining overlaps when it comes finding patterns in data. And all techniques have their underlying statistical principles. I always picture this as a Venn diagram with a huge intersection.

Computer sciences is related to all those fields too. I would say that you need “data science” techniques to do computer-scientific research, but computer science knowledge is not necessarily implied in “data science”. However, programming skills - I see programming and computer science as different professions, where programming is more the tool in order solve problems - are also important to work with the data and to conduct data analysis.

You have a really nice study plan, and it all makes sense. But I am not sure if you “want” to call yourself just “data scientist”, I have the impression that “data scientist” is such a ambiguous term that can mean everything or nothing. What I want to convey is that you will end up being something more - more “specialized” - than “just” a data scientist.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

Q: Should I go for a ‘balanced’ dataset or a ‘representative’ dataset?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Next Q](#))

My ‘machine learning’ task is of separating benign Internet traffic from malicious traffic. In the real world scenario, most (say 90% or more) of Internet traffic is benign. Thus I felt that I should go with the similar kind of data for training my models as well. But then I did come across a research paper or two (in my area of work) which have used a balanced data to train models, implying equal number of instances of benign and malicious traffic.

In general, if I am building ML models, should I go for a dataset which is representative of the real world problem, or is a balanced dataset better suited for building the models (since certain classifiers do not behave well with class imbalance, or due to other reasons not known to me)?

Can someone shed more light on the *pros* and *cons* of both the choices, and how to decide which one to go for?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Next Q](#))

User: [pnp](#) 

[Answer](#)  by [dsea](#) 

I would say the answer depends on your use case. Based on my experience:

- If you’re trying to build a representative model — one that describes the data rather than necessarily predicts — then I would suggest using a representative sample of your data.

- If you want to build a predictive model, particularly one that performs well by measure of AUC or rank-order and plan to use a basic ML framework (i.e. Decision Tree, SVM, Naive Bayes, etc), then I would suggest you feed the framework a balanced dataset. Much of the literature on class imbalance finds that random undersampling (down sampling the majority class to the size of the minority class) can drive performance gains.
 - If you're building a predictive model, but are using a more advanced framework (i.e. something that determines sampling parameters via wrapper or a modification of a bagging framework that samples to class equivalence), then I would suggest again feeding the representative sample and letting the algorithm take care of balancing the data for training.
-

[Answer](#)  by [pasmod-turing](#) 

I think it always depends on the scenario. Using a representative data set is not always the solution. Assume that your training set with 1000 negative examples and 20 positive examples. Without any modeification of the classifier, your algorithm will tend to classify all new examples as negative. In some scenarios this is O.K. But in many cases the costs of missing postive examples is high so you have to find a solution for it.

In such cases you can use a cost sensitive machine learning algorithm. For example in the case of medical diagnosis data analysis.

In summary: Classification erros do not have the same cost!

[Answer](#)  by [damienfrancois](#) 

There always is the solution to try both approaches and keep the one that maximizes the expected performances.

In your case, I would assume you prefer minimizing false negatives at the cost of some false positive, so you want to bias your classifier against the strong negative prior, and address the imbalance by reducing the number of negative examples in your training set.

Then compute the precision/recall, or sensitivity/specificity, or whatever criterion suits you on the full, imbalanced, dataset to make sure you haven't ignored a significant pattern present in the real data while building the model on the reduced data.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Next Q](#))

Q: Data Science Project Ideas

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

I don't know if this is a right place to ask this question, but a community dedicated to Data Science should be the most appropriate place in my opinion.

I have just started with Data Science and Machine learning. I am looking for long term project ideas which I can work on for like 8 months.

A mix of Data Science and Machine learning would be great.

A project big enough to help me understand the core concepts and also implement them at the same time would be very beneficial.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [kevin-desai](#)

[Answer](#) by [aleksandr-blekh](#)

I would try to analyze and solve one or more of the problems published on **Kaggle Competitions** (<https://www.kaggle.com/competitions>). Note that the competitions are grouped by their expected *complexity*, from 101 (bottom of the list) to Research and Featured (top of the list). A color-coded vertical band is a *visual guideline* for grouping. You can **assess time** you could spend on a project by **adjusting** the expected *length* of corresponding competition, based on your *skills* and *experience*.

A number of **data science project ideas** can be found by browsing the following Coursolve webpage: <https://www.coursolve.org/browse-needs?query=Data%20Science>.

If you have skills and desire to work on a **real data science project**, focused on **social impacts**, visit DataKind projects page: <http://www.datakind.org/projects>. More projects with social impacts focus can be found at Data Science for Social Good fellowship webpage: <http://dssg.io/projects>.

Science Project Ideas page at My NASA Data site looks like another place to visit for inspiration: <http://mynasadata.larc.nasa.gov/804-2>.

If you would like to use **open data**, this long list of applications on `data.gov` can provide you with some interesting *data science* project ideas: <http://www.data.gov/applications>.

[Answer](#) by [ffriend](#)

Take something from your everyday life. Create predictor of traffic jams in your region, craft personalised music recommender, analyse car market, etc. Choose **real problem** that you **want to solve** - this will not only keep you motivated, but also make you go through the whole development circle from data collection to hypothesis testing.

[Answer](#) by [alexey-grigorev](#)

[Introduction to Data Science](#) course that is being run on Coursera now includes real-world project assignment where companies post their problems and students are encouraged to solve them. This is done via [coursolve.com](#) (already mentioned here).

More information [here](#) (you have to be enrolled in the course to see that link)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

Q: Predicting next medical condition from past conditions in claims data

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Next Q](#)), [beginner](#) ([Next Q](#))

I am, admittedly, very new to data science. I have spent the last 8 months or so learning as much as I can about the field and its methods. I am having issues choosing which methods to apply.

I am currently working with a large set of health insurance claims data that includes some laboratory and pharmacy claims. The most consistent information in the data set, however, is made up of diagnosis (ICD-9CM) and procedure codes (CPT, HCSPCS, ICD-9CM).

My goals are to:

1. Identify the most influential precursor conditions (comorbidities) for a medical condition like chronic kidney disease;
2. Identify the likelihood (or probability) that a patient will develop a medical condition based on the conditions they have had in the past;
3. Do the same as 1 and 2, but with procedures and/or diagnoses.
4. Preferably, the results would be interpretable by a doctor

I have looked at things like the [Heritage Health Prize Milestone papers](#) and have learned a lot from them, but they are focused on predicting hospitalizations.

I have thrown a number of algorithms at the problem (random forests, logistic regression, CART, Cox regressions) and it's been an amazing learning experience. I have not been able to decide on what "works" or "doesn't work," if you know what I mean. I have enough knowledge and skills to be misled by my own excitement and naivete; what I need is to be able to get excited about something real.

So here are my questions: What methods do you think work well for problems like this? And, what resources would be most useful for learning about data science applications and methods relevant to healthcare and clinical medicine?

EDIT #2 to add plaintext table:

CKD is the target condition, "chronic kidney disease", ".any" denotes that they have acquired that condition at any time, ".isbefore.ckd" means they had that condition before

their first diagnosis of CKD. The other abbreviations correspond with other conditions identified by ICD-9CM code groupings. This grouping occurs in SQL during the import process. Each variable, with the exception of patient_age, is binary.

[Skip code block](#)

	gender	patient_age	anx.any	art.any	ast.any	bpa.any	can.any	cer.any	chf.any	ckd.any	dep.any	dia.any	
1	Male	31	1	0	1	1	0	0	0	0	0	1	
2	Female	29	1	0	1	1	0	0	0	0	0	0	
3	Female	31	0	1	1	1	0	0	0	0	0	1	
4	Female	53	1	1	1	1	1	0	0	0	0	0	
5	Male	47	0	1	0	0	0	0	0	0	0	0	
6	Female	48	0	1	1	1	1	0	0	0	0	0	
	skn.any	tra.any	anx.isbefore.ckd	art.isbefore.ckd	ast.isbefore.ckd	bpa.isbefore.ckd	can.isbefore.ckd	dep.isbefore.ckd	dia.isbefore.ckd	end.isbefore.ckd	flu.isbefore.ckd	hrt.isbefore.ckd	hyp.isbefore.ckd
1	1	1			0		0		0				0
2	0	1			0		0		0				0
3	0	0			0		0		0				0
4	1	0			0		0		0				0
5	0	1			0		0		0				0
6	1	0			0		0		0				0
	dep.isbefore.ckd	dia.isbefore.ckd	end.isbefore.ckd	flu.isbefore.ckd	hrt.isbefore.ckd	hyp.isbefore.ckd							
1	0			0		0		0					0
2	0			0		0		0					0
3	0			0		0		0					0
4	0			0		0		0					0
5	0			0		0		0					0
6	0			0		0		0					0
	sdp.isbefore.ckd	skn.isbefore.ckd	tra.isbefore.ckd										
1	0		0		0								
2	0		0		0								
3	0		0		0								
4	0		0		0								
5	0		0		0								
6	0		0		0								

EDIT to add sample data frame:

[Skip code block](#)

```
structure(list(gender = structure(c(1L, 2L, 2L, 2L, 1L, 2L), .Label = c("Male", "Female"), class = "factor"), patient_age = c(31, 29, 31, 53, 47, 48), anx.any = c(1, 1, 0, 1, 0, 0), art.any = c(0, 0, 1, 1, 1, 1), ast.any = c(1, 1, 1, 0, 1, 1), bpa.any = c(1, 1, 1, 1, 0, 0), can.any = c(0, 0, 1, 0, 1, 0), cer.any = c(0, 0, 0, 0, 0, 0), chf.any = c(0, 0, 0, 0, 0, 0), ckd.any = c(0, 0, 0, 0, 0, 0), dep.any = c(1, 1, 0, 1, 0, 0), dia.any = c(0, 0, 1, 0, 0, 0), end.any = c(1, 1, 0, 1, 0, 1), flu.any = c(1, 0, 0, 0, 0, 0), hrt.any = c(1, 0, 0, 1, 0, 1), hyp.any = c(1, 0, 0, 0, 1, 0), inf.any = c(0, 0, 0, 1, 0, 1), men.any = c(1, 0, 1, 0, 1, 0), ren.any = c(0, 0, 0, 0, 0, 0), sdp.any = c(0, 0, 0, 0, 0, 0), skn.any = c(1, 0, 0, 1, 0, 1), tra.any = c(1, 1, 0, 1, 0, 1), anx.isbefore.ckd = c(0, 0, 0, 0, 0, 0), art.isbefore.ckd = c(0, 0, 0, 0, 0, 0), ast.isbefore.ckd = c(0, 0, 0, 0, 0, 0), bpa.isbefore.ckd = c(0, 0, 0, 0, 0, 0), can.isbefore.ckd = c(0, 0, 0, 0, 0, 0), cer.isbefore.ckd = c(0, 0, 0, 0, 0, 0), chf.isbefore.ckd = c(0, 0, 0, 0, 0, 0), ckd.isbefore.ckd = c(0, 0, 0, 0, 0, 0), dep.isbefore.ckd = c(0, 0, 0, 0, 0, 0), dia.isbefore.ckd = c(0, 0, 0, 0, 0, 0), end.isbefore.ckd = c(0, 0, 0, 0, 0, 0), flu.isbefore.ckd = c(0, 0, 0, 0, 0, 0), hrt.isbefore.ckd = c(0, 0, 0, 0, 0, 0), hyp.isbefore.ckd = c(0, 0, 0, 0, 0, 0), inf.isbefore.ckd = c(0, 0, 0, 0, 0, 0), men.isbefore.ckd = c(0, 0, 0, 0, 0, 0), ren.isbefore.ckd = c(0, 0, 0, 0, 0, 0), sdp.isbefore.ckd = c(0, 0, 0, 0, 0, 0), skn.isbefore.ckd = c(0, 0, 0, 0, 0, 0), tra.isbefore.ckd = c(0, 0, 0, 0, 0, 0)), .Names = c("gender", "patient_age", "anx.any", "art.any", "ast.any", "bpa.any", "can.any", "cer.any", "chf.any", "ckd.any", "dep.any", "dia.any", "end.any", "flu.any", "hrt.any", "hyp.any", "inf.any", "men.any", "ren.any", "sdp.any", "skn.any", "tra.any", "anx.isbefore.ckd", "art.isbefore.ckd", "ast.isbefore.ckd", "bpa.isbefore.ckd", "can.isbefore.ckd", "cer.isbefore.ckd", "chf.isbefore.ckd", "ckd.isbefore.ckd", "dep.isbefore.ckd", "dia.isbefore.ckd", "end.isbefore.ckd", "flu.isbefore.ckd", "hrt.isbefore.ckd", "hyp.isbefore.ckd", "inf.isbefore.ckd", "men.isbefore.ckd", "ren.isbefore.ckd", "sdp.isbefore.ckd", "skn.isbefore.ckd", "tra.isbefore.ckd"), row.names = c(NA, 6L), class = "data.frame")
```

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Next Q](#)), [beginner](#) ([Next Q](#))

User: [hodos](#) 

[Answer](#) by [ffriend](#)

I've never worked with medical data, but from general reasoning I'd say that relations between variables in healthcare are pretty complicated. Different models, such as random forests, regression, etc. could capture only part of relations and ignore others. In such circumstances it makes sense to use general **statistical exploration** and **modelling**.

For example, the very first thing I would do is finding out **correlations** between possible precursor conditions and diagnoses. E.g. in what percent of cases chronic kidney disease was preceded by long flu? If it is high, it [doesn't always mean causality](#), but gives pretty good food for thought and helps to better understand relations between different conditions.

Another important step is data visualisation. Does CKD happens in males more often than in females? What about their place of residence? What is distribution of CKD cases by age? It's hard to grasp large dataset as a set of numbers, plotting them out makes it much easier.

When you have an idea of what's going on, perform [hypothesis testing](#) to check your assumption. If you reject null hypothesis (basic assumption) in favour of alternative one, congratulations, you've made "something real".

Finally, when you have a good understanding of your data, try to create complete **model**. It may be something general like [PGM](#) (e.g. manually-crafted Bayesian network), or something more specific like linear regression or [SVM](#), or anything. But in any way you will already know how this model corresponds to your data and how you can measure its efficiency.

As a good starting resource for learning statistical approach I would recommend [Intro to Statistics](#) course by Sebastian Thrun. While it's pretty basic and doesn't include advanced topics, it describes most important concepts and gives systematic understanding of probability theory and statistics.

[Answer](#) by [dani](#)

While I am not a data scientist, I am an epidemiologist working in a clinical setting. Your research question did not specify a time period (ie odds of developing CKD in 1 year, 10 years, lifetime?).

Generally, I would go through a number of steps before even thinking about modeling (univariate analysis, bivariate analysis, collinearity checks, etc). However, the most commonly used method for trying to predict a binary event (using continuous OR binary variables) is logistic regression. If you wanted to look at CKD as a lab value (urine albumin, eGFR) you would use linear regression (continuous outcome).

While the methods used should be informed by your data and questions, clinicians are used to seeing odds ratios and risk ratios as these the most commonly reported measures of association in medical journals such as NEJM and JAMA.

If you are working on this problem from a human health perspective (as opposed to

Business Intelligence) this Steyerberg's [Clinical Prediction Models](#) is an excellent resource.

[Answer](#) by [andy-blankertz](#)

"Identify the most influential precursor conditions (comorbidities) for a medical condition like chronic kidney disease"

I'm not sure that it's possible to ID *the* most influential conditions; I think it will depend on what model you're using. Just yesterday I fit a random forest and a boosted regression tree to the same data, and the order and relative importance each model gave for the variables were quite different.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Next Q](#)), [beginner](#) ([Next Q](#))

[Q: When is there enough data for generalization?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

Are there any general rules that one can use to infer what can be learned/generalized from a particular data set? Suppose the dataset was taken from a sample of people. Can these rules be stated as functions of the sample or total population?

I understand the above may be vague, so a case scenario: Users participate in a search task, where the data are their queries, clicked results, and the HTML content (text only) of those results. Each of these are tagged with their user and timestamp. A user may generate a few pages - for a simple fact-finding task - or hundreds of pages - for a longer-term search task, like for class report.

Edit: In addition to generalizing about a population, given a sample, I'm interested in generalizing about an individual's overall search behavior, given a time slice. Theory and paper references are a plus!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

User: [matt](#)

[Answer](#) by [aleksandr-blekh](#)

It is my understanding that *random sampling* is a **mandatory condition** for making any *generalization* statements. IMHO, other parameters, such as sample size, just affect probability level (confidence) of generalization. Furthermore, clarifying the @ffriend's comment, I believe that you have to **calculate** needed *sample size*, based on desired values of *confidence interval*, *effect size*, *statistical power* and *number of predictors* (this is based on Cohen's work - see References section at the following link). For multiple regression, you can use the following calculator: <http://www.danielsoper.com/statcalc3/calc.aspx?id=1>.

More information on **how to select, calculate and interpret effect sizes** can be found in the following nice and comprehensive paper, which is freely available:

<http://jpepsy.oxfordjournals.org/content/34/9/917.full> 

If you're using R (and even, if you don't), you may find the following Web page on **confidence intervals and R** interesting and useful:

http://osc.centerforopenscience.org/static/CIs_in_r.html 

Finally, the following **comprehensive guide** to survey **sampling** can be helpful, even if you're not using survey research designs. In my opinion, it contains a wealth of useful information on *sampling methods*, *sampling size determination* (including calculator) and much more: <http://home.ubalt.edu/ntsbarsh/stat-data/Surveys.htm> 

[Answer](#)  by [ssdecontrol](#) 

There are two rules for generalizability:

1. The sample must be **representative**. In expectation, at least, the distribution of features in your sample must match the distribution of features in the population. When you are fitting a model with a response variable, *this includes features that you do not observe, but that affect any response variables in your model*. Since it is, in many cases, impossible to know what you do not observe, **random sampling** is used. The idea with randomization is that a random sample, up to sampling error, *must* accurately reflect the distribution of all features in the population, observed and otherwise. This is why **randomization is the “gold standard,”** but if sample control is available by some other technique, or it is defensible to argue that there are no omitted features, then it isn't always necessary.
2. Your sample must be **large enough** that the effect of **sampling error** on the feature distribution is relatively small. This is, again, to ensure representativeness. But deciding who to sample is different from deciding how many people to sample.

Since it sounds like you're fitting a model, there's the additional consideration that certain important combinations of features could be relatively rare in the population. This is not an issue for generalizability, but it bears heavily on your considerations for sample size. For instance, I'm working on a project now with (non-big) data that was originally collected to understand the experiences of minorities in college. As such, it was critically important to ensure that **statistical power** was high *specifically in the minority subpopulation*. For this reason, blacks and Latinos were deliberately **oversampled**. However, the proportion by which they were oversampled was also recorded. These are used to compute survey weights. These can be used to re-weight the sample so as to reflect the estimated population proportions, in the event that a representative sample is required.

An additional consideration arises if your model is hierarchical. A canonical use for a hierarchical model is one of children's behavior in schools. Children are “grouped” by school and share school-level traits. Therefore a representative sample of schools is required, and within each school a representative sample of children is required. This leads to **stratified sampling**. This and some other sampling designs are reviewed in surprising depth on [Wikipedia](#) .

[Answer](#) by [mrmcgreg](#)

To answer a simpler, but related question, namely ‘How well can my model generalize on the data that I have?’ the method of learning curves might be applicable. [This](#) is a lecture given by Andrew Ng about them.

The basic idea is to plot test set error and training set error vs. the complexity of the model you are using (this can be somewhat complicated). If the model is powerful enough to fully ‘understand’ your data, at some point the complexity of the model will be high enough that performance on the training set will be close to perfect. However, the variance of a complex model will likely cause the test set performance to increase at some point.

This analysis tells you two main things, I think. The first is an upper limit on performance. It’s pretty unlikely that you’ll do better on data that you haven’t seen than on your training data. The other thing it tells you is whether or not getting more data might help. If you can demonstrate that you fully understand your training data by driving training error to zero it might be possible, through the inclusion of more data, to drive your test error further down by getting a more complete sample and then training a powerful model on that.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

[Q: Solving a system of equations with sparse data](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

I am attempting to solve a set of equations which has 40 independent variables (x_1, \dots, x_{40}) and one dependent variable (y). The total number of equations (number of rows) is ~ 300 , and I want to solve for the set of 40 coefficients that minimizes the total sum-of-square error between y and the predicted value.

My problem is that the matrix is very sparse and I do not know the best way to solve the system of equations with sparse data. An example of the dataset is shown below:

y	x_1	x_2	x_3	x_4	x_5	$x_6\dots$	x_{40}	
87169	14	0	1	0	0	2	...	0
46449	0	0	4	0	1	4	...	12
846449	0	0	0	0	0	3	...	0....

I am currently using a Genetic Algorithm to solve this and the results are coming out with roughly a factor of two difference between observed and expected.

Can anyone suggest different methods or techniques which are capable of solving a set of equations with sparse data.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

User: [mike1886](#)

[Answer](#) by [aleksandr-blekh](#)

If I understand you correctly, this is the case of **multiple linear regression with sparse**

data (*sparse regression*). Assuming that, I hope you will find the following **resources** useful.

1) NCSU **lecture slides on sparse regression** with overview of algorithms, notes, formulas, graphics and references to literature:

<http://www.stat.ncsu.edu/people/zhou/courses/st810/notes/lect23sparse.pdf> 

2) R ecosystem offers many **packages**, useful for sparse regression analysis, including:

- **Matrix** (<http://cran.r-project.org/web/packages/Matrix>) 
- **SparseM** (<http://cran.r-project.org/web/packages/SparseM>) 
- **MatrixModels** (<http://cran.r-project.org/web/packages/MatrixModels>) 
- **glmnet** (<http://cran.r-project.org/web/packages/glmnet>) 
- **flare** (<http://cran.r-project.org/web/packages/flare>) 

3) A blog post with an **example of sparse regression solution**, based on SparseM:

<http://aleph-nought.blogspot.com/2012/03/multiple-linear-regression-with-sparse.html> 

4) A blog post on using **sparse matrices in R**, which includes a **primer** on using glmnet:

<http://www.johnmyleswhite.com/notebook/2011/10/31/using-sparse-matrices-in-r> 

5) **More examples and some discussion** on the topic can be found on **StackOverflow**:

<http://stackoverflow.com/questions/3169371/large-scale-regression-in-r-with-a-sparse-feature-matrix> 

UPDATE (based on your comment):

If you're trying to solve an LP problem with constraints, you may find this **theoretical paper** useful: <http://web.stanford.edu/group/SOL/papers/gmsw84.pdf> .

Also, check R package **limSolve**: <http://cran.r-project.org/web/packages/limSolve>. And, in general, check packages in CRAN Task View “**Optimization and Mathematical Programming**”: <http://cran.r-project.org/web/views/Optimization.html> .

Finally, check the book “**Using R for Numerical Analysis in Science and Engineering**” (by Victor A. Bloomfield). It has a section on solving systems of equations, represented by **sparse matrices** (section 5.7, pages 99-104), which includes examples, based on some of the above-mentioned packages: http://books.google.com/books?id=9ph_AwAAQBAJ&pg=PA99&lpg=PA99&dq=r+limsolve+sparse+matrix&source=bl&oi=icjmsATGkYDAAg&ved=0CDUQ6AEwAw#v=onepage&q=r%20limsolve%20sparse%20matrix

[Answer](#)  by [stephan-kolassa](#) 

[Aleksandr's answer](#)  is completely correct.

However, the way the question is posed implies that this is a straightforward ordinary least squares regression question: minimizing the sum of squared residuals between a dependent variable and a linear combination of predictors.

Now, while there may be many zeros in your design matrix, your system as such is not overly large: 300 observations on 40 predictors is no more than medium-sized. You can run such a regression using R without any special efforts for sparse data. Just use the `lm()` command (for “linear model”). Use `?lm` to see the help page. And note that `lm` will by

default silently add a constant column of ones to your design matrix (the intercept) - include a `-1` on the right hand side of your formula to suppress this. Overall, assuming all your data (and nothing else) is in a `data.frame` called `foo`, you can do this:

```
model <- lm(y~.-1,data=foo)
```

And then you can look at parameter estimates etc. like this:

```
summary(model)
residuals(model)
```

If your system is *much* larger, say on the order of 10,000 observations and hundreds of predictors, looking at specialized sparse solvers as per [Aleksandr's answer](#) may start to make sense.

Finally, in your comment to [Aleksandr's answer](#), you mention constraints on your equation. If that is actually your key issue, there are ways to calculate constrained least squares in R. I personally like `pcls()` in the `mgcv` package. Perhaps you want to edit your question to include the type of constraints (box constraints, nonnegativity constraints, integrality constraints, linear constraints, ...) you face?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

[Q: Classifying Java exceptions](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

We have a classification algorithm to categorize Java exceptions in Production. This algorithm is based on hierarchical human defined rules so when a bunch of text forming an exception comes up, it determines what kind of exception is (development, availability, configuration, etc.) and the responsible component (the most inner component responsible of the exception). In Java an exception can have several causing exceptions, and the whole must be analyzed.

For example, given the following example exception:

```
com.myapp.CustomException: Error printing...
... (stack)
Caused by: com.foo.webservice.RemoteException: Unable to communicate...
... (stack)
Caused by: com.acme.PrintException: PrintServer002: Timeout....
... (stack)
```

First of all, our algorithm splits the whole stack in three isolated exceptions. Afterwards it starts analyzing these exceptions starting from the most inner one. In this case, it determines that this exception (the second caused by) is of type `Availability` and that the responsible component is a “print server”. This is because there is a rule that matches containing the word `Timeout` associated to the `Availability` type. There is also a rule that matches `com.acme.PrintException` and determines that the responsible component is a print server. As all the information needed is determined using only the most inner exception, the upper exceptions are ignored, but this is not always the case.

As you can see this kind of approximation is very complex (and chaotic) as a human have to create new rules as new exceptions appear. Besides, the new rules have to be compatible with the current ones because a new rule for classifying a new exception must not change the classification of any of the already classified exceptions.

We are thinking about using Machine Learning to automate this process. Obviously, I am not asking for a solution here as I know the complexity but I'd really appreciate some advice to achieve our goal.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

User: [isidrogh](#) 

[Answer](#)  by [ffriend](#) 

First of all, some basics of classification (and in general any supervised ML tasks), just to make sure we have same set of concepts in mind.

Any supervised ML algorithm consists of at least 2 components:

1. Dataset to train and test on.
2. Algorithm(s) to handle these data.

Training dataset consists of a set of pairs (x, y), where x is a **vector of features** and y is **predicted variable**. Predicted variable is just what you want to know, i.e. in your case it is exception type. Features are more tricky. You cannot just throw raw text into an algorithm, you need to extract meaningful parts of it and organize them as feature vectors first.

You've already mentioned a couple of useful features - exception class name (e.g. `com.acme.PrintException`) and contained words ("Timeout"). All you need is to translate your row exceptions (and human-categorized exception types) into suitable dataset, e.g.:

ex_class	contains_timeout	...	ex_type
<hr/>			
[com.acme.PrintException, 1		, ...]	Availability
[java.lang.Exception, 0		, ...]	Network
<hr/>			

This representation is already much better for ML algorithms. But which one to take?

Taking into account nature of the task and your current approach natural choice is to use **decision trees**. This class of algorithms will compute optimal decision criteria for all your exception types and print out resulting tree. This is especially useful, because you will have possibility to manually inspect how decision is made and see how much it corresponds to your manually-crafted rules.

There's, however, possibility that some exceptions with exactly the same features will belong to different exception types. In this case probabilistic approach may work well. Despite its name, **Naive Bayes** classifier works pretty well in most cases. There's one issue with NB and our dataset representation, though: dataset contains *categorical* variables, and Naive Bayes can work with *numerical* attributes only*. Standard way to overcome this problem is to use [dummy variables](#) . In short, dummy variables are binary variables that simply indicate whether specific category presents or not. For example,

single variable `ex_class` with values `{com.acme.PrintException, java.lang.Exception, ...}`, etc. may be split into several variables `ex_class_printexception`, `ex_class_exception`, etc. with values `{0, 1}`:

<code>ex_class_printexception</code>	<code>ex_class_exception</code>	<code>contains_timeout</code>	<code> ex_type</code>
[1,	,	0	,
[0,	,	1	,
		0]
]
			Availability
			Network

One last algorithm to try is **Support Vector Machines (SVM)**. It neither provides helpful visualisation, nor is probabilistic, but often gives superior results.

* - in fact, neither Bayes theorem, nor Naive Bayes itself state anything about variable type, but most software packages that come to mind rely on numerical features.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

[Q: Do Random Forest overfit?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I have been reading around about Random Forests but I cannot really find a definitive answer about the problem of overfitting. According to the original paper of Breiman, they should not overfit when increasing the number of trees in the forest, but it seems that there is not consensus about this. This is creating me quite some confusion about the issue.

Maybe someone more expert than me can give me a more concrete answer or point me in the right direction to better understand the problem.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [markusian](#) 

[Answer](#)  by [alexey-grigorev](#) 

You may want to check [cross-validated](#)  - a stackexchange website for many things, including machine learning.

In particular, this question (with exactly same title) has already been answered multiple times. Check these links: <http://stats.stackexchange.com/search?q=random+forest+overfit> 

But I may give you the short answer to it: yes, it does overfit, and sometimes you need to control the complexity of the trees in your forest, or even prune when they grow too much - but this depends on the library you use for building the forest. E.g. in `randomForest` in R you can only control the complexity

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Machine learning libraries for Ruby](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Are there any machine learning libraries for Ruby that are relatively complete (including a wide variety of algorithms for supervised and unsupervised learning), robustly tested, and well-documented? I love Python's [scikit-learn](#) for its incredible documentation, but a client would prefer to write the code in Ruby since that's what they're familiar with.

Ideally I am looking for a library or set of libraries which, like scikit and numpy, can implement a wide variety of data structures like sparse matrices, as well as learners.

Some examples of things we'll need to do are binary classification using SVMs, and implementing bag of words models which we hope to concatenate with arbitrary numeric data, as described in [this](#) Stackoverflow post.

Thanks in advance!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [andy](#)

[Answer](#) by [andy](#)

I'll go ahead and post an answer for now; if someone has something better I'll accept theirs.

At this point the most powerful option appears to be accessing WEKA using jRuby. We spent yesterday scouring the 'net, and this combination was even used by a [talk at RailsConf 2012](#), so I would guess if there were a comparable pure ruby package, they would have used it.

Note that if you know exactly what you need, there are plenty of individual libraries that either [wrap standalone packages like libsvm](#) or [re-implement some individual algorithms like Naive Bayes in pure Ruby](#) and will spare you from using jRuby.

But for a general-purpose library, WEKA and jRuby seem to be the best bet at this time.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Quick guide into training highly imbalanced data sets](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

I have a classification problem with approximately 1000 positive and 10000 negative samples in training set. So this data set is quite unbalanced. Plain random forest is just trying to mark all test samples as a majority class.

Some good answers about sub-sampling and weighted random forest are given here: [What are the implications for training a Tree Ensemble with highly biased datasets?](#)

Which classification methods besides RF can handle the problem in the best way?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev](#)

[Q](#)) ([Next Q](#))

User: [ihars](#) 

[Answer](#)  by [charles](#) 

- Max Kuhn covers this well in Ch16 of *Applied Predictive Modeling*.
 - As mentioned in the linked thread, imbalanced data is essentially a cost sensitive training problem. Thus any cost sensitive approach is applicable to imbalanced data.
 - There are a large number of such approaches. Not all implemented in R: C50, weighted SVMs are options. Jous-boost. Rusboost I think is only available as Matlab code.
 - I don't use Weka, but believe it has a large number of cost sensitive classifiers.
 - *Handling imbalanced datasets: A review*: Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas'
 - *On the Class Imbalance Problem*: Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, Guangtong Zhou
-

[Answer](#)  by [alexey-grigorev](#) 

Undersampling the majority class is usually the way to go in such situations.

If you think that you have too few instances of the positive class, you may perform oversampling, for example, sample 5n instances with replacement from the dataset of size n.

Caveats:

- Some methods may be sensitive to changes in the class distribution, e.g. for Naive Bayes - it affects the prior probabilities.
 - Oversampling may lead to overfitting
-

[Answer](#)  by [cwharland](#) 

Gradient boosting is also a good choice here. You can use the gradient boosting classifier in sci-kit learn for example. Gradient boosting is a principled method of dealing with class imbalance by constructing successive training sets based on incorrectly classified examples.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

[Q: Looking for a strong Phd Topic in Predictive Analytics in the context of Big Data](#) 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

I'm going to start a Computer Science phd this year and for that I need a research topic. I am interested in Predictive Analytics in the context of Big Data. I am interested by the area of Education (MOOCs, Online courses...). In that field, what are the unexplored areas that can help me choose a strong topic? Thanks.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

User: [innovismail](#) 

[Answer](#)  by [m.dax](#) 

As a fellow CS Ph.D. defending my dissertation in a Big Data-esque topic this year (I started in 2012), the best piece of material I can give you is in a link:

<http://www.rpajournal.com/dev/wp-content/uploads/2014/10/A3.pdf> 

This is an article written by two Ph.D.s from MIT who have talked about Big Data and MOOCs. Probably, you will find this a good starting point. BTW, along this note, if you really want to come up with a valid topic (that a committee and your adviser will let you propose, research and defend) you need to read LOTS and LOTS of papers. The majority of Ph.D. students make the fatal error of thinking that some 'idea' they have is new, when it's not and has already been done. You'll have to do something truly original to earn your Ph.D. Rather than actually focus on forming an idea right now, you should do a good literature survey and the ideas will 'suggest themselves'. Good luck! It's an exciting time for you.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

[Q: Cosine Similarity for Ratings Recommendations? Why use it?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

Lets say I have a database of users who rate different products on a scale of 1-5. Our recommendation engine recommends products to users based on the preferences of other users who are highly similar. My first approach to finding similar users was to use Cosine Similarity, and just treat user ratings as vector components. The main problem with this approach is that it just measures vector angles and doesn't take rating scale or magnitude into consideration.

My question is this:

Are there any drawbacks to just using the percentage difference between the vector components of two vectors as a measure of similarity? What disadvantages, if any, would I encounter if I used that method, instead of Cosine Similarity or Euclidean Distance?

For Example, why not just do this:

```
n = 5 stars  
a = (1, 4, 4)
```

```
b = (2,3,4)  
similarity(a,b) = 1 - ( (|1-2|/5) + (|4-3|/5) + (|4-4|/5) ) / 3 = .86667
```

Instead of Cosine Similarity :

```
a = (1,4,4)  
b = (2,3,4)  
  
CosSimilarity(a,b) =  
(1^2)+(4^2)+(4^2) / sqrt( (1^2)+(4^2)+(4^2) ) * sqrt( (2^2)+(3^2)+(4^2) ) = .9697
```

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

User: [myclamm](#) 

[Answer](#)  by [emre](#) 

Rating bias and scale can easily be accounted for by standardization. The point of using Euclidean similarity metrics in vector space co-embeddings is that it reduces the recommendation problem to one of finding the nearest neighbors, which can be done efficiently both exactly and approximately. What you don't want to do in real-life settings is to have to compare every item/user pair and sort them according to some expensive metric. That just doesn't scale.

One trick is to use an approximation to cull the herd to a manageable size of tentative recommendations, then to run your expensive ranking on top of that.

edit: Microsoft Research is presenting a paper that covers this very topic at RecSys right now: [Speeding Up the Xbox Recommender System Using a Euclidean Transformation for Inner-Product Spaces](#) 

[Answer](#)  by [buruzaemon](#) 

For ratings, I think you would need to use [Spearman's rank correlation](#)  for your similarity metric.

Cosine similarity is often used when comparing documents, and perhaps would not be a good fit for rank variables. Euclidean distance is fine for lower dimensions, but comparison of rank variables normally call for Spearman.

Here's a [question on CrossValidated regarding Spearman \(vs Pearson\)](#) , which might shed more light for you.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

[Q: Hashing Trick - what actually happens](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

When ML algorithms, e.g. Vowpal Wabbit or some of the factorization machines winning click through rate competitions ([Kaggle](#) ), mention that features are 'hashed', what does that actually mean for the model? Lets say there is a variable that represents the ID of an internet add, which takes on values such as '236BG231'. Then I understand that this

feature is hashed to a random integer. But, my question is:

- Is the integer now used in the model, as an integer (numeric) OR
- is the hashed value actually still treated like a categorical variable and one-hot-encoded? Thus the hashing trick is just to save space somehow with large data?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

User: [b_miner](#) 

[Answer](#)  by [cwharland](#) 

The a second bullet is the value in feature hashing. Hashing and one hot encoding to sparse data saves space. Depending on the hash algo you can have varying degrees of collisions which acts as a kind of dimensionality reduction.

Also, in the specific case of Kaggle feature hashing and one hot encoding help with feature expansion/engineering by taking all possible tuples (usually just second order but sometimes third) of features that are then hashed with collisions that explicitly create interactions that are often predictive whereas the individual features are not.

In most cases this technique combined with feature selection and elastic net regularization in LR acts very similar to a one hidden layer NN so it performs quite well in competitions.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

[Q: Where to start on neural networks](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

First of all I know the question may be not suitable for the website but I'd really appreciate it if you just gave me some pointers.

I'm a 16 years old programmer, I've had experience with many different programming languages, a while ago I started a course at Coursera, titled introduction to machine learning and since that moment i got very motivated to learn about AI, I started reading about neural networks and I made a working perceptron using Java and it was really fun but when i started to do something a little more challenging (building a digit recognition software), I found out that I have to learn a lot of math, I love math but the schools here don't teach us much, now I happen to know someone who is a math teacher do you think learning math (specifically calculus) is necessary for me to learn AI or should I wait until I learn those stuff at school?

Also what other things would be helpful in the path of me learning AI and machine learning? do other techniques (like SVM) also require strong math?

Sorry if my question is long, I'd really appreciate if you could share with me any experience you have had with learning AI.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [ashkan](#) 

[Answer](#)  by [emre](#) 

No, you should go ahead and learn the maths on your own. You will “only” need to learn calculus, statistics, and linear algebra (like the rest of machine learning). The theory of neural networks is pretty primitive at this point — it more of an art than a science — so I think you can understand it if you try. *Ipsso facto*, there are a lot of tricks that you need practical experience to learn. There are lot of complicated extensions, but you can worry about them once you get that far.

Once you can understand the Coursera classes on ML and neural networks (Hinton's), I suggest getting some practice. You might like [this](#)  introduction.

[Answer](#)  by [polmath](#) 

I would say... it really depends. You may need to:

- *use* machine learning algorithms: this will be useful for specific applications you may have. In this situation what you need is some programming skills and the taste for testing (practicing will make you strong). Here maths are not so much required I would say.
- be able to *modify* existing algorithms. Your specific application may be reticent to

regular algorithms, so you may need to adapt them to get maximum efficiency. Here maths come into play.

- understand the theory behind algorithms. Here maths are necessary, and will help you increase your knowledge of the field of machine learning, develop your own algorithms, speak the same language as your peers... NN theory may be primitive as said by @Emre, but for instance this is not the case for SVM (the theory behind SVM requires e.g. to understand [reproducing kernel Hilbert spaces](#)).

On the mid term for sure you will need strong maths. But you don't need to wait for them to come to you, you can start right now with linear algebra, which is beautiful and useful for everything. And in case you encounter (possibly temporary) difficulties of any sort with maths, keep on practicing the way you already do (many people can talk about the perceptron but are not able to make a perceptron in Java), this is very valuable.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

Q: Rough vs Fuzzy vs Granular Computing

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

For my Computational Intelligence class, I'm working on classifying short text. One of the papers that I've found makes a lot of use of *granular computing*, but I'm struggling to find a decent explanation of what exactly it is.

From what I can gather from the paper, it sounds to me like granular computing is very similar to fuzzy sets. So, what exactly is the difference. I'm asking about rough sets as well, because I'm curious about them and how they relate to fuzzy sets. If at all.

Edit: [Here](#) is the paper I'm referencing.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [thad](#)

[Answer](#) by [mrmeritology](#)

“**Granularity**” refers to the **resolution** of the variables under analysis. If you are analyzing *height* of people, you could use *course-grained variables* that have only a few possible values — e.g. “above-average, average, below-average” — or a *fine-grained variable*, with many or an infinite number of values — e.g. integer values or real number values.

A measure is “**fuzzy**” if the distinction between alternative values is not crisp. In the course-grained variable for *height*, a “crisp” measure would mean that any given individual could *only* be assigned one value — e.g. a tall-ish person is either “above-average”, or “average”. In contrast, a “fuzzy” measure allows for *degrees of membership*

for each value, with “membership” taking values from 0 to 1.0. Thus, a tall-ish person could be a value of “0.5 above-average”, “0.5 average”, “0.0 below-average”.

Finally, a measure is “**rough**” when two values are given: upper and lower bounds as an estimate of the “crisp” measure. In our example of a tall-ish person, the rough measure would be {UPPER = above-average, LOWER = average}.

Why use granular, fuzzy, or rough measures at all, you might ask? Why not measure everything in nice, precise real numbers? Because many real-world phenomena don’t have a good, reliable intrinsic measure and measurement procedure that results in a real number. If you ask married couples to rate the quality of their marriage on a scale from 1 to 10, or 1.00 to 10.00, they might give you a number (or range of numbers), but how reliable are those reports? Using a coarse-grained measure (e.g. “happy”, “neutral/mixed”, “unhappy”), or fuzzy measure, or rough measure can be more reliable and more credible in your analysis. Generally, it’s much better to use rough/crude measures well than to use precise/fine-grained measures poorly.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

[Q: Machine learning - features engineering from date/time data](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Prev Q](#)) ([Next Q](#))

What are the common/best practices to handle time data for machine learning application?

For example, if in data set there is a column with timestamp of event, such as “2014-05-05”, how you can extract useful features from this column if any?

Thanks in advance!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Prev Q](#)) ([Next Q](#))

User: [igor-bobriakov](#) 

[Answer](#)  by [nar](#) 

In several cases data and events inside a time series are seasonal. In such cases the month and the year of the event matters a lot. Hence in such scenarios you can use binary variables to represent if the event is during a given month/year or not.

Hope this answers your question. If not kindly be a little more specific on what exactly are you trying to achieve

[Answer](#)  by [gurpreet-mohaar](#) 

Divide the data into windows and find features for those windows like autocorrelation coefficients, wavelets, etc. and use those features for learning.

For example, if you have temperature and pressure data, break it down to individual parameters and calculate features like number of local minima in that window and others, and use these features for your model.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Prev Q](#)) ([Next Q](#))

Q: Mahout Similarity algorithm comparison

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

Which of the following is best (or widely used) for calculating item-item similarity measure in mahout and why ?

Pearson Correlation
Spearman Correlation
Euclidean Distance
Tanimoto Coefficient
LogLikelihood Similarity

Is there any thumb-rule to chose from these set of algorithm also how to differentiate each of them ?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

User: [sreejithc321](#)

Answer by [sheldonkreger](#)

For those not familiar, item-item recommenders calculate similarities between items, as opposed to user-user (or user-based) recommenders, which calculate similarities between users. Although some algorithms can be used for both, this question is in regard to item-item algorithms (thanks for being specific in your question).

Accuracy or effectiveness of recommenders is evaluated based on comparing recommendations to a previously collected data set (training set). For example, I have shopping cart data from the last six months; I'll use the first 5 months as training data, then run my various algorithms, and compare the quality against what really happened during the 6th month.

The reason Mahout ships with so many algorithms is because different algorithms are more or less effective in each data set you may work with. So, ideally, you do some testing as I described with many algorithms and compare the accuracy, then choose the winner.

Interestingly, you can also take other factors into account, such as the need to minimize the data set (for performance reasons), and run your tests only with a certain portion of the training data available. In such a case, one algorithm may work better with the smaller data set, but another may work with the complete set. Then, you get to weigh performance VS accuracy VS challenge of implementation (such as deploying on a Hadoop cluster).

Therefore, different algorithms are suited for different project. However, there are some

general rules:

1. All algorithms always do better with unreduced data sets (more data is better).
2. More complex algorithms aren't necessarily better.

I suggest starting with a simple algorithm and ensuring you have high quality data. If you have additional time, you can implement more complex algorithms and create a comparison which is unique to your data set.

Most of my info comes from [This study](#). You'll find lots of detail about implementation there.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

[Q: Consequence of Feature Scaling](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

I am currently using SVM and scaling my training features to the range of [0,1]. I first fit/transform my training set and then apply the **same** transformation to my testing set. For example:

```
### Configure transformation and apply to training set
min_max_scaler = MinMaxScaler(feature_range=(0, 1))
X_train = min_max_scaler.fit_transform(X_train)

### Perform transformation on testing set
X_test = min_max_scaler.transform(X_test)
```

Let assume that a given feature in the training set has a range of [0,100], and that same feature in the testing set has a range of [-10,120]. In the training set that feature will be scaled appropriately to [0,1], while in the testing set that feature will be scaled to a range outside of that first specified, something like [-0.1,1.2].

I was wondering what the consequences of the testing set features being out of range of those being used to train the model? Is this a problem?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [mike1886](#)

[Answer](#) by [cbeleites](#)

Within each class, you'll have distributions of values for the features. That in itself is not a reason for concern.

From a slightly theoretical point of view, you can ask yourself why you should scale your features and why you should scale them in exactly the chosen way.

One reason may be that your particular training algorithm is known to converge faster (better) with values around 0 - 1 than with features which cover other orders of magnitude. In that case, you're probably fine. My guess is that your SVM is fine: you want to avoid too large numbers because of the inner product, but a max of 1.2 vs. a max of 1.0 won't

make much of a difference.

(OTOH, if you e.g. knew your algorithm to not accept negative values you'd obviously be in trouble.)

The practical question is whether your model performs well for cases that are slightly out of the range covered by training. This I believe can best and possibly only be answered by testing with such cases / inspecting test results for performance drop for cases outside the training domain. It is a valid concern and looking into this would be part of the validation of your model.

Observing differences of the size you describe is IMHO a reason to have a pretty close look at model stability.

[Answer](#)  by [robert-smith](#) 

This was meant as a comment but it is too long.

The fact that your test set has a different range **might** be a sign that the training set is not a good representation of the test set. However, if the difference is really small as in your example, it is likely that it won't affect your predictions. Unfortunately, I don't think I have a good reason to think it won't affect a SVM in any circumstance.

Notice that the rationale for using MinMaxScalar is (according to the documentation):

The motivation to use this scaling include robustness to very small standard deviations of features and preserving zero entries in sparse data.

Therefore, it is important for you to make sure that your data fits that case.

If you are really concerned about having a difference range, you should use a regular standardization (such as `preprocessing.scale`) instead.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to connect data-mining with machine learner process](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#))

I want to write a data-mining service in [Google Go](#)  which collects data through scraping and APIs.

However as Go lacks good ML support I would like to do the ML stuff in Python.

Having a web background I would connect both services with something like RPC but as I believe that this is a common problem in data science I think that there is some better solution.

For example most (web) protocols lack at:

- buffering between processes
- clustering over multiple instances

So what (type of libraries) do data scientists use to connect different languages/processes?

Bodo

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [bodokaiser](#) 

[Answer](#)  by [hack-r](#) 

The Data Science Toolkit is [a powerful library](#)  (or collection of libraries, technically) which are available in a number of languages. For instance, I use the implementation called RDSTK in R.

In the case of your preferred language, Google Go, there's [a list](#)  of web-related libraries here which looks very useful.

[Answer](#)  by [rawkintrevo](#) 

If your only motivation for using Google Go is webscraping, and you want to do you ML in python, I would recommend the following stack:

[Python requests for scraping data](#) 

[MongoDB for caching data](#)  (MongoDB's page oriented format makes it a natural home for storing JSON objects commonly returned by APIs)

[pymongo for interfacing python and mongodb](#) 

[scikit-learn for doing your machine learning](#) 

This all happens in python and you can extend it multiple processors with [multiprocessing](#)  or to multiple nodes with django

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#))

Q: Item based and user based recommendation difference in Mahout

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

I would like to know how exactly mahout user based and item based recommendation differ from each other.

It defines that

[User-based](#) : Recommend items by finding similar users. This is often harder to scale because of the dynamic nature of users.

[Item-based](#) : Calculate similarity between items and make recommendations. Items usually don't change much, so this often can be computed off line.

But though there are two kind of recommendation available, what I understand is that both

these will take some data model (say 1,2 or 1,2,.5 as item1,item2,value or user1,user2,value where value is not mandatory) and will perform all calculation as the similarity measure and recommender build-in function we chose and we can run both user/item based recommendation on the same data (is this a correct assumption ??).

So I would like to know how exactly and in which all aspects these two type of algorithm differ.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

User: [sreejithc321](#) 

[Answer](#)  by [mrmcggreg](#) 

You are correct that both models work on the same data without any problem. Both items operate on a matrix of user-item ratings.

In the user-based approach the algorithm produces a rating for an item i by a user u by combining the ratings of other users u' that are similar to u . Similar here means that the two user's ratings have a high Pearson correlation or cosine similarity or something similar.

In the item-based approach we produce a rating for i by u by looking at the set of items i' that are similar to i (in the same sense as above except now we'd be looking at the ratings that items have received from users) that u has rated and then combines the ratings by u of i' into a predicted rating by u for i .

The item-based approach was invented at Amazon (<http://dl.acm.org/citation.cfm?id=642471>) to address their scale challenges with user-based filtering. The number of things they sell is much less and much less dynamic than the number of users so the item-item similarities can be computed offline and accessed when needed.

[Answer](#)  by [srinath](#) 

Item Based Algorithm

```
for every item i that u has no preference for yet
    for every item j that u has a preference for
        compute a similarity s between i and j
        add u's preference for j, weighted by s, to a running average
return the top items, ranked by weighted average
```

User Based Algorithm

```
for every item i that u has no preference for yet
    for every other user v that has a preference for i
        compute a similarity s between u and v
        add v's preference for i, weighted by s, to a running average
return the top items, ranked by weighted average
```

Item vs User based:

- 1) Recommenders scale with the number of items or users they must deal with, so there are scenarios in which each type can perform better than the other
 - 2) Similarity estimates between items are more likely to converge over time than similarities between users
 - 3) We can compute and cache similarities that converge, which can give item based recommenders a performance advantage
 - 4) Item based recommenders begin with a list of a user's preferred items and therefore do not need a nearest item neighborhood as user based recommenders do
-

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#)), [recommendation](#) ([Prev Q](#)) ([Next Q](#))

Q: Visualizing deep neural network training

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Next Q](#))

I'm trying to find an equivalent of Hinton Diagrams for multilayer networks to plot the weights during training.

The trained network is somewhat similar to a Deep SRN, i.e. it has a high number of multiple weight matrices which would make the simultaneous plot of several Hinton Diagrams visually confusing.

Does anyone know of a good way to visualize the weight update process for recurrent networks with multiple layers?

I haven't found much papers on the topic. I was thinking to display time-related information on the weights per layer instead if I can't come up with something. E.g. the weight-delta over time for each layer (omitting the use of every single connection). PCA is another possibility, though I'd like to not produce much additional computations, since the visualization is done online during training.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Next Q](#))

User: [rundosrun](#) 

[Answer](#)  by [piotr-migdal](#) 

The closes thing I know is [ConvNetJS](#) :

ConvNetJS is a Javascript library for training Deep Learning models (mainly Neural Networks) entirely in your browser. Open a tab and you're training. No software requirements, no compilers, no installations, no GPUs, no sweat.

Demos on this site plot weighs and how do they change with time (bear in mind, its many

parameters, as practical networks do have a lot of neurons). Moreover, if you are not satisfied with their plotting, there is access to networks parameters and you can plot as you wish (since it is JavaScript).

[Answer](#) by [aleksandr-blekh](#)

Based on my cursory understanding of the topics, associated with your question, I think that **Gephi** (<https://gephi.github.io>; the original gephi.org link redirects there) should be able to handle *neural network dynamic visualization*. It seems that, in order to achieve your goal, you need to **stream** your graph(s) with corresponding weights (<https://forum.gephi.org/viewtopic.php?t=1875>). For *streaming*, you most likely will need this *plug-in*: <https://marketplace.gephi.org/plugin/graph-streaming>.

UPDATE: You may also find useful SoNIA software:

<http://web.stanford.edu/group/sonia>.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Next Q](#))

[Q: What are the best practices to anonymize user names in data?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-cleaning](#) ([Next Q](#))

I'm working on a project which asks fellow students to share their original text data for further analysis using data mining techniques, and, I think it would be appropriate to anonymize student names with their submissions.

Setting aside the better solutions of a url where students submit their work and a backend script inserts the anonymized ID, **What sort of solutions could I direct students to implement on their own to anonymized their own names?**

I'm still a noob in this area. I don't know what are the norms. I was thinking the solution could be a hashing algorithm. That sounds like a better solution than making up a fake name as two people could pick the same fake name. possible people could pick the same fake name. **What are some of the concerns I should be aware of?**

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-cleaning](#) ([Next Q](#))

User: [xtian](#)

[Answer](#) by [emre](#)

I suspected you were using the names as identifiers. You shouldn't; they're not unique and they raise this privacy issue. Use instead their student numbers, which you can verify from their IDs, stored in hashed form. Use the student's last name as a salt, for good measure (form the string to be hashed by concatenating the ID number and the last name).

[Answer](#) by [stephan-kolassa](#)

A standard practice in psychology (where you want to code participants in order to link

different measurements together) is to have participants choose their mother's maiden name initials and birthdate, e.g., in the format XX-YYMMDD.

This if course can still run into conflicts. Then again, I don't think there is *any* surefire conflict-free anonymization algorithm your students could do *without knowing all the other students*. Mothers' names and birthdates could be identical, own birthdates could be identical, shoe sizes could be, favorite superhero characters... The only thing I could think of would be (US) Social Security numbers, but you *really don't want to use them* .

Bottom line: anonymize on the backend. Or, as [@Emre suggests](#), think about whether you really need an identifier at all. Maybe the DB-generated index is enough?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-cleaning](#) ([Next Q](#))

Q: What is the difference between feature generation and feature extraction?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

Can anybody tell me what is the purpose of feature generation? and why feature space enrichment is needed before classifying an image? Is it a necessary step?

Is there any method to enrich feature space?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [saratha-priya](#) 

[Answer](#)  by [hack-r](#) 

Feature Generation — This is the process of taking raw, unstructured data and defining features (i.e. variables) for potential use in your statistical analysis. For instance, in the case of text mining you may begin with a raw log of thousands of text messages (e.g. SMS, email, social network messages, etc) and generate features by removing low-value words (i.e. stopwords), using certain size blocks of words (i.e. n-grams) or applying other rules.

Feature Extraction — After generating features, it is often necessary to test transformations of the original features and select a subset of this pool of potential original and derived features for use in your model (i.e. feature extraction and selection). Testing derived values is a common step because the data may contain important information which has a non-linear pattern or relationship with your outcome, thus the importance of the data element may only be apparent in its transformed state (e.g. higher order derivatives). Using too many features can result in multiply colinearity or otherwise confound statistical models, whereas extracting the minimum number of features to suit the purpose of your analysis follows the principal of parsimony.

Enhancing your feature space in this way is often a necessary step in classification of images or other data objects because the raw feature space is typically filled with an overwhelming amount of unstructured and irrelevant data that comprises what's often

referred to as “noise” in the paradigm of a “signal” and “noise” (which is to say that some data has predictive value and other data does not). By enhancing the feature space you can better identify the important data which has predictive or other value in your analysis (i.e. the “signal”) while removing confounding information (i.e. “noise”).

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

Q: Amplifying a Locality Sensitive Hash

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I’m trying to build a cosine locality sensitive hash so I can find candidate similar pairs of items without having to compare every possible pair. I have it basically working, but most of the pairs in my data seem to have cosine similarity in the -0.2 to +0.2 range so I’m trying to dice it quite finely and pick things with cosine similarity 0.1 and above.

I’ve been reading Mining Massive Datasets chapter 3. This talks about increasing the accuracy of candidate pair selection by Amplifying a Locality-Sensitive Family. I think I just about understand the mathematical explanation, but I’m struggling to see how I implement this practically.

What I have so far is as follows

1. I have say 1000 movies each with ratings from some selection of 1M users. Each movie is represented by a sparse vector of user scores (row number = user ID, value = user’s score)
2. I build N random vectors. The vector length matches the length of the movie vectors (i.e. the number of users). The vector values are +1 or -1. I actually encode these vectors as binary to save space, with +1 mapped to 1 and -1 mapped to 0
3. I build sketch vectors for each movie by taking the dot product of the movie and each of the N random vectors (or rather, if I create a matrix R by laying the N random vectors horizontally and layering them on top of each other then the sketch for movie m is R^*m), then taking the sign of each element in the resulting vector, so I end with a sketch vector for each movie of +1s and -1s, which again I encode as binary. Each vector is length N bits.
4. Next I look for similar sketches by doing the following
 1. I split the sketch vector into b bands of r bits
 2. Each band of r bits is a number. I combine that number with the band number and add the movie to a hash bucket under that number. Each movie can be added to more than one bucket.
 3. I then look in each bucket. Any movies that are in the same bucket are candidate pairs.

Comparing this to 3.6.3 of mmmds, my AND step is when I look at bands of r bits - a pair of movies pass the AND step if the r bits have the same value. My OR step happens in the buckets: movies are candidate pairs if they are both in any of the buckets.

The book suggests I can “amplify” my results by adding more AND and OR steps, but I’m

at a loss for how to do this practically as the explanation of the construction process for further layers is in terms of checking pairwise equality rather than coming up with bucket numbers.

Can anyone help me understand how to do this?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [philip-pearl](#) 

[Answer](#)  by [philip-pearl](#) 

I think I've worked something out. Basically I'm looking for an approach that works in a map/reduce type environment and I think this approach does it.

So,

- suppose I have b bands of r rows and I want to add another AND stage, say another c ANDs.
- so instead of $b * r$ bits I need hashes of $b * r * c$ bits
- and I run my previous procedure c times, each time on $b * r$ bits
- If x and y are found to be a candidate pair by any of these procedures it emits a key value pair $((x, y), 1)$, with the tuple of IDs (x, y) as the key and the value 1
- At the end of the c procedures I group these pairs by key and sum
- Any pair (x, y) with a sum equal to c was a candidate pair in each of the c rounds, and so is a candidate pair of the entire procedure.

So now I have a workable solution, and all I need to do is work out whether using 3 steps like this will actually help me get a better result with fewer overall hash bits or better overall performance...

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Q: Machine Learning on financial big data

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#))

Disclaimer: although I know some things about big data and am currently learning some other things about machine learning, the specific area that I wish to study is vague, or at least appears vague to me now. I'll do my best to describe it, but this question could still be categorised as too vague or not really a question. Hopefully, I'll be able to reword it more precisely once I get a reaction.

So,

I have some experience with Hadoop and the Hadoop stack (gained via using CDH), and I'm reading a book about Mahout, which is a collection of machine learning libraries. I also think I know enough statistics to be able to comprehend the math behind the machine learning algorithms, and I have some experience with R. My ultimate goal is making a setup that would make trading predictions and deal with financial data in real time.

I wonder if there're any materials that I can further read to help me understand ways of managing that problem; books, video tutorials and exercises with example datasets are all welcome.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#))

User: [chiffa](#)

[Answer](#) by [aleksandr-blekh](#)

There are tons of materials on financial (big) data analysis that you can read and peruse. I'm not an expert in finance, but am curious about the field, especially in the context of data science and R. Therefore, the following are selected relevant resource suggestions that I have for you. I hope that they will be useful.

Books: Financial analysis (general / non-R)

- [Statistics and Finance: An Introduction](#);
- [Statistical Models and Methods for Financial Markets](#).

Books: Machine Learning in Finance

- [Machine Learning for Financial Engineering](#) (!) - seems to be an edited collection of papers;
- [Neural Networks in Finance: Gaining Predictive Edge in the Market](#).

Books: Financial analysis with R

- [Statistical Analysis of Financial Data in R](#);
- [Statistics and Data Analysis for Financial Engineering](#);

- [Financial Risk Modelling and Portfolio Optimization with R](#)
- [Statistics of Financial Markets: An Introduction](#) (code in R and MATLAB).

Academic Journals

- [Algorithmic Finance](#) (open access)

Web sites

- [RMetrics](#)
- [Quantitative Finance on StackExchange](#)

R Packages

- the above-mentioned *RMetrics* site (see [this page](#) for general description);
- *CRAN Task Views*, including [Finance](#), [Econometrics](#) and several other Task Views.

Competitions

- [MODELOFF \(The Financial Modeling World Championships\)](#)

Educational Programs

- [MS in Financial Engineering - Columbia University](#);
- [Computational Finance - Hong Kong University](#).

Blogs (Finance/R)

- [Timely Portfolio](#);
- [Systematic Investor](#);
- [Money-making Mankind](#).

[Answer](#) by [greg-thatcher](#)

I'm doing some similar research, and have found PluralSight, <http://pluralsight.com>, to be an invaluable resource. They have video courses on Machine Learning, AWS, Azure, Hadoop, Big Data, etc. Personally, I find that these video courses allow me to learn the material much faster and more easily than books.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#))

Q: Data Science in C (or C++)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

I'm an R language programmer. I'm also in the group of people who are considered Data Scientists but who come from academic disciplines other than CS.

This works out well in my role as a Data Scientist, however by starting my career in R and only having basic knowledge of other scripting/web languages, I've felt somewhat inadequate in 2 key areas:

1. Lack of a solid knowledge of programming theory
2. Lack of a competitive level of skill in faster and more widely used languages like C, C++ and Java, which could be utilized to increase the speed of the pipeline and Big Data computations as well as to create DS/data products which can be more readily developed into fast back-end scripts or standalone applications

The solution is simple of course — go learn about programming, which is what I've been doing by enrolling in some classes (currently C programming).

However, now that I'm starting to address problems #1 and #2 above, I'm left asking myself "*Just how viable are languages like c and c++ for Data Science?*".

For instance, I can move data around very quickly and interact with users just fine, but what about advanced regression, Machine Learning, text mining and other more advanced statistical operations?

So. can c do the job — what tools are available for advanced statistics, ML, AI, and other areas of Data Science? Or must I lose most of the efficiency gained by programming in C by calling on R scripts or other languages?

The best resource I've found thusfar in C is a library called [Shark](#) , which gives C/C++ the ability to use Support Vector Machines, linear regression (not non-linear and other advanced regression like multinomial probit, etc) and a shortlist of other (great but) statistical functions.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [hack-r](#) 

[Answer](#)  by [andre-holzner](#) 

Or must I lose most of the efficiency gained by programming in C by calling on R scripts or other languages?

Do the opposite: learn C/C++ to write R extensions. Use C/C++ only for the performance critical sections of your new algorithms, use R to build your analysis, import data, make plots etc.

If you want to go beyond R, I'd recommend learning python. There are many libraries available such as [scikit-learn](#)  for machine learning algorithms or [PyBrain](#)  for building Neural Networks etc. (and use pylab/[matplotlib](#)  for plotting and [iPython](#)

[notebooks](#) to develop your analyses). Again, C/C++ is useful to implement time critical algorithms as python extensions.

[Answer](#) by [d.castro](#)

As Andre Holzner has said, extending R with C/C++ extension is a very good way to take advantage of the best of both sides. Also you can try the inverse , working with C++ and occasionally calling function of R with the RInside package o R. Here you can find how

<http://cran.r-project.org/web/packages/RInside/index.html>

<http://dirk.eddelbuettel.com/code/rinside.html>

Once you're working in C++ you have many libraries , many of them built up for specific problems, other more general

<http://www.shogun-toolbox.org/page/features/>

http://image.diku.dk/shark/sphinx_pages/build/html/index.html

<http://mlpack.org>

[Answer](#) by [servais-daligou](#)

R is one of the key tool for data scientist, what ever you do don't stop using it.

Now talking about C, C++ or even Java. They are good popular languages. Whether you need them or will need them depend on the type of job or projects you have. From personal experience, there are so many tools out there for data scientist that you will always feel like you constantly need to be learning.

You can add Python or Matlab to things to learn if you want and keep adding. The best way to learn is to take on a work project using other tools that you are not comfortable with. If I were you, I would learn Python before C. It is more used in the community than C. But learning C is not a waste of your time.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

Q: How to generate synthetic dataset using machine learning model learnt with original dataset?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

Generally, the machine learning model is built on datasets. I'd like to know if there is any way to generate synthetic dataset using such trained machine learning model preserving original dataset characteristics ?

[original data —> build machine learning model —> use ml model to generate synthetic data....!!!]

Is it possible ? Please point me to related resource if possible.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [hadooper](#)

[Answer](#) by [mrmeritology](#)

The general approach is to do traditional statistical analysis on your data set to define a multidimensional random process that will generate data with the same statistical characteristics. The virtue of this approach is that your synthetic data is independent of your ML model, but statistically “close” to your data. (see below for discussion of your alternative)

In essence, you are estimating the multivariate probability distribution associated with the process. Once you have estimated the distribution, you can generate synthetic data through the Monte Carlo method or similar repeated sampling methods. If your data resembles some parametric distribution (e.g. lognormal) then this approach is straightforward and reliable. The tricky part is to estimate the dependence between variables. See:

https://www.encyclopediaofmath.org/index.php/Multi-dimensional_statistical_analysis.

If your data is irregular, then non-parametric methods are easier and probably more robust. [Multivariate kernel density estimation](#) is a method that is accessible and appealing to people with ML background. For a general introduction and links to specific methods, see: https://en.wikipedia.org/wiki/Nonparametric_statistics.

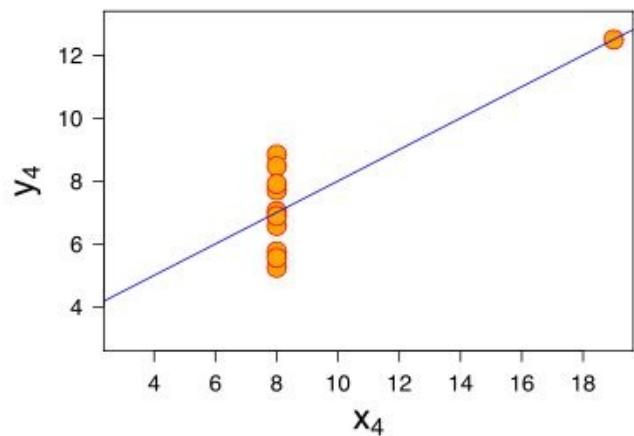
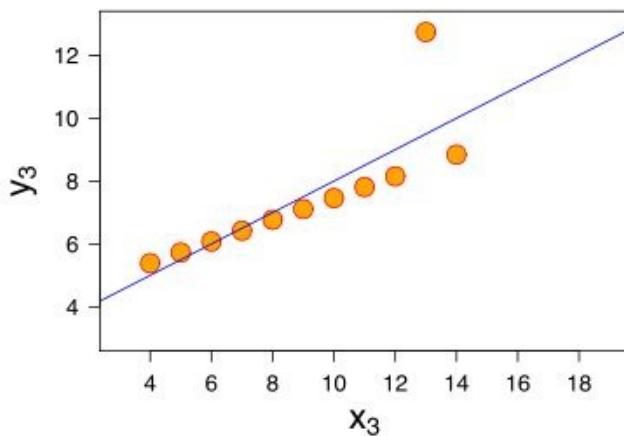
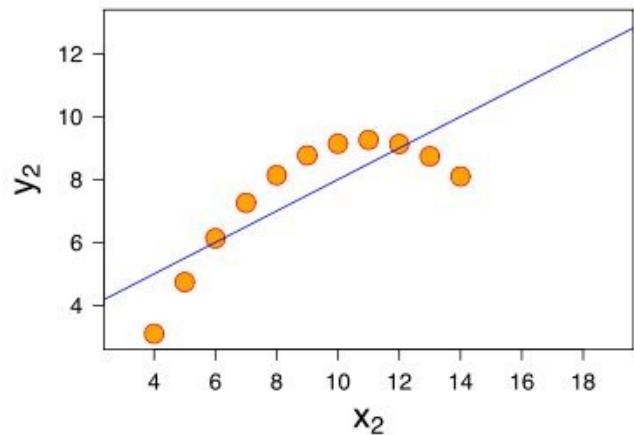
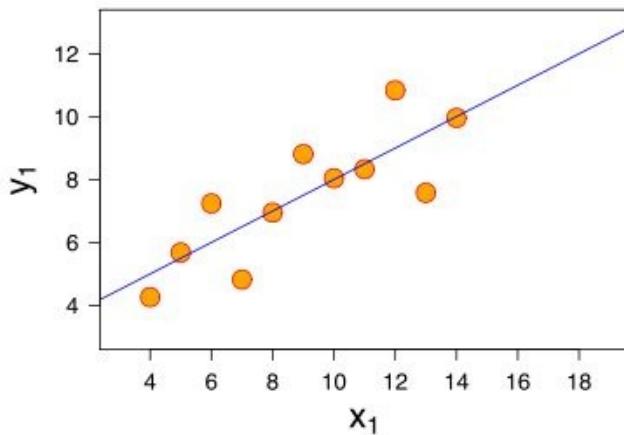
To validate that this process worked for you, you go through the machine learning process again with the synthesized data, and you should end up with a model that is fairly close to your original. Likewise, if you put the synthesized data into your ML model, you should get outputs that have similar distribution as your original outputs.

In contrast, you are proposing this:

[original data —> build machine learning model —> use ml model to generate synthetic data....!!!]

This accomplishes something different than the method I just described. This would solve the [inverse problem](#): “what inputs could generate any given set of model outputs”. Unless your ML model is over-fitted to your original data, this synthesized data *will not* look like your original data in every respect, or even most.

Consider a linear regression model. The same linear regression model can have identical fit to data that have very different characteristics. A famous demonstration of this is through [Anscombe's quartet](#).



Thought I don't have references, I believe this problem can also arise in logistic regression, generalized linear models, SVM, and K-means clustering.

There are some ML model types (e.g. decision tree) where it's possible to inverse them to generate synthetic data, though it takes some work. See: [Generating Synthetic Data to Match Data Mining Patterns](#).

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

Q: What is the term for when a model acts on the thing being modeled and thus changes the concept?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I'm trying to see if there is a conventional term for this concept to help me in my literature research and writing. When a machine learning model causes an action to be taken in the real world that affects future instances, what is that called?

I'm thinking about something like a recommender system that recommends one given product and doesn't recommend another given product. Then, you've increased the likelihood that someone is going to buy the first product and decreased the likelihood that someone is going to buy the second product. So then those sales numbers will eventually become training instances, creating a sort of feedback loop.

Is there a term for this?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [jsmith54](#) 

[Answer](#)  by [mrmeritology](#) 

There are three terms from social science that apply to your situation:

1. [Reflexivity](#)  - refers to circular relationships between cause and effect. In particular, you could use the definition of the term adopted by George Soros to refer to reverse causal loop between share prices (i.e. present value of fundamentals) and business fundamentals. In a way, the share price is a “model” of the fundamental business processes. Usually, people assume that causality is one-way, from fundamentals to share price.
2. [Performativity](#)  - As used by Donald MacKenzie (e.g. [here](#) ), many economic models are not “cameras” — taking pictures of economic reality — but in fact are “engines” — an integral part of the construction of economic reality. He has a book of that title: [An Engine, Not a Camera](#) .
3. [Self-fulfilling Prophecy](#)  - a prediction that directly or indirectly causes itself to become true, by the very terms of the prophecy itself, due to positive feedback between belief and behavior. This is the broadest term, and least specific to the situation you describe.

Of the three terms, I suggest that MacKenzie’s “performativity” is the best fit to your situation. He claims, among other things, that the validity of the economic models (e.g. Black-Scholes option pricing) has been improved by its very use by market participants, and therefore how it reflects in options pricing and trading patterns.

[Answer](#)  by [aleksandr-blekh](#) 

Though it is not specifically a term, focused on *machine learning*, but I would refer to such behavior of a statistical model, using a general term **side effect** (while adding some *clarifying adjectives*, such as expected or unexpected, desired or undesired, and similar). **Modeling outcome** or **transitive feedback loop outcome** might be some of the *alternative terms*.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: What types of features are used in a large-scale click-through rate prediction problem?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

Something that I often see in papers ([example](#) ) about large-scale learning is that click-

through rate (CTR) problems can have up to a billion of features for each example. In [this Google paper](#) the authors mention:

The features used in our system are drawn from a variety of sources, including the query, the text of the ad creative, and various ad-related metadata.

I can imagine a few thousands of features coming from this type of source, I guess through some form of feature hashing.

My question is: how does one get to a billion features? How do companies translate user behavior into features in order to reach that scale of features?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [bar](#)

[Answer](#) by [wojciech-walczak](#)

That really is a nice question, although once you're Facebook or Google etc., you have the opposite problem: how to reduce the number of features from many billions, to let's say, a billion or so.

There really are billions of features out there.

Imagine, that in your feature vector you have billions of possible phrases that the user could type in into search engine. Or, that you have billions of web sites a user could visit. Or millions of locations from which a user could log in to the system. Or billions of mail accounts a user could send mails to or receive mails from.

Or, to switch a bit to social networking site-like problem. Imagine that in your feature vector you have billions of users which a particular user could either know or be in some degree of separation from. You can add billions of links that user could post in his SNS feed, or millions of pages a user could 'like' (or do whatever the SNS allows him to do).

Similar problems may be found in many domains from voice and image recognition, to various branches of biology, chemistry etc. I like your question, because it's a good starting point to dive into the problems of dealing with the abundance of features. Good luck in exploring this area!

UPDATE due to your comment:

Using features other than binary is just one step further in imagining things. You could somehow cluster the searches, and count frequencies of searches for a particular cluster.

In a SNS setting you could build a vector of relations between users defined as degree of separation instead of a mere binary feature of being or not being friends.

Imagine logs that global corporations are holding on millions of their users. There's a whole lot of stuff that can be measured in a more detailed way than binary.

Things become even more complicated once we're considering an online setting. In such a case you do not have time for complicated computations and you're often left with binary features since they are cheaper.

And no, I am not saying, that the problem becomes tractable once it's reduced to a magical number of billion features. I am only saying that a billion of features is something you may end up after a lot of effort in reducing the number of dimensions.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is the “dying ReLU” problem in neural networks?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Referring to the Stanford course notes on [Convolutional Neural Networks for Visual Recognition][1], a paragraph says:

“Unfortunately, ReLU units can be fragile during training and can “die”. For example, a large gradient flowing through a ReLU neuron could cause the weights to update in such a way that the neuron will never activate on any datapoint again. If this happens, then the gradient flowing through the unit will forever be zero from that point on. That is, the ReLU units can irreversibly die during training since they can get knocked off the data manifold. For example, you may find that as much as 40% of your network can be “dead” (i.e. neurons that never activate across the entire training dataset) if the learning rate is set too high. With a proper setting of the learning rate this is less frequently an issue.”

What does dying of neurons here mean?

Could you please provide an intuitive explanation in simpler terms.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [tejas](#) 

[Answer](#)  by [neil-slater](#) 

A “dead” ReLU always outputs the same value (zero as it happens, but that is not important) for any input. Probably this is arrived at by learning a large negative bias term for its weights.

In turn, that means that it takes no role in discriminating between inputs. For classification, you could visualise this as a decision plane *outside* of all possible input data.

Once a ReLU ends up in this state, it is unlikely to recover, because the function gradient at 0 is also 0, so gradient descent learning will not alter the weights. “Leaky” ReLUs with a small positive gradient for negative inputs ($y=0.01x$ when $x < 0$ say) are one attempt to address this issue and give a chance to recover.

The sigmoid and tanh neurons can suffer from similar problems as their values saturate, but there is always at least a small gradient allowing them to recover in the long term.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Q: Why does logistic regression in Spark and R return different models for the same data?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [apache-spark](#) ([Next Q](#))

I've compared the logistic regression models on R (`glm`) and on Spark (`LogisticRegressionWithLBFGS`) on a dataset of 390 obs. of 14 variables.

The results are completely different in the intercept and the weights. How to explain this?

Here is the results of Spark (`LogisticRegressionWithLBFGS`) :

[Skip code block](#)

```
model.intercept :  
1.119830027739959  
model.weights :  
GEST 0.30798496002530473  
DILATE 0.28121771009716895  
EFFACE 0.01780105068588628  
CONSID -0.22782058111362183  
CONTR -0.8094592237248102  
MEMBRAN -1.788173534959893  
AGE -0.05285751197750732  
STRAT -1.6650305527536942  
GRAVID 0.38324952943210994  
PARIT -0.9463956993328745  
DIAB 0.18151162744507293  
TRANSF -0.7413500749909346  
GEMEL 1.5953124037323745
```

Here is the result of R :

[Skip code block](#)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.0682091	3.3944407	0.904	0.366052
GEST	0.0086545	0.1494487	0.058	0.953821
DILATE	0.4898586	0.2049361	2.390	0.016835 *
EFFACE	0.0131834	0.0059331	2.222	0.026283 *
CONSID	0.1598426	0.2332670	0.685	0.493196
CONTR	0.0008504	0.5788959	0.001	0.998828
MEMBRAN	-1.5497870	0.4215416	-3.676	0.000236 ***
AGE	-0.0420145	0.0326184	-1.288	0.197725
STRAT	-0.3781365	0.5860476	-0.645	0.518777
GRAVID	0.1866430	0.1522925	1.226	0.220366
PARIT	-0.6493312	0.2357530	-2.754	0.005882 **
DIAB	0.0335458	0.2163165	0.155	0.876760
TRANSF	-0.6239330	0.3396592	-1.837	0.066219 .
GEMEL	2.2767331	1.0995245	2.071	0.038391 *
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 '	0.1 '	' 1	

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [apache-spark](#) ([Next Q](#))

User: [sparkuser](#)

[Answer](#) by [ryan-j.-smith](#)

A quick glance at the [docs for LogisticRegressionWithLBFGS](#) indicates that it uses feature scaling and L2-Regularization by default. I suspect that R's `glm` is returning a maximum likelihood estimate of the model while Spark's `LogisticRegressionWithLBFGS` is returning a regularized model estimate. Note how the estimated model weights of the

Spark model are all smaller in magnitude than those in the R model.

I'm not sure whether or not `glm` in R is implementing feature scaling, but this would also contribute to different model values.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [apache-spark](#) ([Next Q](#))

Q: I am trying to classify/cluster users profile but don't know how with my attributes

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

I have a dataset about users purchasing product from website. The attributes I have are user id, region(state) of the user, the categories id of product, keywords id of product, keywords id of website, sales amount spent of the product. The goal is to use the information of product and website to identity who the users are, such as "male young gamer"; "stay at home mom". I attached a sample picture as below.

id	website_id	product_id	region	sales_amount	product_category_id	category_name	product_keyword_id	keyword_name	website_keyword_id	keyword_name
23540072	10098	2190	AL	18.89	2	Parts & Accessories	4793	autozone	805	television
23540072	10098	2190	AL	18.89	2	Parts & Accessories	4793	autozone	194	tv
23540198	10098	2190	MD	78.94	2	Parts & Accessories	4762	engine	66	compare
23540198	10098	2190	MD	78.94	2	Parts & Accessories	4762	engine	64	coupons
23540198	10098	2190	MD	78.94	2	Parts & Accessories	4762	engine	65	deals
23540198	10098	2190	MD	78.94	2	Parts & Accessories	4762	engine	3169	android
23540198	10098	2190	MD	78.94	2	Parts & Accessories	4762	engine	63	offers
23540198	10098	2190	MD	78.94	2	Parts & Accessories	4762	engine	86	mobile

There are totally 1940 unique categories and 13845 unique keywords for products. For the website, there are 13063 unique keywords. The whole dataset is huge as that's the daily logging data.

I am thinking of clustering, as those are unsupervised, but those id are ordered number having no numeric meaning, then I don't know how to apply the algorithm. I also think of classification if I add a column of class based on the sales amount of product purchased. I think clustering is more preferred. I don't know what algorithm I should use for this case as the dimensions of the keywords id could be more than 10000 (each product could have many keywords, so does website). I need to use Spark for this project. Can anyone help me out with some ideas,suggestions? Thank you so much!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [sylvia](#) 

[Answer](#)  by [logc](#) 

Right now, I only have time for a very brief answer, but I'll try to expand on it later on.

What you want to do is a **clustering**, since you want to discover some labels for your data. (As opposed to a classification, where you would have labels for at least some of the data and you would like to label the rest).

In order to perform a clustering on your users, you need to have them as some kind of points in an abstract space. Then you will measure distances between points, and say that points that are "near" are "similar", and label them according to their place in that space.

You need to transform your data into something that looks like a user profile, i.e.: a user ID, followed by a vector of numbers that represent the features of this user. In your case, each feature could be a “category of website” or a “category of product”, and the number could be the amount of dollars spent in that feature. Or a feature could be a combination of web and product, of course.

As an example, let us imagine the user profile with just three features:

- dollars spent in “techy” webs,
- dollars spent on “fashion” products,
- and dollars spent on “aggressive” video games on “family-oriented” webs (who knows).

In order to build those profiles, you need to map the “categories” and “keywords” that you have, which are too plentiful, into the features you think are relevant. Look into [topic modeling](#) or [semantic similarity](#) to do so. Once that map is built, it will state that all dollars spent on webs with keywords “gadget”, “electronics”, “programming”, and X others, should all be aggregated into our first feature; and so on.

Do not be afraid of “imposing” the features! You will need to refine them and maybe completely change them once you have clustered the users.

Once you have user profiles, proceed to cluster them using [k-means](#) or whatever else you think is interesting. Whatever technique you use, you will be interested in getting the “representative” point for each cluster. This is usually the geometric “center” of the points in that cluster.

Plot those “representative” points, and also plot how they compare to other clusters. Using a [radar chart](#) is very useful here. Wherever there is a salient feature (something in the representative that is very marked, and is also very prominent in its comparison to other clusters) is a good candidate to help you label the cluster with some catchy phrase (“nerds”, “fashionistas”, “aggressive moms” ...).

Remember that a clustering problem is an open problem, so there is no “right” solution! And I think my answer is quite long already; check also about normalization of the profiles and filtering outliers.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

Q: How do I normalize an array of positive and negative numbers so they are between 0 and 1?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Using [Brain](#) to feed in an array of data with both positive and negative numbers; the output array will be in 0's and 1's and I believe Brain only allows inputs from 0 to 1. So how do I normalize an array of negative and positive numbers so it fits these requirements?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [jonathan](#) 

[Answer](#)  by [stochazesthai](#) 

This is called unity-based normalization. If you have a vector X , you can obtain a normalized version of it, say Z , by doing:

$$Z = \frac{X - \min(X)}{\max(X) - \min(X)}$$

[Answer](#)  by [mcduffee](#) 

Find the largest positive number and the smallest (most negative) number in the array. Add the absolute value of the smallest (most negative) number to every value in the array. Divide each result by the difference between the largest and the smallest number.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to cluster a link traversal dataset](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

I'm using Google Analytics on my mobile app to see how different users use the app. I draw a path based on the pages they move to. Given a list of paths for say a 100 users, how do I go about clustering the users. Which algorithm to use? By the way, I'm thinking of using scikit learn package for the implementation.

My dataset (in csv) would look like this :

```
DeviceID,Pageid,Time_spent_on_Page,Transition.<br>
ABC,Page1, 3s, 1->2.<br>
ABC,Page2, 2s, 2->4.<br>
ABC,Page4,1s,4->1.<br>
```

So the path, here is 1->2->4->1, where 1,2,4 are Pageids.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [gokul1794](#) 

[Answer](#)  by [kasra-manshaei](#) 

@Shagun's answer is right actually. I just expand it!

There are 2 different approaches to your problem:

Graph Approach

- As stated in @Shagun's answer you have a weighted directed graph and you want to cluster the paths. I mention again because it's important to know that your problem is not a *Graph Clustering* or *Community Detection* problem where vertices are

clustered!

- Constructing a Graph in networkx using the last two column of the data, you can add time spent as weight and users who passed that link as an edge attribute. After all you'll have different features for clustering: the set of all vertices an individual ever met in the graph, total, mean and std of time spent, shortest path distribution parameters, ... which can be used for clustering the user behaviors.

Standard Data

- All above can be done by reading data efficiently in a matrix. If you consider each edge for a specified user as a single row (i.e. you'll have $M \times N$ rows where M is the number of users and N the number of edges in case you stick with 100 case!) and add properties as columns you'll probably able to cluster behaviors. if a user passed an edge n times, in the row corresponding to that user and that edge add a count column with value n and same for time spend, etc. Starting and ending edges are also informative. Be careful that node names are categorical variables.

Regarding clustering algorithms you can find enough if you have a quick look at SKlearn. Hope it helped. Good Luck :)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to learn spam email detection?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

I want to learn how a spam email detector is done. I'm not trying to build a commercial product, it'll be a serious learning exercise for me. Therefore, I'm looking for resources, such as existing projects, source code, articles, papers etc that I can follow. I want to learn by examples, I don't think I am good enough to do it from scratch. Ideally, I'd like to get my hand dirty in Bayesian.

Is there anything like that? Programming language isn't a problem for me.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [student-t](#) 

[Answer](#)  by [kasra-manshaei](#) 

First of all check [this](#)  carefully. You'll find a simple dataset and some papers to review.

BUT as you want to start a simple learning project I recommend to not going through papers (which are obviously not *basic*) but try to build your own bayesian learner which is not so difficult.

I personally suggest [Andrew Moore](#) 's lecture slides on Probabilistic Graphical Models

which are freely available and you can learn from them simply and step by step.

If you need more detailed help just comment on this answer and I'll be glad to help :)

Enjoy baysian learning!

[Answer](#) by [mike-wise](#)

In Andrew Ng's Machine Learning Course on Coursera (in someways the flagship course for Coursera) the programmers exercise for Support Vector Machines was an example doing a spam classifier. The lectures are great, famous even, and well worth watching.

There is also this posted course from him:

[http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?
course=MachineLearning&doc=exercises/ex6/ex6.html](http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex6/ex6.html)

[Answer](#) by [sheldonkreger](#)

There is a basic introduction to the Bayesian method for spam detection in the book "Doing Data Science - Straight Talk from the Frontline" by Cathy O'Neil, Rachel Schutt.

The chapter is good, because it explains why other common data science models don't work for spam classifiers. The whole book uses R throughout, so only pick it up if you are interested in working with R.

It uses the Enron email set as training data, since it has emails divided into spam/not spam already.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

Q: Implementing Complementary Naive Bayes in python?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

Problem

I have tried using Naive bayes on a labeled data set of crime data but got really poor results (7% accuracy). Naive Bayes runs much faster than other algorithms I've been using so I wanted to try finding out why the score was so low.

Research

After reading I found that Naive bayes should be used with balanced datasets because it has a bias for classes with higher frequency. Since my data is unbalanced I wanted to try using the Complementary Naive Bayes since it is specifically made for dealing with data skews. In the paper that describes the process, the application is for text classification but I don't see why the technique wouldn't work in other situations. You can find the paper I'm referring to [here](#). In short the idea is to use weights based on the occurrences where a class doesn't show up.

After doing some research I was able to find an implementation in Java but unfortunately I don't know any Java and I just don't understand the algorithm well enough to implement myself.

Question

where I can find an implementation in python? If that doesn't exist how should I go about implementing it myself?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

User: [grasshopper](#)

[Answer](#) by [alexey-grigorev](#)

Naive Bayes should be able to handle imbalanced datasets. Recall that the Bayes formula is

$$P(y | x) = \frac{P(x | y) P(y)}{P(x)} \propto P(x | y) P(y)$$

So $P(x | y) P(y)$ takes the prior $P(y)$ into account.

In your case maybe you overfit and need some smoothing? You can start with +1 smoothing and see if it gives any improvements. In python, when using numpy, I'd implement the smoothing this way:

```
table = # counts for each feature
PT = (table + 1) / (table + 1).sum(axis=1, keepdims=1)
```

Note that this is gives you Multinomial Naive Bayes - which applies only to categorical data.

I can also suggest the following link: <http://www.itshared.org/2015/03/naive-bayes-on-apache-flink.html>. It's about implementing Naive Bayes on Apache Flink. While it's Java, maybe it'll give you some theory you need to understand the algorithm better.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

Q: How to find similarity between different factors in a dataset

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Prev Q](#)) ([Next Q](#))

Introduction

Let's say I have a dataset of different observation of different people and I want to group people together to know which person is closest to the other one. I also want to have a measure to know how close they are to each others and know the statistical significance.

Data

[Skip code block](#)

	eat_rate	drink_rate	sleep_rate	play_rate	name	game
1	0.0542192259	0.13041721	5.013682e-03	1.023533e-06	Paul	Rayman
4	0.0688171511	0.01050611	6.178833e-03	3.238838e-07	Paul	Mario
6	0.0928997660	0.01828468	9.321211e-03	3.525951e-07	Jenn	Mario
7	0.0001631273	0.02212345	7.061524e-05	1.531270e-07	Jean	FIFA
8	0.0028735509	0.05414688	1.341689e-03	4.533366e-07	Mark	FIFA
10	0.0034844717	0.09152440	4.589990e-04	5.802708e-07	Mark	Rayman
11	0.0340738956	0.03384180	1.636508e-02	1.354973e-07	Mark	FIFA
12	0.0266112679	0.20002020	3.380704e-02	4.533366e-07	Mark	Sonic
14	0.0046597056	0.01848672	5.472681e-04	4.034696e-07	Paul	FIFA
15	0.0202715299	0.16365289	2.994086e-02	4.044770e-07	Lucas	SSBM

Reproduce it:

[Skip code block](#)

```
structure(list(eat_rate = c(0.0542192259374624, 0.0688171511010916,
0.0928997659570807, 0.000163127341146237, 0.00287355085557602,
0.00348447171120939, 0.0340738956099744, 0.0266112679045701,
0.00465970561072008, 0.0202715299408583), drink_rate = c(0.130417213859986,
0.0105061117284574, 0.0182846752197192, 0.0221234468128094, 0.0541468835235882,
0.0915243964036772, 0.0338418022022427, 0.200020204061016, 0.0184867158298818,
0.163652894231741), sleep_rate = c(0.00501368170182717, 0.00617883308323771,
0.00932121105128431, 7.06152352370024e-05, 0.00134168946950305,
0.000458999029040516, 0.0163650807661753, 0.0338070438697149,
0.000547268073086768, 0.029940859740489), play_rate = c(1.02353325645595e-06,
3.23883801132467e-07, 3.52595117873603e-07, 1.53127022619393e-07,
4.53336580123204e-07, 5.80270822557701e-07, 1.35497266725713e-07,
4.53336580123204e-07, 4.03469556309652e-07, 4.04476970932148e-07
), name = structure(c(5L, 5L, 2L, 1L, 4L, 4L, 4L, 5L, 3L), .Label = c("Jean",
"Jenn", "Lucas", "Mark", "Paul"), class = "factor"), game = structure(c(3L,
2L, 2L, 1L, 1L, 3L, 1L, 4L, 1L, 5L), .Label = c("FIFA", "Mario",
"Rayman", "Sonic", "SSBM"), class = "factor")), .Names = c("eat_rate",
"drink_rate", "sleep_rate", "play_rate", "name", "game"))
```

```
"drink_rate", "sleep_rate", "play_rate", "name", "game"), row.names = c(1L,  
4L, 6L, 7L, 8L, 10L, 11L, 12L, 14L, 15L), class = "data.frame")
```

Question

Given a dataset as fellow (with continuous and categorical feature), how can I know if a person (a categorical answer) identified by a name is more correlated to another person?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Prev Q](#)) ([Next Q](#))

User: [zipp](#) 

[Answer](#)  by user9424

One way is to normalize your quantitative values (play, eat, drink, sleep rates) so they all have the same range (say, 0 -> 1), then assign each game to its own “dimension”, that takes value 0 or 1. Turn each row into a vector and normalize the length to 1. Now, you can compare the inner product of any two people’s normalized vectors as a measure of similarity. Something like this is used in text mining quite often

R Code for Similarity Matrix

Assumes you've saved your dataframe to the variable "D"

[Skip code block](#)

```
#Get normalization factors for quantitative measures  
maxvect<-apply(D[,1:4],MARGIN=2,FUN=max)  
minvect<-apply(D[,1:4],MARGIN=2,FUN=min)  
rangevect<-maxvect-minvect  
#Normalize quantitative factors  
D_matrix <- as.matrix(D[,1:4])  
NormDMatrix<-matrix(nrow=10,ncol=4)  
colnames(NormDMatrix)<-colnames(D_matrix)  
for (i in 1:4) NormDMatrix[,i]<-(D_matrix[,i]-minvect[i]*rep(1,10))/rangevect[i]  
gamenames<-unique(D[,"game"])  
#Create dimension matrix for games  
Ngames<-length(gamenames)  
GameMatrix<-matrix(nrow=10,ncol=Ngames)  
for (i in 1:Ngames) GameMatrix[,i]<-as.numeric(D[,"game"]==gamenames[i])  
colnames(GameMatrix)<-gamenames  
#combine game matrix with normalized quantitative matrix  
People<-D[, "name"]  
RowVectors<-cbind(GameMatrix, NormDMatrix)  
#normalize each row vector to length of 1 and then store as a data frame with person names  
NormRowVectors<-t(apply(RowVectors,MARGIN=1,FUN=function(x) x/sqrt(sum(x*x))))  
dfNorm<-data.frame(People,NormRowVectors)  
  
#create person vectors via addition of appropriate row vectors  
PersonMatrix<-array(dim=c(length(unique(People)),ncol(RowVectors)))  
rownames(PersonMatrix)<-unique(People)  
for (p in unique(People)){  
  print(p)  
  MatchIndex<-(dfNorm[,1]==p)*seq(1,nrow(NormRowVectors))  
  MatchIndex<-MatchIndex[MatchIndex>0]  
  nclm<-length(MatchIndex)  
  SubMatrix<-matrix(NormRowVectors[MatchIndex, ],nrow=length(MatchIndex),ncol=dim(NormRowVectors)[2])  
  CSUMS<-colSums(SubMatrix)  
  NormSum<-sqrt(sum(CSUMS*CSUMS))  
  PersonMatrix[p, ]<-CSUMS/NormSum  
}  
colnames(PersonMatrix)<-colnames(NormRowVectors)  
#Calculate matrix of dot products  
Similarity<-(PersonMatrix)%*%t(PersonMatrix)
```

[Answer](#) by [thomas-pazur](#)

Despite normalized euclidean distance you can also have a look at the pearson distance as a similarity measure. Here is a neat description :

http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sphilip/pear.htm

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to avoid overfitting in random forest?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

1. I want to avoid overfitting in random forest. In this regard, I intend to use mtry, nodesize and maxnodes etc. Could you please help me how to choose values for these parameters. I use R.
2. Also if possible please tell me how can i use k-fold cross validation for random forest. in R.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

User: [arun](#)

[Answer](#) by [david](#)

Relative to other models, Random Forests are less likely to overfit but it is still something that you want to make an explicit effort to avoid. Tuning model parameters is definitely one element of avoiding overfitting but it isn't the only one. In fact I would say that your training features are more likely to lead to overfitting than model parameters, especially with a Random Forests. So I think the key is really having a reliable method to evaluate your model to check for overfitting more than anything else, which brings us to your second question.

As alluded to above, running cross validation will allow to you avoid overfitting. Choosing your best model based on CV results will lead to a model that hasn't overfit, which isn't necessarily the case for something like out of the bag error. The easiest way to run CV in R is with the caret package. A simple example is below:

[Skip code block](#)

```
> library(caret)
>
> data(iris)
>
> tr <- trainControl(method = "cv", number = 5)
>
> train(Species ~ ., data=iris, method="rf", trControl= tr)
Random Forest

150 samples
  4 predictor
```

```

3 classes: 'setosa', 'versicolor', 'virginica'

No pre-processing
Resampling: Cross-Validated (5 fold)

Summary of sample sizes: 120, 120, 120, 120, 120

Resampling results across tuning parameters:

  mtry  Accuracy   Kappa   Accuracy SD   Kappa SD
  2      0.96       0.94    0.04346135   0.06519202
  3      0.96       0.94    0.04346135   0.06519202
  4      0.96       0.94    0.04346135   0.06519202

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.

```

[Answer](#) by [0xf](#)

Here is a nice link on that on stackexchange

<http://stats.stackexchange.com/questions/111968/random-forest-how-to-handle-overfitting>, however my general experience is the more depth the model has the more it tends to overfit.

[Answer](#) by [moriara](#)

@xof6 is correct in the sense that the more depth the model has the more it tends to overfit, but I wanted to add some more parameters that might be useful to you. I do not know which package you are using with R and I am not familiar with R at all, but I think there must be counterparts of these parameters implemented there.

Number of trees - The bigger this number, the less likely the forest is to overfit. This means that as each decision tree is learning some aspect of the training data, you are getting more options to choose from, so to speak. Number of features - This number constitutes how many features each individual tree learns. As this number grows, the trees get more and more complicated, hence they are learning patterns that might not be there in the test data. It will take some experimenting to find the right value, but such is machine learning. Experiment with the general depth as well, as we mentioned!

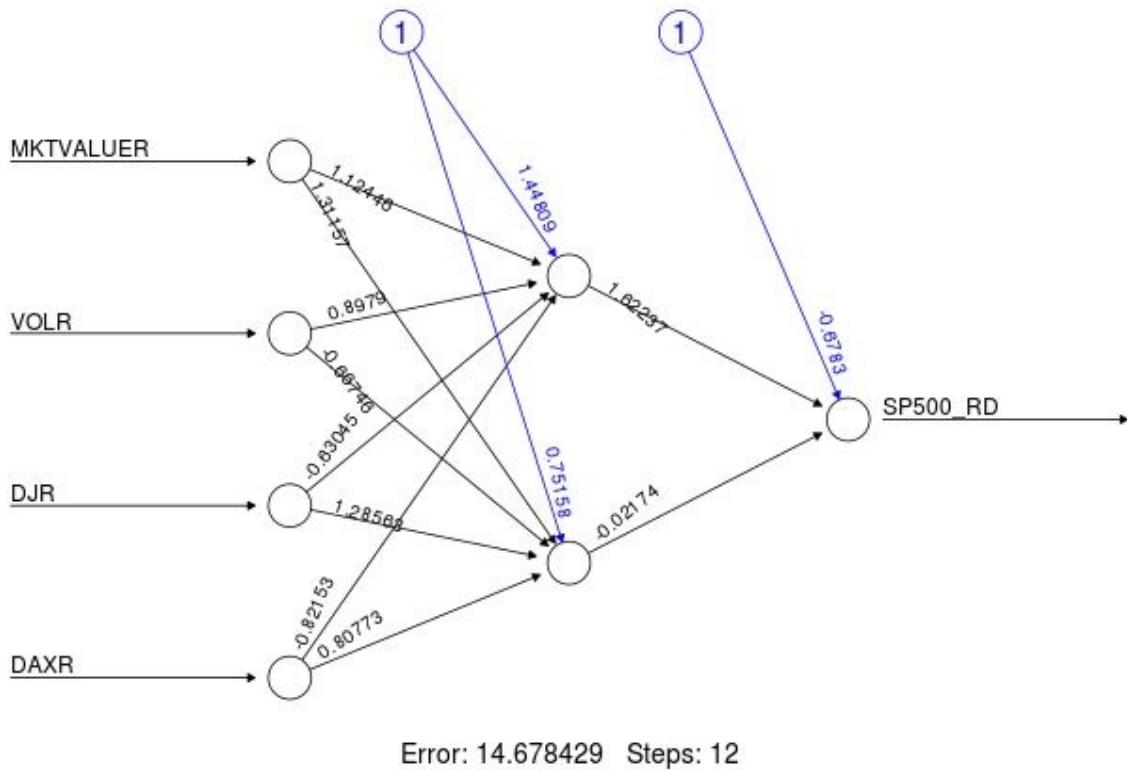
Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

[Q: R - Interpreting neural networks plot](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

I know there are similar question on stats.SE, but I didn't find one that fulfills my request; please, before mark the question as a duplicate, ping me in the comment.

I run a neural network based on `neuralnet` to forecast SP500 index time series and I want to understand how I can interpret the plot posted below:



Particularly, I'm interested to understand what is the interpretation of the hidden layer weight and the input weight; could someone explain me how to interpret that number, please?

Any hint will be appreciated.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

User: [quantopic](#)

[Answer](#) by [cdeterminan](#)

As David states in the comments if you want to interpret a model you likely want to explore something besides neural nets. That said if you want to intuitively understand the network plot it is best to think of it with respect to images (something neural networks are very good at).

1. The left-most nodes (i.e. input nodes) are your raw data variables.
2. The arrows in black (and associated numbers) are the **weights** which you can think of as **how much that variable contributes to the next node**. The blue lines are the bias weights. You can find the purpose of these weights in the excellent answer [here](#) .
3. The middle nodes (i.e. anything between the input and output nodes) are your hidden nodes. This is where the image analogy helps. **Each of these nodes constitute a component that the network is learning to recognize.** For example a nose, mouth, or eye. This is not easily determined and is far more abstract when you are dealing with non-image data.
4. The far-right (output node(s)) node is the final output of your neural network.

Note that this all is omitting the activation function that would be applied at each layer of the network as well.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

Q: What features from sound waves to use for an AI song composer?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

I am planning on making an AI song composer that would take in a bunch of songs of one instrument, extract musical notes (like ABCDEFG) and certain features from the sound wave, preform machine learning (most likely through recurrent neural networks), and output a sequence of ABCDEFG notes (aka generate its own songs / music).

I think that this would be an unsupervised learning problem, but I am not really sure.

I figured that I would use recurrent neural networks, but I have a few questions on how to approach this:

- What features from the sound wave I should extract so that the output music is melodious?
- Is it possible, with recurrent neural networks, to output a vector of sequenced musical notes (ABCDEF)?
- Any smart way I can feed in the features of the soundwaves as well as sequence of musical notes?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

User: [user3377126](#) 

[Answer](#)  by [chad-befus](#) 

First off, ignore the haters. I started working on ML in Music a long time ago and got several degrees using that work. When I started I was asking people the same kind of questions you are. It is a fascinating field and there is always room for someone new. We all have to start somewhere.

The areas of study you are inquiring about are Music Information Retrieval ([Wiki Link](#) ) and Computer Music ([Wiki Link](#) ). You have made a good choice in narrowing your problem to a single instrument (monophonic music) as polyphonic music increases the difficulty greatly.

You're trying to solve two problems really:

- 1) Automatic Transcription of Monophonic Music ([More Readings](#) ) which is the problem of extracting the notes from a single instrument musical piece.
- 2) Algorithmic Composition ([More Readings](#) ) which is the problem of generating new music using a corpus of transcribed music.

To answer your questions directly:

I think that this would be an unsupervised learning problem, but I am not really sure.

Since there are two learning problems here there are two answers. For the Automatic Transcription you will probably want to follow a supervised learning approach, where your classification are the notes you are trying to extract. For the Algorithmic Composition problem it can actually go either way. Some reading in both areas will clear this up a lot.

What features from the sound wave I should extract so that the output music is melodious?

There are a lot of features used commonly in MIR. @abhnj listed MFCC's in his answer but there are a lot more. Feature analysis in MIR takes place in several domains and there are features for each. Some Domains are:

1. The Frequency Domain (these are the values we hear played through a speaker)
2. The Spectral Domain (This domain is calculated via the Fourier function ([Read about the Fast Fourier Transform](#)) and can be transformed using several functions (Magnitude, Power, Log Magnitude, Log Power))
3. The Peak Domain (A domain of amplitude and spectral peaks over the spectral domain)
4. The Harmonic Domain

One of the first problems you will face is how to segment or “cut up” your music signal so that you can extract features. This is the problem of Segmentation ([Some Readings](#)) which is complex in itself. Once you have cut your sound source up you can apply various functions to your segments before extracting features from them. Some of these functions (called window functions) are the: Rectangular, Hamming, Hann, Bartlett, Triangular, Bartlett_hann, Blackman, and Blackman_harris.

Once you have your segments cut from your domain you can then extract features to represent those segments. Some of these will depend on the domain you selected. A few example of features are: Your normal statistical features (Mean, Variance, Skewness, etc.), ZCR, RMS, Spectral Centroid, Spectral Irregularity, Spectral Flatness, Spectral Tonality, Spectral Crest, Spectral Slope, Spectral Rolloff, Spectral Loudness, Spectral Pitch, Harmonic Odd Even Ratio, MFCC's and Bark Scale. There are many more but these are some good basics.

Is it possible, with recurrent neural networks, to output a vector of sequenced musical notes (ABCDEF)?

Yes it is. There have been several works to do this already. ([Here are several readings](#))

Any smart way I can feed in the features of the soundwaves as well as sequence of musical notes?

The standard method is to use the explanation I made above (Domain, Segment, Feature Extract) etc. To save yourself some work I highly recommend starting with a MIR framework such as MARSYAS ([Marsyas](#)). They will provide you with all the basics of feature extraction. There are many frameworks so just find one that uses a language you are comfortable in.

[Answer](#) by [abhnj](#)

I believe the question is, you want to learn from musical pieces and try to generate a tune from the trained instance. Lets see if I can set up a simple model to do this, and then you can extrapolate from there.

So, [MFCC](#) is a good feature when working with sound. You can use that to extract the features from lets say 1-2 second windows of your song. You now have a fingerprint for the audio file. Take a look at [Conditional Restricted Boltzmann Machines](#). They are Neural Networks which use multiple binary states to encode time series information. As you can see in the webpage, they trained on human-gait data and can now generate their own human gait. This is essentially what you want but for music files. So you can train CRBMs on the Audio MFCC vectors that you have.

After the training is done, to generate an audio file you can either “seed” the CRBM with a few seconds of some melody or just randomly initialize it. Then just allow the CRBM to go nuts and record whatever it produces. This is your new audio file. To produce another sample use a different seed.

This solves the question of how you can implement a “melody” generation scheme. There are of course variations. You can add other features to your vector apart from MFCC. You can also use other time series predictors like [LSTM](#) or Markov models.

All of this being said, the problem of generating music might be much more nuanced than it looks at first glance. Machine Learning algorithms just apply previously learned patterns in the data. How does that correspond to “creating” new music , is a philosophical question. If we analyze the aforementioned algorithm, essentially the CRBM will generate a next output based on the probability distribution that it has learnt. It would be very interesting to see what kind of output it generates when the said distribution is that of musical notes.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

[Q: Deriving Confidences from Distribution of Class Probabilities for a Prediction](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

I run into this problem from time to time and have always felt like there should be an obvious answer.

I have probabilities for potential classes (from some classifier). I will offer the prediction

of the class with the highest probability, however, I would also like to attach a confidence for that prediction.

Example: If I have Classes [C1, C2, C3, C4, C5] and my Probabilities are {C1: 50, C2: 12, C3: 13, C4: 12, C5: 13} my confidence in predicting C1 should be higher than if I had Probabilities {C1: 50, C2: 45, C3: 2, C4: 1, C5: 2}.

Reporting that I predict class C1 with 60% probability isn't the whole story. I should be able to derive a confidence from the distribution of probabilities as well. I am certain there is a known method for solving this but I do not know what it is.

EDIT: Taking this to the extreme for clarification: If I had a class C1 with 100% probability (and assuming the classifier had an accurate representation of each class) then I would be extremely confident that C1 was the correct classification. On the other hand if all 5 classes had almost equal probability (Say they are all roughly 20%) than I would be very uncertain claiming that any one was the correct classification. These two extreme cases are more obvious, the challenge is derive a confidence for intermediate examples like the one above.

Any suggestions or references would be of great help.

Thanks in advance.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [chad-befus](#) 

[Answer](#)  by [david](#) 

If I have Classes [C1, C2, C3, C4, C5] and my Probabilities are {C1: 50, C2: 12, C3: 13, C4: 12, C5: 13} my confidence in predicting C1 should be higher than if I had Probabilities {C1: 50, C2: 45, C3: 2, C4: 1, C5: 2}.

Assuming that those probabilities are accurate, this isn't true. In your second case you can be a lot more confident that the ground truth is one of C1 or C2, but in terms of absolute confidence about C1 the probability is the same across both examples. To illustrate this with a more clear example, if you had a 100 sided die that had 50 sides labeled with "C1" then the labels on the other 50 sides are irrelevant to the likelihood that you would roll a "C1".

Now with that said, your probabilities from your model are most certainly not perfect so there may be a way to use the intra-class correlations to improve them. Can you provide some more details about your specific problem and modeling workflow that you have used to get your probabilities?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

Q: Contributions of each feature in classification? 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

I have some features and I am using Weka to classify my instances

For example I have :

Number of adj number of adverb number of punctuation in my feature set

but I want to know the contribution of each feature in the feature set so what metrics or parameters are helpful to get the contribution of features?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [hani](#) 

[Answer](#)  by [franck-dernoncourt](#) 

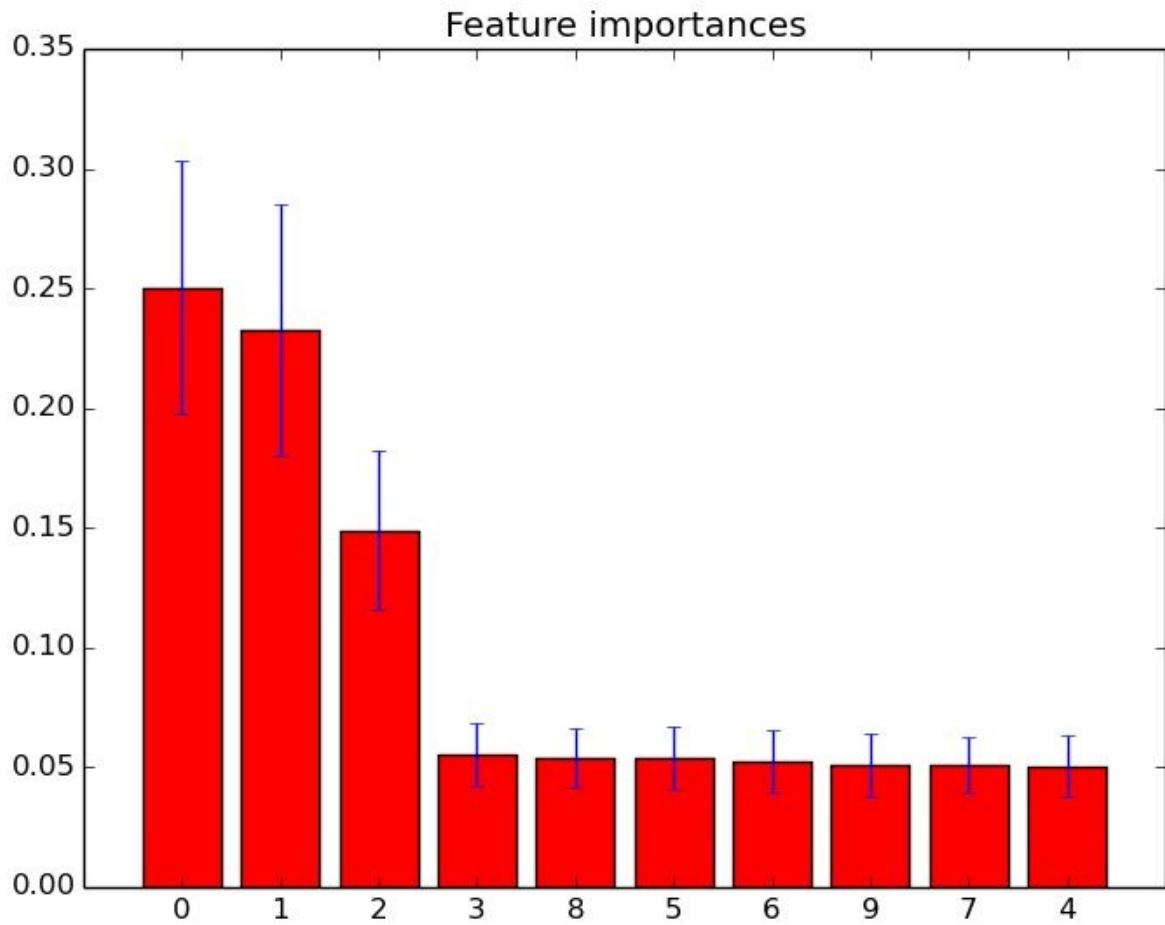
This is called feature ranking, which is closely related to [feature selection](#) .

- feature ranking = determining the importance of any individual feature
- feature selection = selecting a subset of relevant features for use in model construction.

So if you are able to rank features, you can use it to select features, and if you can select a subset of useful features, you've done at least a partial ranking by removing the useless ones.

This [Wikipedia page](#)  and this [Quora post](#)  should give some ideas. The distinction filter methods vs. wrapper based methods vs. embedded methods is the most common one.

One straightforward approximate way is to use [feature importance with forests of trees](#) :



Other common ways:

- [recursive feature elimination](#).
- stepwise regression (or [LARS Lasso](#)).

If you use scikit-learn, check out [module-sklearn.feature_selection](#). I'd guess Weka has some similar functions.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: Books on Reinforcement Learning](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I have been trying to understand reinforcement learning for quite sometime, but somehow I am not able to visualize how to write a program for reinforcement learning to solve a grid world problem. Can you suggest me some text books which would help me build a clear conception of Reinforcement Learning?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [rishika](#)

[Answer](#) by [mtk99](#)

Here you have some good references on Reinforcement Learning:

Classic

Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge, Mass: A Bradford Book; 1998. 322 p.

The draft for the second edition is available for free:

<https://webdocs.cs.ualberta.ca/~sutton/book/the-book.html>

Russell/Norvig Chapter 21:

Russell SJ, Norvig P, Davis E. Artificial intelligence: a modern approach. Upper Saddle River, NJ: Prentice Hall; 2010.

More technical

Szepesvári C. Algorithms for reinforcement learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2010;4(1):1–103.

<http://www.ualberta.ca/~szepesva/RLBook.html>

Bertsekas DP. Dynamic Programming and Optimal Control. 4th edition. Belmont, Mass.: Athena Scientific; 2007. 1270 p. Chapter 6, vol 2 is available for free:

<http://web.mit.edu/dimitrib/www/dpchapter.pdf>

For more recent developments

Wiering M, van Otterlo M, editors. Reinforcement Learning. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 Available from: <http://link.springer.com/10.1007/978-3-642-27645-3>

Kochenderfer MJ, Amato C, Chowdhary G, How JP, Reynolds HJD, Thornton JR, et al. Decision Making Under Uncertainty: Theory and Application. 1 edition. Cambridge, Massachusetts: The MIT Press; 2015. 352 p.

Multi-agent reinforcement learning

Buşoni L, Babuška R, Schutter BD. Multi-agent Reinforcement Learning: An Overview. In: Srinivasan D, Jain LC, editors. Innovations in Multi-Agent Systems and Applications - 1 . Springer Berlin Heidelberg; 2010 p. 183–221. Available from: http://link.springer.com/chapter/10.1007/978-3-642-14435-6_7

Schwartz HM. Multi-agent machine learning : a reinforcement approach. Hoboken, New Jersey: Wiley; 2014.

Videos / Courses

I would also suggest David Silver course in YouTube: https://www.youtube.com/playlist?list=PL5X3mDkKaJrL42i_jhE4N-p6E2Ol62Ofa

[Answer](#) by [sreejithc321](#)

There is a free online course on Reinforcement Learning by Udacity. Check : [Machine Learning: Reinforcement Learning](#)

[Answer](#) by [martin-černý](#)

I really enjoyed [Reinforcement Learning: An introduction](#) by Richard Sutton. It provides a very nice unifying view on RL, although it does not mention the newest approaches (it's from 1998).

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: When to use Random Forest over SVM and vice versa?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

When would one use Random Forest over SVM and vice versa?

I understand that cross-validation and model comparison is an important aspect of choosing a model, but here I would like to learn more about rules of thumb and heuristics of the two methods.

Can someone please explain the subtleties, strengths, and weaknesses of the classifiers as well as problems, which are best suited to each of them?

Thanks for your help!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [rohit](#)

[Answer](#) by [lanenok](#)

I would say, the choice depends very much on what data you have and what is your purpose. A few “rules of thumb”.

Random Forest is intrinsically suited for multiclass problems, while SVM is intrinsically two-class. For multiclass problem you will need to reduce it into multiple binary classification problems.

Random Forest works well with a mixture of numerical and categorical features. When features are on the various scales, it is also fine. Roughly speaking, with Random Forest you can use data as they are. SVM maximizes the “margin” and thus relies on the concept of “distance” between different points. It is up to you to decide if “distance” is meaningful. As a consequence, one-hot encoding for categorical features is a must-do. Further, min-max or other scaling is highly recommended at preprocessing step.

If you have data with n points and m features, an intermediate step in SVM is constructing an $n \times n$ matrix (think about memory requirements for storage) by calculating n^2 dot products (computational complexity). Therefore, as a rule of thumb, SVM is hardly scalable beyond 10^5 points. Large number of features (homogeneous features with meaningful distance, pixel of image would be a perfect example) is generally not a problem.

For a classification problem Random Forest gives you probability of belonging to class. SVM gives you distance to the boundary, you still need to convert it to probability

somewhat if you need probability.

For those problems, where SVM applies, it generally performs better than Random Forest. SVM gives you “support vectors”, that is points in each class closest to the boundary between classes. They may be of interest by themselves for interpretation.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

Q: Connection between Regularization and Gradient Descent

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#))

I would like to understand regularization/shrinkage in the light of MLE/Gradient Descent. I know both concepts but I do not know/understand whether both are used to determine coefficients of a linear model. If so, what are the steps followed?

To further elaborate, regularization is used to reduce variance which is accomplished through penalizing coefficients of a linear model. The tuning parameter, lambda, is determined through cross-validation. Once, lambda is determined the coefficients are automatically determined, right? Hence, why do we need to minimize (RSS + regularization term) to find coefficients? Are the steps the following:

1. Find lambda through cross-validation
2. Minimize (RSS + regularization) through MLE or GD
3. Find coefficients
4. Penalize coefficients to decrease variance
5. We are left with a small subset of coefficients

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [rohit](#) 

[Answer](#)  by [nbartley](#) 

The fitting procedure is the one that actually finds the coefficients of the model. The regularization term is used to indirectly find the coefficients by penalizing big coefficients *during the fitting procedure*. A simple (albeit somewhat biased/naive) example might help illustrate this difference between regularization and gradient descent:

```
x, y <- read input data
for different values of lambda L
  for each fold of cross-validation using X,y,L
    theta <- minimize (RSS + regularization using L) via MLE/GD
    score <- calculate performance of model using theta on the validation set
  if average score across folds for L is better than the current best average score
    L_best <- L
```

As you can see, the fitting procedure (MLE or GD in our case) finds the best coefficients given the specific value of lambda.

As a side note, I would look at this answer [here](#)  about tuning the regularization parameter, because it tends a little bit murky in terms of bias.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#))

Q: How scientists come up with the correct Hidden Markov Model to use?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I understand how a Hidden Markov Model is used in genomic sequences, such as finding a gene. But I don't understand how they came up with a particular Markov model. I mean, how many states should the model have? How many possible transitions? Should the model have a loop?

How would they know that their model is optimal?

Do they imagine, say 10 different models, benchmark those 10 models and publish the best one?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [student-t](#)

[Answer](#) by [matthew-graves](#)

I'm familiar with three main approaches:

1. A priori. You might know that there are four base pairs to pick from, and so allow the HMM to have four states. Or you might know that English has 44 phonemes, and so have 44 states for the hidden phoneme layer in a voice recognition model.
2. Estimation. The number of states can often be estimated beforehand, perhaps by simple clustering on the observed features of the HMM. If the HMM transition matrix is triangular (which is often the case in failure prediction), the number of states determines the shape of the distribution of total time from the start state to the end state.
3. Optimization. Like you suggest, either many models are created and fit and the best model selected. One could also adapt the methodology that learns the HMM to allow the model to add or discard states as needed.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Sentiment Analysis Tutorial](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

I am trying to understand sentiment analysis and how to apply it using any language (R, Python etc). I would like to know if there is a good place on internet for tutorial that I can follow. I googled, but I wasn't very much satisfied because they were not tutorials but more of theory. I want theory and practical examples.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [kurioz7](#) 

[Answer](#)  by [dawny33](#) 

The [NLTK book](#)  is by far the best tutorial on basic NLP I have seen(in Python).

The Coursera [course](#)  on NLP is also fairly good. It takes off from the basics and takes the student to a novice level.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Predicting New Data with Naive Bayes](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

Say I had the following training set for a Naive Bayes algorithm.

Outlook	Person	Play Golf?
Sunny	Joe	Yes
Sunny	Mary	Yes
Raining	Joe	Yes
Raining	Mary	No
Raining	Harry	Yes

If try to predict whether Harry will play golf on a sunny day (which I have no data for). Would it be correct to exclude the person attribute and use the remaining outlook attribute to calculate the probability of this happening? Or could that potentially cause problems with a larger data set that I'm unaware of?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [ryan-king](#) 

[Answer](#)  by [rapaio](#) 

Among Naive Bayes assumptions the main one is that features are conditionally independent. For our problem we would have:

$$P(Play|Outlook, Person) \propto P(Play)P(Outlook|Play)P(Name|Play)$$

To address question *is Harry going to play on a sunny day?*, you have to compute the following:

$$P(Yes|Sunny, Harry) = P(Yes)P(Sunny|Yes)P(Harry|Yes)$$

$$P(No|Sunny, Harry) = P(No)P(Sunny|No)P(Harry|No)$$

and choose the probability with bigger value.

That is what theory says. To address your question I will rephrase the main assumption of Naive Bayes. The assumptions that features are independent given the output means basically that the information given by joint distribution can be obtained by product of marginals. In plain English: assume you can find if Harry plays on sunny days if you only know how much Harry plays in general and how much anybody plays on sunny days. As you can see, *you simply would not use the fact that Harry plays on sunny days even if you would have had that record in your data*. Simply because Naive Bayes assumes there is no useful information in the interaction between the features, and this is the precise meaning of conditional independence, which Naive Bayes relies upon.

That said if you would want to use the interaction of features than you would have either to use a different model, or simply add a new *combined* feature like a concatenation of factors of names and outlook.

As a conclusion when you do not include names in your input features you will have a *general wisdom* classifier like *everybody plays no matter outlook*, since most of the instances have play=yes. If you include the name in your input variables you allow to alter that *general wisdom* with something specific to player. So your classifier *wisdom* would look like *players prefer in general to play, no matter outlook, but Marry like less to play less on Rainy*.

There is however a potential problem with Naive Bayes on your data set. This problem is related with the potential big number of levels for variable Name. In order to approximate the probability there is a general thing that happens: more data, better estimates. This probably would happen with variable Outlook, since there are two levels and adding more data would probably not increase number of levels. So the estimates for Outlook would be probably better with more data. However for name you will not have the same situation. Adding more instances would be possible perhaps only by adding more names. Which means that on average the number of instances for each name would be relatively stable. And if you would have a single instance, like it is the case for Harry, you do not have enough data to estimate $P(Harry|No)$.

As it happens this problem can be alleviated using smoothing. Perhaps Laplace smoothing (or a more general for like Lindstone) is very helpful. The reason is that estimates based on maximum likelihood have big problems with cases like that.

I hope it answers at least partially your question.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

[Q: Bagging vs Dropout in Deep Neural Networks](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Bagging is the generation of multiple predictors that works as ensemble as a single predictor. Dropout is a technique that teach to a neural networks to average all possible subnetworks. Looking at the most important Kaggle's competitions seem that these two techniques are used together very often. I can't see any theoretical difference besides the actual implementation. Who can explain me why we should use both of them in any real application? and why performance improve when we use both of them?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [emanuele](#) 

[Answer](#)  by [jamesmf](#) 

Bagging and dropout do not achieve quite the same thing, though both are types of model averaging.

[Bagging](#)  is an operation across your entire dataset which trains models on a subset of the training data. Thus some training examples are not shown to a given model.

[Dropout](#) , by contrast, is applied to features within each training example. It is true that the result is functionally equivalent to training exponentially many networks (with shared weights!) and then equally weighting their outputs. But dropout works on the feature space, causing certain features to be unavailable to the network, not full examples. Because each neuron cannot completely rely on one input, representations in these networks tend to be more distributed and the network is less likely to overfit.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[**Q: When do I have to use aucPR instead of auROC? \(and vice versa\)**](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [cross-validation](#) ([Next Q](#))

I'm wondering if sometimes, to validate a model, it's not better to use aucPR instead of auROC? Do these cases only depend on the "domain & business understanding" ?

Especially, I'm thinking about the "unbalanced class problem" where, it seems more logical to use the aucPR because recall and precision are well-used metrics for this problem.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [cross-validation](#) ([Next Q](#))

User: [jmvl1t](#) 

[Answer](#)  by [an6u5](#) 

Yes, you are correct that the dominant difference between the area under the curve of a receiver operator characteristic curve ([ROC-AUC](#) ) and the area under the curve of a Precision-Recall curve ([PR-AUC](#) ) lies in its tractability for unbalanced classes. They are very similar and have been shown to contain essentially the same information,

however PR curves are slightly more finicky, but a well drawn curve gives a more complete picture. The issue with PR-AUC is that its difficult to interpolate between points in the PR curve and thus numerical integration to achieve an area under the curve becomes more difficult.

[Check out this discussion of the differences and similarities.](#) 

Quoting Davis' 2006 abstract:

Receiver Operator Characteristic (ROC) curves are commonly used to present results for binary decision problems in machine learning. However, when dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of an algorithm's performance. We show that a deep connection exists between ROC space and PR space, such that a curve dominates in ROC space if and only if it dominates in PR space. A corollary is the notion of an achievable PR curve, which has properties much like the convex hull in ROC space; we show an efficient algorithm for computing this curve. Finally, we also note differences in the two types of curves are significant for algorithm design. For example, in PR space it is incorrect to linearly interpolate between points. Furthermore, algorithms that optimize the area under the ROC curve are not guaranteed to optimize the area under the PR curve.

[This was also discussed on Kaggle recently.](#) 

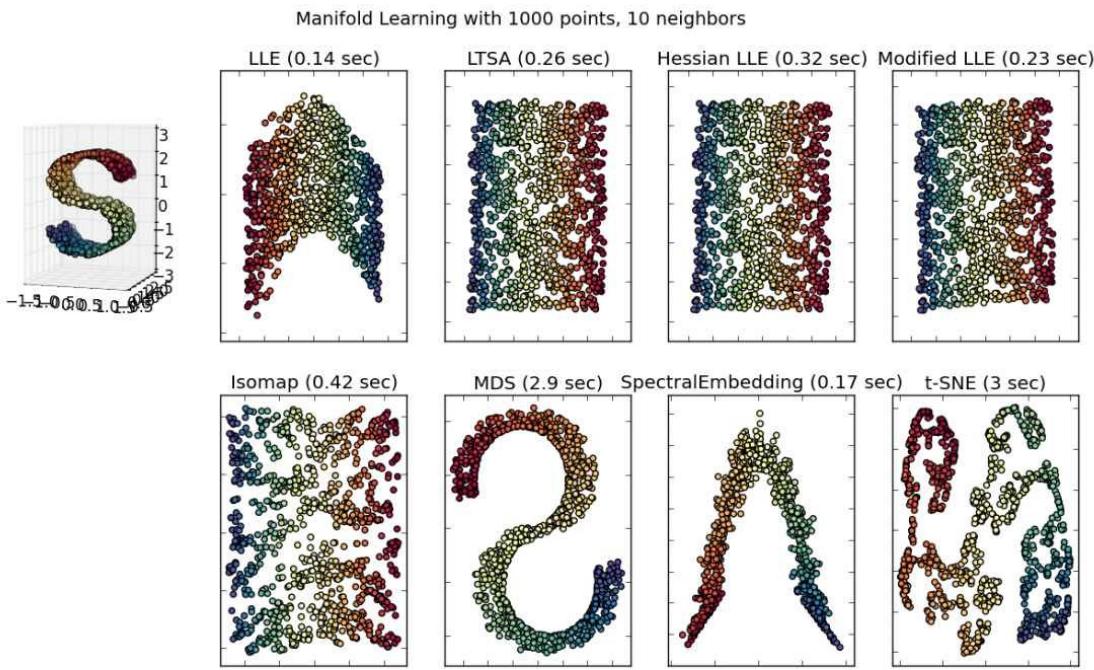
[There is also some useful discussion on Cross Validated.](#) 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [cross-validation](#) ([Next Q](#))

Q: Purpose of visualizing high dimensional data? 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

There are many techniques for visualizing high dimension datasets, such as T-SNE, isomap, PCA, supervised PCA, etc. And we go through the motions of projecting the data down to a 2D or 3D space, so we have a “pretty pictures”. Some of these embedding (manifold learning) methods are described [here](#) .



But is this “pretty picture” actually meaningful? What possible insights can someone grab by trying to visualize this embedded space?

I ask because the projection down to this embedded space is usually meaningless. For example, if you project your data down to principal components generated by PCA, those principal components (eigenvectors) don't correspond to features in the dataset; they're their own feature space.

Similarly, t-SNE projects your data down to a space, where items are near each other if they minimize some KL divergence. This isn't the original feature space anymore. (Correct me if I'm wrong, but I don't even think there is a large effort by the ML community to use t-SNE to aid classification; that's a different problem than data visualization though.)

I'm just very largely confused why people make such a big deal about some of these visualizations.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

User: [hlin117](#)

[Answer](#) by [kasra-manshaei](#)

First of all your explanation about the methods are right. The point is that Embedding algorithms are not to only visualize but basically reducing the dimensionality to cope with two main problems in Statistical Data Analysis, namely **Curse of Dimensionality** and **Low-Sample Size Problem** so that they are not supposed to depict physically understood features and they are not only *meaningful* but also necessary for data analysis!

Actually the visualization is almost the last usage of embedding methods. Projecting high-dimensional data into a lower-dimension space helps to preserve the actual pair-wise distances (mainly Euclidean one) which get distorted in the high dimensions or capturing the most information embedded in the variance of different features.

[Answer](#) by [hariz-naam](#)

Excellent question. In chapter 4 of “Illuminating the Path, The Research and Development Agenda for Visual Analytics” by James J. Thomas and Kristin A. Cook is a discussion on data representations and data transformations. In my research I have approached this question in the context of PCA and factor analysis. My brief answer is that the visualizations are useful if one has the data transformation to move from the visualization space to the original data space. This would additionally be conducted within a visual analytics framework.

[Answer](#) by [marmite-bomber](#)

Based on the statements and the discussions, I think there is an important point to distinct. A transformation to a lower dimensional space may *reduce* the information, which is something different from making the information *meaningless*. Let me use a following analogy:

Observing (2D) pictures of our world (3D) is a usual practice. A visualization method provides only different “glasses” to see a high dimensional space.

A good thing to “trust” a visualization method is to understand the internals. My favourite example is the [MDS](#). It is easy possible to implement this method at your own using some optimization tool (e.g. R *optim*). So you can *see* how the method works, you may *measure the error* of the result etc.

At the end you get a picture preserving the similarity of the original data with some degree of precision. Not more, but not less.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

[Q: Pylearn2 vs TensorFlow](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

I am about to dive into a long NN research project and wanted a push in the direction of Pylearn2 or TensorFlow? As of Dec 2015 has the community started to lean one direction or another?

This [link](#) has given me concern about getting tied to TensorFlow.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

User: [user3155053](#)

[Answer](#) by [franck-dernoncourt](#)

You might want to take into consideration that [Pylearn2 has no more developer](#), and now [points to other Theano-based libraries](#):

There are other machine learning frameworks built on top of Theano that could interest you, such as: Blocks, Keras and Lasagne.

As Dawny33 says, TensorFlow is just getting started, but it is interesting to note that the number of questions on TensorFlow (244) on Stack Overflow already surpasses Torch (166) and will probably catch up with Theano (672) in 2016.

[Answer](#) by [dawny33](#)

As far as I know, the existing (almost all) libraries in Python can handle very complex models of Neural Networks.

TensorFlow is however not polished as of now. It still has a long way to grow before getting accepted as a mainstream library for ML.

So, going ahead with the existing libraries like PyLearn/Keras/Torch, etc makes sense as of now (also as they have wide and dedicated communities already), as you need to concentrate on research rather than worrying on bugs and technical problems of a library.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: Predicting app usage on mobile phone](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

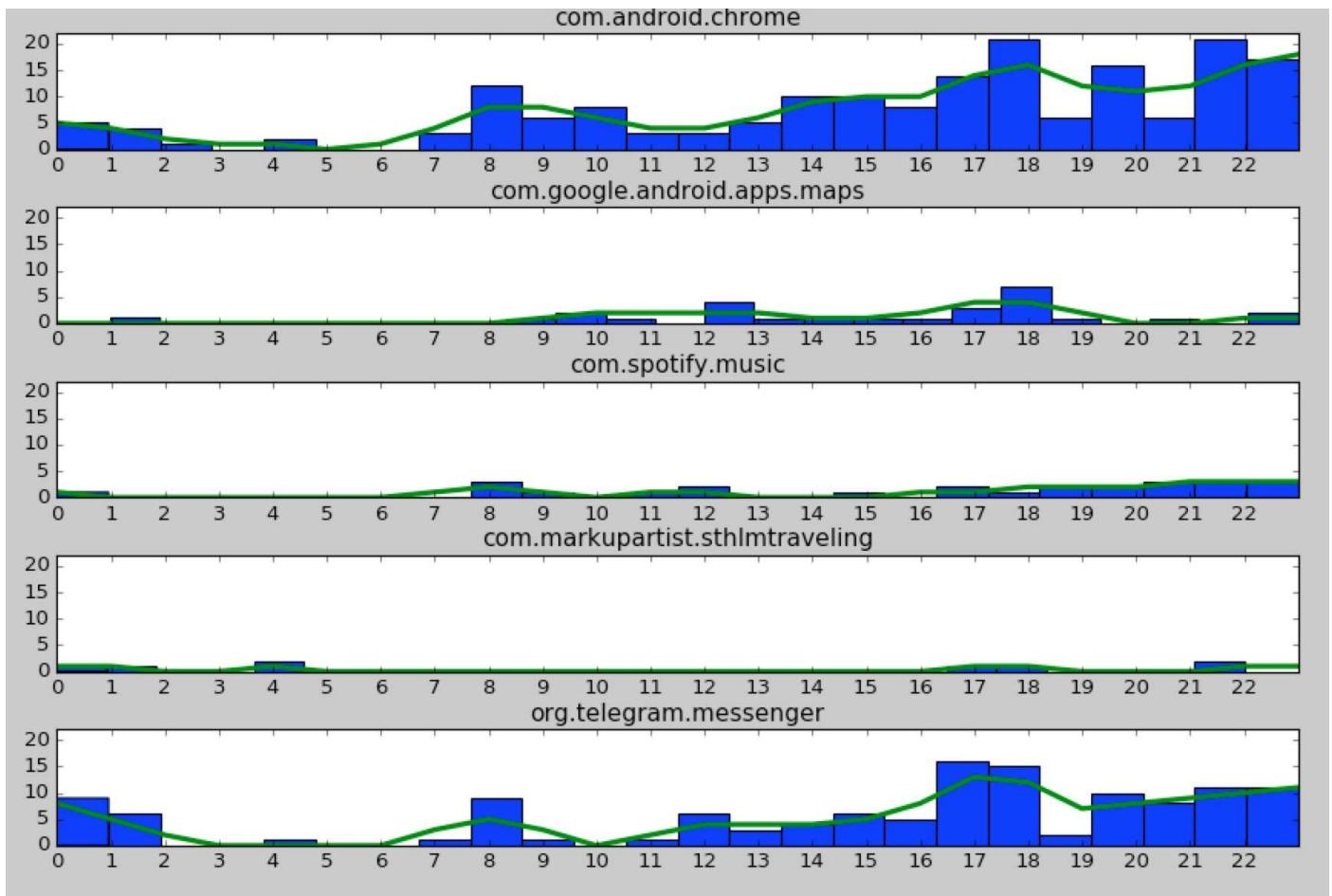
I'm currently building an app that strives to predict how the users uses different apps and give the user a suggestion based on which apps it thinks the user will currently use (a ranked list based on the user's current conditions). I've been collecting some data over the past week now, and I'm not really sure which approach to take. I've been thinking about using multiple seasonalities (correct me if I'm using the wrong terminology) such as time of day, day of week, week of month, month and quarter. I also want to use location, and other sensor data (such as the user state "walking" or "sitting" later on).

I've summarised the usage over the last week on an hourly level for some apps. The bars represent each time the app was opened during that time, and the green line is a weighted moving average which has a weight of 0.5 on its closest neighbours.

Now I see several challenges in front of me and would greatly appreciate some input from others, or some good resources to find further information.

- Do you think my model is a good one for this problem
- How do I account for ageing data?
- How do add up the different seasonalities/states/location? Multiply them?
- Does it make sense smoothening the curve as I do?

Here some data for the last week on an hourly basis:



Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

User: [jonepatr](#)

[Answer](#) by [shawn-mehan](#)

So you have collected data that shows which app is being used at any time, binned into hours in the day. And you have several apps. You mention other dimensions, like user state when using the app (walking, not-walking), (active, not-active - it appears to me that you are not collecting much usage 2-6. is that because the usage is from a self-directed ping from an app while the user is truly away?), location (is this going to be all possible values, or are you going to use something like the fact that this location has been seen often before?). Another interesting relationship could be pairing apps, i.e., mining for a relationship between App A being used after using App B or before App B.

Regardless, then you will definitely have many different dimensions upon which to measure the usage characteristics for any particular usage measurement, and so you are definitely going to have a multiple dimensional problem. You might try to visualize this as an N-space problem, with an axis of measurement for each of your characteristics. Each of your previous measurements represents vectors and you are producing a new vector with your next measurement.

From this, you want to predict future behavior based on measuring the input characteristics from your usage space. You could go for something that classifies as nearest neighbor, and you probably want to do this for your first stab at the problem. You might end up wanting to make the predictive model more sophisticated by adding

probabilities to the classifier and acting on that. This means getting estimates of class membership probability rather than just simple classifications. But I would build the whole thing incrementally. Start simple and add complexity as you require it. The increased complexity will also have effects on performance, so why not baseline with something.

For the aging of data, are you wanting to reduce the predictive power of characteristics that are too long in the tooth? If so, be explicit with yourself about what that means, quantitatively. Do I trust the usage data from last month less than yesterday's data? Perhaps so, but then why? is my usage different because I am different or because last month was special compared to yesterday, or vice-versa? Again, you might benefit from ignoring this at first, but then trying to **search** for "seasonal" or periodicity characteristics from the data. Once you determine if/how it changes, you can weight that contribution compared to your immediate usage in different ways. Perhaps you want to amplify the contribution of a similar period (same time of day && same location && same previous app usage). Perhaps you want to provide an exponential dampening on historical data because the usage is always adapting and changing, and recent usage seems to be a **much** better predictor than 3xcurrent.

For all of this, the proper data science perspective is to let the data lead you.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [predictive-modeling](#) ([Prev Q](#)) ([Next Q](#))

[Q: \(Why\) do activation functions have to be monotonic?](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

I am currently preparing for an exam in neural networks. In several protocols from former exams I read that activation functions of neurons (in multilayer perceptrons) have to be monotonic.

I understand that activation functions should be differentiable, have a derivative which is not 0 on most points and non-linear. I do not understand why being monotonic is important / helpful.

I know the following activation functions and that they are monotonic:

- ReLU
- Sigmoid
- Tanh
- Softmax
- Softplus
- (Identity)

However, I still can't see any reason why for example $\varphi(x) = x^2$.

(Why) do activation functions have to be monotonic?

(Related side question: Is there any reason why the logarithm / exponential function is not

used as an activation function?)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [martin-thoma](#) 

[Answer](#)  by [david-dao](#) 

The monotonicity criterion helps the neural network to converge easier into a more accurate classifier. See this [stackexchange answer](#)  and [wikipedia article](#)  for further details and reasons.

However, the monotonicity criterion is not mandatory for an activation function - It is also possible to train neural nets with non-monotonic activation functions. It just gets harder to optimize the neural network. See [Yoshua Bengio's answer](#) .

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Q: How do AI's learn to act when the problem space is too big

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

I learn best through experimentation and example. I'm learning about neural networks and have (what I think) is a pretty good understanding of classification and regression and also supervised and unsupervised learning, but I've stumbled upon something I can't quiet figure out;

If I wanted to train an AI to play a complicated game; I'm thinking something like a RTS (eg. Age of Empires, Empire Earth etc.). In these types of games there is typically a number of entities controlled by the player (units, buildings) each with different capabilities. It seem like the problem of that the AI does would be classification (eg. choose that unit, and that action), however since the number of units is a variable how does one handle a classification problem in this way?

The only thing I can think of is multiple networks that do different stages (one for overall strategy, one for controlling this type of unit, one for that type of building etc.); but this seems like I'm making the problem to complicated.

Are there any good example of machine learning/neural networks learning complex games (not specifically RTS, but more complicated than [Mario](#) )?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [fraserofsmeg](#) 

[Answer](#)  by [hoap-humanoid](#) 

That is a good question and many scientists around the world are asking the same. Well, first a game like Age of Empires is not considered to have a really big solution space, there are not so many things you can do. It's the same in games like Mario Bros. The problem of learning in easy games like Atari games was solved by the guys of DeepMind

(here the [paper](#)), that was acquired by Google. They used an implementation of Reinforcement Learning with Deep Learning.

Going back to your question. A really big problem is how to imitate the amount of decisions a human being takes every day. Wake up, have breakfast, take a shower, leave your house... All these actions need a really high level of intelligence and many actions to develop.

There are many people working on this problem, I'm one of them. I don't know the solution but I can tell you in which way I'm looking. I follow the theories of Marvin Minsky, he is one of the fathers of AI. This book, the Emotion Machine, tells a very good view of the problem. He suggested that the way to create a machine that imitates the human behavior is not by constructing a unified compact theory of artificial intelligence. On the contrary, he argues that our brain contains resources that compete between each other to satisfy different goals at the same moment. They called this **Ways to Think**.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Q: Where exactly does ≥ 1 come from in SVMs optimization problem constraint?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

I've understood that SVMs are binary, linear classifiers (without the kernel trick). They have training data (x_i, y_i) where x_i is a vector and $y_i \in \{-1, 1\}$ is the class. As they are binary, linear classifiers the task is to find a hyperplane which separates the data points with the label -1 from the data points with the label $+1$.

Assume for now, that the data points are linearly separable and we don't need slack variables.

Now I've read that the training problem is now the following optimization problem:

- $\min_{w,b} \frac{1}{2} \|w\|^2$
- s.t. $y_i(\langle w, x_i \rangle + b) \geq 1$

I think I got that minimizing $\|w\|^2$ means maximizing the margin (however, I don't understand why it is the square here. Would anything change if one would try to minimize $\|w\|$?).

I also understood that $y_i(\langle w, x_i \rangle + b) \geq 0$ means that the model has to be correct on the training data. However, there is a 1 and not a 0. Why?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [martin-thoma](#)

[Answer](#) by [hbaderts](#)

First problem: Minimizing $\|w\|$ or $\|w\|^2$:

It is correct that one wants to maximize the margin. This is actually done by maximizing $\frac{1}{2} \|w\|^2$. This would be the “correct” way of doing it, but it is rather inconvenient. Let’s first drop the 2, as it is just a constant. Now if $\frac{1}{2} \|w\|^2$ is maximal, $\|w\|$ will have to be as small as possible. We can thus find the identical solution by *minimizing* $\|w\|$.

$\|w\|$ can be calculated by $\sqrt{w^T w}$. As the square root is a monotonic function, any point x which maximizes $\sqrt{f(x)}$ will also maximize $f(x)$. To find this point x we thus don’t have to calculate the square root and can minimize $w^T w = \|w\|^2$.

Finally, as we often have to calculate derivatives, we multiply the whole expression by a factor $\frac{1}{2}$. This is done very often, because if we derive $\frac{d}{dx} x^2 = 2x$ and thus $\frac{d}{dx} \frac{1}{2} x^2 = x$. This is how we end up with the problem: minimize $\frac{1}{2} \|w\|^2$.

tl;dr: yes, minimizing $\|w\|$ instead of $\frac{1}{2} \|w\|^2$ would work.

Second problem: ≥ 0 or ≥ 1 :

As already stated in the question, $y_i (\langle w, x_i \rangle + b) \geq 0$ means that the point has to be on the correct side of the hyperplane. However this isn’t enough: we want the point to be at least as far away as the margin (then the point is a support vector), or even further away.

Remember the definition of the hyperplane,

$$H = \{x \mid \langle w, x \rangle + b = 0\} .$$

This description however is not unique: if we scale w and b by a constant c , then we get an equivalent description of this hyperplane. To make sure our optimization algorithm doesn’t just scale w and b by constant factors to get a higher margin, we define that the distance of a support vector from the hyperplane is always 1, i.e. the margin is $\frac{1}{\|w\|}$. A support vector is thus characterized by $y_i (\langle w, x_i \rangle + b) = 1$.

As already mentioned earlier, we want all points to be either a support vector, or even further away from the hyperplane. In training, we thus add the constraint $y_i (\langle w, x_i \rangle + b) \geq 1$, which ensures exactly that.

tl;dr: Training points don’t only need to be correct, they have to be on the margin or further away.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

Q: is neural networks an online algorithm by nature? 

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

I have been doing machine learning for a while, but bits and pieces come together even after some time of practicing.

In neural networks, you adjust the weights by doing one pass (forward pass), and then computing the partial derivatives for the weights (backward pass) after each training example - and subtracting those partial derivatives from the initial weights.

in turn, the calculation of the new weights is mathematically complex (you need to compute the partial derivative of the weights, for which you compute the error at every layer of the neural net - but the input layer).

Is that not by definition an online algorithm, where cost and new weights are calculated after each training example?

Thanks!

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [alejandro-simkievich](#) 

[Answer](#)  by [martin-thoma](#) 

There are three training modes for neural networks

- **stochastic gradient descent:** Adjust the weights after every single training example
- **batch training:** Adjust the weights after going through all data (an epoch)
- **mini-batch training:** Adjust the weights after going through a *mini-batch*. This is usually 128 training examples.

Most of the time, mini-batch training seems to be used.

So the answer is:

No, the neural network learning algorithm is not online algorithm.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Q: What kinds of learning problems are suitable for Support Vector Machines?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

What are the hallmarks or properties that indicate that a certain learning problem can be tackled using support vector machines?

In other words, what is it that, when you see a learning problem, makes you go “oh I should definitely use SVMs for this” rather than Neural networks or Decision trees or anything else?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [ragnar](#)

[Answer](#) by [hoap-humanoid](#)

SVM can be used for classification (distinguishing between several groups or classes) and regression (obtaining a mathematical model to predict something). They can be applied to both linear and non linear problems.

Until 2006 they were the best general purpose algorithm for machine learning. I was trying to find a paper that compared many implementations of the most known algorithms: svm, neural nets, trees, etc. I couldn't find it sorry (you will have to believe me, bad thing). In the paper the algorithm that got the best performance was svm, with the library libsvm.

In 2006 Hinton came up with deep learning and neural nets. He improved the current state of the art by at least 30%, which is a huge advancement. However deep learning only get good performance for huge training sets. If you have a small training set I would suggest to use svm.

Furthermore you can find here a useful infographic about [when to use different machine learning algorithms](#) by scikit-learn. However, to the best of my knowledge there is no agreement among the scientific community about if a problem has X, Y and Z features then it's better to use svm. I would suggest to try different methods. Also, please don't forget that svm or neural nets is just a method to compute a model. It is very important as well the features you use.

[Answer](#) by [pincopallino](#)

Let's assume that we are in a classification setting.

For svm feature engineering is cornerstone:

- the sets have to be linearly separable. Otherwise the data needs to be transformed (eg using Kernels). This is not done by the algo itself and might blow out the number of features.
- I would say that svm performance suffers as we increase the number of dimensions

faster than other methodologies (tree ensemble). This is due to the constrained optimization problem that backs svms. Sometimes feature reduction is feasible, sometimes not and this is when we can't really pave the way for an effective use of svm

- svm will likely struggle with a dataset where the number of features is much larger than the number of observations. This, again, can be understood by looking at the constrained optimization problem.
- categorical variables are not handled out of the box by the svm algorithm.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

Q: Is there any domain where Bayesian Networks outperform neural networks?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Neural networks get top results in Computer Vision tasks (see [MNIST](#) , [ILSVRC](#) , [Kaggle Galaxy Challenge](#) ). They seem to outperform every other approach in Computer Vision. But there are also other tasks:

- [Kaggle Molecular Activity Challenge](#) 
- Regression: [Kaggle Rain prediction](#) , also the [2nd place](#) 
- [Grasp and Lift 2nd](#)  also [third place](#)  - Identify hand motions from EEG recordings

I'm not too sure about ASR (automatic speech recognition) and machine translation, but I think I've also heard that (recurrent) neural networks (start to) outperform other approaches.

I am currently learning about Bayesian Networks and I wonder in which cases those models are usually applied. So my question is:

Is there any challenge / (Kaggle) competition, where the state of the art are Bayesian Networks or at least very similar models?

(Side note: I've also seen [decision trees](#) , [2](#) , [3](#) , [4](#) , [5](#) , [6](#) , [7](#)  win in several recent Kaggle challenges)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [martin-thoma](#) 

[Answer](#)  by [mlgeek-](#) 

One of the areas where Bayesian approaches are often used, is where one needs interpretability of the prediction system. You don't want to give doctors a Neural net and say that it's 95% accurate. You rather want to explain the assumptions your method makes, as well as the decision process the method uses.

Similar area is when you have a strong prior domain knowledge and want to use it in the system.

[Answer](#) by [bayer](#)

Bayesian networks and neural networks are not exclusive of each other. In fact, Bayesian networks are just another term for “directed graphical model”. They can be very useful in designing objective functions neural networks. Yann Lecun has pointed this out here: <https://plus.google.com/+YannLeCunPhD/posts/gWE7Jca3Zoq>.

One example.

The variational auto encoder and derivatives are directed graphical models of the form

$$p(x) = \int_z p(x|z)p(z)dz.$$

A neural networks is used to implemented $p(x|z)$ and an approximation to its inverse: $q(z|x) \approx p(z|x)$.

[Answer](#) by [dawny33](#)

Excellent answers already.

One domain which I can think of, and is working extensively in, is the **customer analytics** domain.

I have to understand and predict the moves and motives of the customers in order to inform and warn both the customer support, the marketing and also the growth teams.

So here, neural networks do a really good job in churn prediction, etc. But, I found and prefer the Bayesian networks style, and here are the reasons for preferring it:

1. Customers always have a pattern. They always have a *reason* to act. And that reason would be something which my team has done for them, or they have learnt themselves. So, everything has a prior here, and in fact that reason is very important as it fuels most of the decision taken by the customer.
2. Every move by the customer and the growth teams in the marketing/sales funnel is cause-effect. So, prior knowledge is vital when it comes to converting a prospective lead into a customer.

So, the concept of *prior* is very important when it comes to customer analytics, which makes the concept of Bayesian networks very important to this domain.

Suggested Learning:

[Bayesian Methods for Neural Networks](#)

[Bayesian networks in business analytics](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Q: Neural networks: which cost function to use?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

I am using [TensorFlow](#) for experiments mainly with neural networks. Although I have done quite some experiments (XOR-Problem, MNIST, some Regression stuff, ...) now, I struggle with choosing the “correct” cost function for specific problems because overall I could be considered a beginner.

Before coming to TensorFlow I coded some fully-connected MLPs and some recurrent networks on my own with [Python](#) and [NumPy](#) but mostly I had problems where a simple squared error and a simple gradient descent was sufficient.

However, since TensorFlow offers quite a lot of cost functions itself as well as building custom cost functions, I would like to know if there is some kind of tutorial maybe specifically for cost functions on neural networks? (I’ve already done like half of the official TensorFlow tutorials but they’re not really explaining **why** specific cost functions or learners are used for specific problems - at least not for beginners)

To give some examples:

```
cost = tf.reduce_mean(tf.nn.softmax_cross_entropy_with_logits(y_output, y_train))
```

I guess it applies the softmax function on both inputs so that the sum of one vector equals 1. But what exactly is cross entropy with logits? I thought it sums up the values and calculates the cross entropy...so some metric measurement?! Wouldn’t this be very much the same if I normalize the output, sum it up and take the squared error? Additionally, why is this used e.g. for MNIST (or even much harder problems)? When I want to classify like 10 or maybe even 1000 classes, doesn’t summing up the values completely destroy any information about *which* class actually was the output?

```
cost = tf.nn.l2_loss(vector)
```

What is this for? I thought l2 loss is pretty much the squared error but TensorFlow’s API tells that its input is just one tensor. Doesn’t get the idea at all?!

Besides I saw this for **cross entropy** pretty often:

```
cross_entropy = -tf.reduce_sum(y_train * tf.log(y_output))
```

...but why is this used? Isn’t the loss in cross entropy mathematically this:

```
-1/n * sum(y_train * log(y_output) + (1 - y_train) * log(1 - y_output))
```

Where is the $(1 - y_{train}) * \log(1 - y_{output})$ part in most TensorFlow examples? Isn’t it missing?

Answers: I know this question is quite open, but I do not expect to get like 10 pages with every single problem/cost function listed in detail. I just need a short summary about when to use which cost function (in general or in TensorFlow, doesn’t matter much to me) and some explanation about this topic. And/or some source(s) for beginners ;)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#)

User: [ascenator](#) 

[Answer](#)  by [winks](#) 

This answer is on the *general* side of cost functions, not related to TensorFlow, and will mostly address the “some explanation about this topic” part of your question.

In most examples/tutorial I followed, the cost function used was somewhat arbitrary. The point was more to introduce the reader to a specific method, not to the cost function specifically. It should not stop you to follow the tutorial to be familiar with the tools, but my answer should help you on how to choose the cost function for your own problems.

If you want answers regarding Cross-Entropy, Logit, L2 norms, or anything specific, I advise you to post multiple, more specific questions. This will increase the probability that someone with the specific knowledge will see your question.

Choosing the right cost function for achieving the desired result is a critical point of machine learning problems. The basic approach, if you do not know exactly what you want out of your method, is to use [Mean Square Error \(Wikipedia\)](#)  for regression problems and Percentage of error for classification problems. However, if you want *good* results out of your method, you need to *define good*, and thus define the adequate cost function. This comes from both domain knowledge (what is your data, what are you trying to achieve), and knowledge of the tools at your disposal.

I do not believe I can guide you through the costs functions already implemented in TensorFlow, as I have very little knowledge of the tool, but I can give you an example on how to write and assess different cost functions.

To illustrate the various differences between cost functions, let us use the example of the binary classification problem, where we want, for each sample x_n , the class $f(x_n) \in \{0, 1\}$.

Starting with **computational properties**; how two functions measuring the “same thing” could lead to different results. Take the following, simple cost function; the percentage of error. If you have N samples, $f(y_n)$ is the predicted class and y_n the true class, you want to minimize

$$\bullet \quad \frac{1}{N} \sum_n \begin{cases} 1 & \text{if } f(x_n) \neq y_n \\ 0 & \text{otherwise} \end{cases} = \sum_n y_n [1 - f(x_n)] + [1 - y_n]f(x_n) .$$

This cost function has the benefit of being easily interpretable. However, it is not smooth; if you have only two samples, the function “jumps” from 0, to 0.5, to 1. This will lead to inconsistencies if you try to use gradient descent on this function. One way to avoid it is to change the cost function to use probabilities of assignment; $p(y_n = 1|x_n)$. The function becomes

$$\bullet \quad \frac{1}{N} \sum_n y_n p(y_n = 0|x_n) + (1 - y_n)p(y_n = 1|x_n) .$$

This function is smoother, and will work better with a gradient descent approach. You will get a ‘finer’ model. However, it has other problem; if you have a sample that is ambiguous, let say that you do not have enough information to say anything better than $p(y_n = 1|x_n) = 0.5$. Then, using gradient descent on this cost function will lead to a model which increases this probability as much as possible, and thus, maybe, overfit.

Another problem of this function is that if $p(y_n = 1|x_n) = 1$ while $y_n = 0$, you are certain to be right, but you are wrong. In order to avoid this issue, you can take the log of the probability, $\log p(y_n|x_n)$. As $\log(0) = \infty$ and $\log(1) = 0$, the following function does not have the problem described in the previous paragraph:

- $\frac{1}{N} \sum_n y_n \log p(y_n = 0|x_n) + (1 - y_n) \log p(y_n = 1|x_n)$.

This should illustrate that in order to optimize the *same thing*, the percentage of error, different definitions might yield different results if they are easier to make sense of, computationally.

It is possible for cost functions A and B to measure the same concept, but A might lead your method to better results than B.

Now let see how different costs function can measure different concepts. In the context of information retrieval, as in google search (if we ignore ranking), we want the returned results to

- have high *precision*, not return irrelevant information
- have high *recall*, return as much relevant results as possible
- [Precision and Recall \(Wikipedia\)](#) 

Note that if your algorithm returns *everything*, it will return every relevant result possible, and thus have high recall, but have very poor precision. On the other hand, if it returns only *one* element, the one that it is the most certain is relevant, it will have high precision but low recall.

In order to judge such algorithms, the common cost function is the [F-score \(Wikipedia\)](#) . The common case is the F_1 -score, which gives equal weight to precision and recall, but the general case it the F_β -score, and you can tweak β to get

- Higher recall, if you use $\beta > 1$
- Higher precision, if you use $\beta < 1$.

In such scenario, **choosing the cost function is choosing what tradeoff your algorithm should do.**

Another example that is often brought up is the case of medical diagnosis, you can choose a cost function that punishes more false negatives or false positives depending on what is preferable:

- More healthy people being classified as sick (But then, we might treat healthy people, which is costly and might hurt them if they are actually not sick)

- More sick people being classified as healthy (But then, they might die without treatment)
-

In conclusion, defining the cost function is defining the goal of your algorithm. The algorithm defines how to get there.

Sidenote: Some cost functions have nice algorithm ways to get to their goals. For example, a nice way to the minimum of the [Hinge loss \(Wikipedia\)](#) exists, by solving the dual problem in [SVM \(Wikipedia\)](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

Q: What's a good machine learning algorithm for low frequency trading?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

I'm trying to train an algorithm to copy some of the top traders on various forex social trading sites. The problem is that the traders only trade around say 10 times per month so even if I only look at minute resolution numbers that's .02% of the time [$10/(60*24*30)*100$].

I've tried using random forest and it gives an error rate of around 2% which is unacceptable and from what I've read most machine learning algorithms have similar errors rates.

Does anyone know of a better approach?

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [charlie](#)

[Answer](#) by [wacax](#)

Random forests, GBM or even the newer and fancier xgboost are not the best candidates for binary classification (predicting ups and down) of stocks predictions or forex trading or at least not as the main algorithm. The reason is that, for this particular problem, they require a huge amount of trees (and tree depth in case of GBM or xgboost) to obtain reasonable accuracy (Breiman suggested using at least 5000 trees and to "not be stingy" and in fact his main ML paper on RF he used 50,000 trees per run).

However, some quants use random forests as feature selectors while others use it to generate new features. It all depends on the characteristics of the data.

I would suggest you read this [question and answers on quant.stackexchange](#) where people discuss what methods are the best and when to use them, among them ISOMAP, Laplacian eigenmaps, ANNs, swarm optimization.

Check out the [machine-learning tag on the same site](#), there you might find information

related to your particular dataset.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

[Q: Implement MLP in tensorflow](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

There are many resources online about how to implement MLP in tensorflow, and most of the samples do work :) But I am interested in a particular one, that I learned from <https://www.coursera.org/learn/machine-learning>. In which, it uses a *cost* function defined as follow:

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \sum_{k=1}^K [-y_k^{(i)} \log((h_\theta(x^{(i)}))_k - (1 - y_k^{(i)}) \log(1 - (h_\theta(x^{(i)}))_k)]$$

h_θ is the *sigmoid* function.

And there's my implementation:

[Skip code block](#)

```
# one hidden layer MLP

x = tf.placeholder(tf.float32, shape=[None, 784])
y = tf.placeholder(tf.float32, shape=[None, 10])

w_h1 = tf.Variable(tf.random_normal([784, 512]))
h1 = tf.nn.sigmoid(tf.matmul(x, w_h1))

w_out = tf.Variable(tf.random_normal([512, 10]))
y_ = tf.matmul(h1, w_out)

# cross_entropy = tf.nn.sigmoid_cross_entropy_with_logits(y_, y)
cross_entropy = tf.reduce_sum(- y * tf.log(y_) - (1 - y) * tf.log(1 - y_), 1)
loss = tf.reduce_mean(cross_entropy)
train_step = tf.train.GradientDescentOptimizer(0.05).minimize(loss)

correct_prediction = tf.equal(tf.argmax(y, 1), tf.argmax(y_, 1))
accuracy = tf.reduce_mean(tf.cast(correct_prediction, tf.float32))

# train
with tf.Session() as s:
    s.run(tf.initialize_all_variables())

    for i in range(10000):
        batch_x, batch_y = mnist.train.next_batch(100)
        s.run(train_step, feed_dict={x: batch_x, y: batch_y})

        if i % 100 == 0:
            train_accuracy = accuracy.eval(feed_dict={x: batch_x, y: batch_y})
            print('step {0}, training accuracy {1}'.format(i, train_accuracy))
```

I think the definition for the layers are correct, but the problem is in the *cross_entropy*. If I use the first one, *the one got commented out*, the model converges quickly; **but if I use the 2nd one, which I think/hope is the translation of the previous equation, the model won't converge**.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [davidshen84](#) 

[Answer](#)  by [emre](#) 

You made three mistakes:

1. You omitted the offset terms before the nonlinear transformations (variables b_1 and b_{out}). This increases the representative power of the neural network.
2. You omitted the softmax transformation at the top layer. This makes the output a probability distributions, so you can calculate the cross-entropy, which is the usual cost function for classification.
3. You used the binary form of the cross-entropy when you should have used the multi-class form.

When I run this I get accuracies over 90%:

[Skip code block](#)

```
import tensorflow as tf
from tensorflow.examples.tutorials.mnist import input_data

mnist = input_data.read_data_sets('/tmp/MNIST_data', one_hot=True)

x = tf.placeholder(tf.float32, shape=[None, 784])
y = tf.placeholder(tf.float32, shape=[None, 10])

w_h1 = tf.Variable(tf.random_normal([784, 512]))
b_1 = tf.Variable(tf.random_normal([512]))
h1 = tf.nn.sigmoid(tf.matmul(x, w_h1) + b_1)

w_out = tf.Variable(tf.random_normal([512, 10]))
b_out = tf.Variable(tf.random_normal([10]))
y_ = tf.nn.softmax(tf.matmul(h1, w_out) + b_out)

# cross_entropy = tf.nn.sigmoid_cross_entropy_with_logits(y_, y)
cross_entropy = tf.reduce_sum(- y * tf.log(y_), 1)
loss = tf.reduce_mean(cross_entropy)
train_step = tf.train.GradientDescentOptimizer(0.05).minimize(loss)

correct_prediction = tf.equal(tf.argmax(y, 1), tf.argmax(y_, 1))
accuracy = tf.reduce_mean(tf.cast(correct_prediction, tf.float32))

# train
with tf.Session() as s:
    s.run(tf.initialize_all_variables())

    for i in range(10000):
        batch_x, batch_y = mnist.train.next_batch(100)
        s.run(train_step, feed_dict={x: batch_x, y: batch_y})

        if i % 1000 == 0:
            train_accuracy = accuracy.eval(feed_dict={x: batch_x, y: batch_y})
            print('step {0}, training accuracy {1}'.format(i, train_accuracy))
```

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Machine Learning Steps](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

Which of the below set of steps options is the correct one when creating a predictive model?

Option 1:

First eliminate the most obviously bad predictors, and preprocess the remaining if needed,

then train various models with cross-validation, pick the few best ones, identify the top predictors each one has used, then retrain those models with those predictors only and evaluate accuracy again with cross-validation, then pick the best one and train it on the full training set using its key predictors and then use it to predict the test set.

Option 2:

First eliminate the most obviously bad predictors, then preprocess the remaining if needed, then use a feature selection technique like recursive feature selection (eg. RFE with rf) with cross-validation for example to identify the ideal number of key predictors and what these predictors are, then train different model types with cross-validation and see which one gives the best accuracy with those top predictors identified earlier. Then train the best one of those models again with those predictors on the full training set and then use it to predict the test set.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

User: [andrew-kostandy](#) 

[Answer](#)  by [dawny33](#) 

I found both of your options slightly faulty. So, this is generally (very broadly) how a predictive modelling:

- **Data Cleaning:** Takes the most time, but every second spent here is worth it. The cleaner your data gets through this step, the lesser would your total time spent be.
- **Splitting the data set:** The data set would be splitted into training and testing sets, which would be used for the modelling and prediction purposes respectively. In addition, an additional split as a cross-validation set would also need to be done.
- **Transformation and Reduction:** Involves processes like transformations, mean and median scaling, etc.
- **Feature Selection:** This can be done in a lot of ways like threshold selection, subset selection, etc.
- **Designing predictive model:** Design the predictive model on the training data depending on the features you have at hand.
- **Cross Validation:**
- **Final Prediction, Validation**

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#))

[**Q: Understanding Reinforcement Learning with Neural Net \(Q-learning\)**](#)

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

I am trying to understand reinforcement learning and markov decision processes (MDP) in the case where a neural net is being used as the function approximator.

I'm having difficulty with the relationship between the MDP where the environment is explored in a probabilistic manner, how this maps back to learning parameters and how the final solution/policies are found.

Am I correct to assume that in the case of Q-learning, the neural-network essentially acts as a function approximator for q-value itself so many steps in the future? How does this map to updating parameters via backpropagation or other methods?

Also, once the network has learned how to predict the future reward, how does this fit in with the system in terms of actually making decisions? I am assuming that the final system would not probabilistically make state transitions.

Thanks

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [catslovejazz](#) 

[Answer](#)  by [mtk99](#) 

In Q-Learning, on every step you will use observations and rewards to update your Q-value function:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[R_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)]$$

You are correct in saying that the neural network is just a function approximation for the q-value function.

In general, the approximation part is just a standard supervised learning problem. Your network uses (s, a) as input and the output is the q-value. As q-values are adjusted, you need to train these new samples to the network. Still, you will find some issues as you are using correlated samples and SGD will suffer.

If you are looking at the DQN paper, things are slightly different. In that case, what they are doing is putting samples in a vector (experience replay). To teach the network, they sample tuples from the vector, bootstrap using this information to obtain a new q-value that is taught to the network. When I say teaching, I mean adjusting the network parameters using stochastic gradient descent or your favourite optimisation approach. By not teaching the samples in the order that are being collected by the policy the decorrelate them and that helps in the training.

Lastly, in order to make a decision on state s , you choose the action that provides the highest q-value:

$$a^*(s) = \operatorname{argmax}_a Q(s, a)$$

If your Q-value function has been learnt completely and the environment is stationary, it is fine to be greedy at this point. However, while learning, you are expected to explore. There are several approaches being ϵ -greedy one of the easiest and most common ways.

Tags: [machine-learning](#) ([Prev Q](#)) ([Next Q](#)), [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[Q: Linear regression with non-symmetric cost function?](#)

Tags: [machine-learning](#) ([Prev Q](#))

I want to predict some value $Y(x)$ and I am trying to get some prediction $\hat{Y}(x)$ that optimizes between being as low as possible, but still being larger than $Y(x)$. In other words:

$$\text{cost}\{Y(x) \gtrsim \hat{Y}(x)\} \gg \text{cost}\{\hat{Y}(x) \gtrsim Y(x)\}$$

I think a simple linear regression should do totally fine. So I somewhat know how to implement this manually, but I guess I'm not the first one with this kind of problem. Are there any packages/libraries (preferably python) out there doing what I want to do? What's the keyword I need to look for?

What if I knew a function $Y_0(x) > 0$ where $Y(x) > Y_0(x)$. What's the best way to implement these restrictions?

Tags: [machine-learning](#) ([Prev Q](#))

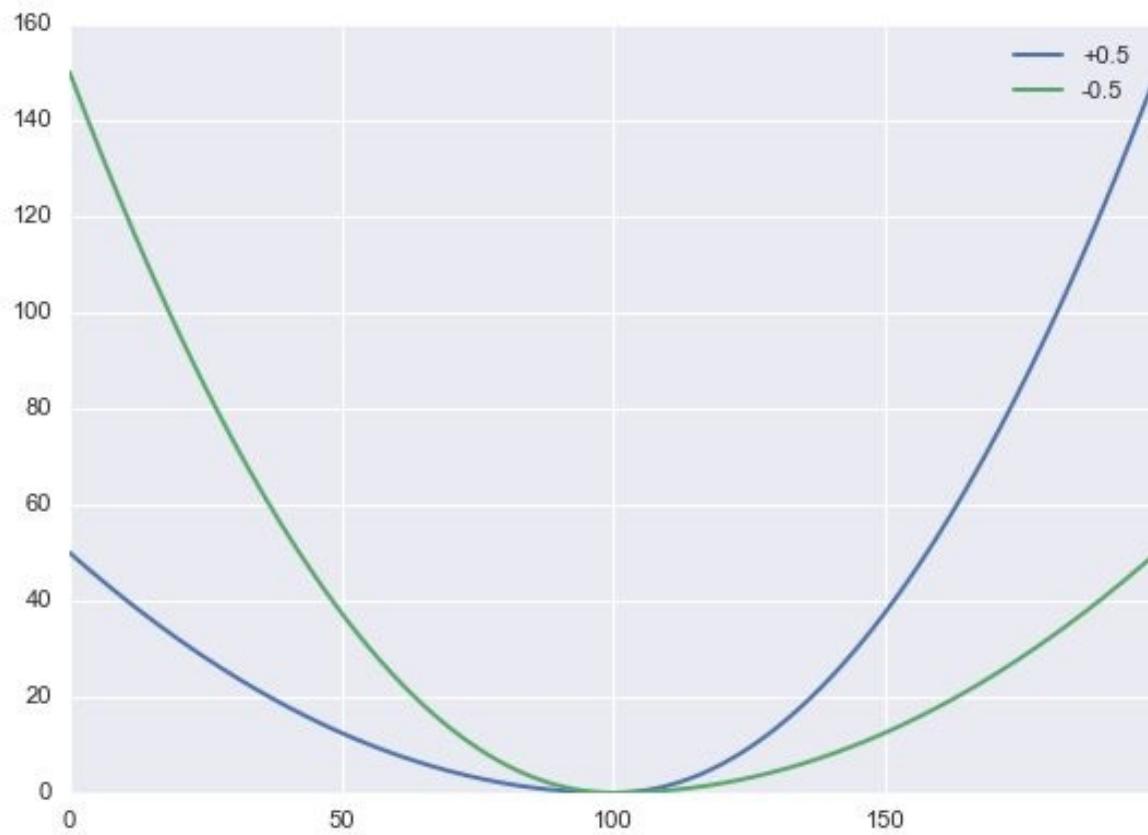
User: [asplankbridge](#)

[Answer](#) by [emre](#)

If I understand you correctly, you want to err on the side of overestimating. If so, you need an appropriate, asymmetric cost function. One simple candidate is to tweak the squared loss:

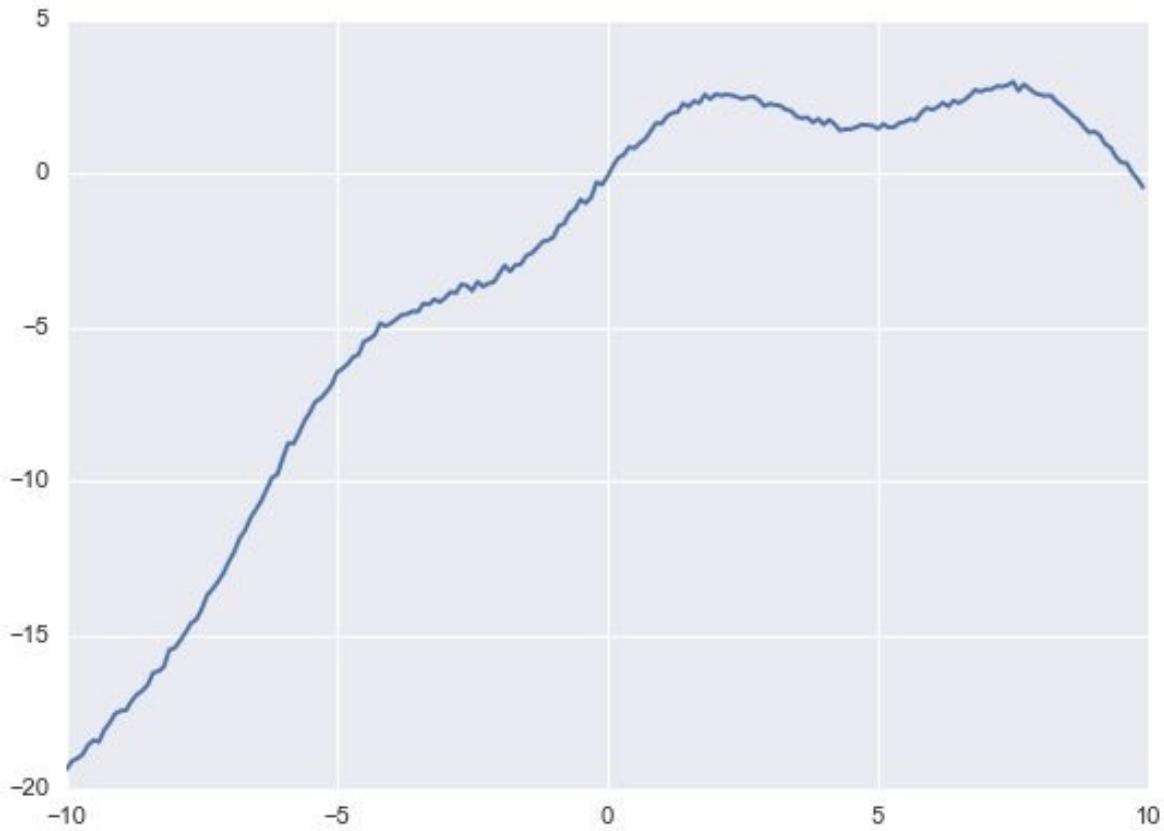
$$L : (x, \alpha) \rightarrow x^2 (\operatorname{sgn} x + \alpha)^2$$

where $-1 < \alpha < 1$ is a parameter you can use to trade off the penalty of underestimation against overestimation. Positive values of α penalize overestimation, so you will want to set α negative. In python this looks like `def loss(x, a): return x**2 * (numpy.sign(x) + a)**2`



Next let's generate some data:

```
import numpy
x = numpy.arange(-10, 10, 0.1)
y = -0.1*x**2 + x + numpy.sin(x) + 0.1*numpy.random.randn(len(x))
```



Finally, we will do our regression in tensorflow, a machine learning library from Google that supports automated differentiation (making gradient-based optimization of such problems simpler). I will use [this example](#) as a starting point.

Skip code block

```
import tensorflow as tf

X = tf.placeholder("float") # create symbolic variables
Y = tf.placeholder("float")

w = tf.Variable(0.0, name="coeff")
b = tf.Variable(0.0, name="offset")
y_model = tf.mul(X, w) + b

cost = tf.pow(y_model-Y, 2) # use sqr error for cost function
def acost(a): return tf.pow(y_model-Y, 2) * tf.pow(tf.sign(y_model-Y) + a, 2)

train_op = tf.train.AdamOptimizer().minimize(cost)
train_op2 = tf.train.AdamOptimizer().minimize(acost(-0.5))

sess = tf.Session()
init = tf.initialize_all_variables()
sess.run(init)

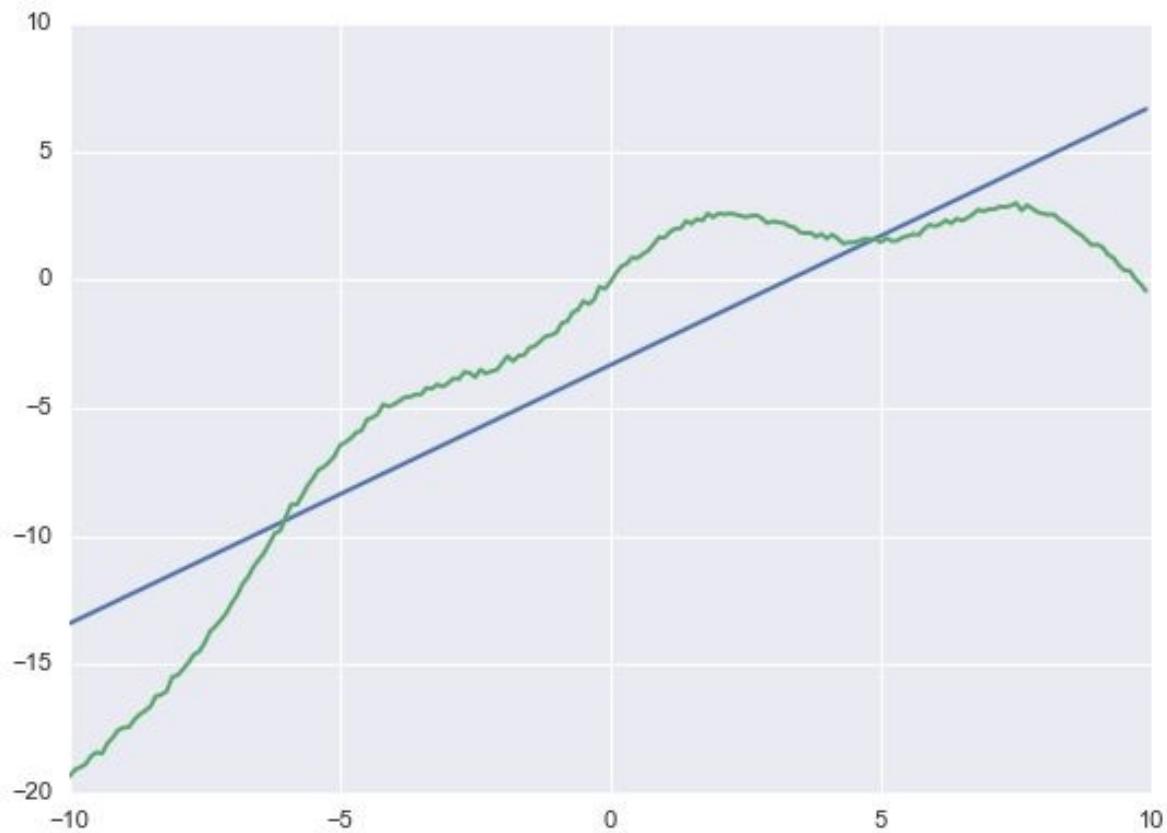
for i in range(100):
    for (xi, yi) in zip(x, y):
#        sess.run(train_op, feed_dict={X: xi, Y: yi})
        sess.run(train_op2, feed_dict={X: xi, Y: yi})

print(sess.run(w), sess.run(b))
```

cost is the regular squared error, while acost is the aforementioned asymmetric loss function.

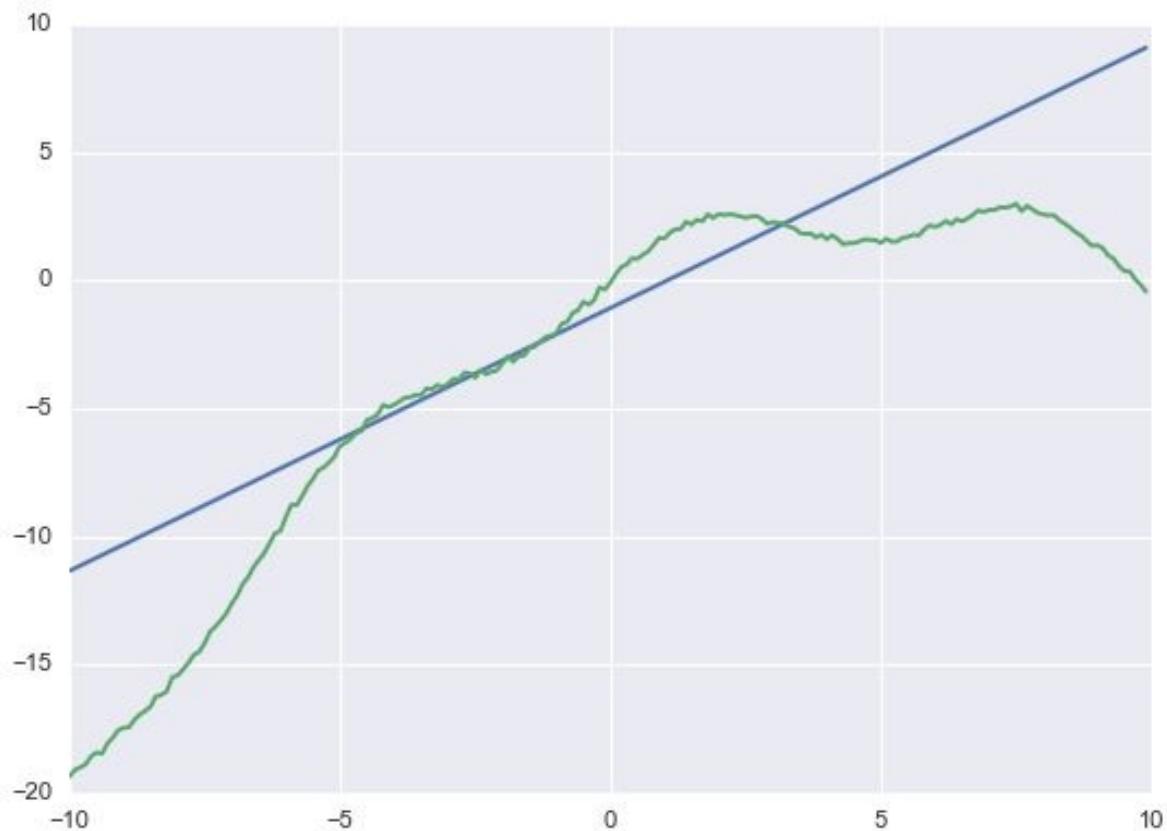
If you use cost you get

1.00764 -3.32445



If you use `acost` you get

1.02604 -1.07742



acost clearly tries not to underestimate. I did not check for convergence, but you get the idea.

Tags: [machine-learning](#) ([Prev Q](#))

Bigdata

Questions

[Q: How big is big data?](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [scalability](#) ([Prev Q](#)) ([Next Q](#))

Lots of people use the term *big data* in a rather *commercial* way, as a means of indicating that large datasets are involved in the computation, and therefore potential solutions must have good performance. Of course, *big data* always carry associated terms, like scalability and efficiency, but what exactly defines a problem as a *big data* problem?

Does the computation have to be related to some set of specific purposes, like data mining/information retrieval, or could an algorithm for general graph problems be labeled *big data* if the dataset was *big enough*? Also, how *big* is *big enough* (if this is possible to define)?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [scalability](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#) 

[Answer](#)  by [rolf](#) 

To me (coming from a relational database background), “Big Data” is not primarily about the data size (which is the bulk of what the other answers are so far).

“Big Data” and “Bad Data” are closely related. Relational Databases require ‘pristine data’. If the data is in the database, it is accurate, clean, and 100% reliable. Relational Databases require “Great Data” and a huge amount of time, money, and accountability is put on to making sure the data is well prepared before loading it in to the database. If the data is in the database, it is ‘gospel’, and it defines the system understanding of reality.

“Big Data” tackles this problem from the other direction. The data is poorly defined, much of it may be inaccurate, and much of it may in fact be missing. The structure and layout of the data is linear as opposed to relational.

Big Data has to have enough volume so that the amount of bad data, or missing data becomes statistically insignificant. When the errors in your data are common enough to cancel each other out, and when the missing data is proportionally small enough to be negligible. When your data access requirements and algorithms are functional even with incomplete and inaccurate data, then you have “Big Data”.

“Big Data” is not really about the volume, it is about the characteristics of the data.

[Answer](#)  by [tim-goodman](#) 

As you rightly note, these days “big data” is something everyone wants to say they’ve got, which entails a certain looseness in how people define the term. Generally, though, I’d say you’re certainly dealing with big data if the scale is such that it’s no longer feasible to manage with more traditional technologies such as RDBMS, at least without complementing them with big data technologies such as Hadoop.

How big your data has to actually be for that to be the case is debatable. Here’s a (somewhat provocative) [blog post](#) that claims that it’s not really the case for less than 5 TB of data. (To be clear, it doesn’t claim “Less than 5 TB isn’t big data”, but just “Less than 5 TB isn’t big enough that you need Hadoop”.)

But even on smaller datasets, big data technologies like Hadoop can have other advantages, including being well suited to batch operations, playing well with unstructured data (as well as data whose structure isn’t known in advance or could change), horizontal scalability (scaling by adding more nodes instead of beefing up your existing servers), and (as one of the commenters on the above-linked post notes) the ability to integrate your data processing with external data sets (think of a map-reduce where the mapper makes a call to another server). Other technologies associated with big data, like NoSQL databases, emphasize fast performance and consistent availability while dealing with large sets of data, as well also being able to handle semi-unstructured data and to scale horizontally.

Of course, traditional RDBMS have their own advantages including ACID guarantees (Atomicity, Consistency, Isolation, Durability) and better performance for certain operations, as well as being more standardized, more mature, and (for many users) more familiar. So even for indisputably “big” data, it may make sense to load at least a portion of your data into a traditional SQL database and use that in conjunction with big data technologies.

So, a more generous definition would be that you have big data so long as it’s big enough that big data technologies provide some added value for you. But as you can see, that can depend not just on the size of your data but on how you want to work with it and what sort of requirements you have in terms of flexibility, consistency, and performance. *How* you’re using your data is more relevant to the question than what you’re using it *for* (e.g. data mining). That said, uses like data mining and machine learning are more likely to yield useful results if you have a big enough data set to work with.

[Answer](#) by [alex-i](#)

Total amount of data in the world: 2.8 zetabytes in 2012, estimated to reach 8 zetabytes by 2015 ([source](#)) and with a doubling time of 40 months. Can’t get bigger than that :)

As an example of a single large organization, Facebook pulls in 500 terabytes per day, into a 100 petabyte warehouse, and runs 70k queries per day on it as of 2012 ([source](#)) Their current warehouse is >300 petabytes.

Big data is probably something that is a good fraction of the Facebook numbers (1/100 probably yes, 1/10000 probably not: it’s a spectrum not a single number).

In addition to size, some of the features that make it “big” are:

- it is actively analyzed, not just stored (quote “If you aren’t taking advantage of big data, then you don’t have big data, you have just a pile of data” Jay Parikh @ Facebook)
- building and running a data warehouse is a major infrastructure project
- it is growing at a significant rate
- it is unstructured or has irregular structure

Gartner definition: “Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing” (The 3Vs) So they also think “bigness” isn’t entirely about the size of the dataset, but also about the velocity and structure and the kind of tools needed.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [scalability](#) ([Prev Q](#)) ([Next Q](#))

[Q: Is the R language suitable for Big Data](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#))

R has many libraries which are aimed at Data Analysis (e.g. JAGS, BUGS, ARULES etc..), and is mentioned in popular textbooks such as: J.Krusche, Doing Bayesian Data Analysis; B.Lantz, “Machine Learning with R”.

I’ve seen a guideline of 5TB for a dataset to be considered as Big Data.

My question is: Is R suitable for the amount of Data typically seen in Big Data problems? Are there strategies to be employed when using R with this size of dataset?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#))

User: [akellyirl](#) 

[Answer](#)  by [mcp_infiltrator](#) 

Actually this is coming around. In the book R in a Nutshell there is even a section on using R with Hadoop for big data processing. There are some work arounds that need to be done because R does all its work in memory, so you are basically limited to the amount of RAM you have available to you.

A mature project for R and Hadoop is [RHadoop](#) 

RHadoop has been divided into several sub-projects, rhdfs, rhbase, rmr2, plyrnr, and quickcheck ([wiki](#) .

[Answer](#)  by [asheeshr](#) 

The main problem with using R for large data sets is the RAM constraint. The reason behind keeping all the data in RAM is that it provides much faster access and data manipulations than would storing on HDDs. If you are willing to take a hit on

performance, then yes, it is quite practical to work with [large datasets in R](#)

- RODBC Package: Allows connecting to external DB from R to retrieve and handle data. Hence, the data being *manipulated* is restricted to your RAM. The overall data set can go much larger.
 - The ff package allows using larger than RAM data sets by utilising memory-mapped pages.
 - BigLM: It builds generalized linear models on big data. It loads data into memory in chunks.
 - bigmemory : An R package which allows powerful and memory-efficient parallel analyses and data mining of massive data sets. It permits storing large objects (matrices etc.) in memory (on the RAM) using external pointer objects to refer to them.
-

[Answer](#) by [statsrus](#)

R is great for “big data”! However, you need a workflow since R is limited (with some simplification) by the amount of RAM in the operating system. The approach I take is to interact with a relational database (see the RSQLite package for creating and interacting with a SQLite database), run SQL-style queries to understand the structure of the data, and then extract particular subsets of the data for computationally-intensive statistical analysis.

This just one approach, however: there are packages that allow you to interact with other databases (e.g., Monet) or run analyses in R with fewer memory limitations (e.g., see pbdR).

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#))

[Q: When are p-values deceptive?](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

What are the data conditions that we should watch out for, where p-values may not be the best way of deciding statistical significance? Are there specific problem types that fall into this category?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [user179](#)

[Answer](#) by [alex-i](#)

You are asking about [Data Dredging](#), which is what happens when testing a very large number of hypotheses against a data set, or testing hypotheses against a data set that were suggested by the same data.

In particular, check out [Multiple hypothesis hazard](#), and [Testing hypotheses suggested by the data](#).

The solution is to use some kind of correction for [False discovery rate](#) or [Familywise error rate](#), such as [Scheffé's method](#) or the (very old-school) [Bonferroni correction](#).

In a somewhat less rigorous way, it may help to filter your discoveries by the confidence interval for the odds ratio (OR) for each statistical result. If the 99% confidence interval for the odds ratio is 10-12, then the OR is ≤ 1 with some *extremely* small probability, especially if the sample size is also large. If you find something like this, it is probably a strong effect even if it came out of a test of millions of hypotheses.

[Answer](#) by [tim-goodman](#)

You shouldn't consider the p-value out of context.

One rather basic point (as illustrated by [xkcd](#)) is that you need to consider how many tests you're actually doing. Obviously, you shouldn't be shocked to see $p < 0.05$ for one out of 20 tests, even if the null hypothesis is true every time.

A more subtle example of this occurs in high-energy physics, and is known as the [look-elsewhere effect](#). The larger the parameter space you search for a signal that might represent a new particle, the more likely you are to see an apparent signal that's really just due to random fluctuations.

[Answer](#) by [dan-c](#)

One thing you should be aware of is the sample size you are using. Very large samples, such as economists using census data, will lead to deflated p-values. This paper "[Too Big to Fail: Large Samples and the p-Value Problem](#)" covers some of the issues.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#))

Q: Which Big Data technology stack is most suitable for processing tweets, extracting/expanding URLs and pushing (only) new links into 3rd party system?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Next Q](#))

(Note: Pulled this question from the [list of questions in Area51](#), but believe the question is self explanatory. That said, believe I get the general intent of the question, and as a result likely able to field any questions on the question that might pop-up.)

Which Big Data technology stack is most suitable for processing tweets, extracting/expanding URLs and pushing (only) new links into 3rd party system?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Next Q](#))

User: [blunders](#)

[Answer](#) by [konstantin-v.-salikhov](#)

I'd suggest [Apache Kafka](#) as message store and any stream processing solution of your choice like [Apache Camel](#) or [Twitter Storm](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Next Q](#))

Q: How does a query into a huge database return with negligible latency?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

For example, when searching something in Google, results return nigh-instantly.

I understand that Google sorts and indexes pages with algorithms etc., but I imagine it infeasible for the results of every single possible query to be indexed (and results are personalized, which renders this even more infeasible)?

Moreover, wouldn't the hardware latency in Google's hardware be huge? Even if the data in Google were all stored in TB/s SSDs, I imagine the hardware latency to be huge, given the sheer amount of data to process.

Does MapReduce help solve this problem?

EDIT: Okay, so I understand that popular searches can be cached in memory. But what about unpopular searches? Even for the most obscure search I have conducted, I don't think the search has ever been reported to be larger than 5 seconds. How is this possible?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

User: [namehere](#)

[Answer](#) by [sean-owen](#)

MapReduce has nothing to do with real-time anything. It is a batch-oriented processing framework suitable for some offline tasks, like ETL and index building. Google has moved off of MapReduce for most jobs now, and even the Hadoop ecosystem is doing the same.

The answer to low latency is generally to keep precomputed indices in memory. Anything that touches disk is hard to make fast and scale. This is how newer-generation Hadoop-based SQL engines like [Impala](#) get so much speed compared to MapReduce-based infrastructure like [Hive](#), for example.

Search infrastructure can't cache the results of every single query. But it sure can cache intermediate results, or, more complete results for top queries. With a little caching you can serve results for a significant minority of all queries.

Search is also split across servers. So one machine can delegate to 100 to each get a part of the result and then combine them.

You can also get away with some degree of approximation. Google does not literally form a thousand pages of search results; it just has to get the first page about right.

Keep in mind that Google has *millions* of computers around the globe. Your queries are going to a data center geographically near to you and that is only serving your geography. This cuts out most of the latency, which is network and not processing time in the data center.

[Answer](#) by [anony-mousse](#)

MapReduce is not used in searching. It was used a long time ago to build the index; but it is a batch processing framework, and most of the web does not change all the time, so the newer architectures are all *incremental* instead of batch oriented.

Search in Google will largely work the same it works in Lucene and Elastic Search, except for a lot of fine tuned extra weighting and optimizations. But at the very heart, they will use some form of an **inverted index**. In other words, they do *not* search several terabytes when you enter a search query (even when it is not cached). They likely don't look at the actual documents at all. But they use a lookup table that lists which documents match your query term (with stemming, misspellings, synonyms etc. all preprocessed). They probably retrieve the *list* of the top 10000 documents for each word (10k integers - just a few kb!) and compute the best matches from that. Only if there aren't good matches in these lists, they expand to the next such blocks etc.

Queries for common words can be easily cached; and via preprocessing you can build a list of the top 10k results and then rerank them according to the user profile. There is nothing to be gained by computing an "exact" answer, too. Looking at the top 10k results is likely enough; there is no correct answer; and if a better result somewhere at position 10001 is missed, nobody will know or notice (or care). It likely was already ranked down in preprocessing and would not have made it into the top 10 that is presented to the user at the end (or the top 3, the user actually looks at)

Rare terms on the other hand aren't much of a challenge either - one of the lists only contains a few matching documents, and you can immediately discard all others.

I recommend reading this article:

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department, Stanford University, Stanford, CA 94305

<http://infolab.stanford.edu/~backrub/google.html>

And yes, that's the Google founders who wrote this. It's not the latest state, but it will already work at a pretty large scale.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

Q: When a relational database has better performance than a no relational

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Next Q](#))

When a relational database like mySQL has better performance than a no relational, like mongo?

I saw a question on Quora other day, about why Quora still uses mySQL as their backend. And how their performance is still good.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Next Q](#))

User: [filipe-ferminiano](#) 

[Answer](#)  by [oleksi](#) 

It depends on your data and what you're doing with it. For example, if the processing you have to do requires transactions to synchronize across nodes, it will likely be faster to use transactions implemented in an RDBMS rather than implementing it yourself on top of NoSQL databases which don't support it natively.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Next Q](#))

[Q: Why is it hard to grant efficiency while using libraries?](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

Any small database processing can be easily tackled by Python/Perl/... scripts, that uses libraries and/or even utilities from the language itself. However, when it comes to performance, people tend to reach out for C/C++/low-level languages. The possibility of tailoring the code to the needs seems to be what makes these languages so appealing for BigData — be it concerning memory management, parallelism, disk access, or even low-level optimizations (via assembly constructs at C/C++ level).

Of course such set of benefits would not come without a cost: writing the code, and sometimes even *reinventing the wheel*, can be quite expensive/tiresome. Although there are lots of libraries available, people are inclined to write the code by themselves whenever they need to *grant* performance. What *disables* performance assertions from using libraries while processing large databases?

For example, consider an entreprise that continuously crawls webpages and parses the data collected. For each sliding-window, different data mining algorithms are run upon the data extracted. Why would the developers ditch off using available libraries/frameworks (be it for crawling, text processing, and data mining)? Using stuff already implemented would not only ease the burden of coding the whole process, but also would save a lot of time.

In a single shot:

- what makes writing the code by oneself a *guarantee* of performance?
- why is it *risky* to rely on a frameworks/libraries when you must **assure** high performance?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#) 

[Answer](#) by [sean-owen](#)

I don't think that everyone reaches for C/C++ when performance is an issue.

The advantage to writing low-level code is using fewer CPU cycles, or sometimes, less memory. But I'd note that higher-level languages can call down to lower-level languages, and do, to get some of this value. Python and JVM languages can do this.

The data scientist using, for example, scikit-learn on her desktop is already calling heavily optimized native routines to do the number crunching. There is no point in writing new code for speed.

In the distributed "big data" context, you are more typically bottleneck on data movement: network transfer and I/O. Native code does not help. What helps is not writing the same code to run faster, but writing smarter code.

Higher-level languages are going to let you implement more sophisticated distributed algorithms in a given amount of developer time than C/C++. At scale, the smarter algorithm with better data movement will beat dumb native code.

It's also usually true that developer time, and bugs, cost loads more than new hardware. A year of a senior developer's time might be \$200K fully loaded; over a year that also rents hundreds of servers worth of computation time. It may just not make sense in most cases to bother optimizing over throwing more hardware at it.

I don't understand the follow up about "grant" and "disable" and "assert"?

[Answer](#) by [sahirbazzz](#)

As all we know, in Digital world there are many ways to do the same work / get expected results..

And responsibilities / risks which comes from the code are on developers' shoulders..

This is small but i guess a very useful example from .NET world..

So Many .NET developers use the built-in BinaryReader - BinaryWriter on their data serialization for performance / get control over the process..

This is CSharp source code of the FrameWork's built in BinaryWriter class' one of the overloaded Write Methods :

[Skip code block](#)

```
// Writes a boolean to this stream. A single byte is written to the stream
// with the value 0 representing false or the value 1 representing true.
//
public virtual void Write(bool value)
{
    // _buffer is a byte array which declared in ctor / init codes of the class
    _buffer = ((byte) (value? 1:0));

    // OutStream is the stream instance which BinaryWriter writes the value(s) into it.
    OutStream.WriteByte(_buffer[0]);
}
```

As you see, this method could written without the extra assigning to _buffer variable:

```
public virtual void Write(bool value)
{
```

```
} OutStream.WriteByte((byte) (value ? 1 : 0));
```

Without assigning we could gain few milliseconds..This few milliseconds can accept as “almost nothing” but what if there are multi-thousands of writing (i.e. in a server process)?

Lets suppose that “few” is 2 (milliseconds) and multi-Thousands instances are only 2.000.. This means 4 seconds more process time..4 seconds later returning..

If we continue to subject from .NET and if you can check the source codes of BCL - .NET Base Class Library- from MSDN you can see a lot of performance losts from the developer decides..

Any of the point from BCL source It's normal that you see developer decided to use while() or foreach() loops which could implement a faster for() loop in their code.

This small gains give us the total performance..

And if we return to the BinaryWriter.Write() Method..

Actually extra assigning to a _buffer implementation is not a developer fault..This is exactly decide to “stay in safe” !

Suppose that we decide to not use _buffer and decided to implement the second method..If we try to send multi-thousands bytes over a wire (i.e. upload / download a BLOB or CLOB data) with the second method, it can fail commonly because of connection lost..Cause we try to send all data without any checks and controlling mechanism.When connection lost, Both the server and Client never know the sent data completed or not.

If the developer decides “stay in safe” then normally it means performance costs depends to implemented “stay in safe” mechanism(s).

But if the developer decides “get risky, gain performance” this is not a fault also..Till there are some discussions about “risky” coding.

And as a small note : Commercial library developers always try to stay in safe because they can't know where their code will use.

[Answer](#)  by [steve-kallestad](#) 

Coming from a programmers perspective, frameworks rarely target performance as the highest priority. If your library is going to be widely leveraged the things people are likely to value most are ease of use, flexibility, and reliability.

Performance is generally valued in secondary competitive libraries. “X library is better because it's faster.” Even then very frequently those libraries will trade off the most optimal solution for one that can be widely leveraged.

By using any framework you are inherently taking a risk that a faster solution exists. I might go so far as to say that a faster solution almost always exists.

Writing something yourself is not a guarantee of performance, but if you know what you are doing and have a fairly limited set of requirements it can help.

An example might be JSON parsing. There are a hundred libraries out there for a variety of languages that will turn JSON into a referable object and vice versa. I know of one

implementation that does it all in CPU registers. It's measurably faster than all other parsers, but it is also very limited and that limitation will vary based on what CPU you are working with.

Is the task of building a high-performant environment specific JSON parser a good idea? I would leverage a respected library 99 times out of 100. In that one separate instance a few extra CPU cycles multiplied by a million iterations would make the development time worth it.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

[Q: Tradeoffs between Storm and Hadoop \(MapReduce\)](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Next Q](#))

Can someone kindly tell me about the trade-offs involved when choosing between Storm and MapReduce in Hadoop Cluster for data processing? Of course, aside from the obvious one, that Hadoop (processing via MapReduce in a Hadoop Cluster) is a batch processing system, and Storm is a real-time processing system.

I have worked a bit with Hadoop Eco System, but I haven't worked with Storm. After looking through a lot of presentations and articles, I still haven't been able to find a satisfactory and comprehensive answer.

Note: The term tradeoff here is not meant to compare to similar things. It is meant to represent the consequences of getting results real-time that are absent from a batch processing system.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Next Q](#))

User: [mb_ce](#)

[Answer](#) by [j_houg](#)

MapReduce: A fault tolerant distributed computational framework. MapReduce allows you to operate over huge amounts of data- with a lot of work put in to prevent failure due to hardware. MapReduce is a poor choice for computing results on the fly because it is slow. (A typical MapReduce job takes on the order of minutes or hours, not microseconds)

A MapReduce job takes a file (or some data store) as an input and writes a file of results. If you want these results available to an application, it is your responsibility to put this data in a place that is accessible. This is likely slow, and there will be a lag between the values you can display, and the values that represent your system in its current state.

An important distinction to make when considering using MapReduce in building realtime systems is that of training your model, and applying your model. If you think your model parameters do not change quickly, you can fit them with MapReduce, and then have a mechanism for accessing these pre-fit parameters when you want to apply your model.

Storm: A real-time, streaming computational system. Storm is online framework, meaning, in this sense, a service that interacts with a running application. In contrast to MapReduce, it receives small pieces of data (not a whole file) as they are processed in your application. You define a DAG of operations to perform on the data. A common and simple use case for Storm is tracking counters, and using that information to populate a real-time dashboard.

Storm doesn't have anything (necessarily) to do with persisting your data. Here, streaming is another way to say keeping the information you care about and throwing the rest away. In reality, you probably have a persistence layer in your application that has already recorded the data, and so this a good and justified separation of concerns.

If you want to know more... If you would like to learn more about real-time systems that

that fit parameters with MR and apply the models a different way [here are slides for a talk I gave on building real-time recommendation engines on HBase.](#) 

An excellent paper that marries real-time counting and persistence in an interesting way is [Google News Personalization: Scalable Online Collaborative Filtering](#) 

Another interesting marriage of MR and Storm is [SummingBird](#)  Summingbird allows you to define data analysis operations that can be applied via Storm or MR.

[Answer](#)  by [sean-owen](#) 

This is kind of like asking about the tradeoffs between frying pan and your drawer of silverware. They are not two things you compare, really. You might use them together as part of a larger project.

Hadoop itself is not one thing, but a name for a federation of services, like HDFS, Hive, HBase, MapReduce, etc. Storm is something you use with some of these services, like HDFS or HBase. It is a stream-processing framework. There are others within the extended Hadoop ecosystem, like Spark Streaming.

When would you choose a stream-processing framework? when you need to react to new data in near-real-time. If you need this kind of tool, you deploy this kind of tool, too.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Next Q](#))

[Q: Cascaded Error in Apache Storm](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

Going through the presentation and material of Summingbird by Twitter, one of the reasons that is mentioned for using Storm and Hadoop clusters together in Summingbird is that processing through Storm results in cascading of error. In order to avoid this cascading of error and accumulation of it, Hadoop cluster is used to batch process the data and discard the Storm results after the same data is processed by Hadoop.

What is the reasons for generation of this accumulation of error? and why is it not present in Hadoop? Since I have not worked with Storm, I do not know the reasons for it. Is it because Storm uses some approximate algorithm to process the data in order to process them in real time? or is the cause something else?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

User: [mb_ce](#) 

[Answer](#)  by [steve-kallestad](#) 

Twitter uses Storm for real-time processing of data. Problems can happen with real-time data. Systems might go down. Data might be inadvertently processed twice. Network connections can be lost. A lot can happen in a real-time system.

They use hadoop to reliably process historical data. I don't know specifics, but for

instance, getting solid information from aggregated logs is probably more reliable than attaching to the stream.

If they simply relied on Storm for everything - Storm would have problems due to the nature of providing real-time information at scale. If they relied on hadoop for everything, there's a good deal of latency involved. Combining the two with Summingbird is the next logical step.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

Q: Do I need to learn Hadoop to be a Data Scientist?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

An aspiring data scientist here. I don't know anything about Hadoop, but as I have been reading about Data Science and Big Data, I see a lot of talk about Hadoop. Is it absolutely necessary to learn Hadoop to be a Data Scientist?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

User: [pensu](#) 

[Answer](#)  by [steve-kallestad](#) 

Different people use different tools for different things. Terms like Data Science are generic for a reason. A data scientist could spend an entire career without having to learn a particular tool like hadoop. Hadoop is widely used, but it is not the only platform that is capable of managing and manipulating data, even large scale data.

I would say that a data scientist should be familiar with concepts like MapReduce, distributed systems, distributed file systems, and the like, but I wouldn't judge someone for not knowing about such things.

It's a big field. There is a sea of knowledge and most people are capable of learning and being an expert in a single drop. The key to being a scientist is having the desire to learn and the motivation to know that which you don't already know.

As an example: I could hand the right person a hundred structured CSV files containing information about classroom performance in one particular class over a decade. A data scientist would be able to spend a year gleaning insights from the data without ever needing to spread computation across multiple machines. You could apply machine learning algorithms, analyze it using visualizations, combine it with external data about the region, ethnic makeup, changes to environment over time, political information, weather patterns, etc. All of that would be "data science" in my opinion. It might take something like hadoop to test and apply anything you learned to data comprising an entire country of students rather than just a classroom, but that final step doesn't necessarily make someone a data scientist. And not taking that final step doesn't necessarily disqualify someone from being a data scientist.

[Answer](#)  by [user9170](#) 

As a former Hadoop engineer, it is not needed but it helps. Hadoop is just one system - the most common system, based on Java, and a ecosystem of products, which apply a particular technique “Map/Reduce” to obtain results in a timely manner. Hadoop is not used at Google, though I assure you they use big data analytics. Google uses their own systems, developed in C++. In fact, Hadoop was created as a result of Google publishing their Map/Reduce and BigTable (HBase in Hadoop) white papers.

Data scientists will interface with hadoop engineers, though at smaller places you may be required to wear both hats. If you are strictly a data scientist, then whatever you use for your analytics, R, Excel, Tableau, etc, will operate only on a small subset, then will need to be converted to run against the full data set involving hadoop.

[Answer](#) by [lgylym](#)

You have to first make it clear what do you mean by “learn Hadoop”. If you mean using Hadoop, such as learning to program in MapReduce, then most probably it is a good idea. But fundamental knowledge (database, machine learning, statistics) may play a bigger role as time goes on.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

[Q: Filtering spam from retrieved data](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

I once heard that filtering spam by using blacklists is not a good approach, since some user searching for entries in your dataset may be looking for particular information from the sources blocked. Also it'd become a burden to continuously validate the *current state* of each spammer blocked, checking if the site/domain still disseminate spam data.

Considering that any approach must be efficient and scalable, so as to support filtering on very large datasets, what are the strategies available to get rid of spam in a non-biased manner?

Edit: if possible, any example of strategy, even if just the intuition behind it, would be very welcome along with the answer.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#)

[Answer](#) by [neone4373](#)

Spam filtering, especially in email, has been revolutionized by neural networks, here are a couple papers that provide good reading on the subject:

On Neural Networks And The Future Of Spam A. C. Cosoi, M. S. Vlad, V. Sgariu
<http://ceai.srait.ro/index.php/ceai/article/viewFile/18/8>

Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks Ann Nosseir, Khaled Nagati and Islam Taj-Eddin <http://www.ijcsi.org/papers/IJCSI-10-2-1-17-21.pdf>

Spam Detection using Adaptive Neural Networks: Adaptive Resonance Theory David Ndumiyana, Richard Gotora, and Tarisai Mupamombe

<http://onlineresearchjournals.org/JPESR/pdf/2013/apr/Ndumiyana%20et%20al.pdf> 

EDIT: The basic intuition behind using a neural network to help with spam filtering is by providing a weight to terms based on how often they are associated with spam.

Neural networks can be trained most quickly in a supervised — you explicitly provide the classification of the sentence in the training set — environment. Without going into the nitty gritty the basic idea can be illustrated with these sentences:

Text = “How is the loss of the Viagra patent going to affect Pfizer”, Spam = false Text = “Cheap Viagra Buy Now”, Spam = true Text = “Online pharmacy Viagra Cialis Lipitor”, Spam = true

For a two stage neural network, the first stage will calculate the likelihood of spam based off of if the word exists in the sentence. So from our example:

viagra => 66% buy => 100% Pfizer => 0% etc..

Then for the second stage the results in the first stage are used as variables in the second stage:

viagra & buy => 100% Pfizer & viagra=> 0%

This basic idea is run for many of the permutations of the all the words in your training data. The end results once trained is basically just an equation that based of the context of the words in the sentence can assign a probability of being spam. Set spamminess threshold, and filter out any data higher then said threshold.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

Q: Is FP-Growth still considered “state of the art” in frequent pattern mining?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

As far as I know the development of algorithms to solve the Frequent Pattern Mining (FPM) problem, the road of improvements have some main checkpoints. Firstly, the [Apriori](#)  algorithm was proposed in 1993, by [Agrawal et al.](#) , along with the formalization of the problem. The algorithm was able to *strip-off* some sets from the $2^n - 1$ sets (powerset) by using a lattice to maintain the data. A drawback of the approach was the need to re-read the database to compute the frequency of each set expanded.

Later, on year 1997, [Zaki et al.](#)  proposed the algorithm [Eclat](#) , which *inserted* the resulting frequency of each set inside the lattice. This was done by adding, at each node of the lattice, the set of transaction-ids that had the items from root to the referred node. The main contribution is that one does not have to re-read the entire dataset to know the frequency of each set, but the memory required to keep such data structure built may exceed the size of the dataset itself.

In 2000, [Han et al.](#) proposed an algorithm named [FPGrowth](#), along with a prefix-tree data structure named FPTree. The algorithm was able to provide significant data compression, while also granting that only frequent itemsets would be yielded (without candidate itemset generation). This was done mainly by sorting the items of each transaction in decreasing order, so that the most frequent items are the ones with the least repetitions in the tree data structure. Since the frequency only descends while traversing the tree in-depth, the algorithm is able to *strip-off* non-frequent itemsets.

Edit:

~~As far as I know, this may be considered a state-of-the-art algorithm, but I'd like to know about other proposed solutions. What other algorithms for FPM are considered "state-of-the-art"? What is the *intuition/main-contribution* of such algorithms?~~

Is the FP-Growth algorithm still considered "state of the art" in frequent pattern mining? If not, what algorithm(s) may extract frequent itemsets from large datasets more efficiently?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#)

[Answer](#) by [anony-mousse](#)

State of the art as in: used in practise or worked on in theory?

APRIORI is used everywhere, except in developing new frequent itemset algorithms. It's easy to implement, and easy to reuse in very different domains. You'll find hundreds of APRIORI implementations of varying quality. And it's easy to get APRIORI wrong, actually.

FP-growth is much harder to implement, but also much more interesting. So from an academic point of view, everybody tries to improve FP-growth - getting work based on APRIORI accepted will be very hard by now.

If you have a good implementation, every algorithm has its good and its bad situations in my opinion. A good APRIORI implementation will *only* need to scan the database k times to find all frequent itemsets of length k . In particular if your data fits into main memory this is cheap. What can kill APRIORI is too many frequent 2-itemsets (in particular when you don't use a Trie and similar acceleration techniques etc.). It works best on large data with a low number of frequent itemsets.

Eclat works on columns; but it needs to read each column much more often. There is some work on diffsets to reduce this work. If your data does not fit into main memory, Eclat suffers probably more than Apriori. By going depth first, it will also be able to return a first interesting result much earlier than Apriori, and you can use these results to adjust parameters; so you need less iterations to find good parameters. But by design, it cannot exploit pruning as neatly as Apriori did.

FP-Growth compresses the data set into the tree. This works best when you have lots of duplicate records. You could probably reap quite some gains for Apriori and Eclat too if you can presort your data and merge duplicates into weighted vectors. FP-Growth does this

at an extreme level. The drawback is that the implementation is much harder; and once this tree does not fit into memory anymore it gets a mess to implement.

As for performance results and benchmarks - don't trust them. There are so many things to implement incorrectly. Try 10 different implementations, and you get 10 very different performance results. In particular for APRIORI, I have the impression that most implementations are broken in the sense of missing some of the main contributions of APRIORI... and of those that have these parts right, the quality of optimizations varies a lot.

There are actually even papers on how to implement these algorithms efficiently:

Efficient Implementations of Apriori and Eclat.

Christian Borgelt

Workshop of Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL, USA).

You may also want to read these surveys on this domain:

- Goethals, Bart. "Survey on frequent pattern mining." Univ. of Helsinki (2003).
- Ferenc Bodon, A Survey on Frequent Itemset Mining, Technical Report, Budapest University of Technology and Economic, 2006,
- Frequent Item Set Mining
Christian Borgelt
Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery
2(6):437-456. 2012

Answer  by [dsea](#) 

Most of the recent Frequent Pattern approaches that I've seen in the literature are based on optimizing FP-Growth. I have to admit, I haven't seen many developments within the literature in FPM in many years.

[This wikibook](#)  highlights many of the variants on FP-Growth that are out there.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#))

[**Q: Is Python suitable for big data**](#) 

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

I read in this post [Is the R language suitable for Big Data](#) that big data constitutes 5TB, and while it does a good job of providing information about the feasibility of working with this type of data in R it provides very little information about Python. I was wondering if Python can work with this much data as well.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

User: [ragingsloth](#) 

[Answer](#)  by [indico](#) 

To clarify, I feel like the original question references by OP probably isn't be best for a SO-type format, but I will certainly represent python in this particular case.

Let me just start by saying that regardless of your data size, python shouldn't be your limiting factor. In fact, there are just a couple main issues that you're going to run into dealing with large datasets:

- **Reading data into memory** - This is by far the most common issue faced in the world of big data. Basically, you can't read in more data than you have memory (RAM) for. The best way to fix this is by making atomic operations on your data instead of trying to read everything in at once.
 - **Storing data** - This is actually just another form of the earlier issue, by the time to get up to about 1TB, you start having to look elsewhere for storage. AWS S3 is the most common resource, and python has the fantastic boto library to facilitate dealing with large pieces of data.
 - **Network latency** - Moving data around between different services is going to be your bottleneck. There's not a huge amount you can do to fix this, other than trying to pick co-located resources and plugging into the wall.
-

[Answer](#)  by [ankit](#) 

There are couple off things you need to understand when dealing with Big data -

What is Big data?

You might be aware of famous V's of Big data - Volume, Velocity, Variety... So, Python may not be suitable for all. And it goes with all data science tools available. You need to know which tool is good for what purpose.

If dealing with large Volume of data:

- Pig/Hive/Shark - Data cleaning and ETL work
- Hadoop/Spark - Distributed parallel computing
- Mahout/ML-Lib - Machine Learning

Now, you can use R/Python in intermediate stages but you'll realize that they become bottleneck in your entire process.

If dealing with Velocity of data:

- Kafka/Storm - High throughput system

People are trying to R/Python here but again it depends on kind of parallelism you want and your model complexity.

What sort of analysis you wish to do?

If your model demands the entire data to be first brought into memory then your model should not be complex because if the intermediate data is large then the code will break. And if you think of writing it into disk then you'll face additional delay because disk read/write is slow as compared to RAM.

Conclusion

You can definitely use Python in Big data space (Definitely, since people are trying with R, why not Python) but know your data and business requirement first. There may be better tools available for same and always remember:

Your tools shouldn't determine how you answer questions. Your questions should determine what tools you use.

[Answer](#) by [theblackcat](#)

Python has some very good tools for working with big data:

numpy

Numpy's memmory-mapped arrays let you access a file saved on disk as though it were an array. Only the parts of the array you are actively working with need to be loaded into memory. It can be used pretty much the same as an ordinary array.

h5py and pytables

These two libraries provide access to HDF5 files. These files allow access to just part of the data. Further, thanks to the underlying libraries used to access the data, many mathematical operations and other manipulations of the data can be done without loading it into a python data structure. Massive, highly structured files are possible, much bigger than 5 TB. It also allows seamless, lossless compression.

databases

There are various types of databases that allow you to store big data sets and load just the parts you need. Many databases allow you to do manipulations without loading the data into a python data structure at all.

pandas

This allows higher-level access to various types of data, including HDF5 data, csv files, databases, even websites. For big data, it provides wrappers around HDF5 file access that makes it easier to do analysis on big data sets.

mpi4py

This is a tool for running your python code in a distributed way across multiple processors or even multiple computers. This allows you to work on parts of your data simultaneously.

task

It provides a version of the normal numpy array that supports many of the normal numpy operations in a multi-core manner that can work on data too large to fit into memory.

blaze

A tool specifically designed for big data. It is basically a wrapper around the above libraries, providing consistent interfaces to a variety of different methods of storing large amounts of data (such as HDF5 or databases) and tools to make it easy to manipulate, do mathematical operations on, and analyze data that is too big to fit into memory.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: Uses of NoSQL database in data science](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

How can [NoSQL](#)  databases like [MongoDB](#)  be used for data analysis? What are the features in them that can make data analysis faster and powerful?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

User: [jithinjustin](#) 

[Answer](#)  by [indico](#) 

To be perfectly honest, most NoSQL databases are not very well suited to applications in big data. For the vast majority of all big data applications, the performance of [MongoDB](#)  compared to a relational database like [MySQL](#)  is [significantly](#)  poor enough to warrant staying away from something like MongoDB entirely.

With that said, there are a couple of really useful properties of NoSQL databases that certainly work in your favor when you're working with large data sets, though the chance of those benefits outweighing the generally poor performance of NoSQL compared to [SQL](#)  for read-intensive operations (most similar to typical big data use cases) is low.

- **No Schema** - If you're working with a lot of unstructured data, it might be hard to actually decide on and rigidly apply a schema. NoSQL databases in general are very supporting of this, and will allow you to insert schema-less documents on the fly, which is certainly not something an SQL database will support.
- [JSON](#)  - If you happen to be working with JSON-style documents instead of with

[CSV](#) files, then you'll see a lot of advantage in using something like MongoDB for a database-layer. Generally the workflow savings don't outweigh the increased query-times though.

- **Ease of Use** - I'm not saying that SQL databases are always hard to use, or that [Cassandra](#) is the easiest thing in the world to set up, but in general NoSQL databases are easier to set up and use than SQL databases. MongoDB is a particularly strong example of this, known for being one of the easiest database layers to use (outside of [SQLite](#)). SQL also deals with a lot of normalization and there's a large legacy of SQL best practices that just generally bogs down the development process.

Personally I might suggest you also check out [graph databases](#) such as [Neo4j](#) that show really good performance for certain types of queries if you're looking into picking out a backend for your data science applications.

[Answer](#) by [emre](#)

One benefit of the schema-free NoSQL approach is that you don't commit prematurely and you can apply the right schema at query time using an appropriate tool like [Apache Drill](#). See [this presentation](#) for details. MySQL wouldn't be my first choice in a big data setting.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

[Q: Scalable Outlier/Anomaly Detection](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

I am trying to setup a Big Data Infra-Structure using Hadoop, Hive, Elastic Search (amongst others) and I would like to run some algorithms over these datasets. I would like the algorithms themselves, to be scalable, so that excludes using tools such as Weka, R, or even RHadoop. The [Apache Mahout Library](#)  seems to be a good option, and it features [algorithms for regression and clustering tasks](#) . What I am struggling to find is a solution for anomaly or outlier detection. Since Mahout features Hidden Markov Models and a variety of clustering techniques (including K-Means) I was wondering if it would be possible to build a model to detect outliers in time-series, using any of this. I would be grateful if somebody experienced on this could advice me a) if it is possible, and in case it is b) how-to do it, plus c) an estimation of the effort involved and d) accuracy/problems of this approach.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

User: [doublebyte](#) 

[Answer](#)  by [prudenko](#) 

I would take a look at [t-digest algorithm](#) . It's [been merged into mahout](#)  and also a part of some other libraries (github.com/addthis/stream-lib/blob/master/src/main/java/com/clearspring/analytics/stream/quantile/TDigest.java) for big data streaming. You can get more about this algorithm particularly and big data anomaly detection in general in next resources:

1. Practical machine learning anomaly detection book.
(info.mapr.com/rs/mapr/images/Practical_Machine_Learning_Anomaly_Detection.pdf)
 2. Webinar: Anomaly Detection When You Don't Know What You Need to Find
(youtube.com/watch?v=i-mSV63Q9rA#t=757)
 3. Anomaly Detection in Elasticsearch (info.prelert.com/anomaly-detection-in-elasticsearch)
 4. Beating Billion Dollar Fraud Using Anomaly Detection: A Signal Processing Approach using Argyle Data on the Hortonworks Data Platform with Accumulo
(oreilly.com/pub/e/3211)
-

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to deal with version control of large amounts of \(binary\) data](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

I am a PhD student of Geophysics and work with large amounts of image data (hundreds of GB, tens of thousands of files). I know svn and git fairly well and come to value a project history, combined with the ability to easily work together and have protection against disk corruption. I find git also extremely helpful for having consistent backups but I know that git cannot handle large amounts of binary data efficiently.

In my masters studies I worked on data sets of similar size (also images) and had a lot of problems keeping track of different version on different servers/devices. Differing 100GB over the network really isn't fun, and cost me a lot of time and effort.

I know that others in science seem to have similar problems, yet I couldn't find a good solution.

I want to use the storage facilities of my institute, so I need something that can use a "dumb" server. I also would like to have an additional backup on a portable hard disk, because I would like to avoid transferring hundreds of GB over the network wherever possible. So, I need a tool that can handle more than one remote location.

Lastly, I really need something that other researcher can use, so it does not need to be super simple, but should be learnable in a few hours.

I have evaluated a lot of different solutions, but none seem to fit the bill:

- [svn](#) is somewhat inefficient and needs a smart server
- hg [bigfile](#)/[largefile](#) can only use one remote
- git [bigfile](#)/[media](#) can also use only one remote, but is also not very efficient
- [attic](#) doesn't seem to have a log, or differencing capabilities
- [bup](#) looks really good, but needs a "smart" server to work

I've tried git-annex, which does everything I need it to do (and much more), but it is very difficult to use and not well documented. I've used it for several days and couldn't get my head around it, so I doubt any other coworker would be interested.

How do researchers deal with large datasets, and what are other research groups using?

To be clear, I am primarily interested in how other researchers deal with this situation, not just this specific dataset. It seems to me that almost everyone should have this problem, yet I don't know anyone who has solved it. Should I just keep a backup of the original data and forget all this version control stuff? Is that what everyone else is doing?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

User: [johann](#)

[Answer](#) by jakebeal

I have dealt with similar problems with very large synthetic biology datasets, where we have many, many GB of [flow cytometry](#) data spread across many, many thousands of files, and need to maintain them consistently between collaborating groups at (multiple) different institutions.

Typical version control like svn and git is not practical for this circumstance, because it's just not designed for this type of dataset. Instead, we have fallen to using "cloud storage"

solutions, particularly [DropBox](#) and [Bittorrent Sync](#). DropBox has the advantage that it does do at least some primitive logging and version control and manages the servers for you, but the disadvantage that it's a commercial service, you have to pay for large storage, and you're putting your unpublished data on a commercial storage; you don't have to pay much, though, so it's a viable option. Bittorrent Sync has a very similar interface, but you run it yourself on your own storage servers and it doesn't have any version control. Both of them hurt my programmer soul, but they're the best solutions my collaborators and I have found so far.

[Answer](#) by [fritz](#)

I have used [Versioning on Amazon S3 buckets](#) to manage 10-100GB in 10-100 files. Transfer can be slow, so it has helped to compress and transfer in parallel, or just run computations on EC2. The [boto](#) library provides a nice python interface.

[Answer](#) by [piotr-migdal](#)

Try looking at [Git Large File Storage \(LFS\)](#). It is new, but might be the thing worth looking at.

As I see, [a discussion on Hacker News](#) mentions a few other ways to deal with large files:

- [git-annex](#) (and e.g. [using it with Amazon S3](#))
 - [Mercurial Largefiles extension](#)
-

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is an ‘old name’ of data scientist?](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#))

Terms like ‘data science’ and ‘data scientist’ are increasingly used these days. Many companies are hiring ‘data scientist’. But I don’t think it’s a completely new job. Data have existed from the past and someone had to deal with data. I guess the term ‘data scientist’ becomes more popular because it sounds more fancy and ‘sexy’ How were data scientists called in the past?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#))

User: [user67275](#)

[Answer](#) by [emre](#)

In reverse chronological order: data miner, statistician, (applied) mathematician.

[Answer](#) by [alex-s-kinman](#)

Terms that covered more or less the same topics that Data Science covers today:

- Pattern Recognition
 - Machine Learning
 - Data Mining
 - Quantitative methods
-

[Answer](#)  by [akavall](#) 

I do think it is new job, basically data scientist has to apply mathematical algorithms on data with considerable constraint in terms 1) Run time of the application 2) Resource use of the application. If these constraints are not present, I would not call the job data science. Moreover, these algorithms are often need to be ran on distributed systems, which is another dimension of the problem.

Of course, this has been done before, in some combination of statistics, mathematics and programming, but it was not wide spread to give rise to the new term. The real rise of data science is from the ability to gather large amounts of data, thus need to process it.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#))

[Q: Reference about social network data-mining](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Next Q](#)), [beginner](#) ([Prev Q](#)) ([Next Q](#))

I am not in the data science field, but I would like to examine in depth this field and, particularly, I would like to start from the analysis of the social networks data.

I am trying to find some good references, both paper, websites and books, in order to start learning about the topic. Browsing on the internet, one can find a lot of sites, forum, papers about the topic, but I'm not able to discriminate among good and bad readings.

I am an R, Matlab, SAS user and I know a little bit of python language.

Could you suggest any references from which I could start studying and deepen the industry?

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Next Q](#)), [beginner](#) ([Prev Q](#)) ([Next Q](#))

User: [quantopic](#) 

[Answer](#)  by [sheldonkreger](#) 

My favorite place to find information about social network analysis is from SNAP, the Stanford Network Analysis Project. Led by Jure Leskovec, this team of students and professors has built software tools, gathered data sets, and published papers on social network analysis.

<http://snap.stanford.edu/> 

The collection of research papers there is outstanding.

They also have a Python tool you could try. <http://snap.stanford.edu/snappy/index.html>

The focus is on graph analysis, because social networks fit this model well. If you are new to graph analysis, I suggest you take a undergraduate level discrete mathematics course, or check out my favorite book on the topic “Graph Theory with Algorithms and its Applications” by Santanu Ray.

For a hands-on approach to social network analysis, check out “Mining the Social Web” by Matthew A Russell. It has examples which cover how to collect and analyze data from the major social networks like Twitter, Facebook, and LinkedIn.

It was Jure Leskovec who initially excited me about this field. He has many great talks on YouTube, for example: https://www.youtube.com/watch?v=LmQ_3nijMCs

[Answer](#) by [stas-prihod'co](#)

I'm going to pursue the following series of online courses on the Coursera: [Become a Social Scientist: Methods and Statistics](#) by *University of Amsterdam*. The good news - it is free, or you can get a nice-looking certificate for \$49 or so. The bad news - the nearest enrollment is Aug 31st 2015. You will have opportunity to get a lot of information in condensed way during a short time frame and you will be enforced to actually apply the knowledge in exercises, quizzes and project assignments. You will also have opportunity to discuss lessons/projects on the forum with many other students and lecturers.

[update] I apologize, I just remembered there is another course [Introduction to Statistics for the Social Sciences](#) by *University of Zurich* - just started April 28th, 2015. If you want to pursue it - do not forget about deadlines for quizzes and homeworks. Good luck!

[Answer](#) by [leo-t](#)

I think Social Media Mining: An Introduction by Zafarani et. al. is an excellent starting point. You can find more about it [here](#). Also a free PDF version is available.

It first goes through the essentials in graph theory and data mining. It covers some more advanced topics in graph mining, social network analysis, recommendation systems, etc.

Besides, I have seen some online courses in coursera ([example](#)). I am not sure about their quality though.

Finally, note that social network analysis is data mining for the social media data like Facebook. It is not social science at all; it is computer science. While you may end up borrowing some ideas from them, what you will end up doing is far from what social science guys are doing. So, going through social science courses and books is likely not a good idea at this point.

P.S. The book is the text book for the social media mining course offered in my university.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Next Q](#)), [beginner](#) ([Prev Q](#)) ([Next Q](#))

Q: What's an efficient way to compare and group millions of store

[names?](#)

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#))

I'm a total amateur as far as data science goes, and I'm trying to figure out a way to do some string comparison on a large dataset.

I've a Google BigQuery table storing merchant transactions, but the store names are all over the board. For example, there can be 'Wal-Mart Super Center' and 'Wal-Mart SC #1234', or 'McDonalds F2222' and 'McDonalds #321'.

What I need to do is group ALL 'Wal-mart' and 'McDonalds' and whatever else. My first approach was doing a recursive reg-ex check, but that took forever and eventually timed-out.

What's the best approach for doing that with a table of 20 million+ rows? I'm open to trying out any technology that would fit this job.

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#))

User: [terrymatula](#)

[Answer](#) by [an6u5](#)

This is an [entity resolution](#) aka [record linkage](#) aka [data matching](#) problem.

I would solve this by removing all of the non-alphabetical characters including numbers, casting into all uppercase and then employing a hierarchical match. First match up the exact cases and then move to a Levenshtein scoring between the fields. Make some sort of a decision about how large you will allow the Levenshtein or normalized Levenshtein score to get before you declare something a non-match.

Assign every row an id and when you have a match, reassign the lower of the IDs to both members of the match.

The [Levenshtein distance](#) algorithm is simple but brilliant ([taken from here](#)):

[Skip code block](#)

```
def levenshtein(a,b):
    "Calculates the Levenshtein distance between a and b."
    n, m = len(a), len(b)
    if n > m:
        # Make sure n <= m, to use O(min(n,m)) space
        a,b = b,a
        n,m = m,n

    current = range(n+1)
    for i in range(1,m+1):
        previous, current = current, [i]+[0]*n
        for j in range(1,n+1):
            add, delete = previous[j]+1, current[j-1]+1
            change = previous[j-1]
            if a[j-1] != b[i-1]:
                change = change + 1
            current[j] = min(add, delete, change)

    return current[n]
```

This [Data Matching](#) book is a good resource and is free for seven days on Amazon.

Nominally, this is an n algorithm without exploiting some sorting efficiencies, so I would expect to have to use multiple cores on 2×10^7 rows. But this should run just fine on an 8 core [AWS instance](#). It will eventually finish on a single core, but might take several hours.

Hope this helps!

[Answer](#) by [image_doctor](#)

I'd be really tempted to be lazy and apply some old technology for a quick and dirty solution, with no programming, using the linux sort command. This will give you a lexicographically sorted list.

If the store names are not the first field, if just reorder them or tell sort to use a different field via the -k switch.

Save the data to a plain CSV text file and then sort them:

```
$sort myStores.csv > sortedByStore.csv
```

You can give sort a hand by allocating it plenty of memory, 16GB in this case:

```
$sort -S16G myStores.csv > sortedByStore.csv
```

You could go further and produce a list of unique store names and counts of instances for them to help you get a handle on what the data looks like:

```
$sort -S16G myStores.csv | cut -f1 -d, | uniq -c > storeIdsAndCounts.csv
```

Or to avoid resorting and have only the unique IDs:

```
$cat sortedByStore.csv | cut -f1 -d, | uniq > storeIds.csv
```

Tags: [bigdata](#) ([Prev Q](#)) ([Next Q](#))

[Q: Original Meaning of “Intelligence” in “Business Intelligence”](#)

Tags: [bigdata](#) ([Prev Q](#)), [definitions](#) ([Next Q](#))

What does the term “**Intelligence**” originally stand for in “**Business Intelligence**”? Does it mean as used in “[Artificial Intelligence](#)” or as used in “[Intelligence Agency](#)”?

In other words, does “[Business Intelligence](#)” mean: “Acting smart & intelligently in business” or “Gathering data and information about the business”?

This question was the topic of a debate among some fellows in our data-science team, so I thought to ask about it from other experts. One might say that both meanings are applicable, but I’m asking for the original intended meaning of the word as proposed in the 1980s.

An acceptable answer should definitely cite original references, and personal opinions are not what I’m seeking.

Tags: [bigdata](#) ([Prev Q](#)), [definitions](#) ([Next Q](#))

User: [seyed-mohammad](#)

[Answer](#) by [laurent-duval](#)

Howard Dresner, in 1989, is believed to have coined the term “business intelligence”, to describe “concepts and methods to improve business decision making by using fact-based support systems.”. When he was at Gartner Group. This is a common mantra, spread over the Web. I have not been able to trace the exact source for this origin yet. Many insist on he was not at Gartner group in 1989, which is confirmed in the [following interview](#). In his 2008 book, Performance Management Revolution: Improving Results Through Visibility and Actionable Insight, the termed is defined as:

BI is knowledge gained through the access and analysis of business information.

He says, at the beginning, that

In 1989, for example, I started-some might say incited-the BI revolution with the premise that all users have a fundamental right to access information without the help of IT.

No apparent claim of the invention of the term on his side. Indeed, one can find older roots in H. P. Luhn, [A Business Intelligence System](#), IBM Journal of Research and Development, 1958, Vol. 2, Issue 4, p. 314—319.

Abstract: An automatic system is being developed to disseminate information to the various sections of any industrial, scientific or government organization. This intelligence system will utilize data-processing machines for auto-abtracting and auto-encoding of documents and for creating interest profiles for each of the “action points” in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points. This paper shows the flexibility of such a system in identifying known information, in finding who needs to know it and in disseminating it efficiently either in abstract form or as a complete document.

The author claims that:

The techniques proposed here to make these things possible are:

1. Auto-abstracting of documents;
2. Auto-encoding of documents;
3. Automatic creation and updating of action-point profiles.

All of these techniques are based on statistical procedures which can be performed on present-day data processing machines. Together with proper communication facilities and input-output equipment a comprehensive system may be assembled to accommodate all information problems of an organization. We call this a Business Intelligence System.

He also gives the explanation of the terms “**business**” and “**intelligence**“:

In this paper, business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, as “the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal.”

So the idea of “linking the facts” is already present in H. P. Luhn paper. To many sources, Howard Dresner has re-invented “Business Intelligence” to re-brand decision support system (DSS) and executive information system (EIS) when at DEC, and the term became famous through the influence of the Gartner group.

Apparently, the term has already been used way before, as in the book [Wholesale Business Intelligence and Southern and Western Merchants' Pocket Directory to the Principal Mercantile Houses in the City of Philadelphia, for the Year 1839](#).

As I could not fetch this source, I will stick to the Luhn/Dresner acceptance. It relates to the [etymology of intelligence](#):

late 14c., “faculty of understanding,” from Old French intelligence (12c.), from Latin intelligentia, intellegentia “understanding, power of discerning; art, skill, taste,” from intelligentem (nominative intelligens) “discerning,” present participle of intelligere “to understand, comprehend,” from inter- “between” (see inter-) + legere “choose, pick out

In Business Intelligence for Dummies (Scheps, 2008), the definition chapter plays on Military Intelligence:

Business Intelligence Defined: No CIA Experience Required So what the heck is business intelligence, anyway? In essence, BI is any activity, tool, or process used to obtain the best information to support the process of making decisions.

For our purposes, BI revolves around putting computing power (highly specialized software in concert with other more common technology assets) to work, to help make the best choices for your organization.

Business intelligence is essentially timely, accurate, high-value, and actionable business insights, and the work processes and technologies used to obtain them.

I would thus bend toward “Gathering data and information about the business”, maybe more “to better conduct business”. Additional historical comments can be found in [Father of BI? Is he having a laugh?](#)

Tags: [bigdata](#) ([Prev Q](#)), [definitions](#) ([Next Q](#))

Data Mining

[Skip to questions,](#)

Wiki by user [clayton](#) 

Conceptually speaking, *data-mining* can be thought of as one item (or set of skills and applications) in the toolkit of the data scientist.

More specifically, data-mining is an activity that seeks patterns in large, complex data sets. It usually emphasizes algorithmic techniques, but may also involve any set of related skills, applications, or methodologies with that goal.

In US-English colloquial speech, data-mining and data-collection are often used interchangeably.

However, a main difference between these two related activities is *intentionality*.

Definition inspired mostly by the contributions of [@statsRus](#)  to Data Science.SE

Questions

[Q: K-Means clustering for mixed numeric and categorical data](#)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#)), [categorical-data](#) ([Next Q](#))

My data set contains a number of numeric attributes and one categorical.

Say, NumericAttr1, NumericAttr2, ..., NumericAttrN, CategoricalAttr,

where CategoricalAttr takes one of three possible values: CategoricalAttrValue1, CategoricalAttrValue2 or CategoricalAttrValue3.

I'm using default k-means clustering algorithm implementation for Octave

<https://blog.west.uni-koblenz.de/2012-07-14/a-working-k-means-code-for-octave/>  It works with numeric data only.

So my question: is it correct to split the categorical attribute CategoricalAttr into three numeric (binary) variables, like IsCategoricalAttrValue1, IsCategoricalAttrValue2, IsCategoricalAttrValue3 ?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#)), [categorical-data](#) ([Next Q](#))

User: [ihars](#) 

[Answer](#)  by [tim-goodman](#) 

The standard k-means algorithm isn't directly applicable to categorical data, for various reasons. The sample space for categorical data is discrete, and doesn't have a natural origin. A Euclidean distance function on such a space isn't really meaningful. As someone put it, "The fact a snake possesses neither wheels nor legs allows us to say nothing about the relative value of wheels and legs." (from [here](#) )

There's a variation of k-means known as k-modes, introduced in [this paper](#)  by Zhexue Huang, which is suitable for categorical data. Note that the solutions you get are sensitive to initial conditions, as discussed [here](#)  (PDF), for instance.

Huang's paper (linked above) also has a section on "k-prototypes" which applies to data with a mix of categorical and numeric features. It uses a distance measure which mixes the Hamming distance for categorical features and the Euclidean distance for numeric features.

A Google search for "k-means mix of categorical data" turns up quite a few more recent papers on various algorithms for k-means-like clustering with a mix of categorical and numeric data. (I haven't yet read them, so I can't comment on their merits.)

Actually, what you suggest (converting categorical attributes to binary values, and then doing k-means as if these were numeric values) is another approach that has been tried before (predating k-modes). (See Ralambondrainy, H. 1995. A conceptual version of the k-means algorithm. Pattern Recognition Letters, 16:1147–1157.) But I believe the k-

modes approach is preferred for the reasons I indicated above.

[Answer](#) by [adesantos](#)

In my opinion, there are solutions to deal with categorical data in clustering. R comes with a specific distance for categorical data. This distance is called Gower (<http://www.inside-r.org/packages/cran/StatMatch/docs/gower.dist>) and it works pretty well.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#)), [categorical-data](#) ([Next Q](#))

[Q: Clustering customer data stored in ElasticSearch](#)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

I have a bunch of customer profiles stored in a [elasticsearch](#) cluster. These profiles are now used for creation of target groups for our email subscriptions.

Target groups are now formed manually using elasticsearch faceted search capabilities (like get all male customers of age 23 with one car and 3 children).

How could I search for interesting groups **automatically** - using data science, machine learning, clustering or something else?

R programming language seems to be a good tool for this task, but I can't form a methodology of such group search. One solution is to somehow find the largest clusters of customers and use them as target groups, so the question is:

How can I automatically choose largest clusters of similar customers (similar by parameters that I don't know at this moment)?

For example: my program will connect to elasticsearch, offload customer data to CSV and using R language script will find that large portion of customers are male with no children and another large portion of customers have a car and their eye color is brown.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [konstantin-v.-salikhov](#)

[Answer](#) by [nick-peterson](#)

One algorithm that can be used for this is the [k-means clustering algorithm](#).

Basically:

1. Randomly choose k datapoints from your set, m_1, \dots, m_k .
2. “Until convergence”:
 1. Assign your data points to k clusters, where cluster i is the set of points for which m_i is the closest of your current means
 2. Replace each m_i by the mean of all points assigned to cluster i .

It is good practice to repeat this algorithm several times, then choose the outcome that minimizes distances between the points of each cluster i and the center m_i .

Of course, you have to know k to start here; you can use cross-validation to choose this parameter, though.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

Q: What are good sources to learn about Bootstrap?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

I think that Bootstrap can be useful in my work, where we have a lot of variables that we don't know the distribution of it. So, simulations could help. What are good sources to learn about Bootstrap/other useful simulation methods?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

User: [filipe-ferminiano](#) 

[Answer](#)  by [iliasfl](#) 

A classic book is by B. Efron who created the technique:

- Bradley Efron; Robert Tibshirani (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC. ISBN 978-0-412-04231-7.
-

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [statistics](#) ([Prev Q](#)) ([Next Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

Q: Are Support Vector Machines still considered “state of the art” in their niche?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

This question is in response to a comment I saw on another question.

The comment was regarding the Machine Learning course syllabus on Coursera, and along the lines of “SVMs are not used so much nowadays”.

I have only just finished the relevant lectures myself, and my understanding of SVMs is that they are a robust and efficient learning algorithm for classification, and that when using a kernel, they have a “niche” covering number of features perhaps 10 to 1000 and number of training samples perhaps 100 to 10,000. The limit on training samples is because the core algorithm revolves around optimising results generated from a square matrix with dimensions based on number of training samples, not number of original features.

So does the comment I saw refer some real change since the course was made, and if so, what is that change: A new algorithm that covers SVM's "sweet spot" just as well, better CPUs meaning SVM's computational advantages are not worth as much? Or is it perhaps opinion or personal experience of the commenter?

I tried a search for e.g. "are support vector machines out of fashion" and found nothing to imply they were being dropped in favour of anything else.

And Wikipedia has this: http://en.wikipedia.org/wiki/Support_vector_machine#Issues ... the main sticking point appears to be difficulty of interpreting the model. Which makes SVM fine for a black-box predicting engine, but not so good for generating insights. I don't see that as a major issue, just another minor thing to take into account when picking the right tool for the job (along with nature of the training data and learning task etc).

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [neil-slater](#)

[Answer](#) by [debasis](#)

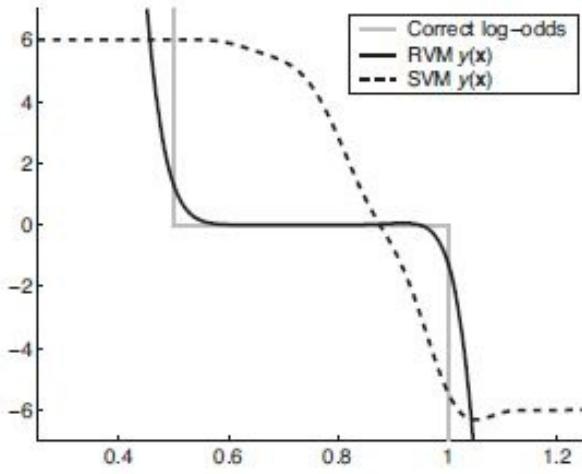
SVM is a powerful classifier. It has some nice advantages (which I guess were responsible for its popularity)... These are:

- Efficiency: Only the support vectors play a role in determining the classification boundary. All other points from the training set needn't be stored in memory.
- The so-called power of kernels: With appropriate kernels you can transform feature space into a higher dimension so that it becomes linearly separable. The notion of kernels work with arbitrary objects on which you can define some notion of similarity with the help of inner products... and hence SVMs can classify arbitrary objects such as trees, graphs etc.

There are some significant disadvantages as well.

- Parameter sensitivity: The performance is highly sensitive to the choice of the regularization parameter C, which allows some variance in the model.
- Extra parameter for the Gaussian kernel: The radius of the Gaussian kernel can have a significant impact on classifier accuracy. Typically a grid search has to be conducted to find optimal parameters. LibSVM has a support for grid search.

SVMs generally belong to the class of "Sparse Kernel Machines". The sparse vectors in the case of SVM are the support vectors which are chosen from the maximum margin criterion. Other sparse vector machines such as the **Relevance Vector Machine** (RVM) perform better than SVM. The following figure shows a comparative performance of the two. In the figure, the x-axis shows one dimensional data from two classes $y=\{0,1\}$. The mixture model is defined as $P(x|y=0)=\text{Unif}(0,1)$ and $P(x|y=1)=\text{Unif}(.5,1.5)$ (Unif denotes uniform distribution). 1000 points were sampled from this mixture and an SVM and an RVM were used to estimate the posterior. The problem of SVM is that the predicted values are far off from the true log odds.



A very effective classifier, which is very popular nowadays, is the **Random Forest**. The main advantages are:

- Only one parameter to tune (i.e. the number of trees in the forest)
- Not utterly parameter sensitive
- Can easily be extended to multiple classes
- Is based on probabilistic principles (maximizing mutual information gain with the help of decision trees)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

[Q: NASDAQ Trade Data](#)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

I am trying to find stock data to practice with, is there a good resource for this? I found this: <ftp://emi.nasdaq.com/ITCH/> but it only has the current year.

I already have a way of parsing the protocol, but would like to have some more data to compare with. It doesn't have to be in the same format, as long as it has price, trades, and date statistics.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [marin](#) 

[Answer](#)  by [mike1886](#) 

You can pull stock data very easily in python and R (probably other languages as well) with the following packages:

In python with: <https://pypi.python.org/pypi/ystockquote> 

This is also a really nice tutorial in iPython which shows you how to pull the stock data and play with it: <http://nbviewer.ipython.org/github/twiecki/financial-analysis-python-tutorial/blob/master/1.%20Pandas%20Basics.ipynb> 

In R with: <http://www.quantmod.com/> 

HTH.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is the use of user data collection besides serving ads?](#)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#))

Well this looks like the most suited place for this question.

Every website collects data of the user, some just for usability and personalization, but the majority like social networks track every move on the web, some free apps on your phone scan text messages, call history and so on.

All this data siphoning is just for selling your profile for advertisers?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [gleissongraca](#) 

[Answer](#)  by [ffriend](#) 

A couple of days ago developers from one product company asked me how they can understand why new users were leaving their website. My first question to them was what these users' profiles looked like and how they were different from those who stayed.

Advertising is only top of an iceberg. User profiles (either filled by users themselves or computed from users' behaviour) hold information about:

- **user categories**, i.e. what kind of people tend to use your website/product
- **paying client portraits**, i.e. who is more likely to use your paid services
- **UX component performance**, e.g. how long it takes people to find the button they need
- **action performance comparison**, e.g. what was more efficient - lower price for a weekend or propose gifts with each buy, etc.

So it's more about improving product and making better user experience rather than selling this data to advertisers.

[Answer](#)  by [steve-kallestad](#) 

Most companies won't sell the data, not on any small scale anyways. Most will use it internally.

User tracking data is important for understanding a lot of things. There's basic A/B testing where you provide different experiences to see which is more effective. There is understanding how your UI is utilized. Categorizing your end users in different ways for a variety of reasons. Figuring out where your end user base is, and within that group where the end users that matter are. Correlating user experiences with social network updates. Figuring out what will draw people to your product and what drives them away. The list of potential for data mining and analysis projects could go on for days.

Data storage is cheap. If you track everything out of the gate, you can figure out what you want to do with that data later.

Scanning text messages is sketchy territory when there isn't a good reason for it. Even when there is a good reason it's sketchy territory. I'd love to say that nobody does it, but there have been instances where big companies have done it and there are a lot of cases where no-name apps at least require access to that kind of data for installation. I generally frown on that kind of thing myself as a consumer, but the data analyst in me would love to see if I could build anything useful from a set of information like that.

[Answer](#)  by [gallamine](#) 

Here's a practical example of using web data for something other than advertising. Distil Networks (disclaimer, I work there) uses network traffic to determine whether page accesses are from humans or bots - scrapers, click fraud, form spam, etc.

Another example is some of the work that Webtrends is doing. They allow site users to build a model for each visitor to predict whether they'll leave, buy, add to cart, etc. Then based on the probability of each action you can change the users experience (e.g. if they're about to leave, give them a coupon). Here's the slides from a talk by them:

<http://www.oscon.com/oscon2014/public/schedule/detail/34809> 

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: Why might several types of models give almost identical results?](#)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

I've been analyzing a data set of ~400k records and 9 variables. The dependent variable is binary. I've fitted a logistic regression, a regression tree, a random forest, and a gradient boosted tree. All of them give virtual identical goodness of fit numbers when I validate them on another data set.

Why is this so? I'm guessing that it's because my observations to variable ratio is so high. If this is correct, at what observation to variable ratio will different models start to give different results?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [andy-blankertz](#)

[Answer](#) by [stask](#)

This results means that whatever method you use, you are able to get reasonably close to the optimal decision rule (aka [Bayes rule](#)). The underlying reasons have been explained in Hastie, Tibshirani and Friedman's "[Elements of Statistical Learning](#)". They demonstrated how the different methods perform by comparing Figs. 2.1, 2.2, 2.3, 5.11 (in my first edition — in section on multidimensional splines), 12.2, 12.3 (support vector machines), and probably some others. If you have not read that book, you need to drop everything **RIGHT NOW** and read it up. (I mean, it isn't worth losing your job, but it is worth missing a homework or two if you are a student.)

I don't think that observations to variable ratio is the explanation. In light of my rationale offered above, it is the relatively simple form of the boundary separating your classes in the multidimensional space that all of the methods you tried have been able to identify.

[Answer](#) by [seanv507](#)

its worth also looking at the training errors.

basically I disagree with your analysis. if logistic regression etc are all giving the same results it would suggest that the 'best model' is a very simple one (that all models can fit equally well - eg basically linear).

So then the question might be why is the best model a simple model?: It might suggest that your variables are not very predictive. Its of course hard to analyse without knowing the data.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [classification](#) ([Prev Q](#)) ([Next Q](#))

Q: K-means: What are some good ways to choose an efficient set of initial centroids?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

When a random initialization of centroids is used, different runs of K-means produce different total SSEs. And it is crucial in the performance of the algorithm. What are some effective approaches toward solving this problem? Recent approaches are appreciated.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [ngub05](#)

[Answer](#) by [ryan-j.-smith](#)

An approach that yields more consistent results is [K-means++](#). This approach acknowledges that there is probably a better choice of initial centroid locations than simple random assignment. Specifically, K-means tends to perform better when centroids are seeded in such a way that doesn't clump them together in space.

In short, the method is as follows:

1. Choose one of your data points at random as an initial centroid.
2. Calculate $D(x)$, the distance between your initial centroid and all other data points, x .
3. Choose your next centroid from the remaining datapoints with probability proportional to $D(x)^2$
4. Repeat until all centroids have been assigned.

Note: $D(x)$ should be updated as more centroids are added. It should be set to be the distance between a data point and the nearest centroid.

You may also be interested to read [this paper](#) that proposes the method and describes its overall expected performance.

[Answer](#) by [jake-c.](#)

I may be misunderstanding your question, but usually k-means chooses your centroids randomly for you depending on the number of clusters you set (i.e. k). Choosing the number for k tends to be a subjective exercise. A good place to start is an Elbow/Scree plot which can be found here:

http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set#The_Elbow_Method

[Answer](#) by [pablo-suau](#)

The usual approach to this problem is to re-run your K-means algorithm several times, with different random initializations of the centroids, and to keep the best solution. You can do that by evaluating the results on your training data or by means of cross validation.

There are many other ways to initialize the centroids, but none of them is going to perform the best for every single problem. You could evaluate these approaches together with random initialization for your particular problem.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [clustering](#) ([Prev Q](#)) ([Next Q](#))

Q: What kind of research can be done with an email data set?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#))

I found a data set called [Enron Email Dataset](#) . It is possibly the only substantial collection of “real” email that is public. I found some prior analysis of this work:

- A paper describing the Enron data was presented at the 2004 CEAS conference.
- Some experiments associated with this data are described on Ron Bekkerman’s home page
- [Parakweet](#)  has released an open source set of Enron sentence data, labeled for speech acts.
- Work at the University of Pennsylvania includes a query dataset for email search as well as a tool for generating spelling errors based on the Enron corpus.

I’m looking for some interesting current trend topics to work with. please give me some suggestions.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [miller](#) 

[Answer](#)  by [student-t](#) 

You’re learning, are you? Try to find something easy and interesting to start. Why don’t you start off something easy like building a Bayesian model to predict which email will get deleted. You should glance over those deleted emails, are they spams? are they just garbage?

Here, you have a simply supervised model where the data-set already labels the emails for you (deleted or not). Think of something easy like words, titles, length of the email etc, see if you can build a model that predicts email deletion.

[Answer](#)  by [sree-harish-venu](#) 

The following are some research that can done on e-mail dataset:

- linguistic analysis to abbreviate an email message
- Categorize e-mail as spam/ham using machine learning techniques.
- identifying concepts expressed in a collection of email messages, and organizing them into an ontology or taxonomy for browsing

[Answer](#) by [mansoor-alam](#)

Wonderful dataset with many opportunities to brush up on text analysis skills!

My first thought would be to try some Topic Modelling on the dataset. If you are using Python there is a library I've used called [gensim](#) which has some fairly thorough tutorials to get you started. A friend of mine [did something similar](#) with the Enron dataset, using parallelized preprocessing and distributed latent Dirichlet allocation to infer topics over the email corpus.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: LinkedIn web scraping](#)

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#))

I recently discovered a [new R package](#) for connecting to the LinkedIn API. Unfortunately the LinkedIn API seems pretty limited to begin with; for example, you can only get basic data on companies, and this is detached from data on individuals. I'd like to get data on all employees of a given company, which you can do [manually on the site](#) but is not possible through the API.

[import.io](#) would be perfect if it [recognised the LinkedIn pagination](#) (see end of page).

Does anyone know any web scraping tools or techniques applicable to the current format of the LinkedIn site, or ways of bending the API to carry out more flexible analysis? Preferably in R or web based, but certainly open to other approaches.

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#))

User: [polyphant](#)

[Answer](#) by [j.a.gartner](#)

Beautiful Soup is specifically designed for web crawling and scraping, but is written for python and not R:

<http://www.crummy.com/software/BeautifulSoup/bs4/doc/>

[Answer](#) by [itdxer](#)

[Scrapy](#) is a great Python library which can help you scrape different sites faster and make your code structure better. Not all sites can be parsed with classic tools, because they can use dynamic JS content building. For this task it is better to use [Selenium](#) (This is a test framework for web sites, but it also a great web scraping tool). There's also a [Python wrapper](#) available for this library. In Google you can find a few tricks which can help you use Selenium inside [Scrapy](#) and make your code clear, organized, and you can use some great tools for [Scrapy](#) library.

I think that Selenium would be a better scraper for LinkedIn than classic tools. There is a lot of javascript and dynamic content. Also, if you want to make authentication in your

account and scrape all available content, you will get a lot of problems with classic authentication using simple libraries like [requests](#) or [urllib](#).

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#))

Q: Web services to mine the social web?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#))

Are there any web services that can be used to analyse data in social networks with respect to a specific research question (e.g. mentioning of certain products in social media discussions)?

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#))

User: [orschiro](#)

[Answer](#) by [wacax](#)

Twitter's API is one of the best sources of social network data. You can extract off twitter pretty much everything you can imagine, you just need an account and a developer ID. The documentation is rather big so I will let you navigate it.

<https://dev.twitter.com/overview/documentation>

As usual there are wrappers that make your life easier.

- [python-twitter](#)
- [twitteR](#)

There are also companies who offer detailed twitter analytics and historic datasets for a fee.

- [Gnip](#)
- [Datasift](#)

Check them out!

[Answer](#) by [dawny33](#)

[Tweepy](#) is one of the best libraries for analyzing and hacking around with the Twitter API. (Being a contributor for tweepy, I can vouch for its stability and quality)

For a Python wrapper for the Facebook graph API, you can use the [Facebook-Insights library](#), which is well-maintained and neat documentation.

There are services out there which can mine you information, but they are limited to the complexity of the query.

For example: "How many people tweeted about banana on Monday?" can be answered from existing tools

“How many people happy people tweeted about banana on Monday” can also be done, but would require more efforts and the software or tool should be able to detect emotion.

So, if you are into research, I would advise you to go with the API’s and good ol’ programming!

Tags: [data-mining](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is a Recurrent Heavy Subgraph?](#)

Tags: [data-mining](#) ([Prev Q](#)), [graphs](#) ([Next Q](#))

I recently came across this term recurrent heavy subgraph in a talk. I don’t seem to understand what it means and Google doesn’t seem to show any good results. Can someone explain what this means in detail.

Tags: [data-mining](#) ([Prev Q](#)), [graphs](#) ([Next Q](#))

User: [mohammad-ghazanfar](#)

[Answer](#) by [thomas-cleberg](#)

The term may best be expressed as a **Recurrent, Heavy Subgraph**. That is, a subgraph which is both **Recurrent and Heavy**.

Heaviness of a subgraph refers to heavily connected vertices- that is, nodes which are connected many times (“many” being relative to the network in question).

Recurrent refers to the propensity of a subgraph to occur more than once.

Thus, a **Recurrent Heavy Subgraph** is a densely connected set of vertices which occurs several times in the overall network.

These subgraphs are often used to determine properties of a network.

For example: In a network of emails interactions within a company organized into 4-person teams with one member acting as the lead, each team’s email activity (if they email between themselves sufficiently to be considered “heavy”) could be described as a **Heavy Subgraph**. The fact that these subgraphs occur many times in the network make them **Recurrent Heavy Subgraphs**. If one was searching for structure in the network, noticing that these recurrent, heavy subgraphs exist would go a long way toward determining the organization of the network as a whole.

Tags: [data-mining](#) ([Prev Q](#)), [graphs](#) ([Next Q](#))

Classification

Questions

Q: Using SVM as a binary classifier, is the label for a data point chosen by consensus? 

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

I'm learning [Support Vector Machines](#) , and I'm unable to understand how a class label is chosen for a data point in a binary classifier. Is it chosen by consensus with respect to the classification in each dimension of the separating hyperplane?

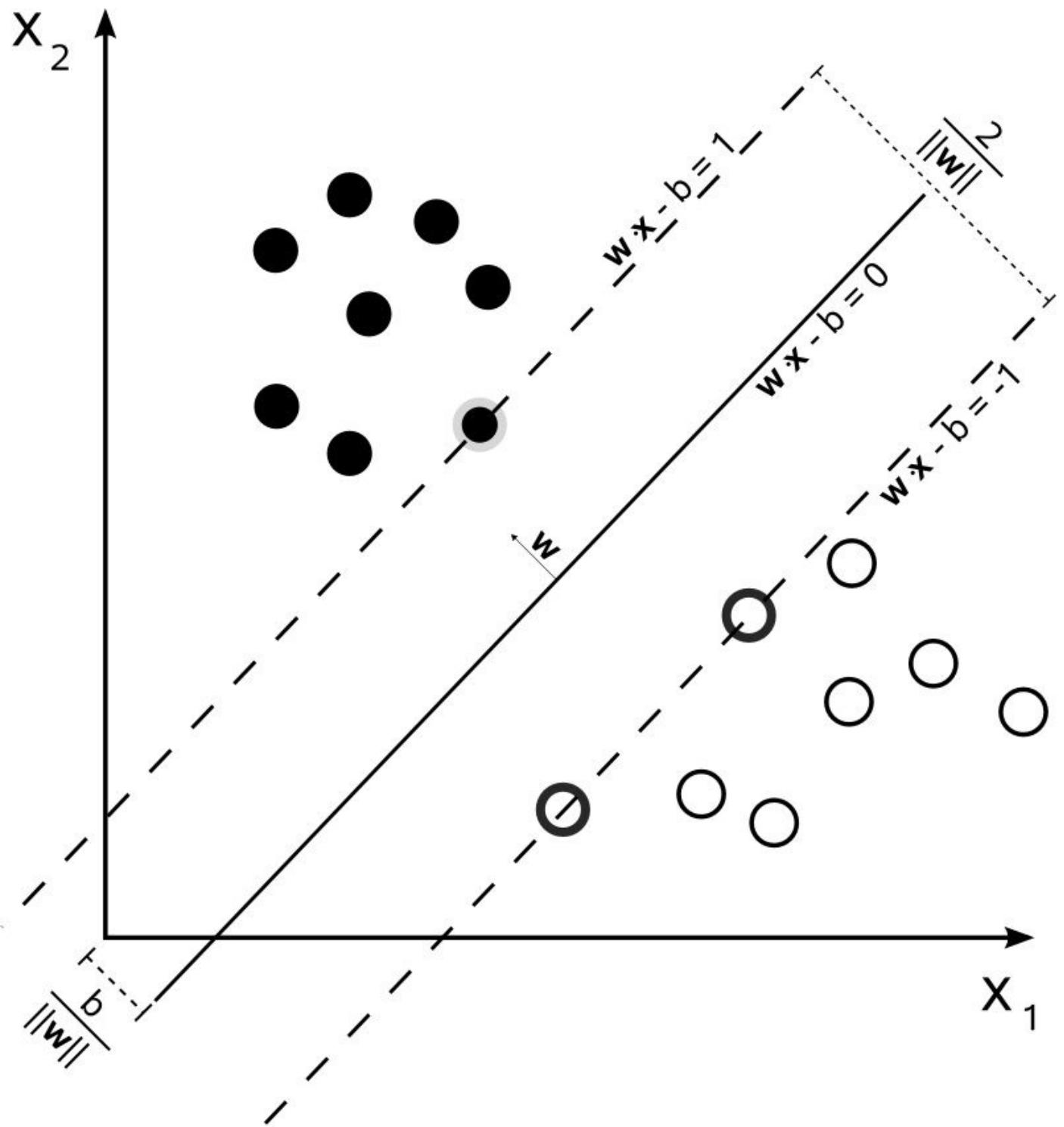
Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [fbrundu](#) 

[Answer](#)  by [rubens](#) 

The term *consensus*, as far as I'm concerned, is used rather for cases when you have more than one source of metric/measure/choice from which to make a decision. And, in order to choose a possible result, you perform some *average evaluation/consensus* over the values available.

This is not the case for SVM. The algorithm is based on a [quadratic optimization](#) , that maximizes the distance from the closest documents of two different classes, using a hyperplane to make the split.



So, the only *consensus* here is the resulting hyperplane, computed from the closest documents of each class. In other words, the classes are attributed to each point by calculating the distance from the point to the hyperplane derived. If the distance is positive, it belongs to a certain class, otherwise, it belongs to the other one.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

Q: Choose binary classification algorithm

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

I have a binary classification problem:

- Approximately 1000 samples in training set
- 10 attributes, including binary, numeric and categorical

Which algorithm is the best choice for this type of problem?

By default I'm going to start with SVM (preliminary having nominal attributes values converted to binary features), as it is considered the best for relatively clean and not noisy data.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [ihars](#) 

[Answer](#)  by [indico](#) 

It's hard to say without knowing a little more about your dataset, and how separable your dataset is based on your feature vector, but I would probably suggest using extreme random forest over standard random forests because of your relatively small sample set.

Extreme random forests are pretty similar to standard random forests with the one exception that instead of optimizing splits on trees, extreme random forest makes splits at random. Initially this would seem like a negative, but it generally means that you have significantly better generalization and speed, though the AUC on your training set is likely to be a little worse.

Logistic regression is also a pretty solid bet for these kinds of tasks, though with your relatively low dimensionality and small sample size I would be worried about overfitting. You might want to check out using K-Nearest Neighbors since it often performs very well with low dimensionalities, but it doesn't usually handle categorical variables very well.

If I had to pick one without knowing more about the problem I would certainly place my bets on extreme random forest, as it's very likely to give you good generalization on this kind of dataset, and it also handles a mix of numerical and categorical data better than most other methods.

[Answer](#)  by [neone4373](#) 

For low parameters, pretty limited sample size, and a binary classifier logistic regression should be plenty powerful enough. You can use a more advanced algorithm but it's probably overkill.

[Answer](#)  by [sean-owen](#) 

When categorical variables are in the mix, I reach for Random Decision Forests, as it handles categorical variables directly without the 1-of-n encoding transformation. This loses less information.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

[Q: Large Scale Personalization - Per User vs Global Models](#)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

I'm currently working on a project that would benefit from personalized predictions. Given an input document, a set of output documents, and a history of user behavior, I'd like to predict which of the output documents are clicked.

In short, I'm wondering what the typical approach to this kind of personalization problem is. Are models trained per user, or does a single global model take in summary statistics of past user behavior to help inform that decision? Per user models won't be accurate until the user has been active for a while, while most global models have to take in a fixed length feature vector (meaning we more or less have to compress a stream of past events into a smaller number of summary statistics).

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [madison-may](#) 

[Answer](#)  by [trey](#) 

The answer to this question is going to vary pretty wildly depending on the size and nature of your data. At a high level, you could think of it as a special case of multilevel models; you have the option of estimating a model with complete pooling (i.e., a universal model that doesn't distinguish between users), models with no pooling (a separate model for each user), and partially pooled models (a mixture of the two). You should really read Andrew Gelman on this topic if you're interested.

You can also think of this as a learning-to-rank problem that either tries to produce point-wise estimates using a single function or instead tries to optimize on some list-wise loss function (e.g., NDCG).

As with most machine learning problems, it all depends on what kind of data you have, the quality of it, the sparseness of it, and what kinds of features you are able to extract from it. If you have reason to believe that each and every user is going to be pretty unique in their behavior, you might want to build a per-user model, but that's going to be unwieldy fast — and what do you do when you are faced with a new user?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

[Q: Which non-training classification methods are available?](#)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

I am trying to find which classification methods, that do not use a training phase, are available.

The scenario is gene expression based classification, in which you have a matrix of gene expression of m genes (features) and n samples (observations). A signature for each class is also provided (that is a list of the features to consider to define to which class belongs a

sample).

An application (non-training) is the [Nearest Template Prediction](#) method. In this case it is computed the cosine distance between each sample and each signature (on the common set of features). Then each sample is assigned to the nearest class (the sample-class comparison resulting in a smaller distance). No already classified samples are needed in this case.

A different application (training) is the [kNN](#) method, in which we have a set of already labeled samples. Then, each new sample is labeled depending on how are labeled the k nearest samples.

Are there any other non-training methods?

Thanks

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [fbrundu](#)

[Answer](#) by [bogatron](#)

What you are asking about is [Instance-Based Learning](#). k-Nearest Neighbors (kNN) appears to be the most popular of these methods and is applicable to a wide variety of problem domains. Another general type of instance-based learning is [Analogical Modeling](#), which uses instances as exemplars for comparison with new data.

You referred to kNN as an application that uses training but that is not correct (the Wikipedia entry you linked is somewhat misleading in that regard). Yes, there are “training examples” (labeled instances) but the classifier doesn’t learn/train from these data. Rather, they are only used whenever you actually want to classify a new instance, which is why it is considered a “lazy” learner.

Note that the Nearest Template Prediction method you mention effectively is a form of kNN with k=1 and cosine distance as the distance measure.

[Answer](#) by [sharon](#)

nsl- I’m a beginner at machine learning, so forgive the lay-like description here, but it sounds like you might be able to use topic modelling, like latent dirichlet analysis (LDA). It’s an algorithm widely used to classify documents, according to what topics they are about, based on the words found and the relative frequencies of those words in the overall corpus. I bring it up mainly because, in LDA it’s not necessary to define the topics in advance.

Since the help pages on LDA are mostly written for text analysis, the analogy I would use, in order to apply it to your question, is:

- Treat each gene expression, or feature, as a ‘word’ (sometimes called a token in typical LDA text-classification applications)
- Treat each sample as a document (ie it contains an assortment of words, or gene expressions)
- Treat the signatures as pre-existing topics

If I’m not mistaken, LDA should give weighted probabilities for each topic, as to how strongly it is present in each document.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

Q: How to define a custom resampling methodology

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#))

I'm using an experimental design to test the robustness of different classification methods, and now I'm searching for the correct definition of such design.

I'm creating different subsets of the full dataset by cutting away some samples. Each subset is created independently with respect to the others. Then, I run each classification method on every subset. Finally, I estimate the accuracy of each method as how many classifications on subsets are in agreement with the classification on the full dataset. For example:

Classification-full	1	2	3	2	1	1	2
Classification-subset1	1	2		2	3	1	
Classification-subset2		2	3		1	1	2...
Accuracy	1	1	1	1	0.5	1	1

Is there a correct name to this methodology? I thought it can fall under [bootstrapping](#)  but I'm not sure about this.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#))

User: [fbrundu](#) 

[Answer](#)  by [rhand](#) 

Random subsampling seems appropriate, bootstrapping is a bit more generic, but also correct.

Here are some references and synonyms: http://www.frank-dieterle.com/phd/2_4_3.html 

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#))

Q: Skewed multi-class data

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

I have a dataset which contains ~100,000 samples of 50 classes. I have been using SVM with an RBF kernel to train and predict new data. The problem though is the dataset is skewed towards different classes.

For example, Class 1 - 30 (~3% each), Class 31 - 45 (~0.6% each), Class 46 - 50 (~0.2% each)

I see that the model tends to very rarely predict the classes which occur less frequent in the training set, even though the test set has the same class distribution as the training set.

I am aware that there are techniques such as ‘undersampling’ where the majority class is scaled down to the minor class. However, is this applicable here where there are so many different classes? Are there other methods to help handle this case?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

User: [mike1886](#) 

[Answer](#)  by [magicharp](#) 

I would suggest you to use libsvm, which already has adjustable class weights implemented in it. Rather than replicating the training samples, one modifies the C parameter for different classes in the SVM optimization. For example if your data has 2 classes, and the first class is only 10% of the data, you would choose class weights to be 10 and 1 for class 1 and 2 respectively. Therefore, margin violations of the first class would cost 10 times more than the margin violations for second class, and per-class accuracies would be more balanced.

[Answer](#)  by [sebastian-raschka](#) 

I am not an expert in using SVMs, but usually (if you are using a machine learning library like Python’s `scikit-learn` or R’s `libsvm`, there is the `class_weight` parameter, or `class.weights`, respectively).

Or if you’d use a Bayes classifier, you would take this “skew” into account via the “prior (class) probabilities” $P(\omega_j)$

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [svm](#) ([Prev Q](#)) ([Next Q](#))

[Q: Cosine similarity versus dot product as distance metrics](#)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

It looks like the cosine similarity of two features is just their dot product scaled by the product of their magnitudes. When does cosine similarity make a better distance metric than the dot product? I.e. do the dot product and cosine similarity have different strengths or weaknesses in different situations?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [ahoffer](#) 

[Answer](#)  by [memming](#) 

Think geometrically. Cosine similarity only cares about angle difference, while dot product cares about angle and magnitude. If you normalize your data to have the same magnitude, the two are indistinguishable. Sometimes it is desirable to ignore the magnitude, hence cosine similarity is nice, but if magnitude plays a role, dot product would be better as a similarity measure. Note that neither of them is a “distance metric”.

[Answer](#) by [sobach](#)

You are right, cosine similarity has a lot of common with dot product of vectors. Indeed, it is a dot product, scaled by magnitude. And because of scaling it is normalized between 0 and 1. CS is preferable because it takes into account variability of data and features' relative frequencies. On the other hand, plain dot product is a little bit "cheaper" (in terms of complexity and implementation).

[Answer](#) by [vimal-chandra-gorijala](#)

I would like to add one more dimension to the answers given above. Usually we use cosine similarity with large text, because using distance matrix on paragraphs of data is not recommended. And also if you intend your cluster to be broad you tend to go with cosine similarity as it captures similarity overall.

For example if you have texts which are two or three words long at max I feel using cosine similarity does not achieve the precision as achieved by distance metric.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

Q: Which cross-validation type best suits to binary classification problem

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [cross-validation](#) ([Prev Q](#)) ([Next Q](#))

Data set looks like:

- 25000 observations
- up to 15 predictors of different types: numeric, multi-class categorical, binary
- target variable is binary

Which cross validation method is typical for this type of problems?

By default I'm using K-Fold. How many folds is enough in this case? (One of the models I use is random forest, which is time consuming...)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [cross-validation](#) ([Prev Q](#)) ([Next Q](#))

User: [ihars](#)

[Answer](#) by [damienfrancois](#)

You will have best results if you care to build the folds so that each variable (and most importantly the target variable) is approximately identically distributed in each fold. This is called, when applied to the target variable, stratified k-fold. One approach is to cluster the inputs and make sure each fold contains the same number of instances from each cluster proportional to their size.

[Answer](#) by [pasmod-turing](#)

I think in your case a 10-fold CV will be O.K.

I think it is more important to randomize the cross validation process than selecting the ideal value for k.

So repeat the CV process several times randomly and compute the variance of your classification result to determine if the results are reliable or not.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [cross-validation](#) ([Prev Q](#)) ([Next Q](#))

Q: what is difference between text classification and topic models?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

I know the difference between clustering and classification in machine learning, but I don't understand the difference between text classification and topic modeling for documents. Can I use topic modeling over documents to identify a topic? Can I use classification methods to classify the text inside these documents?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [ali](#)

[Answer](#) by [sean-owen](#)

Text Classification

I give you a bunch of documents, each of which has a label attached. I ask you to learn why you think the contents of the documents have been given these labels based on their words. Then I give you new documents and ask what you think the label for each one should be. The labels have meaning to me, not to you necessarily.

Topic Modeling

I give you a bunch of documents, without labels. I ask you to explain why the documents have the words they do by identifying some topics that each is “about”. You tell me the topics, by telling me how much of each is in each document, and I decide what the topics “mean” if anything.

You’d have to clarify what you mean by “identify one topic” or “classify the text”.

[Answer](#) by [charlie-greenbacker](#)

But I don’t know what is difference between text classification and topic models in documents

Text classification is a form of supervised learning — the set of possible classes are known/defined in advance and don’t change.

Topic modeling is a form of unsupervised learning (akin to clustering) — the set of possible topics are unknown apriori. They’re defined as part of generating the topic models. With a non-deterministic algorithm like LDA, you’ll get different topics each time you run the algorithm.

Text classification often involves mutually-exclusive classes — think of these as buckets. But it doesn’t have to — given the right kind of labeled input data, you can set of a series of non-mutually-exclusive binary classifiers.

Topic modeling is generally not mutually-exclusive — the same document can have its probability distribution spread across many topics. In addition, there are also hierarchical topic modeling methods, etc.

Also can I use topic model for the documents to identify one topic later on can I use the classification to classify the text inside this documents ?

If you’re asking whether you can take all of the documents assigned to one topic by a topic modeling algorithm and then apply a classifier to that collection, then yes, you certainly can do that. I’m not sure it makes much sense, though — at a minimum, you’d need to pick a threshold for the topic probability distribution above which you’ll include documents in your collection (typically 0.05-0.1). Can you elaborate on your use case?

By the way, there’s a great tutorial on topic modeling using the MALLET library for Java available here: [Getting Started with Topic Modeling and MALLET](#)

[Answer](#) by [erich-schubert](#)

Topic models are usually **unsupervised**. There are “supervised topic models”, too; but even then they try to model **topics within a classes**.

E.g. you may have a class “football”, but there may be topics inside this class that relate to particular matches or teams.

The challenge with topics is that they change over time; consider the matches example above. Such topics may emerge, and disappear again.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: Difference between tf-idf and tf with Random Forests](#)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

I am working on a text classification problem using Random Forest as classifiers, and a bag-of-words approach. I am using the basic implementation of Random Forests (the one present in scikit), that creates a binary condition on a single variable at each split. Given this, is there a difference between using simple tf (term frequency) features, where each word has an associated weight that represents the number of occurrences in the document, or tf-idf (term frequency * inverse document frequency), where the term frequency is also multiplied by a value that represents the ratio between the total number of documents and the number of documents containing the word)?

In my opinion, there should not be any difference between these two approaches, because the only difference is a scaling factor on each feature, but since the split is done at the level of single features this should not make a difference.

Am I right in my reasoning?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [markusian](#)

[Answer](#) by [alexey-grigorev](#)

Decision trees (and hence Random Forests) are insensitive to monotone transformations of input features.

Since multiplying by the same factor is a monotone transformation, I'd assume that for Random Forests there indeed is no difference.

However, you eventually may consider using other classifiers that do not have this property, so it may still make sense to use the entire TF * IDF.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: Document classification: tf-idf prior to or after feature filtering?](#)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

I have a document classification project where I am getting site content and then assigning one of numerous labels to the website according to content.

I found out that [tf-idf](#) could be very useful for this. However, I was unsure as to *when* exactly to use it.

Assuming a website that is concerned with a specific topic makes repeated mention of it, this was my current process:

1. Retrieve site content, parse for plain text
2. Normalize and stem content
3. Tokenize into unigrams (maybe bigrams too)
4. Retrieve a count of each unigram for the given document, filtering low length and low occurrence words
5. Train a classifier such as NaiveBayes on the resulting set

My question is the following: **Where would tf-idf fit in here?** Before normalizing/stemming? After normalizing but before tokenizing? After tokenizing?

Any insight would be greatly appreciated.

Edit:

Upon closer inspection, I think I may have run into a misunderstanding as to how TF-IDF operates. At the above step 4 that I describe, would I have to feed the *entirety* of my data into TF-IDF at once? If, for example, my data is as follows:

```
[({tokenized_content_site1}, category_string_site1),
 ({tokenized_content_site2}, category_string_site2),
 ...
 ({tokenized_content_siten}, category_string_siten)}]
```

Here, the outermost structure is a list, containing tuples, containing a dictionary (or hashmap) and a string.

Would I have to feed the *entirety* of that data into the TF-IDF calculator at once to achieve the desired effect? Specifically, I have been looking at the [scikit-learn](#) `TfidfVectorizer` to do this, but I am a bit unsure as to its use as examples are pretty sparse.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

User: [user991710](#)

[Answer](#) by [charlie-greenbacker](#)

As you've described it, Step 4 is where you want to use TF-IDF. Essentially, TD-IDF will count each term in each document, and assign a score given the relative frequency across the collection of documents.

There's one big step missing from your process, however: annotating a training set. Before you train your classifier, you'll need to manually annotate a sample of your data with the labels you want to be able to apply automatically using the classifier.

To make all of this easier, you might want to consider using the [Stanford Classifier](#). It will perform the feature extraction and build the classifier model (supporting several different machine learning algorithms), but you'll still need to annotate the training data by hand.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#)), [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

[Q: K nearest neighbour](#)

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

Is the k-nearest neighbour algorithm a discriminative or a generative classifier? My first thought on this was that it was generative, because it actually used Bayes's theorem to compute the posterior. Searching further for this, it seems like it is a discriminative model, but I couldn't find the explanation.

So is KNN discriminative first of all? And if it is, is that because it doesn't model the priors or the likelihood?

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

User: [101](#)

[Answer](#) by [nitesh](#)

See a similar answer [here](#). To clarify, k nearest neighbor is a discriminative classifier.

The difference between a generative and a discriminative classifier is that the former models the joint probability where as the latter models the conditional probability (the posterior) starting from the prior.

In the case of nearest neighbors, the conditional probability of a class given a data point is modeled. To do this, one starts with the prior probability on the classes.

Tags: [classification](#) ([Prev Q](#)) ([Next Q](#))

[Q: Decision tree or logistic regression?](#)

Tags: [classification](#) ([Prev Q](#))

I am working on a classification problem. I have a dataset containing equal number of categorical variables and continuous variables. How will i know what technique to use? between a decision tree and a logistic regression?

Is it right to assume that logistic regression will be more suitable for continuous variable and decision tree will be more suitable for continuous + categorical variable?

Tags: [classification](#) ([Prev Q](#))

User: [arun](#)

[Answer](#) by [victor-ma](#)

Long story short: do what @untitledprogrammer said, try both models and cross-validate to help pick one.

Both decision trees (depending on the implementation, e.g. C4.5) and logistic regression should be able to handle continuous and categorical data just fine. For logistic regression, you'll want to [dummy code your categorical variables](#).

As @untitledprogrammer mentioned, it's difficult to know a priori which technique will be better based simply based on the types of features you have, continuous or otherwise. It really depends on your specific problem and the data you have. (See [No Free Lunch Theorem](#))

You'll want to keep in mind though that a logistic regression model is searching for a single linear decision boundary in your feature space, whereas a decision tree is essentially partitioning your feature space into half-spaces using *axis-aligned* linear decision boundaries. The net effect is that you have a non-linear decision boundary, possibly more than one.

This is nice when your data points aren't easily separated by a single hyperplane, but on the other hand, decisions trees are so flexible that they can be prone to overfitting. To combat this, you can try pruning. Logistic regression tends to be less susceptible (but not immune!) to overfitting.

Lastly, another thing to consider is that decision trees can automatically take into account interactions between variables, e.g. xy if you have two independent features x and y . With logistic regression, you'll have to manually add those interaction terms yourself.

So you have to ask yourself:

- what kind of decision boundary makes more sense in your particular problem?
- how do you want to balance bias and variance?
- are there interactions between my features?

Of course, it's always a good idea to just try both models and do cross-validation. This

will help you find out which one is more likely to have better generalization error.

[Answer](#) by [untitledprogrammer](#)

Try using both regression and decision trees. Compare the efficiency of each technique by using a 10 fold cross validation. Stick to the one with higher efficiency. It would be difficult to judge which method would be a better fit just by knowing that your dataset is continuous and, or categorical.

Tags: [classification](#) ([Prev Q](#))

Neuralnetwork

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Neural networks traditionally refer to a network or circuit of biological neurons. The modern usage of the term often refers to [artificial neural networks](#)  (ANN), which are composed of artificial neurons or nodes - programming constructs that mimic the properties of biological neurons. A set of weighted connections between the neurons allows information to propagate through the network to solve artificial intelligence problems without the network designer having had a model of a real system. The goal is good, or human-like, predictive ability.

Questions

[Q: Multi layer back propagation Neural network for classification](#)

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Can someone explain me, how to classify a data like MNIST with MLBP-Neural network if I make more than one output (e.g 8), I mean if I just use one output I can easily classify the data, but if I use more than one, which output should I choose ?

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [aldy-syahdeini](#) 

[Answer](#)  by [rapaio](#) 

Suppose that you need to classify something in K classes, where $K > 2$. In this case the most often setup I use is one hot encoding. You will have K output columns, and in the training set you will set all values to 0, except the one which has the category index, which could have value 1. Thus, for each training data set instance you will have all outputs with values 0 or 1, all outputs sum to 1 for each instance.

This looks like a probability, which reminds me of a technique used often to connect some outputs which are modeled as probability. This is called softmax function, more details [on Wikipedia](#) . This will allow you to put some constraints on the output values (it is basically a logistic function generalization) so that the output values will be modeled as probabilities.

Finally, with or without softmax you can use the output as a discriminant function to select the proper category.

Another final thought would be to avoid to encode your variables in a connected way. For example you can have the binary representation of the category index. This would induce to the learner an artificial connection between some outputs which are arbitrary. The one hot encoding has the advantage that is neutral to how labels are indexed.

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[Q: Any differences in regularisation in MLP between batch and individual updates?](#)

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

I have just learned about regularisation as an approach to control over-fitting, and I would like to incorporate the idea into a simple implementation of backpropagation and [Multilayer perceptron](#)  (MLP) that I put together.

Currently to avoid over-fitting, I cross-validate and keep the network with best score so far on the validation set. This works OK, but adding regularisation would benefit me in that correct choice of the regularisation algorithm and parameter would make my network

converge on a non-overfit model more systematically.

The formula I have for the update term (from Coursera ML course) is stated as a batch update e.g. for each weight, after summing all the applicable deltas for the whole training set from error propagation, an adjustment of `lambda * current_weight` is added as well before the combined delta is subtracted at the end of the batch, where `lambda` is the regularisation parameter.

My implementation of backpropagation uses per-item weight updates. I am concerned that I cannot just copy the batch approach, although it looks OK intuitively to me. *Does a smaller regularisation term per item work just as well?*

For instance `lambda * current_weight / N` where `N` is size of training set - at first glance this looks reasonable. I could not find anything on the subject though, and I wonder if that is because regularisation does not work as well with a per-item update, or even goes under a different name or altered formula.

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [neil-slater](#) 

[Answer](#)  by [insys](#) 

Regularization is relevant in per-item learning as well. I would suggest to start with a basic validation approach for finding out lambda, whether you are doing batch or per-item learning. This is the easiest and safest approach. Try manually with a number of different values. e.g. 0.001, 0.003, 0.01, 0.03, 0.1 etc. and see how your validation set behaves. Later on you may automate this process by introducing a linear or local search method.

As a side note, I believe the value of lambda should be considered in relation to the updates of the parameter vector, rather than the training set size. For batch training you have one parameter update *per dataset pass*, while for online one update *per sample* (regardless of the training set size).

I recently stumbled upon this [Crossvalidated Question](#) , which seems quite similar to yours. There is a link to a paper about [a new SGD algorithm](#) , with some relevant content. It might be useful to take a look (especially pages 1742-1743).

[Answer](#)  by [orelus](#) 

To complement what **insys** said :

Regularization is used when computing the backpropagation for all the weights in your MLP. Therefore, instead of computing the gradient in regard to all the input of the training set (batch) you only use some/one item(s) (stochastic or semi-stochastic). You will end up limiting a result of the update in regard to one item instead of all which is also correct.

Also, if i remember correctly, Andrew NG used L2-regularization. The `/N` in `lambda * current_weight / N` is not mandatory, it just helps rescaling the input. However if you choose not to use it, you will have (in most of the case) to select another value for `lambda`.

You can also use the [Grid-search algorithm](#)  to choose the best value for `lambda` (the

hyperparameter => the one you have to choose).

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to fight underfitting in a deep neural net](#)

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

When I started with artificial neural networks (NN) I thought I'd have to fight overfitting as the main problem. But in practice I can't even get my NN to pass the 20% error rate barrier. I can't even beat my score on random forest!

I'm seeking some very general or not so general advice on what should one do to make a NN start capturing trends in data.

For implementing NN I use Theano Stacked Auto Encoder with [the code from tutorial](#)  that works great (less than 5% error rate) for classifying the MNIST dataset. It is a multilayer perceptron, with softmax layer on top with each hidden later being pre-trained as autoencoder (fully described at [tutorial](#) , chapter 8). There are ~50 input features and ~10 output classes. The NN has sigmoid neurons and all data are normalized to [0,1]. I tried lots of different configurations: number of hidden layers and neurons in them (100->100->100, 60->60->60, 60->30->15, etc.), different learning and pre-train rates, etc.

And the best thing I can get is a 20% error rate on the validation set and a 40% error rate on the test set.

On the other hand, when I try to use Random Forest (from scikit-learn) I easily get a 12% error rate on the validation set and 25%(!) on the test set.

How can it be that my deep NN with pre-training behaves so badly? What should I try?

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [izhak](#) 

[Answer](#)  by [ffriend](#) 

The problem with deep networks is that they have lots of hyperparameters to tune and very small solution space. Thus, finding good ones is more like an art rather than engineering task. I would start with working example from tutorial and play around with its parameters to see how results change - this gives a good intuition (though not formal explanation) about dependencies between parameters and results (both - final and intermediate).

Also I found following papers very useful:

- [Visually Debugging Restricted Boltzmann Machine Training with a 3D Example](#) 
- [A Practical Guide to Training Restricted Boltzmann Machines](#) 

They both describe RBMs, but contain some insights on deep networks in general. For example, one of key points is that networks need to be debugged layer-wise - if previous

layer doesn't provide good representation of features, further layers have almost no chance to fix it.

[Answer](#) by [lmjohns3](#)

While ffriend's answer gives some excellent pointers for learning more about how neural networks can be (extremely) difficult to tune properly, I thought it might be helpful to list a couple specific techniques that are currently used in top-performing classification architectures in the neural network literature.

Rectified linear activations

The first thing that might help in your case is to switch your model's activation function from the [logistic sigmoid](#) — $f(z) = (1 + e^{-z})^{-1}$ — to a [rectified linear \(aka relu\)](#) — $f(z) = \max(0, z)$.

The relu activation has two big advantages:

- its output is a true zero (not just a small value close to zero) for $z \leq 0$ and
- its derivative is constant, either 0 for $z \leq 0$ or 1 for $z > 0$.

A network of relu units basically acts like an ensemble of exponentially many linear networks, because units that receive input $z \leq 0$ are essentially “off” (their output is 0), while units that receive input $z > 0$ collapse into a single linear model for that input. Also the constant derivatives are important because a deep network with relu activations tends to avoid the [vanishing gradient problem](#) and can be trained without layerwise pretraining.

See “Deep Sparse Rectifier Neural Networks” by Glorot, Bordes, & Bengio (<http://jmlr.csail.mit.edu/proceedings/papers/v15/glorot11a/glorot11a.pdf>) for a good paper about these topics.

Dropout

Many research groups in the past few years have been advocating for the use of “dropout” in classifier networks to avoid overtraining. (See for example “Dropout: A simple way to prevent neural networks from overfitting” by Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov <http://www.cs.toronto.edu/~hinton/absps/JMLRdropout.pdf>) In dropout, during training, some constant proportion of the units in a given layer are randomly set to 0 for each input that the network processes. This forces the units that aren't set to 0 to “make up” for the “missing” units. Dropout seems to be an extremely effective regularizer for neural network models in classification tasks. See a blog article about this at <http://fastml.com/regularizing-neural-networks-with-dropout-and-with-dropconnect/>.

[Answer](#) by [martin-thoma](#)

You might be interested in reading the following paper by researchers of Microsoft

Research:

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: [Deep Residual Learning for Image Recognition](#)  on arxiv, 2015.

They had similar problems as you had:

When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. **Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error**, as reported in [11, 42] and thoroughly verified by our experiments.

To solve the problem, they have made use of a skip architecture. With that, they trained very deep networks (1202 layers) and achieved the best result in the ILSVRC 2015 challenge.

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[Q: Neural Network parse string data?](#)

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

So, I'm just starting to learn how a neural network can operate to recognize patterns and categorize inputs, and I've seen how an artificial neural network can parse image data and categorize the images ([demo with convnetjs](#)), and the key there is to downsample the image and each pixel stimulates one input neuron into the network.

However, I'm trying to wrap my head around if this is possible to be done with string inputs? The use-case I've got is a "recommendation engine" for movies a user has watched. Movies have lots of string data (title, plot, tags), and I could imagine "downsampling" the text down to a few key words that describe that movie, but even if I parse out the top five words that describe this movie, I think I'd need input neurons for every english word in order to compare a set of movies? I could limit the input neurons just to the words used in the set, but then could it grow/learn by adding new movies (user watches a new movie, with new words)? Most of the libraries I've seen don't allow adding new neurons after the system has been trained?

Is there a standard way to map string/word/character data to inputs into a neural network? Or is a neural network really not the right tool for the job of parsing string data like this (what's a better tool for pattern-matching in string data)?

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [midnightlightning](#)

[Answer](#) by [madison-may](#)

Using a neural network for prediction on natural language data can be a tricky task, but there are tried and true methods for making it possible.

In the Natural Language Processing (NLP) field, text is often represented using the bag of words model. In other words, you have a vector of length n , where n is the number of words in your vocabulary, and each word corresponds to an element in the vector. In order to convert text to numeric data, you simply count the number of occurrences of each word and place that value at the index of the vector that corresponds to the word. [Wikipedia does an excellent job of describing this conversion process.](#) Because the length of the vector is fixed, its difficult to deal with new words that don't map to an index, but there are ways to help mitigate this problem (lookup [feature hashing](#)).

This method of representation has many disadvantages — it does not preserve the relationship between adjacent words, and results in very sparse vectors. Looking at [n-grams](#) helps to fix the problem of preserving word relationships, but for now let's focus on the second problem: sparsity.

It's difficult to deal directly with these sparse vectors (many linear algebra libraries do a poor job of handling sparse inputs), so often the next step is dimensionality reduction. For that we can refer to the field of [topic modeling](#): Techniques like [Latent Dirichlet Allocation](#) (LDA) and [Latent Semantic Analysis](#) (LSA) allow the compression of

these sparse vectors into dense vectors by representing a document as a combination of topics. You can fix the number of topics used, and in doing so fix the size of the output vector produced by LDA or LSA. This dimensionality reduction process drastically reduces the size of the input vector while attempting to lose a minimal amount of information.

Finally, after all of these conversions, you can feed the outputs of the topic modeling process into the inputs of your neural network.

[Answer](#) by [emre](#)

This is not a problem about neural networks per se, but about representing textual data in machine learning. You can represent the movies, cast, and theme as categorical variables. The plot is more complicated; you'd probably want a [topic model](#) for that, but I'd leave that out until you get the hang of things. It does precisely that textual "downsampling" you mentioned.

Take a look at [this](#) tutorial to learn how to encode categorical variables for neural networks. And good luck!

[Answer](#) by [jamesmf](#)

Both the answers from @Emre and @Madison May make good points about the issue at hand. The problem is one of representing your string as a feature vector for input to the NN.

First, the problem depends on the size of the string you want to process. Long strings containing many tokens (usually words) are often called documents in this setting. There are separate methods for dealing with individual tokens/words.

There are a number of ways to represent documents. Many of them make the [bag-of-words](#) assumption. The simplest types represent the document as a vector of the counts of words, or term frequency (tf). In order to eliminate the effects of document length, usually people prefer to normalize by the number of documents a term shows up in, document frequency ([tf-idf](#)).

Another approach is topic modeling, which learns a latent lower-dimensional representation of the data. [LDA](#) and [LSI/LSA](#) are typical choices, but it's important to remember this is unsupervised. The representation learned will not necessarily be ideal for whatever supervised learning you're doing with your NN. If you want to do topic modeling, you might also try [supervised topic models](#).

For individual words, you can use [word2vec](#), which leverages NNs to embed words into an arbitrary-sized space. Similarity between two word vectors in this learned space tends to correspond to semantic similarity.

A more recently pioneered approach is that of [paragraph vectors](#), which first learns a word2vec-like word model, then builds on that representation to learn a distributed representation of sets of words (documents of any size). This has shown state-of-the-art results in many applications.

When using NNs in NLP, people often use different architectures, like [Recurrent Neural](#)

Nets  (like [Long Short Term Memory](#)  networks). In [some cases](#)  people have even used [Convolutional Neural Networks](#)  on text.

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[Q: How are neural nets related to Fourier transforms?](#)

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

This is an interview question

How are neural nets related to Fourier transforms?

I could find papers that talk about methods to process the Discrete Fourier Transform (DFT) by a single-layer neural network with a linear transfer function. Is there some other correlation that I'm missing?

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [rishi-dua](#) 

[Answer](#)  by [emre](#) 

They are not related in any meaningful sense. Sure, you can use them both to extract features, or do any number of things, but the same can be said about many techniques. I would have asked "what kind of neural network?" to see if the interviewer had something specific in mind.

[Answer](#)  by [rundosrun](#) 

The similarity is regression. NNs can be used for regression and the fourier transform is in its heart just a curve fit of multiple sin and cos functions to some data.

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is conjugate gradient descent?](#)

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

What is Conjugate Gradient Descent of Neural Network? How is it different from Gradient Descent technique?

I came across a [resource](#)  but was unable to understand the difference between the two methods. It has mentioned in the procedure that

the next search direction is determined so that it is conjugate to previous search directions.

What does this sentence mean? And what is line search mentioned in the web page?

Can anyone please explain it with the help of a diagram?

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

User: [rishika](#) 

[Answer](#)  by [emre](#) 

What does this sentence mean?

It means that the next vector should be perpendicular to all the previous ones with respect to a matrix. It's like how the [natural basis](#)  vectors are [perpendicular](#)  to each other, with the added twist of a matrix:

$$x^T A y = 0 \text{ instead of } x^T y = 0$$

And what is line search mentioned in the webpage?

[Line search](#)  is an optimization method that involves guessing how far along a given direction (i.e., along a line) one should move to best reach the local minimum.

Tags: [neuralnetwork](#) ([Prev Q](#)) ([Next Q](#))

Q: Properties for building a Multilayer Perceptron Neural Network using Keras?

Tags: [neuralnetwork](#) ([Prev Q](#))

I am trying to build and train a multilayer perceptron neural network that correctly predicts what president won in what county for the first time. I have the following information for training data.

Total population Median age % BachelorsDeg or higher Unemployment rate Per capita income Total households Average household size % Owner occupied housing % Renter occupied housing % Vacant housing Median home value Population growth House hold growth Per capita income growth Winner

That's 14 columns of training data and the 15th column is what the output should be.

I am trying to use Keras to build a multilayer perceptron neural network, but I need some help understanding a few properties and the pros of cons of choosing different options for these properties.

1. ACTIVATION FUNCTION

I know my first step is to come up with an activation function. I always studied neural networks used sigmoid activation functions. Is a sigmoid activation function the best? How do you know which one to use? Keras additionally gives the options of using a softmax, softplus, relu, tanh, linear, or hard_sigmoid activation function. I'm okay with using whatever, but I just want to be able to understand why and the pros and cons.

2. PROBABILITY INITIALIZAIONS

I know initializations define the probability distribution used to set the initial random weights of Keras layers. The options Keras gives are uniform lecun_uniform, normal, identity, orthogonal, zero, glorot_normal, glorot_uniform, he_normal, and he_uniform. How does my selection here impact my end result or model? Shouldn't it not matter because we are "training" whatever random model we start with and come up with a more optimal weighting of the layers anyways?

Tags: [neuralnetwork](#) ([Prev Q](#))

User: [pr338](#)

[Answer](#) by [jamesmf](#)

1) Activation is an architecture choice, which boils down to a hyperparameter choice. You can make a theoretical argument for using any function, but the best way to determine this is to try several and evaluate on a validation set. It's also important to remember you can mix and match activations of various layers.

2) In theory yes, many random initializations would be the same if your data was extremely well behaved and your network ideal. But in practice initializations seek to

ensure the gradient starts off reasonable and the signal can be backpropagated correctly. Likely in this case any of those initializations would perform similarly, but the best approach is to try them out, switching if you get undesirable results.

Tags: [neuralnetwork](#) ([Prev Q](#))

Statistics

Questions

[Q: Is there a replacement for small p-values in big data?](#)

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

If small p-values are plentiful in big data, what is a comparable replacement for p-values in data with million of samples?

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [blunders](#) 

[Answer](#)  by [christopher-louden](#) 

There is no replacement in the strict sense of the word. Instead you should look at other measures.

The other measures you look at depend on what you type of problem you are solving. In general, if you have a small p-value, also consider the magnitude of the effect size. It may be highly statistically significant but in practice meaningless. It is also helpful to report the confidence interval of the effect size.

I would consider [this paper](#)  as mentioned in DanC's answer to [this question](#).

[Answer](#)  by [alex-i](#) 

See also [When are p-values deceptive?](#)

When there are a lot of variables that can be tested for pair-wise correlation (for example), the replacement is to use any of the corrections for [False discovery rate](#)  (to limit probability that any given discovery is false) or [Familywise error rate](#)  (to limit probability of one or more false discoveries). For example, you might use the Holm–Bonferroni method.

In the case of a large sample rather than a lot of variables, something else is needed. As Christopher said, magnitude of effect a way to treat this. Combining these two ideas, you might use a confidence interval around your magnitude of effect, and apply a false discovery rate correction to the p-value of the confidence interval. The effects for which even the lowest bound of the corrected confidence interval is high are likely to be strong effects, regardless of huge data set size. I am not aware of any published paper that combines confidence intervals with false discovery rate correction in this way, but it seems like a straightforward and intuitively understandable approach.

To make this even better, use a non-parametric way to estimate confidence intervals. Assuming a distribution is likely to give very optimistic estimates here, and even fitting a distribution to the data is likely to be inaccurate. Since the information about the shape of

the distribution past the edges of the confidence interval comes from a relatively small subsample of the data, this is where it really pays to be careful. You can use bootstrapping to get a non-parametric confidence interval.

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

[Q: Data Science as a Social Scientist?](#)

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

as I am very interested in programming and statistics, Data Science seems like a great career path to me - I like both fields and would like to combine them. Unfortunately, I have studied political science with a non-statistical sounding Master. I focused on statistics in this Master, visiting optional courses and writing a statistical thesis on a rather large dataset.

Since almost all job ads are requiring a degree in informatics, physics or some other techy-field, I am wondering if there is a chance to become a data scientist or if I should drop that idea.

I am lacking knowledge in machine learning, sql and hadoop, while having a rather strong informatics and statistics background.

So can somebody tell me how feasible my goal of becoming a data scientist is?

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [christian-sauer](#) 

[Answer](#)  by [steve-kallestad](#) 

The downvotes are because of the topic, but I'll attempt to answer your question as best I can since it's here.

Data science is a term that is thrown around as loosely as Big Data. Everyone has a rough idea of what they mean by the term, but when you look at the actual work tasks, a data scientist's responsibilities will vary greatly from company to company.

Statistical analysis could encompass the entirety of the workload in one job, and not even be a consideration for another.

I wouldn't chase after a job title per se. If you are interested in the field, network (like you are doing now) and find a good fit. If you are perusing job ads, just look for the ones that stress statistical and informatics backgrounds. Hadoop and SQL are both easy to become familiar with given the time and motivation, but I would stick with the areas you are strongest in and go from there.

[Answer](#)  by [mathattack](#) 

I suspect this will get closed since it is very narrow, but my 2 cents...

Data Science requires 3 skills:

- Math/Stats
- Programming
- Domain Knowledge

It can be very hard to show all three. #1 and #2 can be signaled via degrees, but a hiring manager who may not have them doesn't want to trust a liberal arts degree. If you're looking to get into Data Science, position yourself as a domain expert first. Publish election predictions. If you're correct, cite them. That will get you noticed.

If your Domain knowledge is A+ level, you don't need A+ level programming skills, but learn programming enough so that you don't need someone else to fetch data for you.

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

[Q: Data Science oriented dataset/research question for Statistics MSc thesis](#)

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

I'd like to explore 'data science'. The term seems a little vague to me, but I expect it to require:

1. machine learning (rather than traditional statistics);
2. a large enough dataset that you have to run analyses on clusters.

What are some good datasets and problems, accessible to a statistician with some programming background, that I can use to explore the field of data science?

To keep this as narrow as possible, I'd ideally like links to open, well used datasets and example problems.

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

User: [user3279453](#) 

[Answer](#)  by [emre](#) 

Just head to kaggle.com; it'll keep you busy for a long time. For open data there's the [UC Irvine Machine Learning Repository](#). In fact, there's a whole [Stackexchange site](#) devoted to this; look there.

[Answer](#)  by [steve-kallestad](#) 

The [Sunlight Foundation](#) is an organization that is focused on opening up and encouraging non-partisan analysis of government data.

There is a ton of analysis out there in the wild that can be used for comparison, and a wide variety of topics.

They provide [tools](#) and [apis](#) for accessing data, and have helped push to make data available in places like [data.gov](#).

One interesting project is [Influence Explorer](#). You can get [source data here](#) as well as access to real time data.

You might also want to take a look at one of our more popular questions:

[Publicly available datasets](#).

[Answer](#) by [mrmeritology](#)

Is your Masters in Computer Science? Statistics?

Is ‘data science’ going to be at the center of your thesis? Or a side topic?

I’ll assume you’re in Statistics and that you want to focus your thesis on a ‘data science’ problem. If so, then I’m going to go against the grain and suggest that you *should not* start with a data set or an ML method. Instead, you should seek an interesting research problem that’s poorly understood or where ML methods have not yet been proven successful, or where there are many competing ML methods but none seem better than others.

Consider this data source: [Stanford Large Network Dataset Collection](#). While you *could* pick one of these data sets, make up a problem statement, and then run some list of ML methods, that approach really doesn’t tell you very much about what *data science* is all about, and in my opinion doesn’t lead to a very good Masters thesis.

Instead, you might do this: look for all the research papers that use ML on some specific category — e.g. Collaboration networks (a.k.a. co-authorship). As you read each paper, try to find out what they *were* able to accomplish with each ML method and what they weren’t able to address. Especially look for their suggestions for “future research”.

Maybe they all use the same method, but never tried competing ML methods. Or maybe they don’t adequately validate their results, or maybe their data sets are small, or maybe their research questions and hypothesis were simplistic or limited.

Most important: try to find out where this line of research is going. Why are they even bothering to do this? What is significant about it? Where and why are they encountering difficulties?

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

[Q: Best languages for scientific computing](#)

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#))

It seems as though most languages have some number of scientific computing libraries available.

- Python has Scipy
- Rust has SciRust

- C++ has several including `ViennaCL` and `Armadillo`
- Java has `Java Numerics` and `Colt` as well as several other

Not to mention languages like `R` and `Julia` designed explicitly for scientific computing.

With so many options how do you choose the best language for a task? Additionally which languages will be the most performant? Python and R seem to have the most traction in the space, but logically a compiled language seems like it would be a better choice. And will anything ever outperform Fortran? Additionally compiled languages tend to have GPU acceleration, while interpreted languages like R and Python don't. What should I take into account when choosing a language, and which languages provide the best balance of utility and performance? Also are there any languages with significant scientific computing resources that I've missed?

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#))

User: [ragingsloth](#) 

[Answer](#)  by [indico](#) 

This is a pretty massive question, so this is not intended to be a full answer, but hopefully this can help to inform general practice around determining the best tool for the job when it comes to data science. Generally, I have a relatively short list of qualifications I look for when it comes to any tool in this space. In no particular order they are:

- **Performance:** Basically boils down to how quickly the language does matrix multiplication, as that is more or less the most important task in data science.
- **Scalability:** At least for me personally, this comes down to ease of building a distributed system. This is somewhere where languages like Julia really shine.
- **Community:** With any language, you're really looking for an active community that can help you when you get stuck using whichever tool you're using. This is where python pulls very far ahead of most other languages.
- **Flexibility:** Nothing is worse than being limited by the language that you use. It doesn't happen very often, but trying to represent graph structures in Haskell is a notorious pain, and Julia is filled with a lot of code architectures pains as a result of being such a young language.
- **Ease of Use:** If you want to use something in a larger environment, you want to make sure that setup is a straightforward and it can be automated. Nothing is worse than having to set up a finicky build on half a dozen machines.

There are a ton of articles out there about performance and scalability, but in general you're going to be looking at a performance differential of maybe 5-10x between languages, which may or may not matter depending on your specific application. As far as GPU acceleration goes, cudamat is a really seamless way of getting it working with python, and the cuda library in general has made GPU acceleration far more accessible than it used to be.

The two primary metrics I use for both community and flexibility are to look at the language's package manager, and the language questions on a site like SO. If there are a

large number of high-quality questions and answers, it's a good sign that the community is active. Number of packages and the general activity on those packages can also be a good proxy for this metric.

As far as ease of use goes, I am a firm believer that the only way to actually know is to actually set it up yourself. There's a lot of superstition around a lot of Data Science tools, specifically things like databases and distributed computing architecture, but there's no way to really know if something is easy or hard to setup up and deploy without just building it yourself.

[Answer](#) by [marc-claesens](#)

The best language depends on what you want to do. First remark: don't limit yourself to one language. Learning a new language is always a good thing, but at some point you will need to choose. Facilities offered by the language itself are an obvious thing to keep into account *but* in my opinion the following are more important:

- **available libraries:** do you have to implement everything from scratch or can you reuse existing stuff? Note that these libraries need not be in whatever language you are considering, as long as you can interface easily. Working in a language without library access won't help you get things done.
- **number of experts:** if you want external developers or start working in a team, you have to consider how many people actually know the language. As an extreme example: if you decide to work in Brainfuck because you happen to like it, know that you will likely work alone. Many surveys exist that can help assess the popularity of languages, including the number of questions per language on SO.
- **toolchain:** do you have access to *good* debuggers, profilers, documentation tools and (if you're into that) IDEs?

I am aware that most of my points favor established languages. This is from a 'get-things-done' perspective.

That said, I personally believe it is far better to become proficient in a low level language and a high level language:

- low level: C++, C, Fortran, ... using which you can implement certain profiling hot spots *only if you need to* because developing in these languages is typically slower (though this is subject to debate). These languages remain king of the hill in terms of critical performance and are likely to stay on top for a long time.
- high level: Python, R, Clojure, ... to 'glue' stuff together and do non-performance critical stuff (preprocessing, data handling, ...). I find this to be important simply because it is much easier to do rapid development and prototyping in these languages.

[Answer](#) by [armin](#)

First you need to decide what you want to do, then look for the right tool for that task.

A very general approach is to use R for first versions and to see if your approach is

correct. It lacks a little in speed, but has very powerful commands and addon libraries, that you can try almost anything with it: <http://www.r-project.org/>

The second idea is if you want to understand the algorithms behind the libraries, you might wanna take a look at the Numerical Recipies. They are available for different languages and free to use for learning. If you want to use them in commercial products, you need to purchase a licence: http://en.wikipedia.org/wiki/Numerical_Recipes

Most of the time performance will not be the issue but finding the right algorithms and parameters for them, so it is important to have a fast scripting language instead of a monster program that first needs to compile 10 mins before calculating two numbers and putting out the result.

And a big plus in using R is that it has built-in functions or libraries for almost any kind of diagram you might wanna need to visualize your data.

If you then have a working version, it is almost easy to port it to any other language you think is more performant.

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#))

[Q: Standardize numbers for ranking ratios](#)

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

I'm trying to rank some percentages. I have numerators and denominators for each ratio. To give a concrete example, consider ratio as total graduates / total students in a school.

But the issue is that total students vary over a long range (1000-20000). Smaller schools seem to have higher percentage of students graduating, but I want to standardize it, and not let the size of the school affect the ranking. Is there a way to do it?

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

User: [rohit-mittal](#) 

[Answer](#)  by [mrmeritology](#) 

This is relatively simple to do mathematically. First, fit a regression line to the scatter plot of "total graduates" (y) vs. "total students" (x). You will probably see a downward sloping line if your assertion is correct (smaller schools graduate a higher %).

You can identify the slope and y-intercept for this line to convert it into an equation $y = mx + b$, and then do a little algebra to convert the equation into normalized form: " $y / x = m + b / x$ "

Then, with all the ratios in your data , you should *subtract* this RHS:

normalized ratio = (total grads / total students) - (m + b / total students)

If the result is positive, then the ratio is above normal for that size (i.e. above the regression line) and if it is negative it is below the regression line. If you want all positive numbers, you can add a positive constant to move all results above zero.

This is how to do it mathematically, but I suggest that you consider whether it is wise, from a data analysis point of view, to normalize by school size. This depends on the purpose of your analysis and specifically how this ratio is being analyzed in relation to other data.

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#))

[Q: Analyzing A/B test results which are not normally distributed, using independent t-test](#)

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

I have a set of results from an A/B test (one control group, one feature group) which do not fit a Normal Distribution. In fact the distribution resembles more closely the Landau Distribution.

I believe the independent t-test requires that the samples be at least approximately normally distributed, which discourages me using the t-test as a valid method of significance testing.

But my question is: **At what point can one say that the t-test is not a good method of significance testing?**

Or put another way, how can one qualify how reliable the p-values of a t-test are, given only the data set?

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [teebszet](#) 

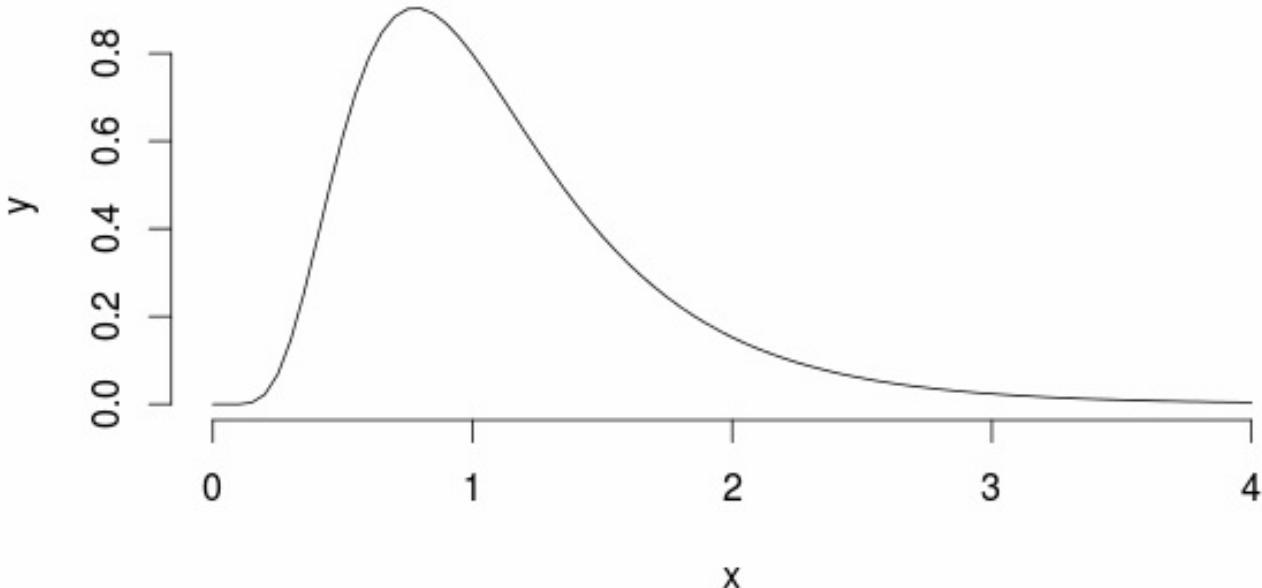
[Answer](#)  by [alexey-grigorev](#) 

The distribution of your data doesn't need to be normal, it's the [Sampling Distribution](#)  that has to be nearly normal. If your sample size is big enough, then the sampling distribution of means from Landau Distribution should be nearly normal, due to the [Central Limit Theorem](#) .

So it means you should be able to safely use t-test with your data.

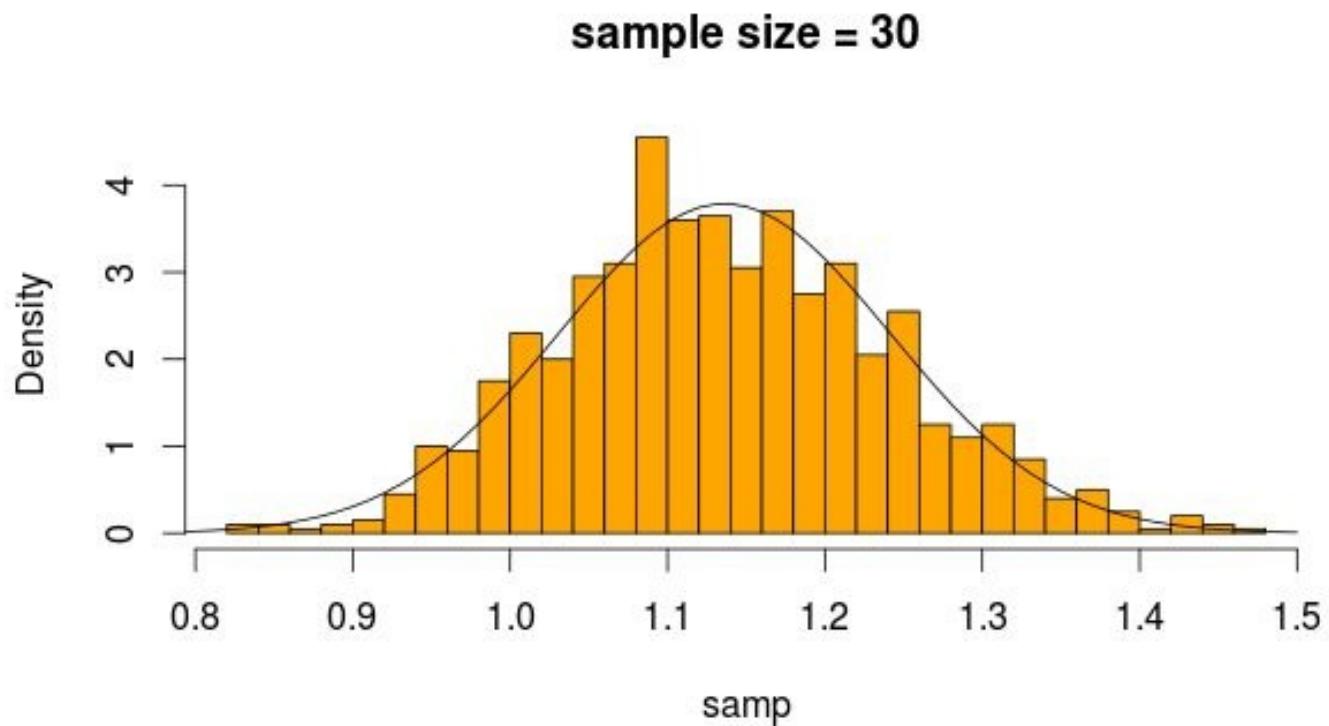
Example

Let's consider this example: suppose we have a population with [Lognormal distribution](#)  with $\mu=0$ and $\sigma=0.5$ (it looks a bit similar to Landau)

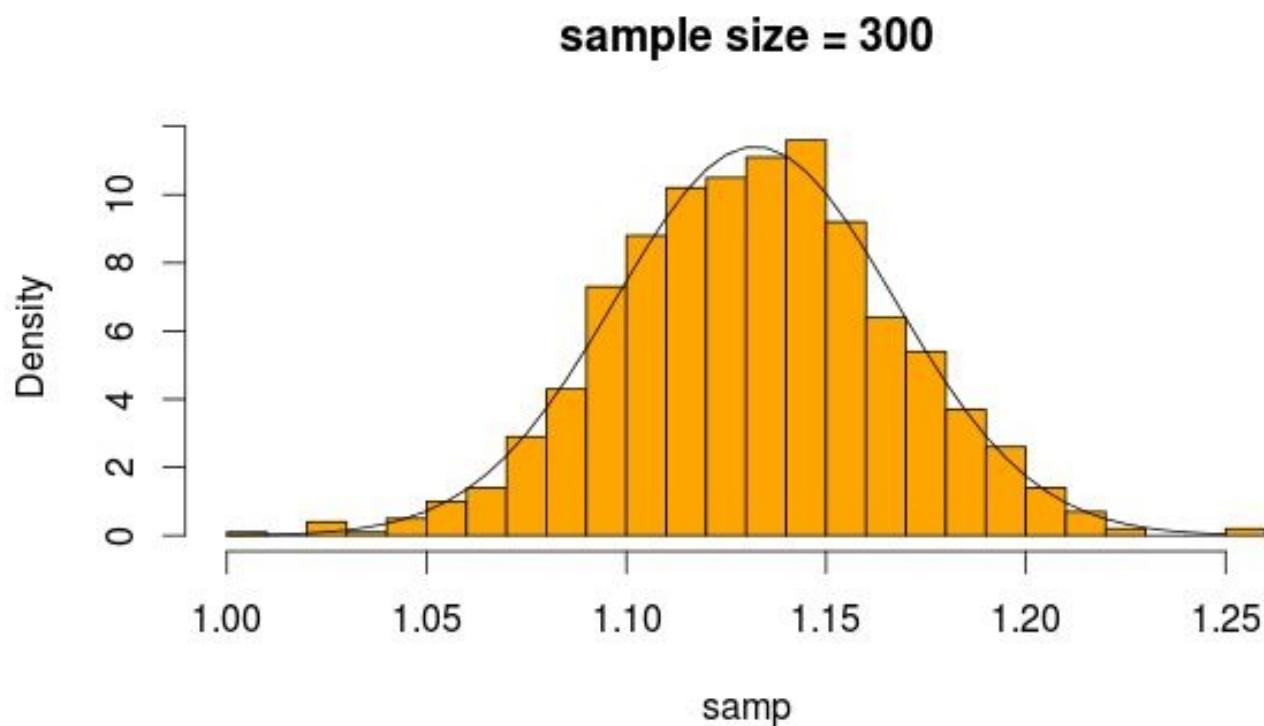


So we sample 30 observations 5000 times from this distribution each time calculating the mean of the sample

And this is what we get



Looks quite normal, doesn't it? If we increase the sample size, it's even more apparent



R code

[Skip code block](#)

```
x = seq(0, 4, 0.05)
y = dlnorm(x, mean=0, sd=0.5)
```

```

plot(x, y, type='l', bty='n')

n = 30
m = 1000

set.seed(0)
samp = rep(NA, m)

for (i in 1:m) {
  samp[i] = mean(rlnorm(n, mean=0, sd=0.5))
}

hist(samp, col='orange', probability=T, breaks=25, main='sample size = 30')
x = seq(0.5, 1.5, 0.01)
lines(x, dnorm(x, mean=mean(samp), sd=sd(samp)))

n = 300
samp = rep(NA, m)

for (i in 1:m) {
  samp[i] = mean(rlnorm(n, mean=0, sd=0.5))
}

hist(samp, col='orange', probability=T, breaks=25, main='sample size = 300')
x = seq(1, 1.25, 0.005)
lines(x, dnorm(x, mean=mean(samp), sd=sd(samp)))

```

Tags: [statistics](#) ([Prev Q](#)) ([Next Q](#)), [dataset](#) ([Prev Q](#)) ([Next Q](#))

Q: Methods for standardizing / normalizing different rank scales

Tags: [statistics](#) ([Prev Q](#))

I know there is the normal *subtract the mean and divide by the standard deviation* for standardizing your data, but I'm interested to know if there are more appropriate methods for this kind of discrete data. Consider the following case.

I have 5 items that have been ranked by customers. First 2 items were ranked on a 1-10 scale. Others are 1-100 and 1-5. To transform everything to a 1 to 10 scale, is there another method better suited for this case?

If the data has a central tendency, then the standard would work fine, but what about when you have more of a halo effect, or some more exponential distribution?

Tags: [statistics](#) ([Prev Q](#))

User: [climbs_lik_a_spyder](#) 

[Answer](#)  by [nitesh](#) 

For item-ratings type of data with the restriction that an item's rating should be between 1 and 10 after transformation, I would suggest using a simple re-scaling, such that the item's transformed rating x_t is given by:

$$x_t = 9 \left(\frac{x_i - x_{min}}{x_{max} - x_{min}} \right) + 1$$

where x_{min} and x_{max} are the minimum and maximum possible rating in the specific scale

for the item, and x_i is the item rating.

In the case of the above scaling, the transformation applied is independent of the data. However, in the normalization, the transformation applied is dependent on the data (through mean and standard deviation), and might change as more data becomes available.

Section 4.3 on page 30 of [this document](#) shows other ways of normalizing in which your restriction (transforming to the same absolute scale) might not be preserved.

Tags: [statistics](#) ([Prev Q](#))

Python

[Skip to questions](#),

Wiki by user [james](#)

[Python](#) is a general-purpose, dynamic, strongly typed language with many 3rd-party libraries for data science applications. There are two versions currently in wide use: 2 and 3. Python 2 is the “old” version, with no new versions being released beyond 2.7, save bugfixes. Python 3 is the “new” version, with active development.

Python syntax is relatively easy to comprehend compared to other languages. For example:

```
numbers = [1, 2, 5, 8, 9]
for number in numbers:
    print("Hello world #", number)
```

Python has a clean look due to its regulatory approach to [whitespace](#). While seemingly restrictive, it allows all Python code to look similar, which makes inspecting code much more predictable. All loops and conditionals (`for`, `while`, `if`, etc) must be indented for the code block that follows.

Popular scientific and data science packages include:

- [Numpy](#) - A fast, N-dimensional array library; the foundation for all things scientific Python.
 - [Scipy](#) - Numerical analysis built on Numpy. Allows for optimization, linear algebra, Fourier Transforms and much else.
 - [Pandas](#) (PArallel DAta) - A fast and extremely flexible package that is very useful for data exploration. It handles NaN data well as well as fast indexing. Handles a wide variety of external data types and file formats.
-

Questions

[Q: What to consider before learning a new language for data analysis](#)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

I'm currently in the very early stages of preparing a new research-project (still at the funding-application stage), and expect that data-analysis and especially visualisation tools will play a role in this project.

In view of this I face the following dilemma: Should I learn Python to be able to use its extensive scientific libraries (Pandas, Numpy, Scipy, ...), or should I just dive into similar packages of a language I'm already acquainted with (Racket, or to a lesser extent Scala)?

(Ideally I would learn Python in parallel with using statistical libraries in Racket, but I'm not sure I'll have time for both)

I'm not looking for an answer to this dilemma, but rather for feedback on my different considerations:

My current position is as follows:

In favour of Python:

- Extensively used libraries
- Widely used (may be decisive in case of collaboration with others)
- A lot of online material to start learning it
- Conferences that are specifically dedicated to Scientific Computing with Python
- Learning Python won't be a waste of time anyway

In favour of a language I already know:

- It's a way to deepen my knowledge of one language rather than getting superficial knowledge of one more language (under the motto: you should at least know one language really well)
- It is feasible. Both Racket and Scala have good mathematics and statistics libraries
- I can start right away with learning what I need to know rather than first having to learn the basics

Two concrete questions:

1. What am I forgetting?
2. How big of a nuisance could the Python 2 vs 3 issue be?

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

User: [user2818584](#) 

[Answer](#)  by [indico](#) 

Personally going to make a strong argument in favor of Python here. There are a large

number of reasons for this, but I'm going to build on some of the points that other people have mentioned here:

1. **Picking a single language:** It's definitely possible to mix and match languages, picking d3 for your visualization needs, FORTRAN for your fast matrix multiplies, and python for all of your networking and scripting. You can do this down the line, but keeping your stack as simple as possible is a good move, especially early on.
2. **Picking something bigger than you:** You never want to be pushing up against the barriers of the language you want to use. This is a huge issue when it comes to languages like Julia and FORTRAN, which simply don't offer the full functionality of languages like python or R.
3. **Pick Community:** The one most difficult thing to find in any language is community. Python is the clear winner here. If you get stuck, you ask something on SO, and someone will answer in a matter of minutes, which is simply not the case for most other languages. If you're learning something in a vacuum you will simply learn much slower.

In terms of the minus points, I might actually push back on them.

Deepening your knowledge of one language is a decent idea, but knowing *only* one language, without having practice generalizing that knowledge to other languages is a good way to shoot yourself in the foot. I have changed my entire favored development stack three time over as many years, moving from MATLAB to Java to haskell to python. Learning to transfer your knowledge to another language is far more valuable than just knowing one.

As far as feasibility, this is something you're going to see again and again in any programming career. Turing completeness means you could technically do everything with HTML4 and css3, but you want to pick the right tool for the job. If you see the ideal tool and decide to leave it by the roadside you're going to find yourself slowed down wishing you had some of the tools you left behind.

A great example of that last point is trying to deploy R code. 'R's networking capabilities are hugely lacking compared to python, and if you want to deploy a service, or use slightly off-the-beaten path packages, the fact that pip has an order of magnitude more packages than CRAN is a huge help.

[Answer](#)  by [little-bobby-tables](#) 

From my experience, the points to keep in mind when considering a data analysis platform are:

1. Can it handle the size of the data that I need? If your data sets fit in memory, there's usually no big trouble, although AFAIK Python is somewhat more memory-efficient than R. If you need to handle larger-than-memory data sets, the platform need to handle it conveniently. In this case, SQL would cover for basic statistics, Python + Apache Spark is another option.
2. Does the platform covers all of my analysis needs? The greatest annoyance I've encountered in data mining projects is having to juggle between several tools,

because tool A handles web connections well, tool B does the statistics and tool C renders nice pictures. You want your weapon-of-choice to cover as many aspects of your projects as possible. When considering this issue, Python is very comprehensive, but R has a lot of build-in statistical tests ready-to-use, if that's what you need.

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

Q: Tools and protocol for reproducible data science using Python

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#))

I am working on a data science project using Python. The project has several stages. Each stage comprises of taking a data set, using Python scripts, auxiliary data, configuration and parameters, and creating another data set. I store the code in git, so that part is covered. I would like to hear about:

1. Tools for data version control.
2. Tools enabling to reproduce stages and experiments.
3. Protocol and suggested directory structure for such a project.
4. Automated build/run tools.

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#))

User: [yuval-f](#)

[Answer](#) by [sebastian-raschka](#)

Since I started doing research in academia I was constantly looking for a satisfactory workflow. I think that I finally found something I am happy with:

1) Put everything under version control, e.g., Git:

For hobby research projects I use GitHub, for research at work I use the private GitLab server that is provided by our university. I also keep my datasets there.

2) I do most of my analyses along with the documentation on IPython notebooks. It is very organized (for me) to have the code, the plots, and the discussion/conclusion all in one document. If I am running larger scripts, I would usually put them into separate script .py files, but I would still execute them from the IPython notebook via the %run magic to add information about the purpose, outcome, and other parameters.

I have written a small cell-magic extension for IPython and IPython notebooks, called “watermark” that I use to conveniently create time stamps and keep track of the different package versions I used and also Git hashes

For example

[Skip code block](#)

```
%watermark  
29/06/2014 01:19:10  
CPython 3.4.1  
IPython 2.1.0  
  
compiler : GCC 4.2.1 (Apple Inc. build 5577)  
system : Darwin  
release : 13.2.0  
machine : x86_64  
processor : i386  
CPU cores : 2  
interpreter: 64bit
```

```
%watermark -d -t  
29/06/2014 01:19:11
```

[Skip code block](#)

```
%watermark -v -m -p numpy,scipy  
  
CPython 3.4.1  
IPython 2.1.0  
  
numpy 1.8.1  
scipy 0.14.0  
  
compiler : GCC 4.2.1 (Apple Inc. build 5577)  
system : Darwin  
release : 13.2.0  
machine : x86_64  
processor : i386  
CPU cores : 2  
interpreter: 64bit
```

For more info, see the [documentation here](#) .

[Answer](#)  by [gaborous](#) 

The best reproducibility tool is to make a log of your actions, something like this:

```
experiment/input ; expected ; observation/output ; current hypothesis and if supported or rejected  
exp1 ; expected1 ; obs1 ; some fancy hypothesis, supported
```

This can be written down on a paper, but, if your experiments fit in a computational framework, you can use computational tools to partly or completely automate that logging process (particularly by helping you track the input datasets which can be huge, and the output figures).

A great reproducibility tool for Python with a low learning curve is of course [IPython/Jupyter Notebook](#)  (don't forget the [%logon](#) and [%logstart](#)  magics).

Another great tool that is very recent (2015) is [recipy](#) , which is very like sumatra (see below), but made specifically for Python. I don't know if it works with Jupyter Notebooks, but I know the author frequently uses them so I guess that if it's not currently supported, it will be in the future.

[Git](#)  is also awesome, and it's not tied to Python. It will help you not only to keep a history of all your experiments, code, datasets, figures, etc. but also provide you with tools

to maintain ([git pickaxe](#)), collaborate ([blame](#)) and debug ([git bisect](#)) using a scientific method of debugging (called [delta debugging](#)). [Here's a story](#) of a fictional researcher trying to make his own experiments logging system, until it ends up being a facsimile of Git.

Another general tool working with any language (with a Python API on [pypi](#)) is [Sumatra](#), which is specifically designed to help you do **replicable** research (*replicable* aims to produce the same results given the exact same code and softwares, whereas *reproducibility* aims to produce the same results given any medium, which is a lot harder and time consuming and not automatable).

Here is how Sumatra works: for each experiment that you conduct through Sumatra, this software will act like a “save game state” often found in videogames. More precisely, it will save:

- all the parameters you provided;
- the exact sourcecode state of your whole experimental application and config files;
- the output/plots/results and also any file produced by your experimental application.

It will then construct a database with the timestamp and other metadatas for each of your experiments, that you can later crawl using the webGUI. Since Sumatra saved the full state of your application for a specific experiment at one specific point in time, you can restore the code that produced a specific result at any moment you want, thus you have replicable research at a low cost (except for storage if you work on huge datasets, but you can configure exceptions if you don't want to save everything everytime).

In the end, you can use either Git or Sumatra, they will provide you with about the same replicability power, but Sumatra is specifically tailored for scientific research so it provides a few fancy tools like a web GUI to crawl your results, while Git is more tailored towards code maintenance (but it has debugging tools like git-bisect so if your experiments involve codes, it may actually be better). Or of course you can use both!

/EDIT: [dsign](#) touched a very important point here: the replicability of your setup is as important as the replicability of your application. In other words, you should at least provide a **full list of the libraries and compilers** you used along with their exact **versions** and the details of your **platform**.

Personally, in scientific computing with Python, I have found that packaging an application along with the libraries is just too painful, thus I now just use an all-in-one scientific python package such as [Anaconda](#) (with the great package manager [conda](#)), and just advise users to use the same package. Another solution could be to provide a script to automatically generate a [virtualenv](#), or to package everything using the commercial [Docker application as cited by dsign](#) or the opensource [Vagrant](#) (with for example [pylearn2-in-a-box](#) which use Vagrant to produce an easily redistributable virtual environment package).

/EDIT2: Here's a [great video](#) summarizing (for debugging but this can also be applied to research) what is fundamental to do reproducible research: logging your experiments and each other steps of the scientific method, a sort of “*explicit experimenting*”.

[Answer](#) by [dsign](#)

Be sure to check out [docker](#)! And in general, all the other good things that software engineering has created along decades for ensuring isolation and reproducibility.

I would like to stress that it is not enough to have *just* reproducible workflows, but also *easy* to reproduce workflows. Let me show what I mean. Suppose that your project uses Python, a database X and Scipy. Most surely you will be using a specific library to connect to your database from Python, and Scipy will be in turn using some sparse algebraic routines. This is by all means a very simple setup, but not entirely simple to setup, pun intended. If somebody wants to execute your scripts, she will have to install all the dependencies. Or worse, she might have incompatible versions of it already installed. Fixing those things takes time. It will also take time to you if you at some moment need to move your computations to a cluster, to a different cluster, or to some cloud servers.

Here is where I find docker useful. Docker is a way to formalize and compile recipes for binary environments. You can write the following in a dockerfile (I'm using here plain English instead of the Dockerfile syntax):

- Start with a basic binary environment, like Ubuntu's
- Install libsparse-dev
- (Pip) Install numpy and scipy
- Install X
- Install libX-dev
- (Pip) Install python-X
- Install IPython-Notebook
- Copy my python scripts/notebooks to my binary environment, these datafiles, and these configurations to do other miscellaneous things. To ensure reproducibility, copy them from a named url instead of a local file.
- Maybe run IPython-Notebook.

Some of the lines will be installing things in Python using pip, since pip can do a very clean work in selecting specific package versions. Check it out too!

And that's it. If after you create your Dockerfile it can be built, then it can be built anywhere, by anybody (provided they also have access to your project-specific files, e.g. because you put them in a public url referenced from the Dockerfile). What is best, you can upload the resulting environment (called an "image") to a public or private server (called a "register") for other people to use. So, when you publish your workflow, you have both a fully reproducible recipe in the form of a Dockerfile, and an easy way for you or other people to reproduce what you do:

```
docker run dockerregistry.thewheezylab.org/nowyouwillbelieveme
```

Or if they want to poke around in your scripts and so forth:

```
docker run -i -t dockerregistry.thewheezylab.org/nowyouwillbelieveme /bin/bash
```

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#))

[Q: Stochastic gradient descent based on vector operations?](#)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

let's assume that I want to train a stochastic gradient descent regression algorithm using a dataset that has N samples. Since the size of the dataset is fixed, I will reuse the data T times. At each iteration or "epoch", I use each training sample exactly once after randomly reordering the whole training set.

My implementation is based on Python and Numpy. Therefore, using vector operations can remarkably decrease computation time. Coming up with a vectorized implementation of batch gradient descent is quite straightforward. However, in the case of stochastic gradient descent I can not figure out how to avoid the outer loop that iterates through all the samples at each epoch.

Does anybody know any vectorized implementation of stochastic gradient descent?

EDIT: I've been asked why would I like to use online gradient descent if the size of my dataset is fixed.

From [1], one can see that online gradient descent converges slower than batch gradient descent to the minimum of the empirical cost. However, it converges faster to the minimum of the expected cost, which measures generalization performance. I'd like to test the impact of these theoretical results in my particular problem, by means of cross validation. Without a vectorized implementation, my online gradient descent code is much slower than the batch gradient descent one. That remarkably increases the time it takes to the cross validation process to be completed.

EDIT: I include here the pseudocode of my on-line gradient descent implementation, as requested by ffriend. I am solving a regression problem.

[Skip code block](#)

```
Method: on-line gradient descent (regression)
Input: X (nxp matrix; each line contains a training sample, represented as a length-p vector), Y (1er
Output: A (length-p+1 vector of coefficients)

Initialize coefficients (assign value 0 to all coefficients)
Calculate outputs F
prev_error = inf
error = sum((F-Y)^2)/n
it = 0
while abs(error - prev_error)>ERROR_THRESHOLD and it<=MAX_ITERATIONS:
    Randomly shuffle training samples
    for each training sample i:
        Compute error for training sample i
        Update coefficients based on the error above
    prev_error = error
    Calculate outputs F
    error = sum((F-Y)^2)/n
    it = it + 1
```

[1] "Large Scale Online Learning", L. Bottou, Y. Le Cunn, NIPS 2003.

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

User: [pablo-suau](#)

[Answer](#) by [ffriend](#)

First of all, word “sample” is normally used to describe [subset of population](#), so I will refer to the same thing as “example”.

Your SGD implementation is slow because of this line:

```
for each training example i:
```

Here you explicitly use exactly one example for each update of model parameters. By definition, vectorization is a technique for converting operations on one element into operations on a vector of such elements. Thus, no, you cannot process examples one by one and still use vectorization.

You can, however, approximate true SGD by using **mini-batches**. Mini-batch is a small subset of original dataset (say, 100 examples). You calculate error and parameter updates based on mini-batches, but you still iterate over many of them without global optimization, making the process stochastic. So, to make your implementation much faster it's enough to change previous line to:

```
batches = split dataset into mini-batches
for batch in batches:
```

and calculate error from batch, not from a single example.

Though pretty obvious, I should also mention vectorization on per-example level. That is, instead of something like this:

```
theta = np.array([...]) # parameter vector
x = np.array([...])    # example
y = 0                  # predicted response
for i in range(len(example)):
    y += x[i] * theta[i]
error = (true_y - y) ** 2 # true_y - true value of response
```

you should definitely do something like this:

```
error = (true_y - sum(np.dot(x, theta))) ** 2
```

which, again, easy to generalize for mini-batches:

```
true_y = np.array([...])    # vector of response values
x = np.array([..., [...]]) # mini-batch
errors = true_y - sum(np.dot(x, theta), 1)
error = sum(e ** 2 for e in errors)
```

[Answer](#) by [ben-allison](#)

Check out the `partial_fit` method of [scikit's SGD classifier](#). You have control over what you call with it: you can do it “true” online learning by passing an instance at a time, or you can batch up instances into mini-batches if all your data are available in an array. If they are, you can slice the array to provide the minibatches.

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: Pandas time series optimization problem: add year](#)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

I have a pandas DataFrame containing a time series column. The years are shifted in the past, so that I have to add a constant number of years to every element of that column.

The best way I found is to iterate through all the records and use

```
x.replace(year=x.year + years) # x = current element, years = years to add
```

It is cythonized as below, but still very slow (proofing)

[Skip code block](#)

```
cdef list _addYearsToTimestamps(list elts, int years):
    cdef cpdatetime x
    cdef int i
    for (i, x) in enumerate(elts):
        try:
            elts[i] = x.replace(year=x.year + years)
        except Exception as e:
            logError(None, "Cannot replace year of %s - leaving value as this: %s" % (str(x), repr(e)))
    return elts

def fixYear(data):
    data.loc[:, 'timestamp'] = _addYearsToTimestamps(list(data.loc[:, 'timestamp']), REAL_YEAR-(list(data['timestamp']).year))
    return data
```

I'm pretty sure that there is a way to change the year without iterating, by using Pandas's Timestamp features. Unfortunately, I don't find how. Could someone elaborate?

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

User: [michael-hooreman](#) 

[Answer](#)  by [pete](#) 

Make a pandas Timedelta object then add with the += operator:

```
x = pandas.Timedelta(days=365)
mydataframe.timestampcolumn += x
```

So the key is to store your time series as timestamps. To do that, use the pandas `to_datetime` function:

```
mydataframe['timestampcolumn'] = pandas.to_datetime(x['epoch'], unit='s')
```

assuming you have your timestamps as epoch seconds in the dataframe x. That's not a requirement of course; see the [to_datetime](#)  documentation for converting other formats.

[Answer](#)  by [michael-hooreman](#) 

Adapted from Pete's answer, here's an implementation of the solution, and the demonstration.

[Skip code block](#)

```
#!/usr/bin/env python3

import random
import pandas
import time
import datetime

def getRandomDates(n):
    tsMin = time.mktime(time.strptime("1980-01-01 00:00:00", "%Y-%m-%d %H:%M:%S"))
    tsMax = time.mktime(time.strptime("2005-12-31 23:59:59", "%Y-%m-%d %H:%M:%S"))
```

```

    return pandas.Series([datetime.datetime.fromtimestamp(tsMin + random.random() * (tsMax - tsMin))]

def setMaxYear(tss, target):
    maxYearBefore = tss.max().to_datetime().year
    # timedelta cannot be given in years, so we compute the number of days to add in the next line
    deltaDays = (datetime.date(target, 1, 1) - datetime.date(maxYearBefore, 1, 1)).days
    return tss + pandas.Timedelta(days=deltaDays)

data = pandas.DataFrame({'t1': getRandomDates(1000)})
data['t2'] = setMaxYear(data['t1'], 2015)
data['delta'] = data['t2'] - data['t1']
print(data)
print("delta min: %s" % str(min(data['delta'])))
print("delta max: %s" % str(max(data['delta'])))

```

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to recognize a two part term when the space is removed? \(“bigdata” and “big data”\)](#)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

I'm not a NLP guy and I have this question.

I have a text dataset containing terms which go like, “big data” and “bigdata”.

For my purpose both of them are the same.

How can I detect them in NLTK (Python)?

Or any other NLP module in Python?

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [kasra-manshaei](#)

[Answer](#) by [will-stanton](#)

There is a nice implementation of this in gensim:

<http://radimrehurek.com/gensim/models/phrases.html>

Basically, it uses a data-driven approach to detect phrases, ie. common collocations. So if you feed the Phrase class a bunch of sentences, and the phrase “big data” comes up a lot, then the class will learn to combine “big data” into a single token “big_data”. There is a more complete tutorial-style blog post about it here:

<http://www.markhneedham.com/blog/2015/02/12/pythongensim-creating-bigrams-over-how-i-met-your-mother-transcripts/>

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: Feature extraction of images in Python](#)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

In my class I have to create an application using two classifiers to decide whether an

object in an image is an example of phylum porifera (seasponge) or some other object. However, I am completely lost when it comes to feature extraction techniques in python. My advisor convinced me to use images which haven't been covered in class. Can anyone direct me towards meaningful documentation or reading or suggest methods to consider?

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

User: [jeremy-barnes](#) 

[Answer](#)  by [jamesmf](#) 

This great tutorial covers the basics of convolutional neural networks, which are currently achieving state of the art performance in most vision tasks:

<http://deeplearning.net/tutorial/lenet.html> 

There are a number of options for CNNs in python, including Theano and the libraries built on top of it (I found keras to be easy to use).

If you prefer to avoid deep learning, you might look into OpenCV, which can learn many other types of features, like Haar cascades and SIFT features.

http://opencv-python-tutorials.readthedocs.org/en/latest/py_tutorials/py_feature2d/py_table_of_contents_feature2.html

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

[Q: Hypertuning XGBoost parameters](#)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [xgboost](#) ([Next Q](#))

XGBoost have been doing a great job, when it comes to dealing with both categorical and continuous dependant variables. But, how do I select the optimized parameters for an XGBoost problem?

This is how I applied the parameters for a recent Kaggle problem:

[Skip code block](#)

```
param <- list( objective           = "reg:linear",
               booster = "gbtree",
               eta      = 0.02, # 0.06, #0.01,
               max_depth        = 10, #changed from default of 8
               subsample       = 0.5, # 0.7
               colsample_bytree = 0.7, # 0.7
               num_parallel_tree = 5
               # alpha = 0.0001,
               # lambda = 1
)

clf <- xgb.train( params           = param,
                  data            = dtrain,
                  nrounds         = 3000, #300, #280, #125, #250, # changed from 300
                  verbose         = 0,
                  early.stop.round = 100,
                  watchlist       = watchlist,
                  maximize        = FALSE,
```

```
) feval=RMPSE
```

All I do to experiment is randomly select (with intuition) another set of parameters for improving on the result.

Is there anyway I automate the selection of optimized(best) set of parameters?

(Answers can be in any language. I'm just looking for the technique)

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [xgboost](#) ([Next Q](#))

User: [dawny33](#) 

[Answer](#)  by [wacax](#) 

Whenever I work with xgboost I often make my own homebrew parameter search but you can do it with the caret package as well like KrisP just mentioned.

1. Caret

See this answer on Cross Validated for a thorough explanation on how to use the caret package for hyperparameter search on xgboost. [How to tune hyperparameters of xgboost trees?](#) 

2. Custom Grid Search

I often begin with a few assumptions based on [Owen Zhang](#) 's slides on [tips for data science](#)  P. 14

GBDT Hyper Parameter Tuning

Hyper Parameter	Tuning Approach	Range	Note
# of Trees	Fixed value	100-1000	Depending on datasize
Learning Rate	Fixed => Fine Tune	[2 - 10] / # of Trees	Depending on # trees
Row Sampling	Grid Search	[.5, .75, 1.0]	
Column Sampling	Grid Search	[.4, .6, .8, 1.0]	
Min Leaf Weight	Fixed => Fine Tune	3/(% of rare events)	Rule of thumb
Max Tree Depth	Grid Search	[4, 6, 8, 10]	
Min Split Gain	Fixed	0	Keep it 0

Best GBDT implementation today: <https://github.com/tqchen/xgboost>

by **Tianqi Chen** (U of Washington)



Here you can see that you'll mostly need to tune row sampling, column sampling and maybe maximum tree depth. This is how I do a custom row sampling and column sampling search for a problem I am working on at the moment:

[Skip code block](#)

```
searchGridSubCol <- expand.grid(subsample = c(0.5, 0.75, 1),
                                colsample_bytree = c(0.6, 0.8, 1))
ntrees <- 100

#Build a xgb.DMatrix object
DMMatrixTrain <- xgb.DMatrix(data = yourMatrix, label = yourTarget)

rmseErrorsHyperparameters <- apply(searchGridSubCol, 1, function(parameterList){

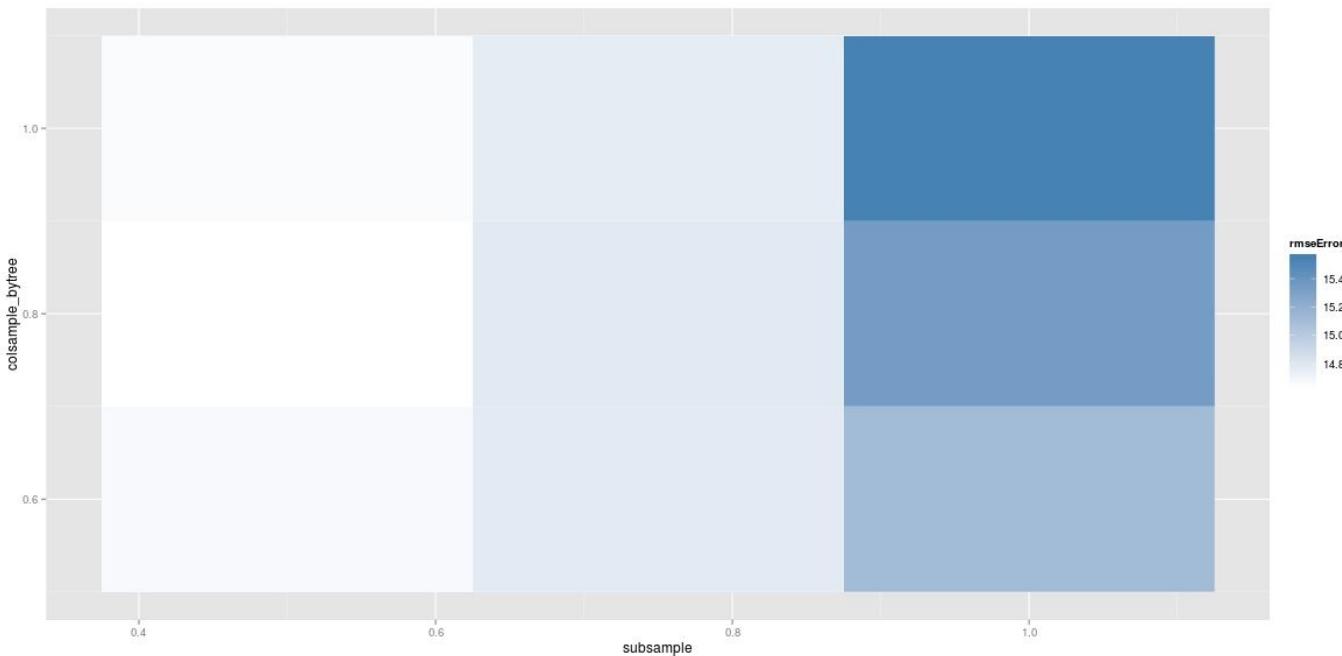
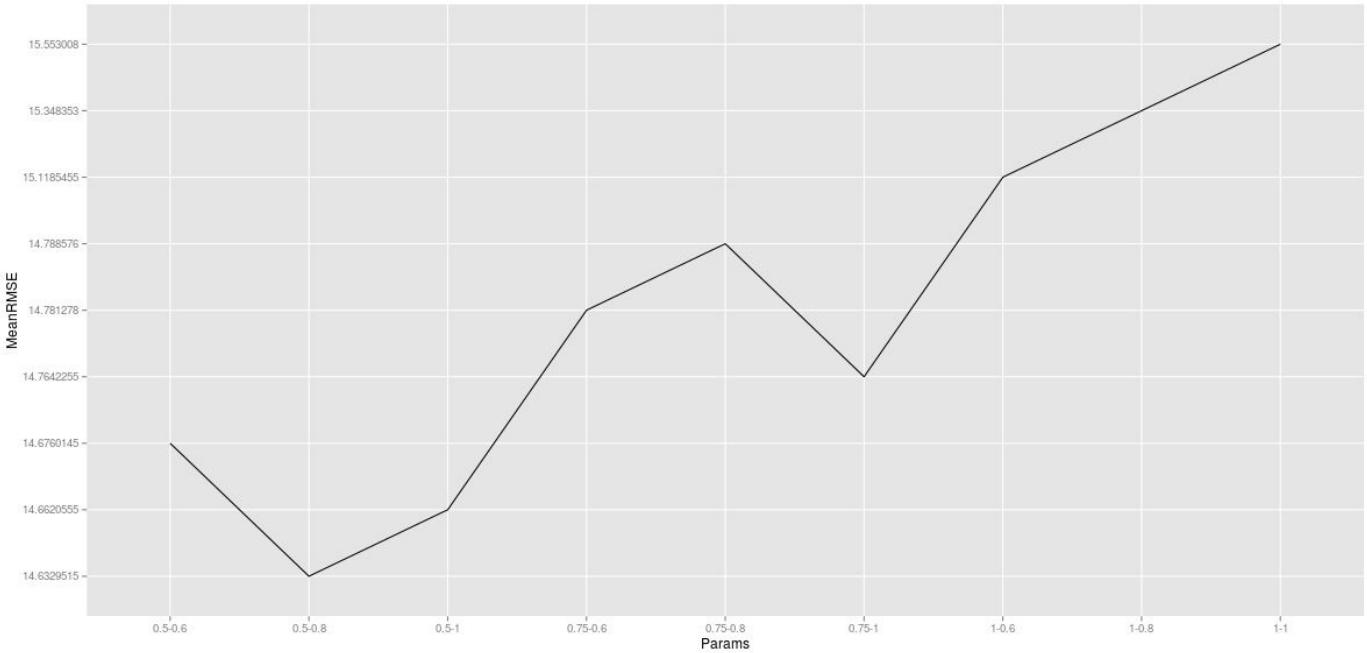
  #Extract Parameters to test
  currentSubsampleRate <- parameterList[["subsample"]]
  currentColsampleRate <- parameterList[["colsample_bytree"]]

  xgboostModelCV <- xgb.cv(data = DMMatrixTrain, nrounds = ntrees, nfolds = 5, showSD = TRUE,
                             metrics = "rmse", verbose = TRUE, "eval_metric" = "rmse",
                             "objective" = "reg:linear", "max_depth" = 15, "eta" = 2/ntrees,
                             "subsample" = currentSubsampleRate, "colsample_bytree" = currentColsampleR

  xvalidationScores <- as.data.frame(xgboostModelCV)
  #Save rmse of the last iteration
  rmse <- tail(xvalidationScores$test.rmse.mean, 1)

  return(c(rmse, currentSubsampleRate, currentColsampleRate))
})
```

And combined with some ggplot2 magic using the results of that apply function you can plot a graphical representation of the search.



In this plot lighter colors represent lower error and each block represents a unique combination of column sampling and row sampling. So if you want to perform an additional search of say eta (or tree depth) you will end up with one of these plots for each eta parameters tested.

I see you have a different evaluation metric (RMPSE), just plug that in the cross validation function and you'll get the desired result. Besides that I wouldn't worry too much about fine tuning the other parameters because doing so won't improve performance too much, at least not so much compared to spending more time engineering features or cleaning the data.

3. Others

Random search and Bayesian parameter selection are also possible but I haven't made/found an implementation of them yet.

Tags: [python](#) ([Prev Q](#)) ([Next Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#)), [xgboost](#) ([Next Q](#))

Q: Building a machine learning model to predict crop yields based on environmental data

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

I have a dataset containing data on temperature, precipitation and soybean yields for a farm for 10 years (2005 - 2014). I would like to predict yields for 2015 based on this data.

Please note that the dataset has DAILY values for temperature and precipitation, but only 1 value per year for the yield, since harvesting of crop happens at end of growing season of crop.

I want to build a regression or some other machine learning based model to predict 2015 yields, based on a regression/some other model derived by studying the relation between yields and temperature and precipitation in previous years.

I am familiar with performing machine learning using scikit-learn. However, not sure how to represent this problem. The tricky part here is that temperature and precipitation are daily but yield is just 1 value per year.

How do I approach this?

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

User: [user308827](#)

Answer by [emre](#)

For starters, you can predict the yield for the upcoming year based on the daily data for the previous year. You can estimate the model parameters by considering each year's worth of data as one "point", then validate the model using cross-validation. You can extend this model by considering more than the past year, but look back too far and you'll have trouble validating your model and overfit.

Tags: [python](#) ([Prev Q](#)) ([Next Q](#))

Q: Improve k-means accuracy

Tags: [python](#) ([Prev Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

Our weapons:

I am experimenting with k-means and Hadoop, where I am chained to these options for various reasons (e.g. [Help me win this war!](#)).

The battlefield:

I have articles, which belong to c categories, where c is fixed. I am vectorizing the contents of the articles to **TF-IDF** features. Now I am running a naive k-means algorithm,

which takes c centroids to begin with and starts, iteratively, grouping articles (i.e. rows of the TF-IDF matrix, where you can see [here](#) how I built it), until convergence occurs.

Special notes:

1. Initial centroids: Tried with random from within each category or with the mean of all the articles from each category.
 2. Distance function: Euclidean.
-

Question(s):

The accuracy is poor, as expected, can I do any better, by making another choice for the initial centroids, or/and pick another distance function?

print "Hello Data Science site!" :)

Tags: [python](#) ([Prev Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

User: [gsamaras](#)

[Answer](#) by [kyle.](#)

Great question, @gsamaras! The way you've set up this experiment makes a lot of sense to me, from a design point of view, but I think there are a couple aspects you can still examine.

First, it's possible that uninformative features are distracting your classifier, leading to poorer results. In text analytics, we often talk about [stop word](#) filtering, which is just the process of removing such text (e.g., the, and, or, etc.). There are standard stop word lists you can easily find online (e.g., [this one](#)), but they can sometimes be heavy-handed. The best approach is to build a table relating feature frequency to class, as this will get at domain-specific features that you won't likely find in such look-up tables. There is varying evidence as to the efficacy of stop word removal in the literature, but I think these findings mostly have to do with classifier-specific (for example, support vector machines tend to be less affected by uninformative features than does a naive bayes classifier. I suspect k-means falls into the latter category).

Second, you might consider a different feature modeling approach, rather than tf-idf. Nothing against tf-idf—it works fine for many problems—but I like to start with binary feature modeling, unless I have experimental evidence showing a more complex approach leads to better results. That said, it's possible that k-means could respond strangely to the switch from a floating-point feature space to a binary one. It's certainly an easily-testable hypothesis!

Finally, you might look at the expected class distribution in your data set. Are all classes equally likely? If not, you may get better results from either a sampling approach, or using a different distance metric. [k-means is known to respond poorly in skewed class situations](#), so this is something to consider as well! There is probably research available in your specific domain describing how others have handled this situation.

Tags: [python](#) ([Prev Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

Clustering

Questions

[Q: Clustering unique visitors by useragent, ip, session_id](#)

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

Given website access data in the form `session_id`, `ip`, `user_agent`, and optionally timestamp, following the conditions below, how would you best cluster the sessions into unique visitors?

`session_id`: is an id given to every new visitor. It does not expire, however if the user doesn't accept cookies/clears cookies/changes browser/changes device, he will not be recognised anymore

`IP` can be shared between different users (Imagine a free wi-fi cafe, or your ISP reassigning IPs), and they will often have at least 2, home and work.

`user_agent` is the browser+OS version, allowing to distinguish between devices. For example a user is likely to use both phone and laptop, but is unlikely to use windows+apple laptops. It is unlikely that the same session id has multiple useragents.

Data might look as the fiddle here: <http://sqlfiddle.com/#!2/c4de40/1> 

Of course, we are talking about assumptions, but it's about getting as close to reality as possible. For example, if we encounter the same ip and useragent in a limited time frame with a different session_id, it would be a fair assumption that it's the same user, with some edge case exceptions.

Edit: Language in which the problem is solved is irrelevant, it's mostly about logic and not implementation. Pseudocode is fine.

Edit: due to the slow nature of the fiddle, you can alternatively read/run the mysql:

[Skip code block](#)

```
select session_id, floor(rand()*256*256*256*256) as ip_num , floor(rand()*1000) as user_agent_id
from
  (select 1+a.nr+10*b.nr as session_id, ceil(rand()*3) as nr
  from
    (select 1 as nr union all select 2 union all select 3 union all select 4 union all select 5
     union all select 6 union all select 7 union all select 8 union all select 9 union all select
   join
    (select 1 as nr union all select 2 union all select 3 union all select 4 union all select 5
     union all select 6 union all select 7 union all select 8 union all select 9 union all select
     order by 1
  )d
inner join
  (select 1 as nr union all select 2 union all select 3 union all select 4 union all select 5
   union all select 6 union all select 7 union all select 8 union all select 9 )e
on d.nr>=e.nr
```

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [adrianbr](#) 

[Answer](#)  by [cwharland](#) 

One possibility here (and this is really an extension of what Sean Owen posted) is to define a “stable user.”

For the given info you have you can imagine making a user_id that is a hash of ip and some user agent info (pseudo code):

```
uid = MD5Hash(ip + UA.device + UA.model)
```

Then you flag these ids with “stable” or “unstable” based on usage heuristics you observe for your users. This can be a threshold of # of visits in a given time window, length of time their cookies persist, some end action on your site (I realize this wasn’t stated in your original log), etc...

The idea here is to separate the users that don’t drop cookies from those that do.

From here you can attribute session_ids to stable uids from your logs. You will then have “left over” session_ids for unstable users that you are relatively unsure about. You may be over or under counting sessions, attributing behavior to multiple people when there is only one, etc... But this is at least limited to the users you are now “less certain” about.

You then perform analytics on your stable group and project that to the unstable group. Take a user count for example, you know the total # of sessions, but you are unsure of how many users generated those sessions. You can find the # sessions / unique stable user and use this to project the “estimated” number of unique users in the unstable group since you know the number of sessions attributed to that group.

```
projected_num_unstable_users = num_sess_unstable / num_sess_per_stable_uid
```

This doesn’t help with per user level investigation on unstable users but you can at least get some mileage out of a cohort of stable users that persist for some time. You can, by various methods, project behavior and counts into the unstable group. The above is a simple example of something you might want to know. The general idea is again to define a set of users you are confident persist, measure what you want to measure, and use certain ground truths (num searches, visits, clicks, etc...) to project into the unknown user space and estimate counts for them.

This is a longstanding problem in unique user counting, logging, etc... for services that don’t require log in.

[Answer](#)  by [sean-owen](#) 

There’s not much you can do with just this data, but what little you can do does not rely on machine learning.

Yes, sessions from the same IP but different User-Agents are almost certainly distinct users. Sessions with the same IP and User-Agent are usually the same user, except in the case of proxies / wi-fi access points. Those you might identify by looking at the distribution of session count per IP to identify likely ‘aggregate’ IPs. Sessions from the same IP / User-Agent that overlap in time are almost surely distinct.

To further distinguish users you would need more info. For example, the sites or IP

addresses that the user is connecting to would be a very strong basis for differentiating sessions. Then you could get into more sophisticated learning to figure out when sessions are the same or different users.

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

[Q: K-means vs. online K-means](#)

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

[K-means](#)  is a well known algorithm for clustering, but there is also an online variation of such algorithm (online K-means). What are the pros and cons of these approaches, and when should each be preferred?

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#) 

[Answer](#)  by [christopher-louden](#) 

Online k-means (more commonly known as [sequential k-means](#) ) and traditional k-means are very similar. The difference is that online k-means allows you to update the model as new data is received.

Online k-means should be used when you expect the data to be received one by one (or maybe in chunks). This allows you to update your model as you get more information about it. The drawback of this method is that it is dependent on the order in which the data is received ([ref](#) .

[Answer](#)  by [anony-mousse](#) 

The original MacQueen k-means publication (the first to use the name “kmeans”) is an online algorithm.

MacQueen, J. B. (1967). “Some Methods for classification and Analysis of Multivariate Observations”. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297

After assigning each point, the mean is incrementally updated using a simple weighted-average formula (old mean is weighted with n, the new observation is weighted with 1, if the mean had n observations before).

As far as I can tell, it was also meant to be a single pass over the data only, although it can be trivially repeated multiple times to reassign points until convergence.

MacQueen usually takes fewer iterations than Lloyds to converge if your data is shuffled (because it updates the mean faster!). On ordered data, it can have problems. On the downside, it requires more computation for each object, so each iteration takes slightly longer (additional math operations, obviously).

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#)), [algorithms](#) ([Prev Q](#)) ([Next Q](#))

[Q: Binning long-tailed / pareto data before clustering](#)

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

I want to cluster a set of long-tailed /pareto-alike data into several bins (Actually the bin number is not determined yet). Is there any algorithms or models I can use?

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [rossini](#)

[Answer](#) by [ihars](#)

There are several approaches. You can start from the second one.

Equal-width (distance) partitioning:

- It divides the range into N intervals of equal size: uniform grid
- if A and B are the lowest and highest values of the attribute, the width of intervals will be: $w = (B-A)/N$.
- The most straightforward - Outliers may dominate presentation - Skewed data is not handled well.

Equal-depth (frequency) partitioning:

- It divides the range into N intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky.

Other Methods

- **Rank:** The rank of a number is its size relative to other values of a numerical variable. First, we sort the list of values, then we assign the position of a value as its rank. Same values receive the same rank but the presence of duplicate values affects the ranks of subsequent values (e.g., 1,2,3,3,5). Rank is a solid binning method with one major drawback, values can have different ranks in different lists.
- **Quantiles** (median, quartiles, percentiles, . . .): Quantiles are also very useful binning methods but like Rank, one value can have different quantile if the list of values changes.
- **Math functions:** For example, logarithmic binning is an effective method for the numerical variables with highly skewed distribution (e.g., income).

Entropy-based Binning

[Entropy based method](#) uses a split approach. The entropy (or the information content) is calculated based on the class label. Intuitively, it finds the best split so that the bins are as pure as possible that is the majority of the values in a bin correspond to have the same

class label. Formally, it is characterized by finding the split with the maximal information gain.

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

Q: What is the best Data Mining algorithm for prediction based on a single variable?

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

I have a variable whose value I would like to predict, and I would like to use only one variable as predictor. For instance, predict traffic density based on weather.

Initially, I thought about using [Self-Organizing Maps](#) (SOM), which performs unsupervised clustering + regression. However, since it has an important component of dimensionality reduction, I see it as more appropriated for a large number of variables.

Does it make sense to use it for a single variable as predictor? Maybe there are more adequate techniques for this *simple* case: I used “Data Mining” instead of “machine learning” in the title of my question, because I think maybe a linear regression could do the job...

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [doublebyte](#)

Answer by [ffriend](#)

Common rule in machine learning is to **try simple things first**. For predicting continuous variables there's nothing more basic than **simple linear regression**. “Simple” in the name means that there's only one predictor variable used (+ intercept, of course):

```
y = b0 + x*b1
```

where b_0 is an intercept and b_1 is a slope. For example, you may want to predict lemonade consumption in a park based on temperature:

```
cons = b0 + temp * b1
```

Temperature is a well-defined **continuous** variable. But if we talk about something more abstract like “weather”, then it's harder to understand how we measure and encode it. It's ok if we say that the weather takes values {terrible, bad, normal, good, excellent} and assign values numbers from -2 to +2 (implying that “excellent” weather is twice as good as “good”). But what if the weather is given by words {shiny, rainy, cool, ...}? We can't give an order to these variables. We call such variables **categorical**. Since there's no natural order between different categories, we can't encode them as a single numerical variable (and linear regression expects numbers only), but we can use so-called **dummy encoding**: instead of a single variable `weather` we use 3 variables - [weather_shiny, weather_rainy, weather_cool], only one of which can take value 1, and others should take value 0. In fact, we will have to drop one variable because of [collinearity](#). So model for predicting traffic from weather may look like this:

```
traffic = b0 + weather_shiny * b1 + weather_rainy * b2 # weather_cool dropped
```

where either b1 or b2 is 1, or both are 0.

Note that you can also encounter non-linear dependency between predictor and predicted variables (you can easily check it by plotting (x, y) pairs). Simplest way to deal with it without refusing linear model is to use polynomial features - simply add polynomials of your feature as new features. E.g. for temperature example (for dummy variables it doesn't make sense, cause 1^n and 0^n are still 1 and 0 for any n):

```
traffic = b0 + temp * b1 + temp^2 * b2 [+ temp^3 * b3 + ...]
```

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

[Q: Using Clustering in text processing](#)

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

Hi this is my first question in the Data Science stack. I want to create an algorithm for text classification. Suppose i have a large set of text and articles. Lets say around 5000 plain texts. I first use a simple function to determine the frequency of all the four and above character words. I then use this as the feature of each training sample. Now i want my algorithm to be able to cluster the training sets according to their features, which here is the frequency of each word in the article. (Note that in this example, each article would have its own unique feature since each article has a different feature, for example an article has 10 "water" and 23 "pure" and another has 8 "politics" and 14 "leverage"). Can you suggest the best possible clustering algorithm for this example?

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [rashid](#) 

[Answer](#)  by [jc-r](#) 

I don't know if you ever read SenseCluster by Ted Pedersen :

<http://senseclusters.sourceforge.net/> . Very good paper for sense clustering.

Also, when you analyze words, think that "computer", "computers", "computering", ... represent one concept, so only one feature. Very important for a correct analysis.

To speak about the clustering algorithm, you could use a [hierarchical clustering](#) . At each step of the algo, you merge the 2 most similar texts according to their features (using a measure of dissimilarity, euclidean distance for example). With that measure of dissimilarity, you are able to find the best number of clusters and so, the best clustering for your texts and articles.

Good luck :)

[Answer](#)  by [emre](#) 

If you want to proceed on your existing path I suggest normalizing each term's frequency by its popularity in the entire corpus, so rare and hence predictive words are promoted.

Then use random projections to reduce the dimensionality of these very long vectors down to size so your clustering algorithm will work better (you don't want to cluster in high dimensional spaces).

But there are other ways of topic modeling. Read [this](#) tutorial to learn more.

[Answer](#) by [chen-guo](#)

Cannot say it is the best one, but Latent Semantic Analysis could be one option. Basically it is based on co-occurrence, you need to weight it first.

http://en.wikipedia.org/wiki/Latent_semantic_analysis

<http://lsa.colorado.edu/papers/dp1.LSAintro.pdf>

The problem is that LSA does not have firm statistic support.

Have fun

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: Fast k-means like algorithm for \$10^{10}\$ points?](#)

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

I am looking to do k-means clustering on a set of 10-dimensional points. The catch: **there are 10^{10} points.**

I am looking for just the center and size of the largest clusters (let's say 10 to 100 clusters); I don't care about what cluster each point ends up in. Using k-means specifically is not important; I am just looking for a similar effect, any approximate k-means or related algorithm would be great (minibatch-SGD means, ...). Since GMM is in a sense the same problem as k-means, doing GMM on the same size data is also interesting.

At this scale, subsampling the data probably doesn't change the result significantly: the odds of finding the same top 10 clusters using a 1/10000th sample of the data are very good. But even then, that is a 10^6 point problem which is on/beyond the edge of tractable.

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

User: [alex-i](#)

[Answer](#) by [anony-mousse](#)

k-means is based on **averages**.

It models clusters using means, and thus **the improvement by adding more data is marginal**. The error of the average estimation reduces with $1/\sqrt{n}$; so adding more data pays off less and less...

Strategies for such large data always revolve around sampling:

If you want sublinear runtime, you have to do sampling!

In fact, Mini-Batch-Kmeans etc. do exactly this: repeatedly sample from the data set.

However, sampling (in particular unbiased sampling) isn't exactly free either... usually, you will have to read your data linearly to sample, because you don't get random access to individual records.

I'd go with MacQueen's algorithm. It's online; by default it does a single pass over your data (although it is popular to iterate this). It's not easy to distribute, but I guess you can afford to linearly read your data say 10 times from a SSD?

[Answer](#)  by [kasra-manshaei](#) 

As a side comment note that using K-means for 10D data **might** end up in nowhere according to the curse of dimensionality. Of course it varies a bit according to the nature of the data but once I tried to determine the threshold in which K-Means starts behaving strange regarding the dimensionality, I got something like 7D. After 7 dimensions it started to miss correct clusters (my data was manually generated according to 4 well-separated Gaussian distributions and I used MATLAB *kmeans* function for my little experiment).

Tags: [clustering](#) ([Prev Q](#)) ([Next Q](#))

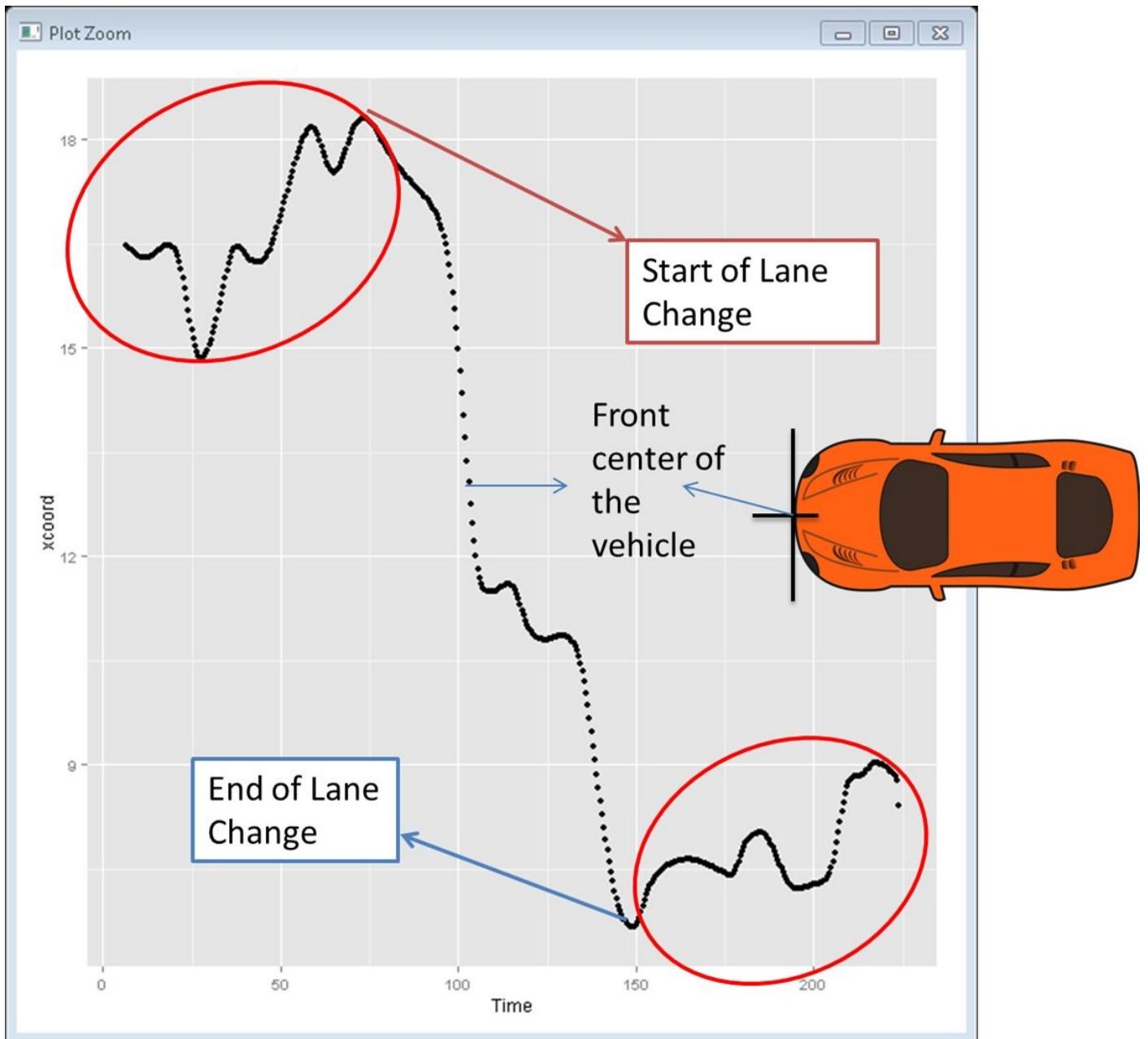
[Q: How to create clusters of position data?](#)

Tags: [clustering](#) ([Prev Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#))

I am asking this question because the [previous](#)  one wasn't very helpful and I asked about a different solution for the same problem.

The Problem

I have lateral positions, `xcoord`, of vehicles over time which were recorded as the distances from the right edge of the road. This can be seen for one vehicle in the following plot:



Each point on the plot represents the position of the front center of the vehicle. When the vehicle changes the lane (lane numbers not shown) there is a drastic change in the position as seen after the 'Start of Lane Change' on the plot.

The data behind this plot are like below:

	Vehicle.ID	Frame.ID	xcoord	Lane
1	2	13	16.46700	2
2	2	14	16.44669	2
3	2	15	16.42600	2
4	2	16	16.40540	2
5	2	17	16.38486	2
6	2	18	16.36433	2

I want to identify the start and end data points of a lane change by clustering the data as shown in the plot. The data points in the plot circled in red are more similar to each other because the variation between them is smaller compared to the data points in the middle which see large variation in position (xcoord).

My questions are: Is it possible to apply any clustering technique to segment these data so that I could identify the start and end point of a lane change? If yes, which technique would be most suitable?

I use R. I have tried Hierarchical clustering before but don't know how to apply it in this context. Please help.

Tags: [clustering](#) ([Prev Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#))

User: [umair-durrani](#) 

[Answer](#)  by [anony-mousse](#) 

I doubt any of the clustering algorithms will work well.

Instead, you should look into:

- segmentation (yes, this is something different), specifically time series segmentation
- change detection (as you said, there is a rather constant distribution first, then a change, then a rather constant distribution again)
- segment-wise regression may also work: try to find the best fit that is constant, linearly changing, and constant again. It's essentially four parameters to optimize in this restricted model: average before and after + beginning and end of transition.

Tags: [clustering](#) ([Prev Q](#)), [r](#) ([Prev Q](#)) ([Next Q](#))

R

[Skip to questions,](#)

Wiki by user [tasos](#) 

[R](#)  is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R was created by [Ross Ihaka](#)  and [Robert Gentleman](#)  and is now developed by the [R Development Core Team](#) . The R environment is easily extended through a packaging system on [CRAN](#) .

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and Mac OS.

Questions

[Q: Running an R script programmatically](#)

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

I have an R script that generates a report based on the current contents of a database. This database is constantly in flux with records being added/deleted many times each day. How can I ask my computer to run this every night at 4 am so that I have an up to date report waiting for me in the morning? Or perhaps I want it to re-run once a certain number of new records have been added to the database. How might I go about automating this? I should mention I'm on Windows, but I could easily put this script on my Linux machine if that would simplify the process.

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

User: [rnorberg](#) 

[Answer](#)  by [adrianbr](#) 

For windows, use the task scheduler to set the task to run for example daily at 4:00 AM

It gives you many other options regarding frequency etc.

http://en.wikipedia.org/wiki/Windows_Task_Scheduler 

[Answer](#)  by [asheeshr](#) 

How can I ask my computer to run this every night at 4 am so that I have an up to date report waiting for me in the morning?

You can set up a cronjob on a Linux system. These are run at the set time, if the computer is on. To do so, open a terminal and type:

```
crontab -e
```

And add:

```
00 4 * * * r source(/home/FilePath/.../myRscript.R)
```

Source: [Stack Overflow](#) 

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#)), [databases](#) ([Prev Q](#)) ([Next Q](#))

[Q: What are R's memory constraints?](#)

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

In reviewing “[Applied Predictive Modeling](#)” a [reviewer states](#) :

One critique I have of statistical learning (SL) pedagogy is the absence of computation performance considerations in the evaluation of different modeling techniques. With its emphases on bootstrapping and cross-validation to tune/test models, SL is quite compute-intensive. Add to that the re-sampling that's embedded in techniques like bagging and boosting, and you have the specter of computation hell for supervised learning of large data sets. **In fact, R's memory constraints impose pretty severe limits on the size of models that can be fit by top-performing methods like random forests.** Though SL does a good job calibrating model performance against small data sets, it'd sure be nice to understand performance versus computational cost for larger data.

What are R's memory constraints, and do they impose severe limits on the size of models that can be fit by top-performing methods like [random forests](#)?

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

User: [blunders](#)

[Answer](#) by [binga](#)

As Konstantin has pointed, R performs all its computation in the system's memory i.e. RAM. Hence, RAM capacity is a very important constraint for computation intensive operations in R. Overcoming this constraint, data is being stored these days in HDFS systems, where data isn't loaded onto memory and program is run instead, program goes to the data and performs the operations, thus overcoming the memory constraints.

RHadoop (<https://github.com/RevolutionAnalytics/RHadoop/wiki>) is the connector you are looking for.

Coming to the impact on algorithms which are computation intensive, Random Forests/Decision Trees/Ensemble methods on a considerable amount of data (minimum 50,000 observations in my experience) take up a lot of memory and are considerably slow. To speed up the process, parallelization is the way to go and parallelization is inherently available in Hadoop! That's where, Hadoop is really efficient.

So, if you are going for ensemble methods which are compute intensive and are slow, you would want to try out on the HDFS system which gives a considerable performance improvement.

[Answer](#) by [konstantin-v.-salikhov](#)

R performs all computation in-memory so you can't perform operation on a dataset that is larger than available RAM amount. However there are some libraries that allow bigdata processing using R and one of popular libraries for bigdata processing like Hadoop.

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [hadoop](#) ([Prev Q](#)) ([Next Q](#))

Q: R error using package tm (text-mining)

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

I am attempting to use the tm package to convert a vector of text strings to a corpus element.

My code looks something like this

Corpus(d1\$Yes)

where d1\$Yes is a factor with 124 levels, each containing a text string.

For example, d1\$Yes[246] = "So we can get the boat out!"

I'm receiving the following error: "Error: inherits(x, "Source") is not TRUE"

I'm not sure how to remedy this.

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [ivoire](#) 

[Answer](#)  by [nfmclure](#) 

You have to tell Corpus what kind of source you are using. Try:

```
corpus(VectorSource(d1$Yes))
```

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [text-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: Do you have to normalize data when building decision trees using R?](#)

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [beginner](#) ([Prev Q](#)) ([Next Q](#))

So, our data set this week has 14 attributes and each column has very different values. One column has values below 1 while another column has values that go from three to four whole digits.

We learned normalization last week and it seems like you're supposed to normalize data when they have very different values. For decision trees, is the case the same?

I'm not sure about this but would normalization affect the resulting decision tree from the same data set? It doesn't seem like it should but...

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [beginner](#) ([Prev Q](#)) ([Next Q](#))

User: [jae](#) 

[Answer](#)  by [rapaio](#) 

Most common types of decision trees you encounter are not affected by any monotonic transformation. So, as long as you preserve order, the decision trees are the same (obviously by the same tree here I understand the same decision structure, not the same values for each test in each node of the tree).

The reason why it happens is because how usual impurity functions work. In order to find the best split it searches on each dimension (attribute) a split point which is basically an if

clause which groups target values corresponding to instances which has test value less than split value, and on the right the values greater than equal. This happens for numerical attributes (which I think is your case because I do not know how to normalize a nominal attribute). Now you might note that the criteria is less than or greater than. Which means that the real information from the attributes in order to find the split (and the whole tree) is only the order of the values. Which means that as long as you transform your attributes in such a way that the original ordering is reserved, you will get the same tree.

Not all models are insensitive to such kind of transformation. For example linear regression models give the same results if you multiply an attribute with something different than zero. You will get different regression coefficients, but the predicted value will be the same. This is not the case when you take a log of that transformation. So for linear regression, for example, normalizing is useless since it will provide the same result.

However this is not the case with a penalized linear regression, like ridge regression. In penalized linear regressions a constraint is applied to coefficients. The idea is that the constraint is applied to the sum of a function of coefficients. Now if you inflate an attribute, the coefficient will be deflated, which means that in the end the penalization for that coefficient it will be artificially modified. In such kind of situation, you normalize attributes in order that each coefficient to be constraint ‘fairly’.

Hope it helps

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [beginner](#) ([Prev Q](#)) ([Next Q](#))

Q: visualize a horizontal box plot in R

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

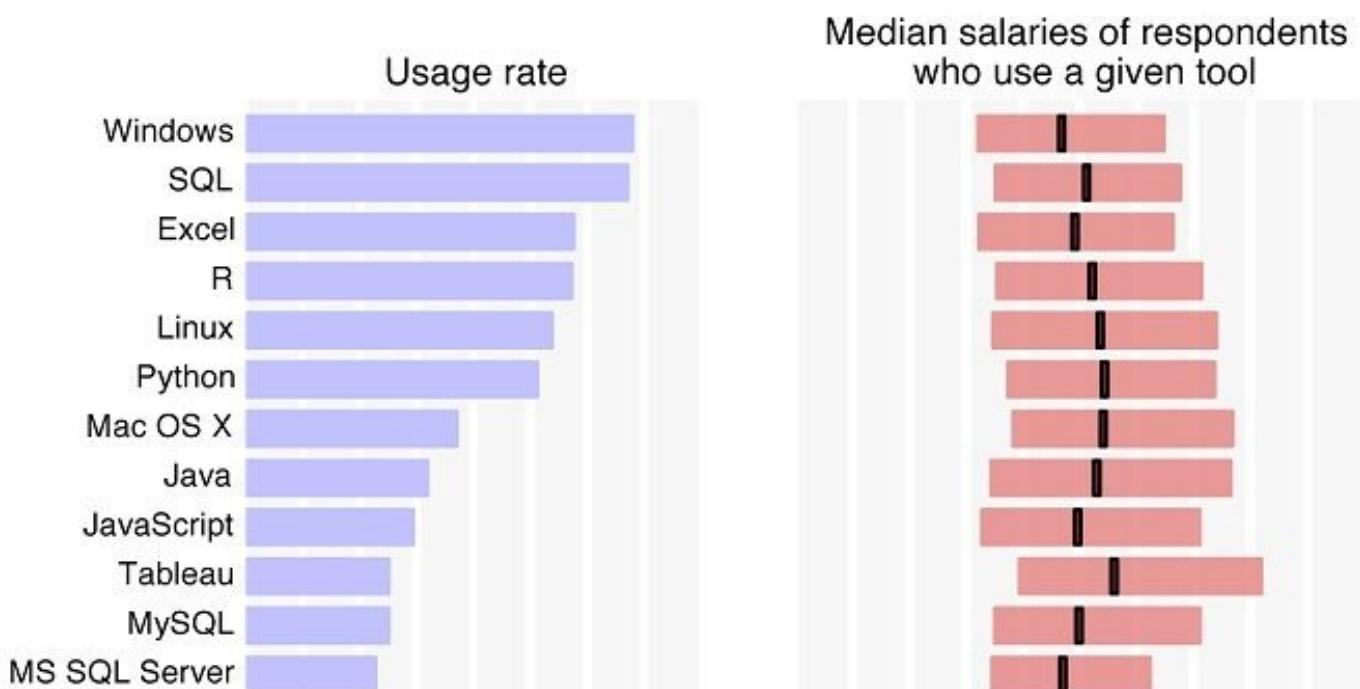
I have a dataset like this. The data has been collected through a questionnaire and I am going to do some exploratory data analysis.

```
windows <- c("yes", "no", "yes", "yes", "no")
sql     <- c("no", "yes", "no", "no", "no")
excel   <- c("yes", "yes", "yes", "no", "yes")
salary  <- c(100, 200, 300, 400, 500)

test<- as.data.frame (cbind(windows,sql,excel,salary),stringsAsFactors=TRUE)
test[,"salary"] <- as.numeric(as.character(test[,"salary"]))
```

I have an outcome variable (salary) in my dataset and a couple of input variables (tools). How can I visualize a horizontal box plot like this:

Most commonly used tools (used by at least 10% of sample)



Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

User: [hamideh-iraj](#)

[Answer](#) by [nitesh](#)

Let's start by creating some fake dataset.

```
software = sample(c("Windows", "Linux", "Mac"), n=100, replace=T)
salary = runif(n=100, min=1, max=100)
test = data.frame(software, salary)
```

This should create a dataframe test that will look like somewhat like:

[skip code block](#)

```
software    salary
1   Windows  96.697217
2       Linux  29.770905
```

```

3   Windows 94.249612
4       Mac 71.188701
5     Linux 94.028326
6     Linux 7.482632
7       Mac 98.841689
8       Mac 81.152623
9   Windows 54.073761
10  Windows 1.707829

```

EDIT based on comment Note that if the data does not already exist in the above format, it can be changed to this format. Let's take a data frame provided in the original question and lets assume the dataframe is called `raw_test`.

```

windows sql excel salary
1   yes  no  yes    100
2   no   yes yes    200
3   yes  no  yes    300
4   yes  no   no    400
5   no   no  yes    500

```

Now, using the `melt` function/ method from the `reshape` package in R, first create the dataframe `test` (that will be used for final plotting) as follows:

```

# use melt to convert from wide to long format
test = melt(raw_test,id.vars=c("salary"))
# subset to only select where value is "yes"
test = subset(test, value == 'yes')
# replace column name from "variable" to "software"
names(test)[2] = "software"

```

Now, you will get a datframe `test` that looks like:

```

salary software value
1      100 windows  yes
3      300 windows  yes
4      400 windows  yes
7      200     sql  yes
11     100 excel   yes
12     200 excel   yes
13     300 excel   yes
15     500 excel   yes

```

Having created the dataset. We will now generate the plot.

First, create the bar plot on the left based on the counts of software that represents usage rate.

```
p1 <- ggplot(test, aes(factor(software))) + geom_bar() + coord_flip()
```

Next, create the boxplot on the right.

```
p2 <- ggplot(test, aes(factor(software), salary)) + geom_boxplot() + coord_flip()
```

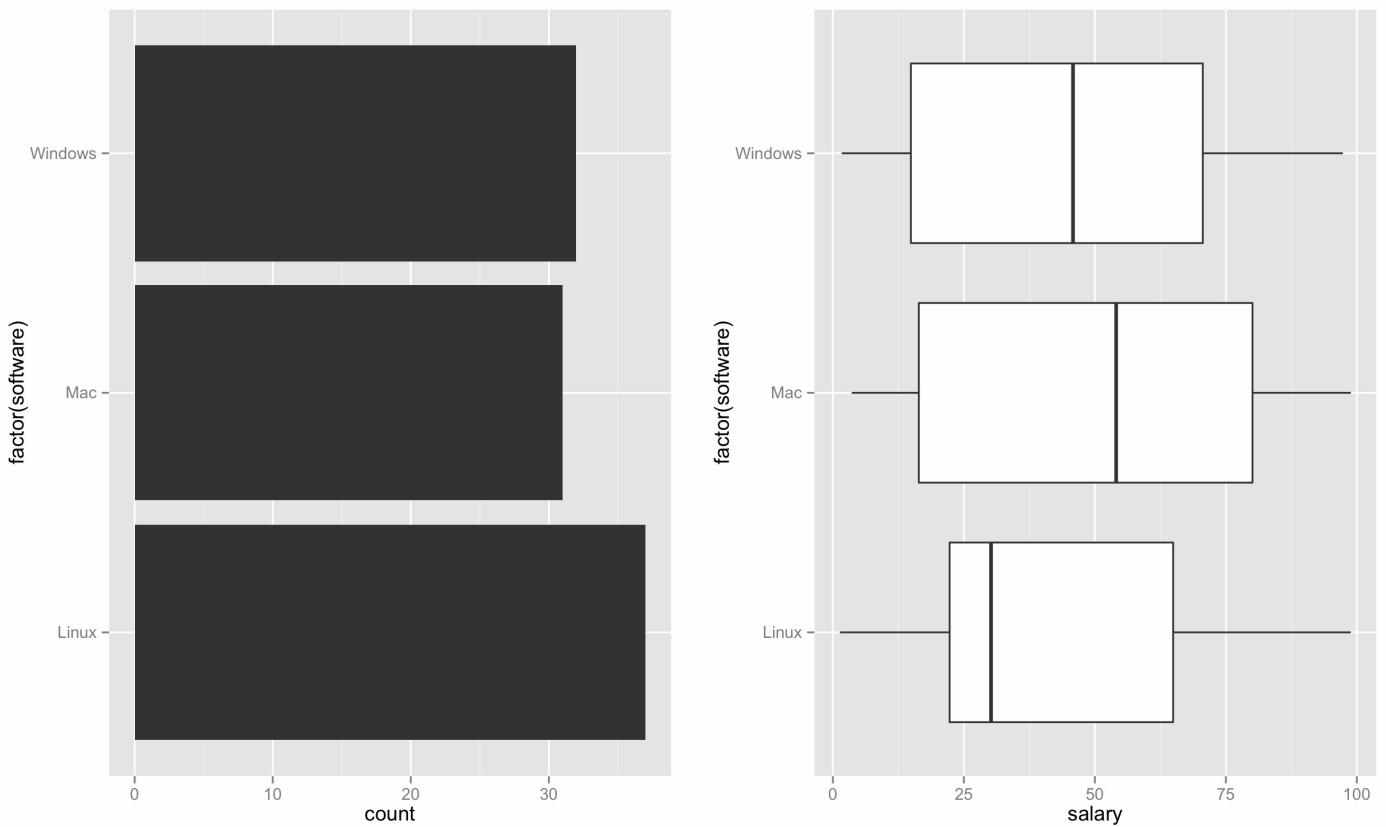
Finally, place both these plots next to each other.

```

require('gridExtra')
grid.arrange(p1,p2,nrow=1)

```

This should create a plot like:



[Answer](#) by [lauren-goodwin](#)

You are going to need to make a column that contains software info— for example name it software and the salary column has the corresponding salary so something like

Software	Salary
Microsoft	100
Microsoft	300
Microsoft	400
SQL	200

and so on...then you can plot with the code below

```
p <- ggplot(test, aes(factor(software), salary))
p + geom_boxplot() + coord_flip()
```

Tags: [r](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

[Q: Software Testing for Data Science in R](#)

Tags: [r](#) ([Prev Q](#))

I often use [Nose, Tox or Unittest](#) when testing my python code, specially when it has to be integrated with other modules or other pieces of code. However, now that I've found myself using R more than python for ML modelling and development. I realized that I don't really test my R code (And more importantly I really don't know how to do it well). So my question is, what are good packages that allow you to test R code in a similar manner as Nose, Tox or Unittest do in Python. Additional references such as tutorials will be greatly appreciated as well.

Bonus points for packages in R similar to

1. [Hypothesis](#)

or

2. [Feature Forge](#)

Related Talk:

[Trey Causey: Testing for Data Scientists](#)

Tags: [r](#) ([Prev Q](#))

User: [wacax](#)

[Answer](#) by [phiver](#)

Packages for unit testing and assertive testing that are actively maintained: Packages for unit testing

1. testthat: more information on how to use you can find [here](#) or on [github](#)
2. Runit: [Cran page](#)

Packages for assertions:

1. assertthat: info on [github](#)
2. assertive: Assertive has a lot of subpackages available in case you do not need all of them. check on cran
3. assertr: info on [github](#)
4. ensurer: info on [github](#)
5. tester: info on [github](#)

It is a matter of preference what you want to use for assertions. Read [this bioconductor](#) page for more info on the difference between RUnit and testthat.

Tags: [r](#) ([Prev Q](#))

Text Mining

Questions

[Q: How to grow a list of related words based on initial keywords?](#)

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

I recently saw a cool feature that [was once available](#)  in Google Sheets: you start by writing a few related keywords in consecutive cells, say: “blue”, “green”, “yellow”, and it automatically generates similar keywords (in this case, other colors). See more examples in [this YouTube video](#) .

I would like to reproduce this in my own program. I’m thinking of using Freebase, and it would work like this intuitively:

1. Retrieve the list of given words in Freebase;
2. Find their “common denominator(s)” and construct a distance metric based on this;
3. Rank other concepts based on their “distance” to the original keywords;
4. Display the next closest concepts.

As I’m not familiar with this area, my questions are:

- Is there a better way to do this?
- What tools are available for each step?

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [cafe876](#) 

[Answer](#)  by [joews](#) 

The [word2vec algorithm](#)  may be a good way to retrieve more elements for a list of similar words. It is an unsupervised “deep learning” algorithm that has previously been demonstrated with Wikipedia-based training data (helper scripts are provided on the Google code page).

There are currently [C](#)  and [Python](#)  implementations. This [tutorial](#)  by [Radim Řehůřek](#) , the author of the [Gensim topic modelling library](#) , is an excellent place to start.

The “[single topic](#)”  demonstration on the tutorial is a good example of retrieving similar words to a single term (try searching on ‘red’ or ‘yellow’). It should be possible to extend this technique to find the words that have the greatest overall similarity to a set of input words.

[Answer](#)  by [charlie-greenbacker](#) 

Have you considered a frequency-based approach exploiting simple word co-occurrence in corpora? At least, that's what I've seen most folks use for this. I think it might be covered briefly in Manning and Schütze's book, and I seem to remember something like this as a homework assignment back in grad school...

More background here: <http://nlp.stanford.edu/IR-book/html/htmledition/automatic-thesaurus-generation-1.html> 

For this step:

Rank other concepts based on their “distance” to the original keywords;

There are several semantic similarity metrics you could look into. Here's a link to some slides I put together for a class project using a few of these similarity metrics in WordNet: <http://www.eecis.udel.edu/~trnka/CISC889-11S/lectures/greenbacker-WordNet-Similarity.pdf> 

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

Q: Unsupervised Feature Learning for NER

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

I have implemented NER system with the use of CRF algorithm with my handcrafted features that gave quite good results. The thing is that I used lots of different features including POS tags and lemmas.

Now I want to make the same NER for different language. The problem here is that I can't use POS tags and lemmas. I started reading articles about deep learning and unsupervised feature learning.

My question is, if it's possible to use methods for unsupervised feature learning with CRF algorithm. Did anyone try this and got any good result? Is there any article or tutorial about this matter.

I still don't completely understand this way of feature creation so I don't want to spend too much time for something that won't work. So any information would be really helpful. To create whole NER system based on deep learning is a bit too much for now.

Thank you in advance.

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [maticdiba](#) 

[Answer](#)  by [madison-may](#) 

Yes, it is entirely possible to combine unsupervised learning with the CRF model. In particular, I would recommend that you explore the possibility of using [word2vec](#)  features as inputs to your CRF.

Word2vec trains a to distinguish between words that are appropriate for a given context

and words that are randomly selected. Select weights of the model can then be interpreted as a dense vector representation of a given word.

These dense vectors have the appealing property that words that are semantically or syntactically similar have similar vector representations. Basic vector arithmetic even reveals some interesting learned relationships between words.

For example, $\text{vector}(\text{"Paris"}) - \text{vector}(\text{"France"}) + \text{vector}(\text{"Italy"})$ yields a vector that is quite similar to $\text{vector}(\text{"Rome"})$.

At a high level, you can think of word2vec representations as being similar to LDA or LSA representations, in the sense that you can convert a sparse input vector into a dense output vector that contains word similarity information.

For that matter, LDA and LSA are also valid options for unsupervised feature learning — both attempt to represent words as combinations of “topics” and output dense word representations.

For English text Google distributes word2vec models pretrained on a huge 100 billion word Google News dataset, but for other languages you’ll have to train your own model.

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: Clustering strings inside strings?](#)

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

I am not sure whether I formulated the question correctly. Basically, what I want to do is:

Let’s suppose I have a list of 1000 strings which look like this:

cvzxcvzx**string**cvzcxvz

ototorotr**string**grptprt

vmvmvmeop**string**2vmrprp

vccermpqp**string**2rowerm

proororor**string**3potrprt

mprto2435**string**3famerpaer

etc.

I’d like to extract these reoccurring strings that occur on the list. What solution should I use? Does anyone know about algorithm that could do this?

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [gggggggg5555](#) 

[Answer](#)  by [emre](#) 

Interesting question! I have not encountered it before so here is a solution I just made up, inspired by the approach taken by the word2vec paper:

1. Define the pair-wise similarity based on the longest common substring (LCS), or the LCS normalized by the products of the string lengths. Cache this in a matrix for any pair of strings considered since it is expensive to calculate. Also consider approximations.
2. Find a Euclidean (hyperspherical, perhaps?) embedding that minimizes the error (Euclidean distance if using the ball, and the dot product if using the sphere). Assume random initialization, and use a gradient-based optimization method by taking the Jacobian of the error.
3. Now you have a Hilbert space embedding, so cluster using your algorithm of choice!

Response to deleted comment asking how to cluster multiple substrings: The bulk of the complexity lies in the first stage; the calculation of the LCS, so it depends on efficiently you do that. I've had luck with genetic algorithms. Anyway, what you'd do in this case is define a similarity *vector* rather than a scalar, whose elements are the k-longest pair-wise LCS; see [this](#) discussion for algorithms. Then I would define the error by the sum of the errors corresponding to each substring.

Something I did not address is how to choose the dimensionality of the embedding. The word2vec paper might provide some heuristics; see [this](#) discussion. I recall they used pretty big spaces, on the order of a 1000 dimensions, but they were optimizing something more complicated, so I suggest you start at R^2 and work your way up. Of course, you will want to use a higher dimensionality for the multiple LCS case.

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: Ethically and Cost-effectively Scaling Data Scraps](#)

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#))

Few things in life give me pleasure like scraping structured and unstructured data from the Internet and making use of it in my models.

For instance, the Data Science Toolkit (or `RDSTK` for R programmers) allows me to pull lots of good location-based data using IP's or addresses and the `tm.webmining.plugin` for R's `tm` package makes scraping financial and news data straightforward. When going beyond such (semi-) structured data I tend to use `XPath`.

However, I'm constantly getting throttled by limits on the number of queries you're allowed to make. I think Google limits me to about 50,000 requests per 24 hours, which is a problem for Big Data.

From a *technical* perspective getting around these limits is easy — just switch IP addresses and purge other identifiers from your environment. However, this presents both ethical and financial concerns (I think?).

Is there a solution that I'm overlooking?

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#))

User: [hack-r](#)

[Answer](#) by [rawkintrevo](#)

For many APIs (most I've seen) ratelimiting is a function of your API Key or OAuth credentials. (Google, Twitter, NOAA, Yahoo, Facebook, etc.) The good news is you won't need to spoof your IP, you just need to swap out credentials as they hit their rate limit.

A bit of shameless self promotion here but I wrote a python package specifically for handling this problem.

<https://github.com/rawkintrevo/angemilner>

<https://pypi.python.org/pypi/angemilner/0.2.0>

It requires a `mongodb` daemon and basically you make a page for each one of your keys. So you have 4 email addresses each with a separate key assigned. When you load the key in you specify the maximum calls per day and minimum time between uses.

Load keys:

```
from angemilner import APIKeyLibrarian
l= APIKeyLibrarian()
l.new_api_key("your_assigned_key1", 'noaa', 1000, .2)
l.new_api_key("your_assigned_key2", 'noaa', 1000, .2)
```

Then when you run your scraper for instance the NOAA api:

```
url= 'http://www.ncdc.noaa.gov/cdo-web/api/v2/stations'
payload= { 'limit': 1000,
          'datasetid': 'GHCND',
          'startdate': '1999-01-01' }
```

```
r = requests.get(url, params=payload, headers= {'token': 'your_assigned_key'})
```

becomes:

```
url= 'http://www.ncdc.noaa.gov/cdo-web/api/v2/stations'  
payload= { 'limit': 1000,  
          'datasetid': 'GHCND',  
          'startdate': '1999-01-01' }  
  
r = requests.get(url, params=payload, headers= {'token': l.check_out_api_key('noaa')['key']})
```

so if you have 5 keys, `l.check_out_api_key` returns the key that has the least uses and waits until enough time has elapsed for it to be used again.

Finally to see how often your keys have been used / remaining useage available:

```
pprint(l.summary())
```

I didn't write this for R because most scraping is done in python (most of MY scraping). It could be easily ported.

Thats how you can *technically* get around rate limiting. *Ethically* ...

UPDATE The example uses Google Places API [here](#) 

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#))

[Q: Extract most informative parts of text from documents](#)

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

Are there any articles or discussions about extracting part of text that holds the most of information about current document.

For example, I have a large corpus of documents from the same domain. There are parts of text that hold the key information what single document talks about. I want to extract some of those parts and use them as kind of a summary of the text. Is there any useful documentation about how to achieve something like this.

It would be really helpful if someone could point me into the right direction what I should search for or read to get some insight in work that might have already been done in this field of Natural language processing.

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [maticdiba](#) 

[Answer](#)  by [charlie-greenbacker](#) 

What you're describing is often achieved using a simple combination of [TF-IDF](#)  and [extractive summarization](#) .

In a nutshell, TF-IDF tells you the relative importance of each word in each document, in comparison to the rest of your corpus. At this point, you have a score for each word in each document approximating its "importance." Then you can use these individual word scores to compute a composite score for each sentence by summing the scores of each

word in each sentence. Finally, simply take the top-N scoring sentences from each document as its summary.

Earlier this year, I put together an iPython Notebook that culminates with an implementation of this in Python using NLTK and Scikit-learn: [A Smattering of NLP in Python](#).

Tags: [text-mining](#) ([Prev Q](#)) ([Next Q](#)), [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is an alternative name for “Unstructured Data”?](#)

Tags: [text-mining](#) ([Prev Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#))

I'm writing my thesis at the moment, and for some time - due to a lack of a proper alternative - I've stuck with "unstructured data" for referring to natural, free flowing text, e.g. Wikipedia articles.

This nomenclature has bothered me from the very beginning, since it opens a debate that I don't want to get into. Namely, that "unstructured" implies that natural language lacks structure, which it does not - the most obvious being syntax. It also gives a negative impression, since it is the opposite of "structured", which is accepted as being positive. This is not the focus of my thesis, though the "unstructured" part itself plays an important role.

I completely agree with the writer of [this article](#), but he proposes no alternative except for "rich data", which doesn't cover my point. The point I'm trying to make that the text lacks a traditional database-like (e.g. tabular) structure of the data, with every piece of data having a clear data type and semantics that is easy to interpret using computer programs. Of course I'd like to condense this definition into a term, but so far I've been unsuccessful coming up with, or discovering an acceptable taxonomy in literature.

Tags: [text-mining](#) ([Prev Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#))

User: [benjamin-b.](#) 

[Answer](#)  by [victor](#) 

It is a bad idea to counterpose "unstructure data" to, say, tabular data (as in "non-tabular data"), as you will have to elliminate other alternatives as well (e.g., "non-tabular and non-graph and ... data"). "Plain text" (— my choice) or "raw text" or "raw data" sound fine.

[Answer](#)  by [l.-amber-o'hearn](#) 

"Raw data" is what we say in NLP.

Tags: [text-mining](#) ([Prev Q](#)), [definitions](#) ([Prev Q](#)) ([Next Q](#))

NLP

Questions

[Q: Latent Dirichlet Allocation vs Hierarchical Dirichlet Process](#)

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Latent Dirichlet Allocation \(LDA\)](#)  and [Hierarchical Dirichlet Process \(HDP\)](#)  are both topic modeling processes. The major difference is LDA requires the specification of the number of topics, and HDP doesn't. Why is that so? And what are the differences, pros, and cons of both topic modelling methods?

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [alvas](#) 

[Answer](#)  by [tim-goodman](#) 

HDP is an extension of LDA, designed to address the case where the number of mixture components (the number of “topics” in document-modeling terms) is not known a priori. So that's the reason why there's a difference.

Using LDA for document modeling, one treats each “topic” as a distribution of words in some known vocabulary. For each document a mixture of topics is drawn from a Dirichlet distribution, and then each word in the document is an independent draw from that mixture (that is, selecting a topic and then using it to generate a word).

For HDP (applied to document modeling), one also uses a Dirichlet process to capture the uncertainty in the number of topics. So a common base distribution is selected which represents the countably-infinite set of possible topics for the corpus, and then the finite distribution of topics for each document is sampled from this base distribution.

As far as pros and cons, HDP has the advantage that the maximum number of topics can be unbounded and learned from the data rather than specified in advance. I suppose though it is more complicated to implement, and unnecessary in the case where a bounded number of topics is acceptable.

[Answer](#)  by [charlie-greenbacker](#) 

Anecdotally, I've never been impressed with the output from hierarchical LDA. It just doesn't seem to find an optimal level of granularity for choosing the number of topics. I've gotten much better results by running a few iterations of regular LDA, manually inspecting the topics it produced, deciding whether to increase or decrease the number of topics, and continue iterating until I get the granularity I'm looking for.

Remember: hierarchical LDA can't read your mind... it doesn't know what you actually intend to use the topic modeling for. Just like with k-means clustering, you should choose

the k that makes the most sense for your use case.

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: What is generative and discriminative model? How are they used in Natural Language Processing?](#)

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

[This question](#)  asks about generative vs. discriminative algorithm, but can someone give an example of the difference between these forms when applied to Natural Language Processing? **How are generative and discriminative models used in NLP?**

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [alvas](#) 

[Answer](#)  by [sean-owen](#) 

Let's say you are predicting the topic of a document given its words.

A generative model describes how likely each topic is, and how likely words are given the topic. This is how it says documents are actually “generated” by the world — a topic arises according to some distribution, words arise because of the topic, you have a document. Classifying documents of words W into topic T is a matter of maximizing the joint likelihood: $P(T,W) = P(W|T)P(T)$

A discriminative model operates by only describing how likely a topic is given the words. It says nothing about how likely the words or topic are by themselves. The task is to model $P(T|W)$ directly and find the T that maximizes this. These approaches do not care about $P(T)$ or $P(W)$ directly.

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: Coreference Resolution for German Texts](#)

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

does anybody know a library for performing coreference resolution on German texts?

As far as I know OpenNLP and Standord NLP are not able to perform coreference resolution for German Texts.

The only tool that I know is [CorZu](#) which is a python library.

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

User: [pasmod-turing](#)

[Answer](#) by [sylvain-peyronnet](#)

Here is a couple of tools that may be worth a look:

- Bart, an open source tool that have been used for several languages, including German. [Available from the website](#)
 - Sucre is a tool developed at the University of Stuttgart. I don't know if it's available easily. [You can see this paper about it](#).
-

Tags: [nlp](#) ([Prev Q](#)) ([Next Q](#))

[Q: Accuracy of Stanford NER](#)

Tags: [nlp](#) ([Prev Q](#))

I am performing Named Entity Recognition using Stanford NER. I have successfully trained and tested my model. Now I want to know:

- 1) What is the general way of measuring accuracy of NER model ?? For example what techniques or approaches are used ??
- 2) Is there any built-in method in STANFORD NER for evaluating the accuracy ??

Tags: [nlp](#) ([Prev Q](#))

User: [sarmad](#)

[Answer](#) by [franck-dernoncourt](#)

http://en.wikipedia.org/wiki/Named-entity_recognition#Formal_evaluation :

To evaluate the quality of a NER system's output, several measures have been defined. While accuracy on the token level is one possibility, it suffers from two problems: the vast majority of tokens in real-world text are not part of entity names as usually defined, so the baseline accuracy (always predict "not an entity") is extravagantly high, typically >90%; and mispredicting the full span of an entity name

is not properly penalized (finding only a person's first name when their last name follows is scored as $\frac{1}{2}$ accuracy).

In academic conferences such as CoNLL, a variant of the F1 score has been defined as follows:

- Precision is the number of predicted entity name spans that line up exactly with spans in the gold standard evaluation data. I.e. when [Person Hans] [Person Blick] is predicted but [Person Hans Blick] was required, precision for the predicted name is zero. Precision is then averaged over all predicted entity names.
- Recall is similarly the number of names in the gold standard that appear at exactly the same location in the predictions.
- F1 score is the harmonic mean of these two.

It follows from the above definition that any prediction that misses a single token, includes a spurious token, or has the wrong class, "scores no points", i.e. does not contribute to either precision or recall.

Tags: [nlp](#) ([Prev Q](#))

Dataset

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Datasets are structured data files in any format, collected together with the documentation that explains their production or use.

Questions

[Q: Publicly Available Datasets](#)

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#))

One of the common problems in data science is gathering data from various sources in a somehow cleaned (semi-structured) format and combining metrics from various sources for making a higher level analysis. Looking at the other people's effort, especially other questions on this site, it appears that many people in this field are doing somewhat repetitive work. For example analyzing tweets, facebook posts, Wikipedia articles etc. is a part of a lot of big data problems.

Some of these data sets are accessible using public APIs provided by the provider site, but usually some valuable information or metrics are missing from these APIs and everyone has to do the same analyses again and again. For example, although clustering users may depend on different use cases and selection of features, but having a base clustering of Twitter/Facebook users can be useful in many Big Data applications, which is neither provided by the API, nor available publicly in independent data sets.

Is there any index or publicly available data set hosting site containing valuable data sets that can be reused in solving other big data problems? I mean something like GitHub (or a group of sites/public data sets or at least a comprehensive listing) for the data science. If not, what are the reasons of not having such a platform for data science? Commercial value of data, need to frequently update data sets, ...? Can we not have an open-source model for sharing data sets devised for data scientists?

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [amir-ali-akbari](#) 

[Answer](#)  by [rubens](#) 

There is, in fact, a very reasonable list of publicly-available datasets, supported by different enterprises/sources. Here are some of them:

- [Public Datasets on Amazon WebServices](#) 
- [Frequent Itemset Mining Implementation Repository](#) 
- [UCI Machine Learning Repository](#) 
- [KDnuggets](#)  — big list of lots of public repositories.

Now, two considerations on your question. First one, regarding policies of database sharing. From personal experience, there are some databases that can't be made publicly available, either for involving privacy restraints (as for some social network informations), or for concerning government information (like health system databases).

Another point concerns the usage/application of the dataset. Although some bases can be reprocessed to suit the needs of the application, it would be great to have some *nice organization* of the datasets by purpose. The *taxonomy* should involve social graph analysis, itemset mining, classification, and lots of other research areas there may be.

[Answer](#) by [ihars](#)

Update:

Kaggle.com, a home of modern data science & machine learning enthusiasts:), opened [it's own repository of the data sets](#).

In addition to the listed sources.

Some social network data sets:

- [Stanford University large network dataset collection \(SNAP\)](#)
- [A huge twitter dataset that includes followers + large collection of twitter datasets here](#)
- [LastFM data set](#)

There are plenty of sources listed at Stats SE:

- [Locating freely available data samples](#)
- [Data APIs/feeds available as packages in R](#)
- [Free data set for very high dimensional classification](#)

[Answer](#) by [mcp_infiltrator](#)

There are many openly available data sets, one many people often overlook is [data.gov](#). As mentioned previously Freebase is great, so are all the examples posted by @Rubens

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#))

[Q: Publicly available social network datasets/APIs](#)

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#))

As an extension to our great list of [publicly available datasets](#), I'd like to know if there is any list of publicly available social network datasets/crawling APIs. It would be very nice if alongside with a link to the dataset/API, characteristics of the data available were added. Such information should be, and is not limited to:

- the name of the social network;
- what kind of user information it provides (posts, profile, friendship network, ...);
- whether it allows for crawling its contents via an API (and rate: 10/min, 1k/month, ...);
- whether it simply provides a snapshot of the whole dataset.

Any suggestions and further characteristics to be added are very welcome.

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#)

[Answer](#) by [sobach](#)

A couple of words about social networks APIs. About a year ago I wrote a review of popular social networks' APIs for researchers. Unfortunately, it is in Russian. Here is a summary:

Twitter (<https://dev.twitter.com/docs/api/1.1>)

- almost all data about tweets/texts and users is available;
- lack of sociodemographic data;
- great streaming API: useful for real time text processing;
- a lot of wrappers for programming languages;
- getting network structure (connections) is possible, but time-expensive (1 request per 1 minute).

Facebook (<https://developers.facebook.com/docs/reference/api/>)

- rate limits: about 1 request per second;
- well documented, sandbox present;
- FQL (SQL-like) and «regular Rest» Graph API;
- friendship data and sociodemographic features present;
- a lot of data is beyond *event horizon*: only friends' and friends' of friends data is more or less complete, almost nothing could be investigated about random user;
- some strange API bugs, and looks like nobody cares about it (e.g., some features available through FQL, but not through Graph API synonym).

Instagram (<http://instagram.com/developer/>)

- rate limits: 5000 requests per hour;
- real-time API (like Streaming API for Twitter, but with photos) - connection to it is a little bit tricky: callbacks are used;
- lack of sociodemographic data;
- photos, filters data available;
- unexpected imperfections (e.g., it's possible to collect only 150 comments to post/photo).

Foursquare (<https://developer.foursquare.com/overview/>)

- rate limits: 5000 requests per hour;
- kingdom of geosocial data :)
- quite closed from researches because of privacy issues. To collect checkins data one need to build composite parser working with 4sq, bit.ly, and twitter APIs at once;
- again: lack of sociodemographic data.

Google+ (<https://developers.google.com/+/api/latest/>)

- about 5 requests per second (try to verify);
- main methods: activities and people;

- like on Facebook, a lot of personal data for random user is hidden;
- lack of user connections data.

And out-of-competition: I reviewed social networks for Russian readers, and #1 network here is vk.com. It's translated to many languages, but popular only in Russia and other CIS countries. API docs link: <http://vk.com/dev/>. And from my point of view, it's the best choice for homebrew social media research. At least, in Russia. That's why:

- rate limits: 3 requests per second;
- public text and media data available;
- sociodemographic data available: for random user availability level is about 60-70%;
- connections between users are also available: almost all friendships data for random user is available;
- some special methods: e.g., there is a method to get online/offline status for exact user in realtime, and one could build schedule for his audience.

[Answer](#) by [little-bobby-tables](#)

It's not a social network per se, but Stackexchange publish their entire database dump periodically:

- [Stackexchange data dump hosted on the archive.org](#)
- [Post describing the database dump schema](#)

You can extract some social information by analyzing which users ask and answer to each other. One nice thing is that since posts are tagged, you can analyze sub-communities easily.

[Answer](#) by [christian-sauer](#)

An example from germany: Xing a site similar to linkedin but limited to german speaking countries.

Link to it's developer central: <https://dev.xing.com/overview>

Provides access to: User profiles, Conversations between users (limited to the user itself), Job advertisings, Contacts and Contacts of Contacts, news from the network and some geolocation api.

Yes it has an api, but I did not find information about the rate. But it seems to me, that some information is limited to the consent of the user.

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#))

[Q: Interactive Graphing while logging data](#)

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

I'm looking to graph and interactively explore live/continuously measured data. There are quite a few options out there, with plot.ly being the most user-friendly. Plot.ly has a fantastic and easy to use UI (easily scalable, pannable, easily zoomable/fit to screen), but cannot handle the large sets of data I'm collecting. Does anyone know of any alternatives?

I have MATLAB, but don't have enough licenses to simultaneously run this and do development at the same time. I know that LabVIEW would be a great option, but it is currently cost-prohibitive.

Thanks in advance!

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

User: [clayton-pipkin](#)

[Answer](#) by [aleksandr-blekh](#)

For this answer, I have assumed that you prefer **open source solutions** to *big data visualization*. This assumption is based on budgetary details from your question. However, there is one *exclusion* to this - below I will add a reference to one commercial product, which I believe might be beneficial in your case (provided that you could afford that). I also assume that *browser-based solutions are acceptable* (I would even prefer them, unless you have specific contradictory requirements).

Naturally, the first candidate as a solution to your problem I would consider **D3.js JavaScript library**: <http://d3js.org>. However, despite *flexibility* and other *benefits*, I think that this solution is *too low-level*.

Therefore, I would recommend you to take a look at the following **open source projects for big data visualization**, which are *powerful* and *flexible* enough, but operate at a *higher level of abstraction* (some of them are based on D3.js foundation and sometimes are referred to as D3.js [visualization stack](#)).

- **Bokeh** - Python-based interactive visualization library, which supports big data and streaming data: <http://bokeh.pydata.org>
- **Flot** - JavaScript-based interactive visualization library, focused on jQuery: <http://www.flotcharts.org>
- **NodeBox** - unique rapid data visualization system (not browser-based, but multi-language and multi-platform), based on generative design and visual functional programming: <https://www.nodebox.net>
- **Processing** - complete software development system with its own programming language, libraries, plug-ins, etc., oriented to visual content: <https://www.processing.org> (allows executing Processing programs in a browser via <http://processingjs.org>)
- **Crossfilter** - JavaScript-based interactive visualization library for big data by Square (very fast visualization of large multivariate data sets):

<http://square.github.io/crossfilter>

- **bigvis** - an R package for big data exploratory analysis (not a visualization library per se, but could be useful to process large data sets /aggregating, smoothing/ prior to visualization, using various R graphics options): <https://github.com/hadley/bigvis>
- **prefuse** - Java-based interactive visualization library: <http://prefuse.org>
- **Lumify** - big data integration, analysis and visualization platform (interesting feature: supports Semantic Web): <http://lumify.io>

Separately, I'd like to mention two open source *big data analysis and visualization projects*, focused on **graph/network data** (with some support for *streaming data* of that type): [Cytoscape](#) and [Gephi](#). If you are interested in some other, *more specific (maps support, etc.)* or *commercial* (basic free tiers), projects and products, please see this **awesome compilation**, which I thoroughly *curated* to come up with the main list above and *analyzed*: <http://blog.profitbricks.com/39-data-visualization-tools-for-big-data>.

Finally, as I promised in the beginning, **Zoomdata** - a commercial product, which I thought you might want to take a look at: <http://www.zoomdata.com>. The reason I made an exclusion for it from my open source software compilation is due to its **built-in support for big data platforms**. In particular, Zoomdata provides *data connectors* for Cloudera Impala, Amazon Redshift, MongoDB, Spark and Hadoop, plus search engines, major database engines and streaming data.

Disclaimer: I have no affiliation with Zoomdata whatsoever - I was just impressed by their *range of connectivity options* (which might **cost** you dearly, but that's another *aspect* of this topic's analysis).

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#)), [visualization](#) ([Prev Q](#)) ([Next Q](#))

[Q: Where did this NY Times op-ed get his Google Search data?](#)

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

I hope this is a question appropriate for SO.

The article in question: <http://www.nytimes.com/2015/01/25/opinion/sunday/seth-stephens-davidowitz-searching-for-sex.html>

As far as I can tell, the only publicly available data from Google Search is through their Trends API. The help page states that

The numbers on the graph reflect how many searches have been done for a particular term, relative to the total number of searches done on Google over time. They don't represent absolute search volume numbers, because the data is normalized and presented on a scale from 0-100.

However in the article, the author reports (absolute) "average monthly searches". The source is stated as:

All monthly search numbers are approximate and derived from anonymous and

aggregate web activity.

Source: analysis of Google data by (author)

So, how did he get this “anonymous and aggregate web activity”?

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

User: [alexw](#) 

[Answer](#)  by [seth-stephens-davidowitz](#) 

Google AdWords. That has absolute search volumes.

Tags: [dataset](#) ([Prev Q](#)) ([Next Q](#)), [search](#) ([Prev Q](#)) ([Next Q](#))

[Q: Data available from industry](#)

Tags: [dataset](#) ([Prev Q](#))

I'm going to start my degree thesis and I want to do a fault detector system using machine learning techniques. I need datasets for my thesis but I don't know where I can get that data. I'm looking for historical operation/maintenance/fault datasets of any kind of machine in the oil&gas industry (Drills, steam injectors etc) or electrical companies (transformators, generators etc).

Thank you for your help.

Tags: [dataset](#) ([Prev Q](#))

User: [juan-david](#) 

[Answer](#)  by [ihars](#) 

A huge list of open data sets is listed here:

- [Publicly available datasets](#)

Including Amazon, KDnuggets, Stanford, Twitter, Freebase, Google Public and more.

Tags: [dataset](#) ([Prev Q](#))

Efficiency

Questions

[Q: What is the most efficient data indexing technique](#)

Tags: [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

As we all know, there are some data indexing techniques, using by well-known indexing apps, like Lucene (for java) or Lucene.NET (for .NET), MurMurHash, B+Tree etc. For a No-Sql / Object Oriented Database (which I try to write/play a little around with C#), which technique you suggest?

I read about MurMurhash-2 and specially v3 comments say Murmur is very fast. Also Lucene.Net has good comments on it. But what about their memory footprints in general? Is there any efficient solution which uses less footprint (and of course if faster is preferable) than Lucene or Murmur? Or should I write a special index structure to get the best results?

If I try to write my own, then is there any accepted scale for a good indexing, something like 1% of data-node, or 5% of data-node? Any useful hint will be appreciated.

Tags: [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

User: [sahirbazzz](#) 

[Answer](#)  by [rapaio](#) 

I think you messed up some things in your question. Lucene (I know nothing about Lucene,.NET, but I suppose is the same) is a library used to analyze, split in tokens, and store documents in order to be able to query and retrieve them later. Lucene has a pretty old but effective model, it uses inverted trees to find and retrieve documents. Without further details, all documents are split in tokens (terms), and for each term is maintained a data structure, which stores all the documents which contains the given term. As a data structure could be used a BTree, a hash table and in the latest major revisions you can even plug in your own data structures.

A BTree (see [Wikipedia page](#)  for further details), is a kind of a tree data structure, which is appropriate for working with big chunks of data and is often used for storing tree-like ordered structures on disk. For in-memory other trees performs better.

Murmur hash (see [Wikipedia page](#)  for further details), is a family of hash functions used in hash table. The implementation of the hash table is not important, it could be a standard chained implementation or more advanced open hash addressing scheme. The idea is that the hash tables allows one to get fast a key, from an unordered set of keys, and can answer to tasks like: is this key part of this set of keys? which is the value associated with this key?

Now back to your main problem. You have one library (Lucene) and two data structures, both data structures are used in Lucene. Now you see that it is not possible to answer your question in these terms since they are not comparable.

However, regarding your footprint and performance part of the question. First of all you have to know which kind of operations you need to implement.

Do you need only get value for key, or do you need to find all elements in a range? In other words do you need order or not? If you do, than a tree can help. If you do not, than a hash table, which is faster could be used instead.

Do you have a lot of data which does not fit the memory? If yes than a disk-based solution would help (like BTree). If your data fit the memory, than use the fastest in-memory solution and use disk only as a storage (with a different structure, much simpler).

Tags: [efficiency](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to speedup message passing between computing nodes](#)

Tags: [efficiency](#) ([Prev Q](#)) ([Next Q](#))

I'm developing a distributed application, and as it's been designed, there'll be a great load of communication during the processing. Since the communication is already as much *spread* along the entire process as possible, I'm wondering if there any standard solutions to improve the performance of the message passing layer of my application.

What changes/improvements could I apply to my code to reduce the time spent sending messages? For what it's worth, I'm communicating up to 10GB between 9 computing nodes, and the framework I'm using is implemented with OpenMPI.

Tags: [efficiency](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#) 

[Answer](#)  by [indico](#) 

Firstly, I would generally agree with everything that AirThomas suggested. Caching things is generally good if you can, but I find it slightly brittle since that's very dependent on exactly what your application is. Data compression is another very solid suggestion, but my impression on both of these is that the speedups you're looking at are going to be relatively marginal. Maybe as high as 2-5x, but I would be very surprised if they were any faster than that.

Under the assumption that pure I/O (writing to/reading from memory) is *not* your limiting factor (if it is, you're probably not going to get a lot faster), I would make a strong plug for [zeromq](#) .

We took a normal TCP socket, injected it with a mix of radioactive isotopes stolen from a secret Soviet atomic research project, bombarded it with 1950-era cosmic rays, and put it into the hands of a drug-addled comic book author with a badly-disguised fetish for bulging muscles clad in spandex. Yes, ØMQ sockets are the world-saving superheroes of the networking world.

While that may be a little dramatic, zeromq sockets in my opinion are one of the most amazing pieces of software that the world of computer networks has put together in several years. I'm not sure what you're using for your message-passing layer right now, but if you're using something traditional like rabbitmq, you're liable to see speedups of multiple orders of magnitude (personally noticed about 500x, but depends a lot of architecture)

Check out some basic benchmarks [here](#). 

[Answer](#)  by [air](#) 

If you expect (or find) that nodes are requesting the same data more than once, perhaps you could benefit from a caching strategy? Especially where some data is used much more often than others, so you can target only the most frequently-used information.

If the data is mutable, you also need a way to confirm that it hasn't changed since the last request that's less expensive than repeating the request.

This is further complicated if each node has its own separate cache. Depending on the nature of your system and task(s), you could consider adding a node dedicated to serving information between the processing nodes, and building a single cache on that node.

For an example of when that *might* be a good idea, let's suppose I retrieve some data from a remote data store over a low-bandwidth connection, and I have some task(s) requiring that data, which are distributed exclusively among local nodes. I definitely wouldn't want each node requesting information separately over that low-bandwidth connection, which another node might have previously requested. Since my local I/O is much less expensive than my I/O over the low-bandwidth connection, I might add a node between the processing nodes and the remote source that acts as an intermediate server. This node would take requests from the processing nodes, communicate with the remote data store, and cache frequently-requested data to minimize the use of that low-bandwidth connection.

The core concepts here that *may* be applicable to your specific case are:

- Eliminate or reduce redundant I/O;
- Take advantage of trade-offs between memory use and computation time;
- Not all I/O is created equal.

Tags: [efficiency](#) ([Prev Q](#)) ([Next Q](#))

[Q: Scikit Learn Logistic Regression Memory Leak](#)

Tags: [efficiency](#) ([Prev Q](#))

I'm curious if anyone else has run into this. I have a data set with about 350k samples, each with 4k sparse features. The sparse fill rate is about 0.5%. The data is stored in a `scipy.sparse.csr.csr_matrix` object, with `dtype='numpy.float64'`.

I'm using this as an input to sklearn's Logistic Regression classifier. The [documentation](#)  indicates that sparse CSR matrices are acceptable inputs to this classifier. However, when I train the classifier, I get extremely bad memory performance; the memory usage of my process explodes from ~150 MB to fill all the available memory and then everything grinds to a halt as memory swapping to disk takes over.

Does anyone know why this classifier might expand the sparse matrix to a dense matrix? I'm using the default parameters for the classifier at the moment, within an updated anaconda distribution. Thanks!

```
scipy.__version__ = '0.14.0'  
sklearn.__version__ = '0.15.2'
```

Tags: [efficiency](#) ([Prev Q](#))

User: [davea](#) 

[Answer](#) by [davea](#)

Ok, this ended up being an RTFM situation, although in this case it was RTF error message.

While running this, I kept getting the following error:

```
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the §
```

I assumed that, since this had to do with the target vector, and since it was a warning only, that it would just silently change my target vector to 1-D.

However, when I explicitly converted my target vector to 1-D, my memory problems went away. Apparently having the target vector in an incorrect form caused it to convert my input vectors into dense vectors from sparse vectors.

Lesson learned: **follow the recommendations** when sklearn ‘suggests’ you do something.

Tags: [efficiency](#) ([Prev Q](#))

Algorithms

Questions

[Q: Is logistic regression actually a regression algorithm?](#)

Tags: [algorithms](#) ([Prev Q](#)) ([Next Q](#))

The usual definition of regression (as far as I am aware) is *predicting a continuous output variable from a given set of input variables.*

Logistic regression is a binary classification algorithm, so it produces a categorical output.

Is it really a regression algorithm? If so, why?

Tags: [algorithms](#) ([Prev Q](#)) ([Next Q](#))

User: [joeys](#) 

[Answer](#)  by [christopher-louden](#) 

Short Answer

Yes, logistic regression is a regression algorithm and it does predict a continuous outcome: the probability of an event. That we use it as a binary classifier is due to the interpretation of the outcome.

Detail

Logistic regression is a type of generalize linear regression model.

In an ordinary linear regression model, a continuous outcome, y , is modeled as the sum of the product of predictors and their effect:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n + e$$

where e is the error.

Generalized linear models do not model y directly. Instead, they use transformations to expand the domain of y to all real numbers. This transformation is called the link function. For logistic regression the link function is the logit function (usually, see note below).

The logit function is defined as

$$\ln(y/(1 + y))$$

Thus the form of logistic regression is:

$$\ln(y/(1 + y)) = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n + e$$

where y is the probability of an event.

The fact that we use it as a binary classifier is due to the interpretation of the outcome.

Note: probit is another link function used for logistic regression but logit is the most widely used.

Tags: [algorithms](#) ([Prev Q](#)) ([Next Q](#))

[Q: Are there any unsupervised learning algorithms for time sequenced data?](#)

Tags: [algorithms](#) ([Prev Q](#)) ([Next Q](#))

Each observation in my data was collected with a difference of 0.1 seconds. I don't call it a time series because it doesn't have a date and time stamp. In the examples of clustering algorithms (I found online) and PCA the sample data have 1 observation per case and are not timed. But my data have hundreds of observations collected every 0.1 seconds per vehicle and there are many vehicles.

Note: I have asked this question on quora as well.

Tags: [algorithms](#) ([Prev Q](#)) ([Next Q](#))

User: [umair-durrani](#)

[Answer](#) by [kasra-manshaei](#)

What you have is a sequence of events according to time so do not hesitate to call it Time Series!

Clustering in time series has 2 different meanings:

1. **Segmentation of time series** i.e. you want to segment an individual time series into different time intervals according to internal similarities.
2. **Time series clustering** i.e. you have several time series and you want to find different clusters according to similarities between them.

I assume you mean the second one and here is my suggestion:

You have many vehicles and many observations per vehicle i.e. you have many vehicles. So you have several matrices (each vehicle is a matrix) and each matrix contains N rows (Nr of observations) and T columns (time points). One suggestion could be applying PCA to each matrix to reduce the dimensionality and observing data in PC space and see if there is meaningful relations between *different observations within a matrix (vehicle)*. Then you can put each observation for all vehicles on each other and make a matrix and apply PCA to that to see relations of a single observation between different vehicles.

If you do not have negative values **Matrix Factorization** is strongly recommended for dimension reduction of matrix form data.

Another suggestion could be putting all matrices on top of each other and build a $N \times M \times T$ tensor where N is the number of vehicles, M is the number of observations and T is the time sequence and apply **Tensor Decomposition** to see relations globally.

A very nice approach to Time Series Clustering is shown in [this](#) paper where the implementation is quite straight forward.

I hope it helped!

Good Luck :)

EDIT

As you mentioned you mean Time Series Segmentation I add this to the answer.

Time series segmentation is the only clustering problem that has a ground truth for evaluation. Indeed you consider the generating distribution behind the time series and analyze it I strongly recommend [this](#), [this](#), [this](#), [this](#), [this](#) and [this](#) where your problem is comprehensively studied. Specially the last one and the PhD thesis.

Good Luck!

Tags: [algorithms](#) ([Prev Q](#)) ([Next Q](#))

Q: Which algorithms or methods can be used to detect an outlier from this data set?

Tags: [algorithms](#) ([Prev Q](#))

Suppose I have a data set : Amount of money (100, 50, 150, 200, 35, 60 ,50, 20, 500). I have [Googled](#) the web looking for techniques that can be used to find a possible outlier in this data set but I ended up confused.

My question is: Which algorithms, techniques or methods can be used to detect possible outlier in this data set?

PS: Consider that the data does not follow a normal distribution. Thanks.

Tags: [algorithms](#) ([Prev Q](#))

User: [giovanrich](#)

[Answer](#) by [michael-hooreman](#)

A simple approach would be using the same thing as box plots does: away than 1.5 (median-q1) or 1.5 (q3-median) = outlier.

I find it useful in lots of cases even it not perfect and maybe too simple.

It has the advantage to not suppose normality.

[Answer](#) by [tristan-reid](#)

One way of thinking of outlier detection is that you're creating a predictive model, then you're checking to see if a point falls within the range of predictions. From an information-theoretic point of view, you can see how much each observation increases the entropy of your model.

If you are treating this data as just a collection of numbers, and you don't have some proposed model for how they're generated, you might as well just look at the average. If you're certain the numbers aren't normally distributed, you can't make statements as to

how far ‘off’ a given number is from the average, but you can just look at it in absolute terms.

Applying this, you can take the average of all the numbers, then exclude each number and take the average of the others. Whichever average is most different from the global average is the biggest outlier. Here’s some python:

[Skip code block](#)

```
def avg(a):
    return sum(a)/len(a)

l = [100, 50, 150, 200, 35, 60 ,50, 20, 500]
m = avg(l)
for idx in range(len(l)):
    print("outlier score of {0}: {1}".format(l[idx], abs(m - avg([elem for i, elem in enumerate(l) if
>>
outlier score of 100: 4
outlier score of 50: 10
outlier score of 150: 3
outlier score of 200: 9
outlier score of 35: 12
outlier score of 60: 9
outlier score of 50: 10
outlier score of 20: 14
outlier score of 500: 46
```

Tags: [algorithms](#) ([Prev Q](#))

Hadoop

[Skip to questions](#),

Wiki by user [dawny33](#)

The [Apache™ Hadoop™](#) project develops open-source software for reliable, scalable, distributed computing.

“[Hadoop](#)” typically refers to the software in the project that implements the [map-reduce](#) data analysis framework, plus the distributed file system (HDFS) that underlies it.

Since version 0.23, Hadoop disposes of a standalone resource manager : [yarn](#).

This resource manager makes it easier to use other modules alongside with the MapReduce engine, such as :

- [Ambari](#), A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually along with features to diagnose their performance characteristics in a user-friendly manner.
- [Avro](#), a data serialization system based on JSON schemas.
- [Cassandra](#), a replicated, fault-tolerant, decentralized and scalable database system.
- [Chukwa](#): A data collection system for managing large distributed systems.
- [HBase](#), A scalable, distributed database that supports structured data storage for large tables.
- [Hive](#), A data warehouse infrastructure that provides data summarization and ad hoc querying.
- [Mahout](#), a library of machine learning algorithms compatible with M/R paradigm.
- [Pig](#), a platform/programming language for authoring parallelizable jobs
- [Storm](#), a system for real-time and stream processing
- [ZooKeeper](#), a system for coordinating distributed nodes, similar to Google’s Chubby
- [Oozie](#), a workflow scheduler system to manage Apache Hadoop jobs.
- [Spark](#), a fast and general engine for large-scale data processing.
- [Flink](#), a fast and reliable large-scale data processing engine.

Recommended reference sources:

- [Hive Language Reference](#)

Commercial support is available from [a variety of companies](#).

Questions

[Q: What is the difference between Hadoop and noSQL](#)

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

I heard about many tools / frameworks for helping people to process their data (big data environment).

One is called Hadoop and the other is the noSQL concept. What is the difference in point of processing?

Are they complementary?

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

User: [ruffp](#) 

[Answer](#)  by [mcp_infiltrator](#) 

NoSQL is a way to store data that does not require there to be some sort of relation. The simplicity of its design and horizontal scale-ability, one way they store data is the key : value pair design. This lends itself to processing that is similar to Hadoop. The use of a NoSQL db really depends on the type of problem that one is after.

Here is a good wikipedia link [NoSQL](#) 

Hadoop is a system that is meant to store and process huge chunks of data. It is a distributed file system dfs. The reason it does this is that central to its design it makes the assumption that hardware failures are common, thus making multiple copies of the same piece of information and spreading it across multiple machines and racks, so if one goes down, no problem, we have two more copies. Here is a great link for Hadoop from wikipedia as well, you will see that it is, in my opinion more than just storage, but also processing: [Hadoop](#) 

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [tools](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

[Q: Can map-reduce algorithms written for MongoDB be ported to Hadoop later?](#)

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [scalability](#) ([Prev Q](#)) ([Next Q](#))

In our company, we have a MongoDB database containing a lot of unstructured data, on which we need to run map-reduce algorithms to generate reports and other analyses. We have two approaches to select from for implementing the required analyses:

1. One approach is to extract the data from MongoDB to a Hadoop cluster and do the analysis completely in Hadoop platform. However, this requires considerable investment on preparing the platform (software and hardware) and educating the team to work with Hadoop and write map-reduce tasks for it.

2. Another approach is to just put our effort on designing the map-reduce algorithms, and run the algorithms on MongoDB map-reduce functionalities. This way, we can create an initial prototype of final system that can generate the reports. I know that the MongoDB's map-reduce functionalities are much slower compared to Hadoop, but currently the data is not that big that makes this a bottleneck yet, at least not for the next six months.

The question is, using the second approach and writing the algorithms for MongoDB, can them be later ported to Hadoop with little needed modification and algorithm redesign? MongoDB just supports JavaScript but programming language differences are easy to handle. However, is there any fundamental differences in the map-reduce model of MongoDB and Hadoop that may force us to redesign algorithms substantially for porting to Hadoop?

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [scalability](#) ([Prev Q](#)) ([Next Q](#))

User: [amir-ali-akbari](#) 

[Answer](#)  by [steve-kallestad](#) 

There will definitely be a translation task at the end if you prototype using just mongo.

When you run a MapReduce task on mongodb, it has the data source and structure built in. When you eventually convert to hadoop, your data structures might not look the same. You could leverage the mongodb-hadoop connector to access mongo data directly from within hadoop, but that won't be quite as straightforward as you might think. The time to figure out how exactly to do the conversion most optimally will be easier to justify once you have a prototype in place, IMO.

While you will need to translate mapreduce functions, the basic pseudocode should apply well to both systems. You won't find anything that can be done in MongoDB that can't be done using Java or that is significantly more complex to do with Java.

[Answer](#)  by [damian-melniczuk](#) 

You can use map reduce algorithms in Hadoop without programming them in Java. It is called streaming and works like Linux piping. If you believe that you can port your functions to read and write to terminal, it should work nicely. [Here](#)  is example blog post which shows how to use map reduce functions written in Python in Hadoop.

[Answer](#)  by [phyrox](#) 

You also can create a MongoDB-Hadoop [connection](#) .

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [scalability](#) ([Prev Q](#)) ([Next Q](#))

[Q: Does Amazon RedShift replace Hadoop for ~1XTB data?](#)

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [aws](#) ([Next Q](#))

There is plenty of hype surrounding Hadoop and its eco-system. However, in practice, where many data sets are in the terabyte range, is it not more reasonable to use [Amazon RedShift](#) for querying large data sets, rather than spending time and effort building a Hadoop cluster?

Also, how does Amazon Redshift compare with Hadoop with respect to setup complexity, cost, and performance?

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [aws](#) ([Next Q](#))

User: [trienism](#)

[Answer](#) by [enno-shioji](#)

tl;dr: They markedly differ in many aspects and I can't think Redshift will replace Hadoop.

-Function

You can't run anything other than SQL on Redshift. Perhaps most importantly, you can't run any type of custom functions on Redshift. In Hadoop you can, using many languages (Java, Python, Ruby.. you name it). For example, NLP in Hadoop is easy, while it's more or less impossible in Redshift. I.e. there are lots of things you can do in Hadoop but not on Redshift. This is probably the most important difference.

-Performance Profile

Query execution on Redshift is in most cases significantly more efficient than on Hadoop. However, this efficiency comes from the indexing that is done when the data is loaded into Redshift (I'm using the term indexing very loose here). Therefore, it's great if you load your data once and execute multiple queries, but if you want to execute only one query for example, you might actually lose out in performance overall.

-Cost Profile

Which solution wins out in cost depends on the situation (like performance), but you probably need quite a lot of queries in order to make it cheaper than Hadoop (more specifically Amazon's Elastic Map Reduce). For example, if you are doing OLAP, it's very likely that Redshift comes out cheaper. If you do daily batch ETLs, Hadoop is more likely to come out cheaper.

Having said that, we've replaced part of our ETL that was done in Hive to Redshift, and it was a pretty great experience; mostly for the ease of development. Redshift's Query Engine is based on PostgreSQL and is very mature, compared to Hive's. Its ACID characteristics make it easier to reason about it, and the quicker response time allows more testing to be done. It's a great tool to have, but it won't replace Hadoop.

EDIT: As for setup complexity, I'd even say it's easier with Hadoop if you use AWS's EMR. Their tools are so mature that it's ridiculously easy to have your Hadoop job running. Tools and mechanisms surrounding Redshift's operation aren't that mature yet.

For example, Redshift can't handle trickle loading and thus you have to come up with something that turns that into a batched load, which can add some complexity to your ETL.

[Answer](#) by [anurag-gupta](#)

Current size limit for Amazon Redshift is 128 nodes or 2 PBs of compressed data. Might be circa 6PB uncompressed though mileage varies for compression. You can always let us know if you need more. anurag@aws (I run Amazon Redshift and Amazon EMR)

[Answer](#) by [steve-kallestad](#)

Personally, I don't think it's all that difficult to set up a hadoop cluster, but I know that it is sometimes painful when you are getting started.

HDFS size limitations well exceed a TB (or did you mean exabyte?). If I'm not mistaken it scales to yottabytes or some other measurement that I don't even know the word for. Whatever it is, it's really big.

Tools like Redshift have their place, but I always worry about vendor specific solutions. My main concern is always "what do I do when I am dissatisfied with their service?" - I can go to google and shift my analysis work into their paradigm or I can go to hadoop and shift that same work into that system. Either way, I'm going to have to learn something new and do a lot of work translating things.

That being said, it's nice to be able to upload a dataset and get to work quickly - especially if what I'm doing has a short lifecycle. Amazon has done a good job of answering the data security problem.

If you want to avoid hadoop, there will always be an alternative. But it's not all that difficult to work with once you get going with it.

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [aws](#) ([Next Q](#))

[Q: What are the use cases for Apache Spark vs Hadoop](#)

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#))

With Hadoop 2.0 and YARN Hadoop is supposedly no longer tied only map-reduce solutions. With that advancement, what are the use cases for Apache Spark vs Hadoop considering both sit atop of HDFS? I've read through the introduction documentation for Spark, but I'm curious if anyone has encountered a problem that was more efficient and easier to solve with Spark compared to Hadoop.

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#))

User: [idclark](#)

[Answer](#) by [sean-owen](#)

Hadoop means HDFS, YARN, MapReduce, and a lot of other things. Do you mean Spark

vs *MapReduce*? Because Spark runs on/with Hadoop, which is rather the point.

The primary reason to use Spark is for speed, and this comes from the fact that its execution can keep data in memory between stages rather than always persist back to HDFS after a Map or Reduce. This advantage is very pronounced for iterative computations, which have tens of stages each of which is touching the same data. This is where things might be “100x” faster. For simple, one-pass ETL-like jobs for which MapReduce was designed, it’s not in general faster.

Another reason to use Spark is its nicer high-level language compared to MapReduce. It provides a functional programming-like view that mimics Scala, which is far nicer than writing MapReduce code. (Although you have to either use Scala, or adopt the slightly-less-developed Java or Python APIs for Spark). [Crunch](#) and [Cascading](#) already provide a similar abstraction on top of MapReduce, but this is still an area where Spark is nice.

Finally Spark has as-yet-young but promising subprojects for ML, graph analysis, and streaming, which expose a similar, coherent API. With MapReduce, you would have to turn to several different other projects for this (Mahout, Giraph, Storm). It’s nice to have it in one package, albeit not yet ‘baked’.

Why would you not use Spark? [paraphrasing](#) myself:

- Spark is primarily Scala, with ported Java APIs; MapReduce might be friendlier and more native for Java-based developers
- There is more MapReduce expertise out there now than Spark
- For the data-parallel, one-pass, ETL-like jobs MapReduce was designed for, MapReduce is lighter-weight compared to the Spark equivalent
- Spark is fairly mature, and so is YARN now, but Spark-on-YARN is still pretty new. The two may not be optimally integrated yet. For example until recently I don’t think Spark could ask YARN for allocations based on number of cores? That is: MapReduce might be easier to understand, manage and tune

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#))

[Q: Processing data stored in Redshift](#)

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [aws](#) ([Prev Q](#)) ([Next Q](#))

We’re currently using Redshift as our data warehouse, which we’re very happy with. However, we now have a requirement to do machine learning against the data in our warehouse. Given the volume of data involved, ideally I’d want to run the computation in the same location as the data rather than shifting the data around, but this doesn’t seem possible with Redshift. I’ve looked at MADlib, but this is not an option as Redshift does not support UDFs (which MADlib requires). I’m currently looking at shifting the data over to EMR and processing it with the Apache Spark machine learning library (or maybe H2O, or Mahout, or whatever). So my questions are:

1. is there a better way?
2. if not, how should I make the data accessible to Spark? The options I've identified so far include: use Sqoop to load it into HDFS, use DBInputFormat, do a Redshift export to S3 and have Spark grab it from there. What are the pros/cons for these different approaches (and any others) when using Spark?

Note that this is off-line batch learning, but we'd like to be able to do this as quickly as possible so that we can iterate experiments quickly.

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [aws](#) ([Prev Q](#)) ([Next Q](#))

User: [deanj](#) 

[Answer](#)  by [rahul-pathak](#) 

The new Amazon Machine Learning Service may work for you. It works directly with Redshift and might be a good way to start. <http://aws.amazon.com/machine-learning/> 

If you're looking to process using EMR, then you can use Redshift's UNLOAD command to land data on S3. Spark on EMR can then access it directly without you having to pull it into HDFS.

Spark on EMR: <https://aws.amazon.com/articles/Elastic-MapReduce/4926593393724923> 

Tags: [hadoop](#) ([Prev Q](#)) ([Next Q](#)), [aws](#) ([Prev Q](#)) ([Next Q](#))

[Q: Good books for Hadoop, Spark, and Spark Streaming](#)

Tags: [hadoop](#) ([Prev Q](#))

Can anyone suggest any good books to learn hadoop and map reduce basics?

Also something for Spark, and Spark Streaming?

Thanks

Tags: [hadoop](#) ([Prev Q](#))

User: [tsar2512](#) 

[Answer](#)  by [hack-r](#) 

There's such an overwhelming amount of literature that with programming, databases, and Big Data I like to stick to the O'reilly series as my go-to source. O'reilly books are extremely popular in the industry and I've been very satisfied.

A current version of

1. [Hadoop: The Definitive Guide](#) ,
2. [MapReduce Design Patterns](#) , and
3. [Learning Spark](#) 

might suit your needs by providing high quality, immediately useful information and avoiding information overload — all are published by O'reilly.

Spark Streaming is covered in Chapter 13 of “Learning Spark”.

Tags: [hadoop](#) ([Prev Q](#))

SVM

Questions

[Q: What's the relationship between an SVM and hinge loss?](#)

Tags: [svm](#) ([Prev Q](#))

My colleague and I are trying to wrap our heads around the difference between logistic regression and an SVM. Clearly they are optimizing different objective functions. Is an SVM as simple as saying it's a discriminative classifier that simply optimizes the hinge loss? Or is it more complex than that? How do the support vectors come into play? What about the slack variables? Why can't you have deep SVM's the way you can't you have a deep neural network with sigmoid activation functions?

Tags: [svm](#) ([Prev Q](#))

User: [simon](#)

[Answer](#) by [sean-owen](#)

They are both discriminative models, yes. The logistic regression loss function is conceptually a function of all points. Correctly classified points add very little to the loss function, adding more if they are close to the boundary. The points near the boundary are therefore more important to the loss and therefore deciding how good the boundary is.

SVM uses a hinge loss, which conceptually puts the emphasis on the boundary points. Anything farther than the closest points contributes nothing to the loss because of the “hinge” (the max) in the function. Those closest points are the support vectors, simply. Therefore it actually reduces to picking a boundary that creates the largest margin — distance to closest point. The theory is that the boundary case is all that really matters to generalization.

The downside is that hinge loss is not differentiable, but that just means it takes more math to discover how to optimize it via Lagrange multipliers. It doesn't really handle the case where data isn't linearly separable. Slack variables are a trick that lets this possibility be incorporated cleanly into the optimization problem.

You can use hinge loss with “deep learning”, e.g. <http://arxiv.org/pdf/1306.0239.pdf>

Tags: [svm](#) ([Prev Q](#))

Tools

Questions

[Q: Google prediction API: What training/prediction methods Google Prediction API employs?](#)

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

The details of the Google Prediction API are on this [page](#) , but I am not able to find any details about the prediction algorithms running behind the API.

So far I have gathered that they let you provide your preprocessing steps in PMML format.

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

User: [tahir-akhtar](#) 

[Answer](#)  by [rapaio](#) 

If you take a look over the specifications of PMML which you can find [here](#)  you can see on the left menu what options you have (like ModelTree, NaiveBayes, Neural Nets and so on).

[Answer](#)  by [brent-blazek](#) 

A variety of methods are available to the user. The support documentation gives walkthroughs and tips for when one or another model is most appropriate.

[This page](#)  shows the following learning methods:

- “AssociationModel”
- “ClusteringModel”
- “GeneralRegressionModel”
- “MiningModel”
- “NaiveBayesModel”
- “NeuralNetwork”
- “RegressionModel”
- “RuleSetModel”
- “SequenceModel”
- “SupportVectorMachineModel”
- “TextModel”
- “TimeSeriesModel”
- “TreeModel”

EDIT: I don't see any specific information about the algorithms, though. For example, does the tree model use information gain or gini index for splits?

[Answer](#) by [steve-kallestad](#)

Google does not publish the models they use, but they specifically do not support models from the PMML specification.

If you look closely at the documentation on [this page](#), you will notice that the model selection within the schema is greyed out indicating that it is an unsupported feature of the schema.

The [documentation does spell out](#) that by default it will use a regression model for training data that has numeric answers, and an unspecified categorization model for training data that results in text based answers.

The Google Prediction API also supports hosted models (although only a few demo models are currently available), and models specified with a PMML transform. The documentation does contain an [example of a model defined by a PMML transform](#). (There is also a note on that page stating that PMML ...Model elements are not supported).

The PMML standard that google partially supports is [version 4.0.1](#).

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

[Q: Do you need a virtual machine as an instrument for your data science practice?](#)

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

I am brand new to the field of data science, want to break into it, and there are so many tools out there. These VMs have a lot of software on them, but I haven't been able to find any side-by-side comparison.

Here's a start from my research, but if someone could tell me that one is objectively more rich-featured, with a larger community of support, and useful to get started then that would help greatly:

datasciencetoolKIT.org -> vm is on vagrant cloud (4 GB) and seems to be more "hip" with R, iPython notebook, and other useful command-line tools (html->txt, json->xml, etc). There is a book being released in August with detail.

datasciencetoolBOX.org -> vm is a vagrant box (24 GB) downloadable from their website. There seems to be more features here, and more literature.

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

User: [user3659451](#)

[Answer](#) by [asheeshr](#)

Do you need a VM?

You need to keep in mind that a virtual machine is a software emulation of your own or another machine hardware configuration that can run an operating systems. In most basic terms, it acts as a layer interfacing between the virtual OS, and your own OS which then communicates with the lower level hardware to provide support to the virtual OS. What this means for you is:

Cons

Hardware Support

A drawback of virtual machine technology is that it supports only the hardware that both the virtual machine hypervisor and the guest operating system support. Even if the guest operating system supports the physical hardware, it sees only the virtual hardware presented by the virtual machine. The second aspect of virtual machine hardware support is the hardware presented to the guest operating system. No matter the hardware in the host, the hardware presented to the guest environment is usually the same (with the exception of the CPU, which shows through). For example, VMware GSX Server presents an AMD PCnet32 Fast Ethernet card or an optimized VMware-proprietary network card, depending on which you choose. The network card in the host machine does not matter. VMware GSX Server performs the translation between the guest environment's network card and the host environment's network card. This is great for standardization, but it also means that host hardware that VMware does not understand will not be present in the guest environment.

Performance Penalty

Virtual machine technology imposes a performance penalty from running an additional layer above the physical hardware but beneath the guest operating system. The performance penalty varies based on the virtualization software used and the guest software being run. This is significant.

Pros

Isolation

One of the key reasons to employ virtualization is to isolate applications from each other. Running everything on one machine would be great if it all worked, but many times it results in undesirable interactions or even outright conflicts. The cause often is software problems or business requirements, such as the need for isolated security. Virtual machines allow you to isolate each application (or group of applications) in its own sandbox environment. The virtual machines can run on the same physical machine (simplifying IT hardware management), yet appear as independent machines to the software you are running. For all intents and purposes—except performance, the virtual machines are independent machines. If one virtual machine goes down due to application or operating system error, the others continue running, providing services your business needs to function smoothly.

Standardization

Another key benefit virtual machines provide is standardization. The hardware that is presented to the guest operating system is uniform for the most part, usually with the CPU being the only component that is “pass-through” in the sense that the guest sees what is on the host. A standardized hardware platform reduces support costs and increases the share of IT resources that you can devote to accomplishing goals that give your business a competitive advantage. The host machines can be different (as indeed they often are when hardware is acquired at different times), but the virtual machines will appear to be the same across all of them.

Ease of Testing

Virtual machines let you test scenarios easily. Most virtual machine software today provides snapshot and rollback capabilities. This means you can stop a virtual machine, create a snapshot, perform more operations in the virtual machine, and then roll back again and again until you have finished your testing. This is very handy for software development, but it is also useful for system administration. Admins can snapshot a system and install some software or make some configuration changes that they suspect may destabilize the system. If the software installs or changes work, then the admin can commit the updates. If the updates damage or destroy the system, the admin can roll them back. Virtual machines also facilitate scenario testing by enabling virtual networks. In VMware Workstation, for example, you can set up multiple virtual machines on a virtual network with configurable parameters, such as packet loss from congestion and latency. You can thus test timing-sensitive or load-sensitive applications to see how they perform under the stress of a simulated heavy workload.

Mobility

Virtual machines are easy to move between physical machines. Most of the virtual machine software on the market today stores a whole disk in the guest environment as a single file in the host environment. Snapshot and rollback capabilities are implemented by storing the change in state in a separate file in the host information. Having a single file represent an entire guest environment disk promotes the mobility of virtual machines. Transferring the virtual machine to another physical machine is as easy as moving the virtual disk file and some configuration files to the other physical machine. Deploying another copy of a virtual machine is the same as transferring a virtual machine, except that instead of moving the files, you copy them.

Which VM should I use if I am starting out?

The Data Science Box or the Data Science Toolbox are your best bets if you just getting into data science. They have the basic software that you will need, with the primary difference being the virtual environment in which each of these can run. The DSB can run on AWS while the DST can run on Virtual Box (which is the most common tool used for

VMs).

Sources

- <http://www.devx.com/vmspecialreport/Article/30383> 
 - <http://jeroenjanssens.com/2013/12/07/lean-mean-data-science-machine.html> 
-

[Answer](#)  by [tomaskazemekas](#) 

In most cases a practicing data scientist creates his own working environment on personal computer installing preferred software packages. Normally it is sufficient and efficient use of computing resources, because to run a virtual machine (VM) on your main machine you have to allocate a significant portion of RAM for it. The software will run noticeably slower on both the main and the virtual machine unless a lot of RAM.

Due to this impact on speed it is not common to use VMs as main working environment but they are a good solution in several cases when there is a need of additional working environment.

The VMs be considered when:

1. There is a need to easily replicate a number of identical computing environments when teaching a course or doing a presentation on a conference.
2. There is a need to save and recreate an exact environment for an experiment or a calculation.
3. There is a need to run a different OS or to test a solution on a tool that runs on a different OS.
4. One wants to try out a bundle of software tools before installing them on the main machine. E.g. there is an opportunity to instal an instance of Hadoop (CDH) on a VM during an [Intro to Hadoop](#)  course on Udacity.
5. VMs are sometimes used for fast deployment in the cloud like AWS EC, Rackspace etc.

The VMs mentioned in the original question are made as easily installable data science software bundles. There are more than these two. This [blog post](#)  by Jeroen Janssens gives a comparison of at least four:

1. Data Science Toolbox
 2. Mining the Social Web
 3. Data Science Toolkit
 4. Data Science Box
-

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

Q: Book keeping of experiment runs and results 

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

I am a hands on researcher and I like testing out viable solutions, so I tend to run a lot of experiments. For example, if I am calculating a similarity score between documents, I might want to try out many measures. In fact, for each measure I might need to make several runs to test the effect of some parameters.

So far, I've been tracking the runs inputs and their results by writing out the results into files with as much info about the inputs. The problem is that retrieving a specific result becomes a challenge sometimes, even if I try to add the input info to the filename. I tried using a spreadsheet with links to results but this isn't making a huge difference.

What tools/process do you use for the book keeping of your experiments?

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

User: [timeleft](#) 

[Answer](#)  by [seanv507](#) 

you might want to look at <http://deeplearning.net/software/jobman/intro.html> 

it was designed for deep learning (I guess), but it is application agnostic. It is effectively an API version of SeanEasters approach

[Answer](#)  by [sean-easter](#) 

I recently ran into a similar problem: How to manage extracting a variety of features from a large dataset, without knowing up front what all of them would be. (Even calculating mean values repeatedly would be computationally expensive.) Further, how would I manage predictions based on different feature sets? Meaning, if I added a new feature, how would I know which models to train on new features? It could quickly snowball into a huge mess.

My current solution is to track it all in a local NoSQL database (MongoDB). For example, I might have a collection `features`, each entry of which has a name, a description of how the feature was calculated, the python file that ran the extraction, etc.

Likewise, a collection `models` includes models run on the data. Each entry might have a name, a list of features that were used to train the model, its eventual parameters, predicted values on a held-out test set, metrics for how the model performed, etc.

From my vantage point, this has a number of benefits:

- By saving predictions, I can use them later in ensemble predictions.
- Because I keep track of which features were used, I know which ones need retraining as I extract more features.
- By saving model descriptions, I ensure that I always know what I've tried. I never have to wonder, "Have I tried LASSO with regularization parameters set by grid-search CV?" I can always look it up, and see how successful it was.

From your question, it sounds like you could adapt this approach to your problem's

workflow. Install Mongo or another database of choice, and then save each experimental run, its inputs, its results, and anything else you might wish to track over the course of the project. This should be much easier to query than a spreadsheet, at the least.

Tags: [tools](#) ([Prev Q](#)) ([Next Q](#))

[Q: Do data scientists use Excel?](#)

Tags: [tools](#) ([Prev Q](#))

I would consider myself a journeyman data scientist. Like most (I think), I made my first charts and did my first aggregations in high school and college, using Excel. As I went through college, grad school and ~7 years of work experience, I quickly picked up what I consider to be more advanced tools, like SQL, R, Python, Hadoop, LaTeX, etc.

We are interviewing for a data scientist position and one candidate advertises himself as a “senior data scientist” (a very buzzy term these days) with 15+ years experience. When asked what his preferred toolset was, he responded that it was Excel.

I took this as evidence that he was not as experienced as his resume would claim, but wasn't sure. After all, just because it's not my preferred tool, doesn't mean it's not other people's. **Do experienced data scientists use Excel? Can you assume a lack of experience from someone who does primarily use Excel?**

Tags: [tools](#) ([Prev Q](#))

User: [jhowix](#) 

[Answer](#)  by [robert-smith](#) 

Most non-technical people often use Excel as a database replacement. I think that's wrong but tolerable. However, someone who is supposedly experienced in data analysis simply can not use Excel as his main tool (excluding the obvious task of looking at the data for the first time). That's because Excel was never intended for that kind of analysis and as a consequence of this, it is incredibly easy to make mistakes in Excel (that's not to say that it is not incredibly easy to make another type of mistakes when using other tools, but Excel aggravates the situation even more.)

To summarize what Excel doesn't have and is a must for any analysis:

1. Reproducibility. A data analysis needs to be reproducible.
2. Version control. Good for collaboration and also good for reproducibility. Instead of using xls, use csv (still very complex and has lots of edge cases, but csv parsers are fairly good nowadays.)
3. Testing. If you don't have tests, your code is broken. If your code is broken, your analysis is worse than useless.
4. Maintainability.
5. Accuracy. Numerical accuracy, accurate date parsing, among others are really lacking in Excel.

More resources:

[European Spreadsheet Risks Interest Group - Horror Stories](#) 

[You shouldn't use a spreadsheet for important work \(I mean it\)](#) 

[Microsoft's Excel Might Be The Most Dangerous Software On The Planet](#) 

[Destroy Your Data Using Excel With This One Weird Trick!](#) 

[Excel spreadsheets are hard to get right](#) 

[Answer](#)  by [aleksandr-blekh](#) 

Do experienced data scientists use Excel?

I've seen some experienced data scientists, who use Excel - either due to their preference, or due to their workplace's business and IT environment specifics (for example, many financial institutions use Excel as their major tool, at least, for modeling). However, I think that most experienced data scientists recognize the need to use tools, which are optimal for particular tasks, and adhere to this approach.

Can you assume a lack of experience from someone who does primarily use Excel?

No, you cannot. This is the corollary from my above-mentioned thoughts. Data science does not automatically imply big data - there is plenty of data science work that Excel can handle quite well. Having said that, if a data scientist (even experienced one) does not have knowledge (at least, basic) of modern data science tools, including big data-focused ones, it is somewhat disturbing. This is because experimentation is deeply ingrained into the nature of data science due to exploratory data analysis being a essential and, even, a crucial part of it. Therefore, a person, who does not have an urge to explore other tools within their domain, could rank lower among candidates in the overall fit for a data science position (of course, this is quite fuzzy, as some people are very quick in learning new material, plus, people might have not had an opportunity to satisfy their interest in other tools due to various personal or workplace reasons).

Therefore, in conclusion, I think that the best answer an experienced data scientist might have to a question in regard to their preferred tool is the following: **My preferred tool is the optimal one, that is the one that best fits the task at hand.**

[Answer](#)  by [gerenuk](#) 

Excel allows only very small data and doesn't have anything that is sufficiently useful and flexible for machine learning or even just plotting. All I would do in Excel, is stare at a subset of the data for a first glance over the values to make sure I don't miss anything visible by eye.

So, if his favourite tool is Excel, this might suggest he rarely deals with machine learning, statistics, larger data sizes or any advanced plotting. Someone like this I wouldn't call a Data Scientist. Of course titles don't matter and it depends a lot on your requirements.

In any case, don't make a judgement by statements of experience or CV. I've seen CVs

and known the people behind it.

Don't assume. **Test him!** You should be good enough to set up a test. It has been shown that interviews alone are close to useless to determine skills (they only show personality). Set up a very simple supervised learning test and let him use any tool he wants.

And if you want to screen people at an interview first, then ask him about very basic but important insights about statistics or machine learning. Something that every single of your current employees knows.

Tags: [tools](#) ([Prev Q](#))

Recommendation

Questions

[Q: Does click frequency account for relevance?](#)

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#))

While building a rank, say for a search engine, or a recommendation system, is it valid to rely on click frequency to determine the relevance of an entry?

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#) 

[Answer](#)  by [johnny000](#) 

For my part I can say that I use click frequency on i.e. eCommerce products. When you combine it with the days of the year it can even bring you great suggestions.

i.e.: We have historical data from 1 year over 2 products (Snowboots[], Sandalettes[])

```
Snowboots[1024, 1253, 652, 123, 50, 12, 8, 4, 50, 148, 345, 896]  
Sandalettes[23, 50, 73, 100, 534, 701, 1053, 1503, 1125, 453, 213, 25]
```

where [0] = January

As you can see, snowboots are much more searched in January than sandalettes, so you should suggest snowboots to someone searching shoes on your site or /we on january.

You can also see if something is “fresh” at this time, like when people often click a unknown product it could be an insight for a new comming trend or something.

That are just some examples where you could use click frequency as an insight. I think there are **no rules** for what you can use or not in recommendations, **as long as it makes sense.**

[Answer](#)  by [rolf](#) 

Is it valid to use click frequency, then **yes**. Is it valid to use **only** the click frequency, then probably **no**.

Search relevance is much more complicated than just one metric. [There are entire books on the subject](#)  Extending this answer beyond a simple yes/no would likely make the answer far too broad (and opinionated)

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#))

[Q: How should one deal with implicit data in recommendation](#)

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#))

A recommendation system keeps a log of what recommendations have been made to a particular user and whether that user accepts the recommendation. It's like

user_id	item_id	result
1	4	1
1	7	-1
5	19	1
5	80	1

where 1 means the user accepted the recommendation while -1 means the user did not respond to the recommendation.

Question: If I am going to make recommendations to a bunch of users based on the kind of log described above, and I want to maximize MAP@3 scores, how should I deal with the implicit data (1 or -1)?

My idea is to treat 1 and -1 as ratings, and predict the rating using factorization machines-type algorithms. But this does not seem right, given the asymmetry of the implicit data (-1 does not mean the user does not like the recommendation).

Edit 1 Let us think about it in the context of a matrix factorization approach. If we treat -1 and 1 as ratings, there will be some problem. For example, user 1 likes movie A which scores high in one factor (e.g. having glorious background music) in the latent factor space. The system recommends movie B which also scores high in “glorious background music”, but for some reason user 1 is too busy to look into the recommendation, and we have a -1 rating movie B. If we just treat 1 or -1 equally, then the system might be discouraged to recommend movie with glorious BGM to user 1 while user 1 still loves movie with glorious BGM. I think this situation is to be avoided.

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#))

User: [wdg](#) 

[Answer](#)  by [sean-owen](#) 

Your system isn't just trained on items that are recommended right? if so you have a big feedback loop here. You want to learn from all clicks/views, I hope.

You suggest that not-looking at an item is a negative signal. I strongly suggest you do not treat it that way. Not interacting with something is almost always best treated as no information. If you have an explicit signal that indicates a dislike, like a down vote (or, maybe watched 10 seconds of a video and stopped), maybe that's valid.

I would not construe this input as rating-like data. (Although in your case, you may get away with it.) Instead think of them as weights, which is exactly the treatment in the Hu Koren Volinsky paper on ALS that @Trey mentions in a comment. This lets you record relative strength of positive/negative interactions.

Finally I would note that this paper, while is very likely to be what you're looking for, does not provide for negative weights. It is simple to extend in this way. If you get that far I can point you to the easy extension, which exists already in two implementations that I know of, in [Spark](#)  and [Oryx](#) .

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#))

[Q: Create most “average” cosine similarity observation](#)

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Prev Q](#)) ([Next Q](#))

For a recommendation system I'm using cosine similarity to compute similarities between items. However, for items with small amounts of data I'd like to bin them under a general “average” category (in the general not mathematical sense). To accomplish this I'm currently trying to create a synthetic observation to represent that middle of the road point.

So for example if these were my observations (rows are observations, cols are features):

```
[[0, 0, 0, 1, 1, 1, 0, 1, 0],  
 [1, 0, 1, 0, 0, 0, 1, 0, 0],  
 [1, 1, 1, 1, 0, 1, 0, 1, 1],  
 [0, 0, 1, 0, 0, 1, 0, 1, 0]]
```

A strategy where I'd simply take the actual average of all features across observations would generate a synthetic datapoint such as follows, which I'd then append to the matrix before doing the similarity calculation.

```
[ 0.5 ,  0.25,  0.75,  0.5 ,  0.25,  0.75,  0.25,  0.75,  0.25]
```

While this might work well with certain similarity metrics (e.g. L1 distance) I'm sure there are much better ways for cosine similarity. Though, at the moment, I'm having trouble reasoning my way through angles between lines in high dimensional space.

Any ideas?

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Prev Q](#)) ([Next Q](#))

User: [eric-chiang](#) 

[Answer](#)  by [debasis](#) 

You are doing the correct thing. Technically, this averaging leads to computing the [centroid](#)  in the Euclidean space of a set of N points. The centroid works pretty well with cosine similarities (cosine of the angles between normalized vectors), e.g. [the Rocchio algorithm](#) .

Tags: [recommendation](#) ([Prev Q](#)) ([Next Q](#)), [similarity](#) ([Prev Q](#)) ([Next Q](#))

[Q: Price optimization for tiered and seasonal products](#)

Tags: [recommendation](#) ([Prev Q](#))

Assuming I can collect the demand of the purchase of a certain product that are of different market tiers. Example: Product A is low end goods. Product B is another low end goods. Product C and D are middle-tier goods and product E and F are high-tier goods.

We have collected data the last year on the following 1. Which time period (season - festive? non-festive?) does the different tier product reacts based on the price set? Reacts

refer to how many % of the product is sold at certain price range 2. How fast the reaction from the market after marketing is done? Marketing is done on 10 June and the products are all sold by 18 June for festive season that slated to happen in July (took 8 days at that price to finish selling)

How can data science benefit in terms of recommending 1. If we should push the marketing earlier or later? 2. If we can higher or lower the price? (Based on demand and sealing rate?)

Am I understanding it right that data science can help a marketer in this aspect? Which direction should I be looking into if I am interested to learn about it.

Tags: [recommendation](#) ([Prev Q](#))

User: [guo-hong-lim](#) 

[Answer](#)  by [sheldonkreger](#) 

You should be able to use [linear regression](#)  to find correlation between the factors which cause your products to sell better (or worse).

There are many correlations you can test against in this data set. Some examples are:

1. If a product has been marketed aggressively, does it sell more quickly?
2. If a low tier item is available, do fewer high-tier items sell?
3. If multiple high-tier items are available, are fewer sold of each item?

Keep in mind that correlation does not necessarily imply causation. Always think about other factors which may cause sales to go up and down. For example, you may sell more high tier items in a season one year than another year. But, this could be due to changes in the overall economy, rather than changes in your pricing.

The second thing you can do is perform [A/B tests](#)  on your product sales pages. This gives you clear feedback right away. Some example tests could be:

1. Show the user one high-tier product and one low-tier product (A). Show the user two high-tier products and no low-tier products(B). Which page generates more revenue?
2. Send out marketing emails for a seasonal sale 5 days in advance to one group of users (A). Send the same email to a different set of users 1 day in advance (B).

There are many possibilities. Use your intuition and think about previous knowledge you have about your products.

Tags: [recommendation](#) ([Prev Q](#))

Visualization

[Skip to questions](#),

Wiki by user [dawny33](#) 

Overview

Data visualization refers to techniques for presenting results in graphical form, such as histograms, scatterplots, or boxplots. Data visualization is a special challenge for data with high dimensionality.

If your question is only about how to get particular software to produce a specific effect, then it is likely not on topic here. Programming questions (for example, in Python, or in R with ggplot, etc.) for which you can supply a reproducible example are usually welcomed on [StackOverflow](#) .

References

The following question contains references to data visualization resources:

- [Modern successor to Exploratory Data Analysis by Tukey?](#) 
-

Questions

[Q: How to animate growth of a social network?](#)

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Prev Q](#)) ([Next Q](#))

I am seeking for a library/tool to visualize how social network changes when new nodes/edges are added to it.

One of the existing solutions is [SoNIA: Social Network Image Animator](#). It let's you make movies like [this one](#).

SoNIA's documentation says that it's broken at the moment, and besides this I would prefer JavaScript-based solution instead. So, my question is: are you familiar with any tools or are you able to point me to some libraries which would make this task as easy as possible?

Right after posting this question I'll dig into [sigma.js](#), so please consider this library covered.

In general, my input data would be something like this:

```
time_elapsed; node1; node2
1; A; B
2; A; C
3; B; C
```

So, here we have three points in time (1, 2, 3), three nodes (A, B, C), and three edges, which represent a triadic closure between the three considered nodes.

Moreover, every node will have two attributes (age and gender), so I would like to be able to change the shape/colour of the nodes.

Also, after adding a new node, it would be perfect to have some ForceAtlas2 or similar algorithm to adjust the layout of the graph.

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Prev Q](#)) ([Next Q](#))

User: [wojciech-walczak](#)

[Answer](#) by [graeme-stuart](#)

Fancy animations are cool

I was very impressed when I saw [this animation](#) of the [discourse](#) git repository. They used [Gource](#) which is specifically for git. But it may give ideas about how to represent the dynamics of growth.

You can create animations with matplotlib

[This stackoverflow answer](#) seems to point at a python/networkx/matplotlib solution.

But D3.js provides interaction

If you're looking for a web-based solution then d3.js is excellent. See [this](#) and [this](#) for example. See also [this stackoverflow question](#), the accepted answer points to D3.js again.

Conclusion

I would be drawn towards the python/networkx options for network analysis (possibly to add attributes to your raw data file for example). Then, for visualisation and dissemination D3.js is perfect. You might be surprised how easy it can be to write d3.js once you get into it. I believe [it even works within an ipython notebook!](#)

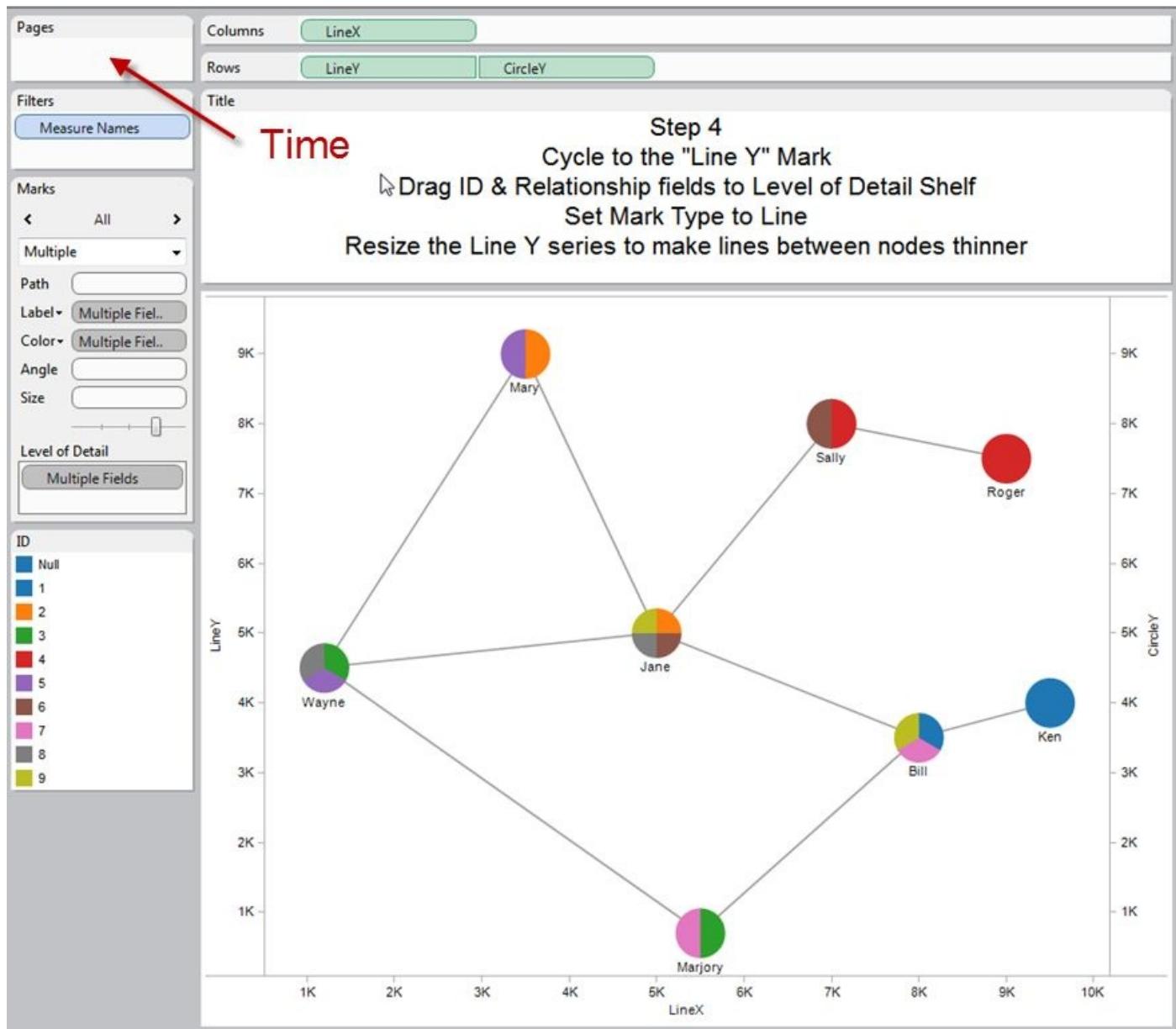
[Answer](#) by [ihars](#)

My first guess is to [visualize social network in Tableau](#).

And particularly: [building network graphs in Tableau](#).

What you need is to add time dimension to the “Pages” section to be able to see network change dynamics.

This is screen from the link above.



[Answer](#) by [wojciech-walczak](#)

It turned out that this task was quite easy to accomplish using [vis.js](#). This was the best example code which I have found.

The example of what I have built upon this is [here](#) (scroll to the bottom of this post). This graph represents the growth of a subnetwork of Facebook friends. Green dots are females, blue ones are males. The darker the colour, the older the user. By clicking “Dodaj węzły” you can add more nodes and edges to the graph.

Anyway, I am still interested in other ways to accomplish this task, so I won't accept any answer as for now.

Thanks for your contributions!

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#)), [social-network-analysis](#) ([Prev Q](#)) ([Next Q](#)), [time-series](#) ([Prev Q](#)) ([Next Q](#))

Q: What visualization technique to best describe a recommendation

[dataset?](#)

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#))

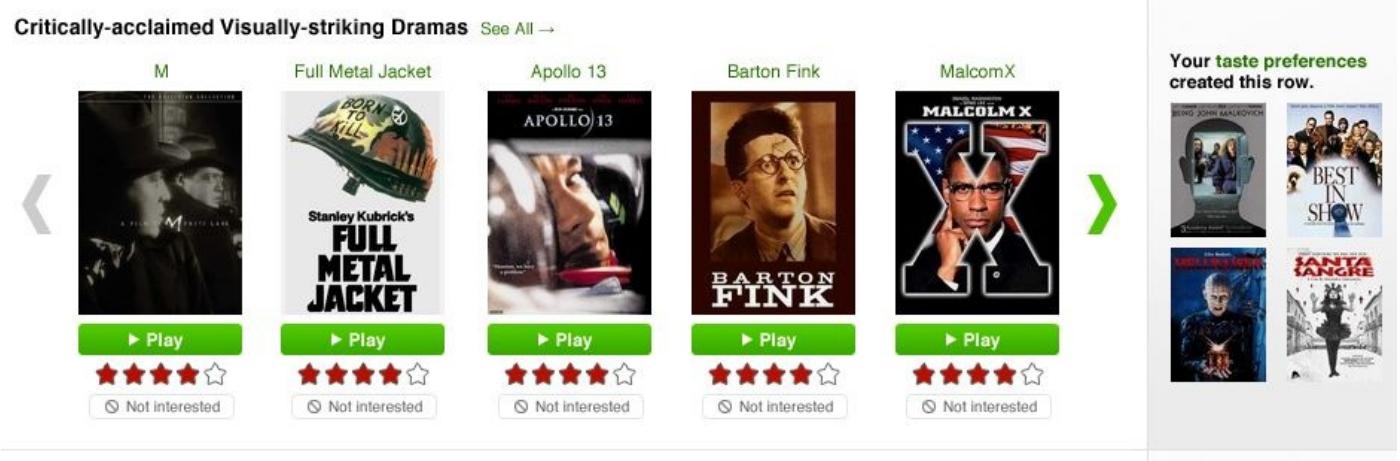
I've written a simple recommender which generates recommendations for users based on what they have clicked. The recommender generates a data file with the following format:

```
userid,userid,similarity (between 0 and 1 - closer to 0 the more similar the users)
a,b,.2
a,c,.3
a,d,.4
a,e,.1
e,b,.3
e,c,.5
e,d,.8
```

I've looked at [some graphs](#) , but I'm not sure which one to use, or if there are other ones that will better display the user similarities from the dataset above. Any suggestions?

I'm aiming this visualization at business users who are not at all technical. I would just like to show them an easy to understand visual that details how similar some users are and so convince the business that for these users the recommendation system is useful.

@Steve Kallestad do you mean something like this :



Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#))

User: [blue-sky](#) 

[Answer](#)  by [phonon](#) 

I think you're looking for a [similarity matrix](#)  (see bottom of the page). If you don't have data on similarity between certain pairs, you can always leave them as grey or white. Also, this will only work for data sets small enough to actually make out what's going on. I'd say 25 rows / columns maximum.

In a similarity matrix, x, and y coordinates correspond to the two things you're comparing, while a [colormap](#)  magnitude represents similarity

EDIT: One thing you could do to replace the colormap is the insert, say, circles of different sizes according to the similarity metric. Or you could insert the numbers themselves, again, varying the size of the number as the magnitude of that number varies. Size usually works best in business visualizations.

[Answer](#) by [steve-kallestad](#)

Personally, I think Netflix got it right. Break it down into a confidence rating from 1-5 and show your recommendations based on the number of yellow stars.

It doesn't have to be stars, but those icon based graphs are very easy to interpret and get the point across clearly.

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#))

Q: Visualizing a graph with a million vertices

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#)), [graphs](#) ([Prev Q](#)) ([Next Q](#))

What is the best tool to use to visualize (draw the vertices and edges) a graph with 1000000 vertices? There are about 50000 edges in the graph. And I can compute the location of individual vertices and edges.

I am thinking about writing a program to generate a svg. Any other suggestions?

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#)), [graphs](#) ([Prev Q](#)) ([Next Q](#))

User: [cici](#)

[Answer](#) by [aleksandr-blekh](#)

I also suggest Gephi software (<https://gephi.github.io>), which seems to be quite powerful. Some additional information on using Gephi with **large networks** can be found [here](#) and, more generally, [here](#). Cytoscape (<http://www.cytoscape.org>) is an alternative to Gephi, being another popular platform for complex network analysis and visualization.

If you'd like to work with networks **programmatically** (including visualization) in R, Python or C/C++, you can check igraph collection of libraries. Speaking of R, you may find interesting the following blog posts: on **using R with Cytoscape** (<http://www.vesnam.com/Rblog/viznets1>) and on **using R with Gephi** (<http://www.vesnam.com/Rblog/viznets2>).

For **extensive lists** of *network analysis and visualization software*, including some comparison and reviews, you might want to check the following pages: 1) http://wiki.cytoscape.org/Network_analysis_links; 2) <http://www.kdnuggets.com/software/social-network-analysis.html>; 3) <http://www.activatenetworks.net/social-network-analysis-sna-software-review>.

[Answer](#) by [spacedman](#)

<https://gephi.github.io> says it can handle a million edges. If your graph has 1000000 vertices and only 50000 edges then most of your vertices won't have any edges anyway.

In fact the Gephi spec is the dual of your example: "Networks up to 50,000 nodes and 1,000,000 edges"

[Answer](#) by [sobach](#)

I think, that Gephi could face with lack-of-memory issues, you will need at least 8Gb of RAM. Though number of edges is not extremely huge.

Possibly, more appropriate tool in this case would be [GraphViz](#). It's a command line tool for network visualizations, and presumably would be more tolerant to graph size. Moreover, as I remember, in GraphViz it is possible to use precomputed coordinates to facilitate computations.

I've tried to find a real-world examples of using GraphViz with huge graphs, but didn't succeed. Though I found similar discussion on [Computational Science](#).

Tags: [visualization](#) ([Prev Q](#)) ([Next Q](#)), [graphs](#) ([Prev Q](#)) ([Next Q](#))

[Q: How to plot large web-based heatmaps?](#)

Tags: [visualization](#) ([Prev Q](#))

I want to plot large heatmaps (say a matrix 500×500). I can do it in Python/matplotlib.pyplot with pcolor, but it is not interactive (and I need an interactive heatmap). I have tried with D3.js but what I found is aiming at displaying small heatmaps: <http://bl.ocks.org/tjdecke/5558084> Naively extending this example with a bigger matrix (e.g. 500×500) can crash the web-browser.

So, can anyone point me toward a good way of displaying and interacting with large heatmaps with a web-based technology: I want to be able to interact on a web-page or a ipython notebook.

Tags: [visualization](#) ([Prev Q](#))

User: [mic](#)

[Answer](#) by [magsol](#)

[Plotly](#) and [Lightning](#) are [supposedly] able to visualize extremely large data sets.

Tags: [visualization](#) ([Prev Q](#))

Databases

[Skip to questions](#),

Wiki by user [dawny33](#) 

From [Wikipedia](#) :

A database is an organized collection of data. The data is typically organized to model relevant aspects of reality (for example, the availability of rooms in hotels), in a way that supports processes requiring this information (for example, finding a hotel with vacancies).

A large proportion of websites and applications rely on databases. They are a crucial component of telecommunications systems, banking systems, video games, and just about any other software system or electronic device that maintains some amount of persistent information. In addition to persistence, database systems provide a number of other properties that make them exceptionally useful and convenient: reliability, efficiency, scalability, concurrency control, data abstraction, and high-level query languages.

Databases are so ubiquitous and important that computer science graduates frequently cite their database class as the one most useful to them in their industry or graduate-school careers.² 

The term *database* should not be confused with [Database Management System](#) (DBMS). A DBMS is the system software used to create and manage databases and provide users and applications with access to the database(s). A database is to a DBMS as a document is to a word processor.

Some useful references:

- [What is database?](#)  from Database Administrators SE.
 - [design](#) 
 - [managing a database](#) 
 - [tuning of a database](#) 
 - [database security](#) 
 - [Database connections](#) 
 - [Database normalization](#) 
-

Questions

[Q: Is this Neo4j comparison to RDBMS execution time correct?](#)

Tags: [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

Background: Following is from the book [Graph Databases](#) , which covers a performance test mentioned in the book [Neo4j in Action](#) :

Relationships in a graph naturally form paths. Querying, or traversing, the graph involves following paths. Because of the fundamentally path-oriented nature of the datamodel, the majority of path-based graph database operations are highly aligned with the way in which the data is laid out, making them extremely efficient. In their book Neo4j in Action, Partner and Vukotic perform an experiment using a relational store and Neo4j.

The comparison shows that the graph database is substantially quicker for connected data than a relational store. Partner and Vukotic's experiment seeks to find friends-of-friends in a social network, to a maximum depth of five. Given any two persons chosen at random, is there a path that connects them which is at most five relationships long? For a social network containing 1,000,000 people, each with approximately 50 friends, the results strongly suggest that graph databases are the best choice for connected data, as we see in Table 2-1.

Table 2-1. Finding extended friends in a relational database versus efficient finding in Neo4j

Depth	RDBMS Execution time (s)	Neo4j Execution time (s)	Records returned
2	0.016	0.01	~2500
3	30.267	0.168	~110,000
4	1543.505	1.359	~600,000
5	Unfinished	2.132	~800,000

At depth two (friends-of-friends) both the relational database and the graph database perform well enough for us to consider using them in an online system. While the Neo4j query runs in two-thirds the time of the relational one, an end-user would barely notice the the difference in milliseconds between the two. By the time we reach depth three (friend-of-friend-of-friend), however, it's clear that the relational database can no longer deal with the query in a reasonable timeframe: the thirty seconds it takes to complete would be completely unacceptable for an online system. In contrast, Neo4j's response time remains relatively flat: just a fraction of a second to perform the query—definitely quick enough for an online system.

At depth four the relational database exhibits crippling latency, making it practically useless for an online system. Neo4j's timings have deteriorated a little too, but the latency here is at the periphery of being acceptable for a responsive online system. Finally, at depth five, the relational database simply takes too long to complete the query. Neo4j, in contrast, returns a result in around two seconds. At depth five, it transpires almost the entire network is our friend: for many real-world use cases, we'd likely trim the results, and the timings.

Questions are:

- Is this a reasonable test to emulate what one might expect to find in a social network?
(Meaning do real social networks normally have nodes with approximately 50 friends for example; seems like the “rich get richer” model would be more natural for social networks, though might be wrong.)
- Regardless of the naturalness of the emulation, is there any reason to believe the results are off, or unreplicable?

Tags: [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

User: [blunders](#) 

[Answer](#)  by [rapaio](#) 

Looking at this document called [Anatomy of Facebook](#)  I note that the median is 100. Looking at the cumulative function plot I can bet that the average is higher, near 200. So 50 seems to not be the best number here. However I think that this is not the main issue here.

The main issue is the lack of information on how the database was used.

It seems reasonable that a data storage designed specially for graph structures to be more efficient than traditional RDBMs. However, even if the RDBMs are not in the latest trends as a data storage of choice, these systems evolved continuously in a race with the data set dimensions. There are various types of possible designs, various ways of indexing data, improvements related with concurrency and so on.

To conclude I think that regarding reproducibility, the study lack a proper description of how the database schema was designed. I do not expect that a database to dominate on such kind of interrogations, however I would expect that with a well-tuned design the differences to not be such massive.

[Answer](#)  by [dvk](#) 

There are good/fast ways to model graphs in RDBMS, and dumb/slow ways.

- Some use clever indexing and Stored Procs, trading CPU load and tuned temp tables on RAM disks for faster graph retrieval speed.
- Some use precomputed graph paths (this may be less feasible in social network scenario, but in a tree with majority of nodes being leaf nodes, it's a pretty good tradeoff space-for-time)
- Some simply compute in a loop, using un-tuned in-indexed temp table. From the #s thrown in the article, that smells like what they did (30 second- performance on fairly smallish data-set)

For example, I have my own tree computation.

- It is encapsulated in a highly-tuned stored proc
- While it's running in an enterprise-sized-hardware Sybase ASE15 dataserver,

that server is shared with a couple terabytes of data from **all other** enterprise apps, some much more data hungry than mine; and isn't dedicated solely to executing my queries.

- I did **not** have access to the main speedup tool, a temp table on a RAM disk.
- A representative set of data I was retrieving that seems to somewhat match theirs was getting a 150,000 node subtree out of 2.5M node full forest dataset (unlimited depth of tree, which varies between 5 and 15, but smaller *average* arity of a given node than the 50 friends listed in the experiment)
- I tuned it to the point that this query ~30-45 seconds. It most certainly does NOT exhibit the exponential slowdown that the figures in the question seem to indicate on their RDBMS performance, which is extra double strange given there is no exponential growth in the result set (which to me reeks of un-tuned index on a temp table from personal experience).

So, this comparison is **most likely** incorrect and based on poor RDBMS side design, although as the previous answer noted, it is impossible to ascertain without them open sourcing 100% of their code and table definitions.

Tags: [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

[Q: is this a good case for NOSQL?](#)

Tags: [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

I'm currently facing a project that I could solve with a relational database in a relatively painful way. Having heard so much about NOSQL, I'm wondering if there is not a more appropriate way of tackling it:

Suppose we are tracking a group of animals in a forest ($n \sim 500$) and would like to keep a record of a set of observations (this is a fictional scenario).

We would like to store the following information in a database:

- a unique identifier for each animal
- a description of the animal with structured fields: Species, Genus, Family, ...
- a free text field with additional information
- each time-point at which it was detected close to a reference point
- a picture of the animal
- an indication whether two given animals are siblings

And:

- there might be additional features appearing later as more data comes in

We would like to be able to execute the following types of queries:

- return all the animals spotted between in a given time interval
- return all the animals of a given Species or Family
- perform a text search on the free text field

Which particular database system would you recommend ? Is there any tutorial / examples that I could use as a starting point ?

Tags: [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

User: [cafe876](#) 

[Answer](#)  by [spacedman](#) 

Three tables: **animal**, **observation**, and **sibling**. The observation has an **animal_id** column which links to the animal table, and the sibling table has **animal_1_id** and **animal_2_id** columns that indicates two animals are siblings for each row.

Even with 5000 animals and 100000 observations I don't think query time will be a problem for something like PostgreSQL for most reasonable queries (obviously you can construct unreasonable queries but you can do that in any system).

So I don't see how this is "relatively painful". Relative to what? The only complexity is the sibling table. In NOSQL you might store the full list of siblings in the record for each animal, but then when you add a sibling relationship you have to add it to both sibling's animal records. With the relational table approach I've outlined, it only exists once, but at the expense of having to test against both columns to find an animal's siblings.

I'd use PostgreSQL, and that gives you the option of using PostGIS if you have location data - this is a geospatial extension to PostgreSQL that lets you do spatial queries (point in polygon, points near a point etc) which might be something for you.

I really don't think the properties of NOSQL databases are a problem here for you - you aren't changing your schema every ten minutes, you probably **do** care that your database is ACID-compliant, and you don't need something web-scale.

<http://www.mongodb-is-web-scale.com/>  [warning: strong language]

Tags: [databases](#) ([Prev Q](#)) ([Next Q](#)), [nosql](#) ([Prev Q](#)) ([Next Q](#))

[Q: What makes columnar databases suitable for data science?](#)

Tags: [databases](#) ([Prev Q](#))

What are some of the advantages of columnar data-stores which make them more suitable for data science and analytics?

Tags: [databases](#) ([Prev Q](#))

User: [dawny33](#) 

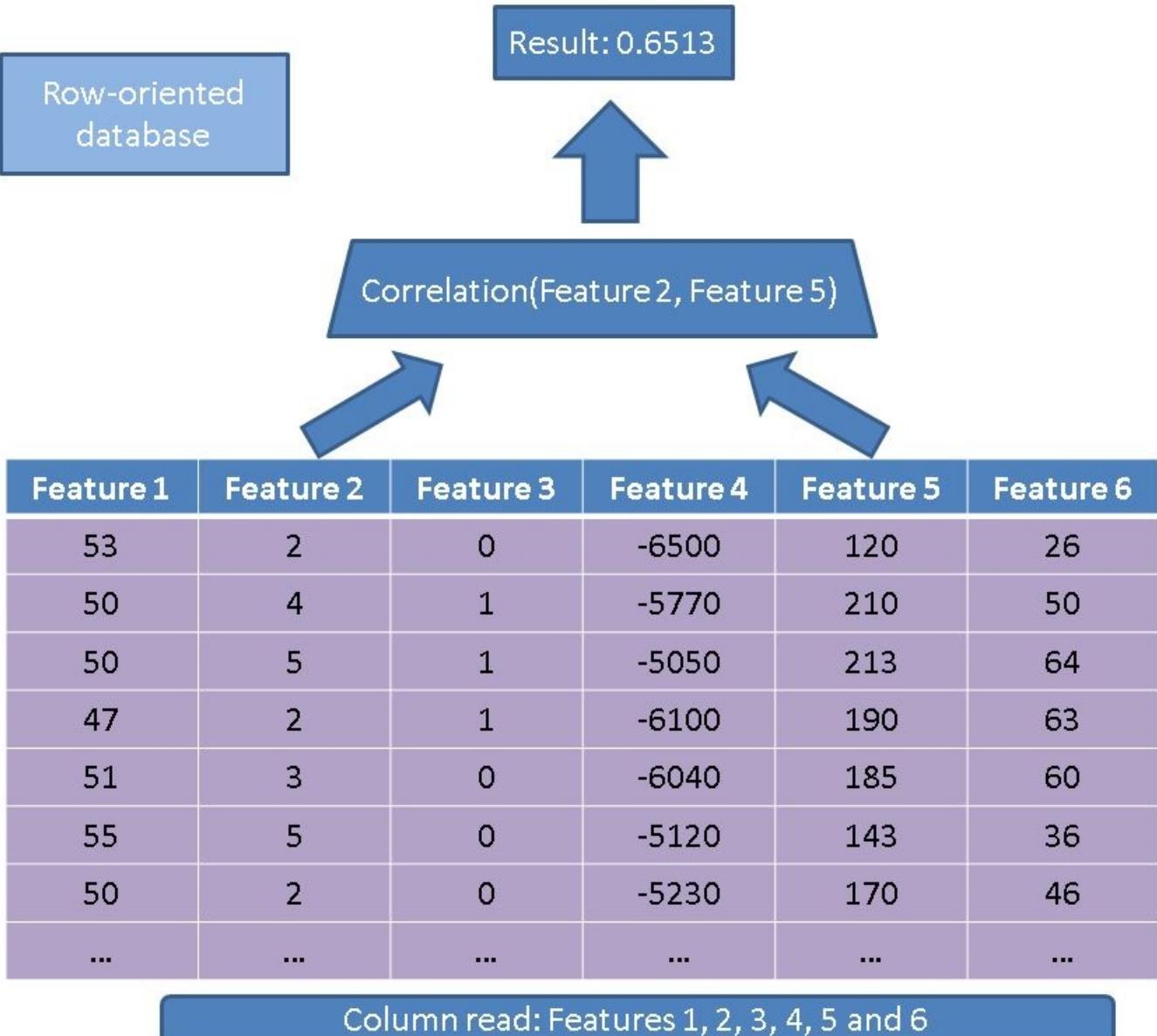
[Answer](#)  by [franck-dernoncourt](#) 

A column-oriented database (=columnar data-store) stores the data of a table column by column on the disk, while a row-oriented database stores the data of a table row by row.

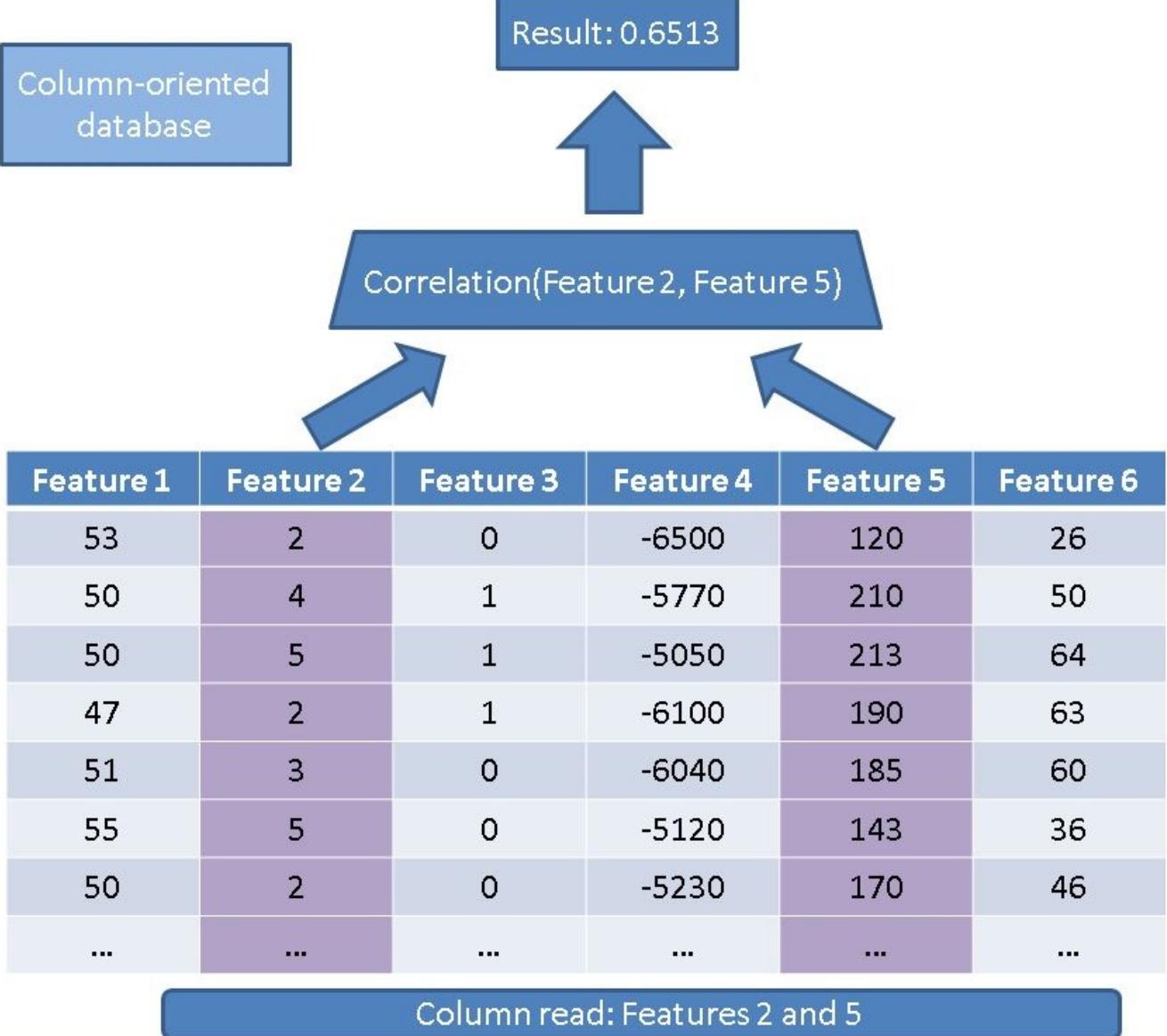
There are two main advantages of using a column-oriented database in comparison with a row-oriented database. The first advantage relates to the amount of data one's need to read in case we perform an operation on just a few features. Consider a simple query:

```
SELECT correlation(feature2, feature5)  
FROM records
```

A traditional executor would read the entire table (i.e. all the features):



Instead, using our column-based approach we just have to read the columns which are interested in:



The second advantage, which is also very important for large databases, is that column-based storage allows better compression, since the data in one specific column is indeed homogeneous than across all the columns.

The main drawback of a column-oriented approach is that manipulating (lookup, update or delete) an entire given row is inefficient. However the situation should occur rarely in databases for analytics (“warehousing”), which means most operations are read-only, rarely read many attributes in the same table and writes are only appends.

Some RDMS offer a column-oriented storage engine option. For example, PostgreSQL has natively no option to store tables in a column-based fashion, but Greenplum has created a closed-source one (DBMS2, 2009). Interestingly, Greenplum is also behind the open-source library for scalable in-database analytics, MADlib (Hellerstein et al., 2012), which is no coincidence. More recently, CitusDB, a startup working on high speed, analytic database, released their own open-source columnar store extension for PostgreSQL, CSTORE (Miller, 2014). Google’s system for large scale machine learning Sibyl also uses column-oriented data format (Chandra et al., 2010). This trend reflects the growing interest around column-oriented storage for large-scale analytics. Stonebraker et al. (2005) further discuss the advantages of column-oriented DBMS.

Two concrete use cases: [How are most datasets for large-scale machine learning stored?](#)

(most of the answer comes from Appendix C of: [BeatDB: An end-to-end approach to unveil saliences from massive signal data sets. Franck Demorcourt, S.M, thesis, MIT Dept of EECS](#))

[Answer](#) by [anony-mousse](#)

It depends on *what* you do.

Column stores have two key benefits:

- whole columns can be skipped
- run-length compression works better on columns (for certain data types; in particular with few distinct values)

However they also have drawbacks:

- many algorithms will need all columns, and only record at a time (e.g. k-means) or may even need to compute a pairwise distance matrix
- compression techniques only work well on sparse data types and factors, but not well on double-valued continuous data
- appends on column stores are expensive, so it is not ideal for streaming / changing data

Columnar storage is really popular for OLAP aka “stupid analytics” (Michael Stonebraker) and of course for *preprocessing* where you may indeed be interested in discarding whole columns (but you would need to have structured data first - you don’t store JSONs in columnar format). Because the columnar layout is really nice for e.g. counting how many apples you have sold last week.

For much of the scientific / data science use cases, **array databases** appear to be the way to go (plus, of course, unstructured input data). E.g. SciDB and RasDaMan.

In many cases (e.g. deep learning), matrixes and arrays are the data types you need, not columns. MapReduce etc. can still be useful in preprocessing, of course. Maybe even column data (but array database usually support a column-like compression, too).

[Answer](#) by [user554481](#)

I haven’t used a columnar database, but I’ve used an open source columnar file format called Parquet, and I think the benefits are probably the same — faster processing of data when you only need to query a small subset of a large number of columns. I had a query running on about 50 terabytes of Avro files (a row oriented file format) with 673 columns that took about an hour and a half on a 140 node Hadoop cluster. With Parquet, the same query took about 22 minutes because I only needed 5 columns.

If you had a small number of columns or were using a large proportion of your columns, I don’t think a columnar database would make much of a difference vs a row oriented one because you would still have to basically scan all of your data. I believe columnar

databases store columns separately whereas row oriented databases store rows separately. Your query will be faster any time you're able to read less data from disk.

[This link](#) explains more of the details.

Tags: [databases](#) ([Prev Q](#))

Feature Selection

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Feature selection, also called **attribute selection** or **feature reduction**, refers to techniques for identifying a subset of features of a data set that are relevant to a given problem. By removing irrelevant and redundant features, successful feature selection can avoid the [curse of dimensionality](#)  and improve the performance, speed, and interpretability of subsequent models.

Feature selection includes **manual methods** (such those based on domain knowledge) and automatic methods. **Automatic methods** are often categorized into **filter**, **wrapper**, and **embedded** approaches.

Filter approaches perform feature selection as a separate preprocessing step before the learning algorithm. Filter approaches thus look only at the intrinsic properties of the data. Filter methods include Wilcoxon rank sum tests and [Correlation based tests](#) .

Wrapper approaches uses performance of a learning algorithm to select features. A search algorithm is “wrapped” around the learning algorithm to ensure the space of feature subsets is adequately searched. As such, wrapper methods can be seen as conducting the model hypothesis search within the feature subset search. Examples of wrapper approaches are [simulated annealing](#)  and [beam search](#) .

Embedded approaches incorporate variable selection as a part of the training process, with feature relevance obtained analytically from the objective of the learning model. Embedded methods can be seen as a search in the combined space of feature subsets and hypotheses. Examples of embedded approaches are [boosting](#)  and recursive ridge regression.

Questions

Q: What is dimensionality reduction? What is the difference between feature selection and extraction?

Tags: [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

From wikipedia,

dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

What is the difference between feature selection and feature extraction?

What is an example of dimensionality reduction in a Natural Language Processing task?

Tags: [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

User: [alvas](#)

[Answer](#) by [damienfrancois](#)

Simply put:

- feature selection: you select a subset of the original feature set; while
- feature extraction: you build a new set of features from the original feature set.

Examples of feature extraction: extraction of contours in images, extraction of digrams from a text, extraction of phonemes from recording of spoken text, etc.

Feature extraction involves a transformation of the features, which often is not reversible because some information is lost in the process of dimensionality reduction.

[Answer](#) by [ragingsloth](#)

Dimensionality reduction is typically choosing a basis or mathematical representation within which you can describe most but not all of the variance within your data, thereby retaining the relevant information, while reducing the amount of information necessary to represent it. There are a variety of techniques for doing this including but not limited to PCA, ICA, and Matrix Feature Factorization. These will take existing data and reduce it to the most discriminative components. These all allow you to represent most of the information in your dataset with fewer, more discriminative features.

Feature Selection is hand selecting features which are highly discriminative. This has a lot more to do with feature engineering than analysis, and requires significantly more work on the part of the data scientist. It requires an understanding of what aspects of your dataset are important in whatever predictions you're making, and which aren't. Feature extraction usually involves generating new features which are composites of existing features. Both

of these techniques fall into the category of feature engineering. Generally feature engineering is important if you want to obtain the best results, as it involves creating information that may not exist in your dataset, and increasing your signal to noise ratio.

[Answer](#)  by [iliasfl](#) 

As in @damienfrancois answer feature selection is about selecting a subset of features. So in NLP it would be selecting a set of specific words (the typical in NLP is that each word represents a feature with value equal to the frequency of the word or some other weight based on TF/IDF or similar).

Dimensionality reduction is the introduction of new feature space where the original features are represented. The new space is of lower dimension than the original space. In case of text an example would be the [hashing trick](#)  where a piece of text is reduced to a vector of few bits (say 16 or 32) or bytes. The amazing thing is that the geometry of the space is preserved (given enough bits), so relative distances between documents remain the same as in the original space, so you can deploy standard machine learning techniques without having to deal with unbound (and huge number of) dimensions found in text.

Tags: [feature-selection](#) ([Prev Q](#)) ([Next Q](#))

[Q: Does scikit-learn have forward selection/stepwise regression algorithm?](#)

Tags: [feature-selection](#) ([Prev Q](#))

I'm working on the problem with too many features and training my models takes way too long. I implemented forward selection algorithm to choose features.

However, I was wondering does scikit-learn have forward selection/stepwise regression algorithm?

Tags: [feature-selection](#) ([Prev Q](#))

User: [maksud](#) 

[Answer](#)  by [brentlance](#) 

No, sklearn doesn't seem to have a forward selection algorithm. However, it does provide recursive feature elimination, which is a greedy feature elimination algorithm similar to sequential backward selection. See the documentation here:

http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html 

Tags: [feature-selection](#) ([Prev Q](#))

NoSQL

[Skip to questions](#),

Wiki by user [tasos](#) 

NoSQL (sometimes expanded to “not only [sql](#) rdbms 

NoSQL systems:

- Specifically designed for high load
- Natively support horizontal scalability
- Fault tolerant
- Store data in denormalised manner
- Do not usually enforce strict database schema
- Do not usually store data in a table
- Sometimes provide eventual consistency instead of ACID transactions

In contrast to RDBMS, NoSQL systems:

- Do not guarantee data consistency
- Usually support a limited query language (subset of SQL or another custom query language)
- May not provide support for transactions/distributed transactions
- Do not usually use some advanced concepts of RDBMS, such as triggers, views, stored procedures

NoSQL implementations can be categorised by their manner of implementation:

- [Column-oriented](#) Document store Graph Key-value store Multivalue databases Object databases Tripplestore Tuple store 

Free NoSQL Books

- [CouchDB: The Definitive Guide](#) The Little MongoDB Book The Little Redis Book



Questions

[Q: What is the Best NoSQL backend for a mobile game](#)

Tags: [nosql](#) ([Prev Q](#))

What is the best noSQL backend to use for a mobile game? Users can make a lot of servers requests, it needs also to retrieve users' historical records (like app purchasing) and analytics of usage behavior.

Tags: [nosql](#) ([Prev Q](#))

User: [filipe-ferminiano](#) 

[Answer](#)  by [alex-i](#) 

Some factors you might consider:

Developer familiarity: go with whatever you or your developers are familiar with. Mongo, Couch, Riak, DynamoDB etc all have their strengths but all should do ok here, so rather than going for an unfamiliar solution that might be slightly better go for familiar and save a bunch of development time.

Ease of cloud deployment: for example, if you are using Amazon AWS, then DynamoDB is likely an excellent choice. Sure, you could use Mongo on AWS, but why bother? Other cloud providers have their own preferred db, for example if you are using Google AppEngine, it makes sense to use BigTable or Cloud Datastore.

Your use case seems both well suited to NoSQL and not very challenging since your data has a natural partition by user. I think you'd be technically ok with anything, which is why I'm mainly covering other factors.

Tags: [nosql](#) ([Prev Q](#))

Predictive Modeling

Questions

[Q: Best regression model to use for sales prediction](#)

Tags: [predictive-modeling](#) ([Prev Q](#))

I have the following variables along with sales data going back a few years:

- date # simple date, can be split in year, month etc
- shipping_time (0-6 weeks) # 0 weeks means in stock, more weeks means the product is out of stock but a shipment is on the way to the warehouse. Longer shipping times have a significant impact on sales.
- sales # amount of products sold

I need to predict the sales (which vary seasonally) while taking into account the shipping time. What would be a simple regression model that would produce reasonable results? I tried linear regression with only date and sales, but this does not account for seasonality, so the prediction is rather weak.

Edit: As a measure of accuracy, I will withhold a random sample of data from the input and compare against the result.

Extra points if it can be easily done in python/scipy

Data can look like this

date	delivery_time	sales
2015-01-01	0	10
2015-01-01	7	12
2015-01-02	7	13
...		

Tags: [predictive-modeling](#) ([Prev Q](#))

User: [noobstats](#) 

[Answer](#)  by [thomas-cleberg](#) 

This is a pretty classic **ARIMA** dataset. **ARIMA** is implemented in the StatsModels package for Python, the documentation for which is available [here](#) .

An **ARIMA** model with seasonal adjustment may be the simplest reasonably successful forecast for a complex time series such as sales forecasting. It may (probably will) be that you need to combine the method with an additional model layer to detect additional fluctuation beyond the auto-regressive function of your sales trend.

Unfortunately, simple linear regression models tend to fare quite poorly on time series data.

[Answer](#) by [dawny33](#)

Did you try time series modelling? If not, then you should.

I tried linear regression with only date and sales, but this does not account for seasonality

The [moving average model](#) is something which would fit nicely to your dataset.

However, as you say that your model is exhibiting seasonality, you need to adjust the moving averages so that it takes the seasonality into account.

So, the best model for your dataset would be the SARIMA model. It is just the Auto-Regressive Integrated Moving Average (ARIMA) model but with seasonal adjustments.

[Here](#) is one of the questions which I have answered which further helps you understand minor seasonality and trend adjustments, along with the R code.

[[Further reading](#)]

Tags: [predictive-modeling](#) ([Prev Q](#))

Definitions

[Skip to questions,](#)

Wiki by user [clayton](#) 

Use the [definitions](#) tag when:

You think we should create an official definition.

An existing Tag Wiki needs a more precise definition to avoid confusion and we need to create consensus before an edit.

(rough draft - needs filling out)

Questions

[Q: Parallel and distributed computing](#)

Tags: [definitions](#) ([Prev Q](#)) ([Next Q](#))

What is(are) the difference(s) between parallel and distributed computing? When it comes to scalability and efficiency, it is very common to see solutions dealing with computations in clusters of machines, and sometimes it is referred to as a parallel processing, or as distributed processing.

In a certain way, the computation seems to be always parallel, since there are things running concurrently. But is the distributed computation simply related to the use of more than one machine, or are there any further specificities that distinguishes these two kinds of processing? Wouldn't it be redundant to say, for example, that a computation is *parallel AND distributed*?

Tags: [definitions](#) ([Prev Q](#)) ([Next Q](#))

User: [rubens](#) 

[Answer](#)  by [damienfrancois](#) 

Simply set, ‘parallel’ means running concurrently on distinct resources (CPUs), while ‘distributed’ means running across distinct computers, involving issues related to networks.

Parallel computing using for instance [OpenMP](#)  is not distributed, while parallel computing with [Message Passing](#)  is often distributed.

Being in a ‘distributed but not parallel’ setting would mean under-using resources so it is seldom encountered but it is conceptually possible.

[Answer](#)  by [marco13](#) 

The terms “parallel computing” and “distributed computing” certainly have a large overlap, but can be differentiated further. Actually, you already did this in your question, by later asking about “parallel processing” and “distributed processing”.

One could consider “distributed computing” as the more general term that involves “distributed processing” as well as, for example, “distributed storage”. The common term, “distributed”, usually refers to some sort of [Message Passing](#)  over a network, between machines that are physically separated.

The term “parallel computing” is also in the process of being further defined, e.g. by explicitly differentiating between the terms “parallel” and “concurrent”, where - roughly - the first one refers [data parallelism](#)  and the latter to [task parallelism](#) , although there are hardly really strict and binding definitions.

So one could say that

- “distributed processing” usually (although not necessarily) means that it also is

“parallel processing”

- “distributed computing” is more general, and also covers aspects that are not related to parallelism
 - and obviously, “parallel computing”/“parallel processing” does not imply that it is “distributed”
-

[Answer](#)  by [rubens](#) 

The answers presented so far are very nice, but I was also expecting an emphasis on a particular difference between parallel and distributed processing: the code executed. Considering parallel processes, the code executed is the same, regardless of the level of parallelism (instruction, data, task). You write a *single code*, and it will be executed by different threads/processors, e.g., while computing matrices products, or generating permutations.

On the other hand, distributed computing involves the execution of different algorithms/programs at the same time in different processors (from one or more machines). Such computations are later merged into a intermediate/final results by using the available means of data communication/synchronization (shared memory, network). Further, distributed computing is very appealing for BigData processing, as it allows for exploiting disk parallelism (usually the bottleneck for large databases).

Finally, for the level of parallelism, it may be taken rather as a constraint on the synchronization. For example, in GPGPU, which is single-instruction multiple-data (SIMD), the parallelism occurs by having different inputs for a single instruction, each pair (*data_i, instruction*) being executed by a different thread. Such is the restraint that, in case of divergent branches, it is necessary to discard lots of unnecessary computations, until the threads reconverge. For CPU threads, though, they commonly diverge; yet, one may use synchronization structures to grant concurrent execution of specific sections of the code.

Tags: [definitions](#) ([Prev Q](#)) ([Next Q](#))

[Q: Starting my career as Data Scientist, is Software Engineering experience required?](#)

Tags: [definitions](#) ([Prev Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

I am an MSc student at the University of Edinburgh, specialized in machine learning and natural language processing. I had some practical courses focused on data mining, and others dealing with machine learning, bayesian statistics and graphical models. My background is a BSc in Computer Science.

I did some software engineering and I learnt the basic concepts, such as design patterns, but I have never been involved in a large software development project. However, I had a data mining project in my MSc. My question is, if I want to go for a career as Data Scientist, should I apply for a graduate data scientist position first, or should I get a

position as graduate software engineer first, maybe something related to data science, such as big data infrastructure or machine learning software development?

My concern is that I might need good software engineering skills for data science, and I am not sure if these can be obtained by working as a graduate data scientist directly.

Moreover, at the moment I like Data Mining, but what if I want to change my career to software engineering in the future? It might be difficult if I specialised so much in data science.

I have not been employed yet, so my knowledge is still limited. Any clarification or advice are welcome, as I am about to finish my MSc and I want to start applying for graduate positions in early October.

Tags: [definitions](#) ([Prev Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

User: [keiszn](#) 

[Answer](#)  by [aleksandr-blekh](#) 

1) I think that there's no need to question whether your background is adequate for a career in data science. CS degree IMHO is **more than enough** for data scientist from software engineering point of view. Having said that, theoretical knowledge is not very helpful without matching **practical experience**, so I would definitely try to **enrich** my experience through participating in *additional school projects, internships or open source projects* (maybe ones, focused on data science / machine learning / artificial intelligence).

2) I believe your concern about **focusing** on data science **too early** is unfounded, as long as you will be practicing software engineering either as a part of your data science job, or additionally in your spare time.

3) I find the following **definition of a data scientist** rather accurate and hope it will be helpful in your future career success:

A *data scientist* is someone who is better at statistics than any software engineer and better at software engineering than any statistician.

P.S. Today's **enormous** number of various resources on data science topics is mind-blowing, but this **open source curriculum for learning data science** might fill some gaps between your BSc/MSc respective curricula and reality of the data science career (or, at least, provide some direction for further research and maybe answer some of your concerns): <http://datasciencemasters.org> , or on GitHub: <https://github.com/datasciencemasters/go> .

[Answer](#)  by [christian-sauer](#) 

From the job ads I have seen, the answer depends: There are jobs which are more technical in nature (designing big data projects, doing some analysis) or the exact opposite (doing analysis, storage etc. is someone else's job).

So I would say that **SOME** software design skills are extremely useful, but you don't need the ability to build a huge program in C# / Java or whatever. Why I like some SW skills is

simply that your code probably looks way better than code from someone who never programmed for the sake of programming. Most of the time, the latter code is very hard to understand / debug for outsiders. Also, sometimes your analysis needs to be integrated in a bigger program, an understanding of the needs of the programs certainly helps.

[Answer](#)  by [emre](#) 

Absolutely. Keep your software skills sharp. You can do this in an academic program if you simply implement by yourself all the algorithms you learn about.

Good selection of courses, btw. Consider getting an internship too.

Tags: [definitions](#) ([Prev Q](#)), [education](#) ([Prev Q](#)) ([Next Q](#))

Education

Questions

[Q: Qualifications for PhD Programs](#)

Tags: [education](#) ([Prev Q](#)) ([Next Q](#))

Yann LeCun mentioned in his [AMA](#)  that he considers having a PhD very important in order to get a job at a top company.

I have a masters in statistics and my undergrad was in economics and applied math, but I am now looking into ML PhD programs. Most programs say there are no absolutely necessary CS courses; however I tend to think most accepted students have at least a very strong CS background. I am currently working as a data scientist/statistician but my company will pay for courses. Should I take some intro software engineering courses at my local University to make myself a stronger candidate? What other advice you have for someone applying to PhD programs from outside the CS field?

edit: I have taken a few MOOCs (Machine Learning, Recommender Systems, NLP) and code R/python on a daily basis. I have a lot of coding experience with statistical languages and implement ML algorithms daily. I am more concerned with things that I can put on applications.

Tags: [education](#) ([Prev Q](#)) ([Next Q](#))

User: [bstockton](#) 

[Answer](#)  by [emre](#) 

If I were you I would take a MOOC or two (e.g., [Algorithms, Part I](#) , [Algorithms, Part II](#) , [Functional Programming Principles in Scala](#) ), a good book on data structures and algorithms, then just code as much as possible. You could implement some statistics or ML algorithms, for example; that would be good practice for you and useful to the community.

For a PhD program, however, I would also make sure I were familiar with the type of maths they use. If you want to see what it's like at the deep end, browse the papers at the [JMLR](#) . That will let you calibrate yourself in regards to theory; can you sort of follow the maths?

Oh, and you don't need a PhD to work at top companies, unless you want to join research departments like his. But then you'll spend more time doing development, and you'll need good coding skills...

[Answer](#)  by [laurik](#) 

Your time would probably be better spent on Kaggle than in a PhD program. When you

read the stories by winners ([Kaggle blog](#)) you'll see that it takes a large amount of practice and the winners are not just experts of one single method.

On the other hand, being active and having a plan in a PhD program can get you connections that you otherwise would probably not get.

I guess the real question is for you - what are the reasons for wanting a job at a top company?

[Answer](#) by [mrmeritology](#)

You already have a Masters in Statistics, which is great! In general, I'd suggest to people to take as much statistics as they can, especially Bayesian Data Analysis.

Depending on what you want to do with your PhD, you would benefit from foundational courses in the discipline(s) in your application area. You already have Economics but if you want to do Data Science on social behavior, then courses in Sociology would be valuable. If you want to work in fraud prevention, then a courses in banking and financial transactions would be good. If you want to work in information security, then taking a few security courses would be good.

There are people who argue that it's not valuable for Data Scientists to spend time on courses in sociology or other disciplines. But consider the recent case of the Google Flu Trends project. In [this article](#) their methods were strongly criticized for making avoidable mistakes. The critics call it "Big Data hubris".

There's another reason for building strength in social science disciplines: personal competitive advantage. With the rush of academic degree programs, certificate programs, and MOOCs, there is a mad rush of students into the Data Science field. Most will come out with capabilities for core Machine Learning methods and tools. PhD graduates will have more depth and more theoretical knowledge, but they are all competing for the same sorts of jobs, delivering the same sorts of value. With this flood of graduates, I expect that they won't be able to command premium salaries.

But if you can differentiate yourself with a combination of formal education and practical experience in a particular domain and application area, then you should be able to set yourself apart from the crowd.

(Context: I'm in a PhD program in Computational Social Science, which has a heavy focus on modeling, evolutionary computation, and social science disciplines, and less emphasis on ML and other empirical data analysis topics).

Tags: [education](#) ([Prev Q](#)) ([Next Q](#))

Q: What do you think of Data Science certifications?

Tags: [education](#) ([Prev Q](#))

I've now seen two data science certification programs - the [John Hopkins one available at Coursera](#) and the [Cloudera one](#).

I'm sure there are others out there.

The John Hopkins set of classes is focused on R as a toolset, but covers a range of topics:

- R Programming
- cleaning and obtaining data
- Data Analysis
- Reproducible Research
- Statistical Inference
- Regression Models
- Machine Learning
- Developing Data Products
- And what looks to be a Project based completion task similar to Cloudera's Data Science Challenge

The Cloudera program looks thin on the surface, but looks to answer the two important questions - "Do you know the tools", "Can you apply the tools in the real world". Their program consists of:

- Introduction to Data Science
- Data Science Essentials Exam
- Data Science Challenge (a real world data science project scenario)

I am not looking for a recommendation on a program or a quality comparison.

I am curious about other certifications out there, the topics they cover, and how seriously DS certifications are viewed at this point by the community.

EDIT: These are all great answers. I'm choosing the correct answer by votes.

Tags: [education](#) ([Prev Q](#))

User: [steve-kallestad](#) 

[Answer](#)  by [patlaf](#) 

I did the first 2 courses and I'm planning to do all the others too. If you don't know R, it's a really good program. There are assignments and quizzes every week. Many people find some courses very difficult. You are going to have hard time if you don't have any programming experience (even if they say it's not required).

Just remember.. it's not because you can drive a car that you are a F1 pilot ;)

[Answer](#)  by [neone4373](#) 

As a former analytics manager and a current lead data scientist, I am very leery of the need for data science certificates. The term data scientist is pretty vague and the field of data science is in its infancy. A certificate implies some sort of uniform standard which is just lacking in data science, it is still very much the wild west.

While a certificate is probably not going to hurt you, I think your time would be better spent developing the experience to know when to use a certain approach, and depth of

understanding to be able to explain that approach to a non-technical audience.

[Answer](#)  by [stanpol](#) 

The certification programs you mentioned are really entry level courses. Personally, I think these certificates show only person's persistence and they can be only useful to those who is applying for internships, not the real data science jobs.

Tags: [education](#) ([Prev Q](#))

Search

Questions

[Q: How can we effectively measure the impact of our data decisions](#)

Tags: [search](#) ([Prev Q](#))

Apologies if this is very broad question, what I would like to know is how effective is A/B testing (or other methods) of effectively measuring the effects of a design decision.

For instance we can analyse user interactions or click results, purchase/ browse decisions and then modify/tailor the results presented to the user.

We could then test the effectiveness of this design change by subjecting 10% of users to the alternative model randomly but then how objective is this?

How do we avoid influencing the user by the model change, for instance we could decided that search queries for 'David Beckham' are probably about football so search results become biased towards this but we could equally say that his lifestyle is just as relevant but this never makes it into the top 10 results that are returned.

I am curious how this is dealt with and how to measure this effectively.

My thoughts are that you could be in danger of pushing a model that you think is correct and the user obliges and this becomes a self-fulfilling prophecy.

I've read an article on this: <http://techcrunch.com/2014/06/29/ethics-in-a-data-driven-world/>  and also the book: <http://shop.oreilly.com/product/0636920028529.do>  which discussed this so it piqued my interest.

Tags: [search](#) ([Prev Q](#))

User: [edchum](#) 

[Answer](#)  by [christopher-louden](#) 

In A/B testing, bias is handled very well by ensuring visitors are randomly assigned to either version A or version B of the site. This creates independent samples drawn from the same population. Because the groups are independent and, on average, only differ in the version of the site seen, the test measures the effect of the design decision.

Slight aside: Now you might argue that the A group or B group may differ in some demographic. That commonly happens by random chance. To a certain degree this can be taken care of by covariate adjusted randomization. It can also be taken care of by adding covariates to the model that tests the effect of the design decision. It should be noted that there is still some discussion about the proper way to do this within the statistics community. Essentially A/B testing is an application of a [Randomized Control Trial](#)  to website design. Some people disagree with adding covariates to the test. Others, such as

Frank Harrel (see [Regression Modeling Strategies](#)) argue for the use of covariates in such models.

I would offer the following suggestions:

- Design the study in advance so as to take care of as much sources of bias and variation as possible.
 - Let the data speak for itself. As you get more data (like about searches for David Beckham), let it dominate your assumptions about how the data should be (as how the posterior dominates the prior in Bayesian analysis when the sample size becomes large).
 - Make sure your data matches the assumptions of the model.
-

Tags: [search](#) ([Prev Q](#))

Similarity

Questions

[Q: Similarity measure based on multiple classes from a hierarchical taxonomy?](#)

Tags: [similarity](#) ([Prev Q](#)) ([Next Q](#))

Could anyone recommend a good similarity measure for objects which have multiple classes, where each class is part of a hierarchy?

For example, let's say the classes look like:

[Skip code block](#)

```
1 Produce
  1.1 Eggs
    1.1.1 Duck eggs
    1.1.2 Chicken eggs
  1.2 Milk
    1.2.1 Cow milk
    1.2.2 Goat milk
2 Baked goods
  2.1 Cakes
    2.1.1 Cheesecake
    2.1.2 Chocolate
```

An object might be tagged with items from the above at any level, e.g.:

```
Omelette: eggs, milk (1.1, 1.2)
Duck egg omelette: duck eggs, milk (1.1.1, 1.2)
Goat milk chocolate cheesecake: goat milk, cheesecake, chocolate (1.2.2, 2.1.1, 2.1.2)
Beef: produce (1)
```

If the classes weren't part of a hierarchy, I'd probably look at cosine similarity (or equivalent) between classes assigned to an object, but I'd like to use the fact that different classes with the same parents also have some similarity value (e.g. in the example above, beef has some small similarity to omelette, since they both have items from the class '1 produce').

If it helps, the hierarchy has ~200k classes, with a maximum depth of 5.

Tags: [similarity](#) ([Prev Q](#)) ([Next Q](#))

User: [dave-challis](#) 

[Answer](#)  by [aleksandr-blekh](#) 

While I don't have enough expertise to advise you on selection of the best *similarity measure*, I've seen a number of them in various papers. The following **collection** of research papers hopefully will be useful to you in determining the optimal measure for your research. Please note that I intentionally included papers, using both *frequentist* and *Bayesian* approaches to hierarchical classification, including class information, for the sake of more *comprehensive* coverage.

Frequentist approach:

- [Semantic similarity based on corpus statistics and lexical taxonomy](#)
- [Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example](#) (also see [additional results and data](#))
- [Learning hierarchical similarity metrics](#)
- [A new similarity measure for taxonomy based on edge counting](#)
- [Hierarchical classification of real life documents](#)
- [Hierarchical document classification using automatically generated hierarchy](#)
- [Split-Order distance for clustering and classification hierarchies](#)
- [A hierarchical k-NN classifier for textual data](#)

Bayesian approach:

- [Improving classification when a class hierarchy is available using a hierarchy-based prior](#)
- [Bayesian aggregation for hierarchical genre classification](#)
- [Hierarchical classification for multiple, distributed web databases](#)

Tags: [similarity](#) ([Prev Q](#)) ([Next Q](#))

Q: Applications and differences for Jaccard similarity and Cosine Similarity

Tags: [similarity](#) ([Prev Q](#))

Jaccard similarity and cosine similarity are two very common measurements while comparing item similarities. However, I am not very clear in what situation which one should be preferable than another.

Can somebody help clarify the differences of these two measurements (the difference in concept or principle, not the definition or computation) and their preferable applications?

Tags: [similarity](#) ([Prev Q](#))

User: [shihpeng](#)

Answer by [saq7](#)

Jaccard Similarity is given by $s_{ij} = \frac{p}{p+q+r}$

where,

p = # of attributes positive for both objects

q = # of attributes 1 for i and 0 for j

r = # of attributes 0 for i and 1 for j

Whereas, cosine similarity = $\frac{A \cdot B}{\|A\| \|B\|}$ where A and B are object vectors.

Simply put, in cosine similarity, the number of common attributes is divided by the total number of possible attributes. Whereas in Jaccard Similarity, the number of common attributes is divided by the number of attributes that exists in at least one of the two objects.

And there are many other measures of similarity, each with its own eccentricities. When deciding which one to use, try to think of a few representative cases and work out which index would give the most usable results to achieve your objective.

For example, if you have two objects both with 10 attributes, out of a possible 100 attributes. Further they have all 10 attributes in common. In this case, the Jaccard index will be 1 and the cosine index will be 0.001. Now consider another scenario where object A has 10 attributes, and object B has 50 attributes, but B has all 10 attributes that A has. Here, Jaccard index will be 0.2 and cosine index will still be 0.001. So the key question is if that extra bit of information reflected, in this case, in the Jaccard index useful or does it hurt or does it not matter. Your choice will depend on which is the best solution for your problem.

The Cosine index could be used to identify plagiarism, but will not be a good index to identify mirror sites on the internet. Whereas the Jaccard index, will be a good index to identify mirror sites, but not so great at catching copy pasta plagiarism (within a larger document).

Of course, these are toy examples to illustrate a point. When applying these indices, you must think about your problem thoroughly and figure out how to define similarity. Once you have a definition in mind, you can go about shopping for an index.

[Answer](#)  by [vikramnath-venkatasubramani](#) 

Jaccard similarity is used for two types of binary cases:

1. Symmetric, where 1 and 0 has equal importance (gender, marital status,etc)
2. Asymmetric, where 1 and 0 have different levels of importance (testing positive for a disease)

Cosine similarity is usually used in the context of text mining for comparing documents or emails. If the cosine similarity between two document term vectors is higher, then both the documents have more number of words in common

Another difference is $1 - \text{Jaccard Coefficient}$ can be used as a dissimilarity or distance measure, whereas the cosine similarity has no such constructs. A similar thing is the Tonimoto distance, which is used in taxonomy.

Tags: [similarity](#) ([Prev Q](#))

Social Network Analysis

Questions

[Q: How should ethics be applied in data science](#)

Tags: [social-network-analysis](#) ([Prev Q](#))

There was a recent furore with [facebook experimenting on their users to see if they could alter user's emotions](#)  and now [okcupid](#) .

Whilst I am not a professional data scientist I read about [data science ethics](#)  from [Cathy O'Neill's book 'Doing Data Science'](#)  and would like to know if this is something that professionals are taught at academic level (I would expect so) or something that is ignored or is lightly applied in the professional world. Particularly for those who ended up doing data science *accidentally*.

Whilst the linked article touched on data integrity, the book also discussed the moral ethics behind understanding the impact of the data models that are created and the impact of those models which can have adverse effects when used inappropriately (sometimes unwittingly) or when the models are inaccurate, again producing adverse results.

The article discusses a code of practice and mentions the [Data Science Association's Code of conduct](#) , is this something that is in use? Rule 7 is of particular interest (quoted from their website):

- (a) A person who consults with a data scientist about the possibility of forming a client-data scientist relationship with respect to a matter is a prospective client.
- (b) Even when no client-data scientist relationship ensues, a data scientist who has learned information from a prospective client shall not use or reveal that information.
- (c) A data scientist subject to paragraph (b) shall not provide professional data science services for a client with interests materially adverse to those of a prospective client in the same or a substantially related industry if the data scientist received information from the prospective client that could be significantly harmful to that person in the matter

Is this something that is practiced professionally? Many users blindly accept that we get some free service (mail, social network, image hosting, blog platform etc..) and agree to an EULA in order to have ads pushed at us.

Finally how is this regulated, I often read about users being up in arms when the terms of a service change but it seems that it requires some liberty organisation, class action or a [senator](#)  to react to such things before something happens.

By the way I am not making any judgements here or saying that all data scientists behave like this, I'm interested in what is taught academically and practiced professionally.

Tags: [social-network-analysis](#) ([Prev Q](#))

User: [edchum](#) 

[Answer](#)  by [climbs_lika_spyder](#) 

I think ethics in Data Science is important. There is a fundamental difference in using user data to better their experience and show relevant ads and using user data to trick people into clicking on ads for the sake of monetary profit. Personally I like ads that give me relevant information like deals on things I would buy anyway. However, showing me weight loss ads because I got dumped is creepy and unethical. As my friend Peter always says, “don’t be creepy with data”.

Tags: [social-network-analysis](#) ([Prev Q](#))

Time Series

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Overview

Time series are data observed over time (either in continuous time or at discrete time periods).

Time series analysis includes trend identification, temporal pattern recognition, spectral analysis, and forecasting future values based on the past.

The salient characteristic of methods of time series analysis (as opposed to more general methods to analyze relationships among data) is accounting for the possibility of *serial correlation* (also known as temporal correlation) among the data. Positive serial correlation means successive observations in time tend to be close to one another, whereas negative serial correlation means successive observations tend to oscillate between extremes. Time series analysis also differs from analyses of more general stochastic processes by focusing on the *inherent direction* of time, creating a potential asymmetry between past and future.

Questions

[Q: How to merge monthly, daily and weekly data?](#)

Tags: [time-series](#) ([Prev Q](#)) ([Next Q](#))

Google Trends returns weekly data so I have to find a way to merge them with my daily/monthly data.

What I have done so far is to break each serie into daily data, for exemple:
from:

2013-03-03 - 2013-03-09 37

to:

2013-03-03 37 2013-03-04 37 2013-03-05 37 2013-03-06 37 2013-03-07 37 2013-03-08 37 2013-03-09 37

But this is adding a lot of complexity to my problem. I was trying to predict google searchs from the last 6 months values, or 6 values in monthly data. Daily data would imply a work on 180 past values. (I have 10 years of data so 120 points in monthly data / 500+ in weekly data/ 3500+ in daily data)

The other approach would be to “merge” daily data in weekly/monthly data. But some questions arise from this process. Some data can be averaged because their sum represent something. Rainfall for example, the amount of rain in a given week will be the sum of the amounts for each days composing the weeks.

In my case I am dealing with prices, financial rates and other things. For the prices it is common in my field to take volume exchanged into account, so the weekly data would be a weighted average. For financial rates it is a bit more complex a some formulas are involved to build weekly rates from daily rates. For the other things i don't know the underlying properties. I think those properties are important to avoid meaningless indicators (an average of fiancial rates would be a non-sense for example).

So three questions:

For known and unknown properties, how should I proceed to go from daily to weekly/monthly data ?

I feel like breaking weekly/monthly data into daily data like i've done is somewhat wrong because I am introducing quantities that have no sense in real life. So almost the same question:

For known and unknown properties, how should I proceed to go from weekly/monthly to daily data ?

Last but not least: **when given two time series with different time steps, what is better: Using the Lowest or the biggest time step ?** I think this is a compromise between the number of data and the complexity of the model but I can't see any strong argument to choose between those options.

Edit: if you know a tool (in R Python even Excel) to do it easily it would be very appreciated.

Tags: [time-series](#) ([Prev Q](#)) ([Next Q](#))

User: [were_cat](#) 

[Answer](#)  by [gchaks](#) 

when given two time series with different time steps, what is better: Using the Lowest or the biggest time step ?

For your timeseries analysis you should do both: get to the highest granularity possible with the daily dataset, and also repeat the analysis with the monthly dataset. With the monthly dataset you have 120 data points, which is sufficient to get a timeseries model even with seasonality in your data.

For known and unknown properties, how should I proceed to go from daily to weekly/monthly data ?

To obtain say weekly or monthly data from daily data, you can use smoothing functions. For financial data, you can use moving average or exponential smoothing, but if those do not work for your data, then you can use the spline smoothing function “smooth.spline” in R: <https://stat.ethz.ch/R-manual/R-patched/library/stats/html/smooth.spline.html> 

The model returned will have less noise than the original daily dataset, and you can get values for the desired time points. Finally, these data points can be used in your timeseries analysis.

For known and unknown properties, how should I proceed to go from weekly/monthly to daily data ?

To obtain daily data when you have monthly or weekly data, you can use interpolation. First, you should find an equation to describe the data. In order to do this you should plot the data (e.g. price over time). When factors are known to you, this equation should be influenced by those factors. When factors are unknown, you can use a best fit equation. The simplest would be a linear function or piecewise linear function, but for financial data this won't work well. In that case, you should consider piecewise cubic spline interpolation. This link goes into more detail on possible interpolation functions: <http://people.math.gatech.edu/~meyer/MA6635/chap2.pdf> .

In R, there is a method for doing interpolation of timeseries data. Here you would create a vector with say weekly values and NAs in the gaps for the daily values, and then use the “interpNA” function to get the interpolated values for the NAs. However, this function uses the “approx” function to get the interpolated values, which applies either a linear or constant interpolation. To perform cubic spline interpolation in R, you should use the “splinefun” function instead.

Something to be aware of is that timeseries models typically do some sort of averaging to forecast future values whether you are looking at exponential smoothing or Auto-

Regressive Integrated Moving Average (ARIMA) methods amongst others. So a timeseries model to forecast daily values may not be the best choice, but the weekly or monthly models may be better.

[Answer](#) by [aleksandr-blekh](#)

I'm not an expert in this area, but I believe that your question is concerned with *time series aggregation and disaggregation*. If that is the case, here are some hopefully relevant resources, which might be helpful in solving your problem (first five items are main, but representative, and last two are supplementary):

- [Temporal Aggregation and Economic Time Series](#)
 - [Temporal Disaggregation of Time Series](#) (IMHO, an excellent overview paper)
 - [CRAN Task View: Time Series Analysis](#) (R-focused)
 - [Introduction to R's Time Series Facilities](#)
 - [Working with Financial Time Series Data in R](#)
 - [Notes on chapters contents for the book “Time Series Analysis and Forecasting”](#)
 - [Discussion on Cross Validated](#) on daily to monthly data conversion (Python-focused)
-

[Answer](#) by [charlie-greenbacker](#)

This won't be a very satisfying answer, but here's my take...

For known and unknown properties, how should I proceed to go from daily to weekly/monthly data ?

For known and unknown properties, how should I proceed to go from weekly/monthly to daily data ?

Same answer for both: you can't do this for unknown properties, and for known properties it will depend on how the values were computed.

As you alluded to:

(an average of financial rates would be a non-sense for example)

There is no single transformation that will be appropriate in all cases, whether the properties/values are known or unknown. Even with known properties, you'll likely need a unique transformation for each type: mean, median, mode, min, max, boolean, etc.

when given two time series with different time steps, what is better: Using the Lowest or the biggest time step ?

Whenever possible, try to preserve the full granularity of the smallest possible step. Assuming you know how to transform the values, you can always roll-up the steps (e.g., day to month, month to year)... but you won't necessarily be able to reconstruct smaller steps from larger ones following a lossy conversion.

Tags: [time-series](#) ([Prev Q](#)) ([Next Q](#))

[Q: Finding unpredictability or uncertainty in a time series](#)

Tags: [time-series](#) ([Prev Q](#))

I am interested in finding a statistic that tracks the unpredictability of a time series. For simplicity sake, assume that each value in the time series is either 1 or 0. So for example, the following two time series are entirely predictable TS1: 1 1 1 1 1 1 1 1 TS2: 0 1 0 1 0 1 0 1 0 1

However, the following time series is not that predictable: TS3: 1 1 0 1 0 0 1 0 0 0 0 0 1 1 0 1 1 1

I am looking for a statistic that given a time series, would return a number between 0 and 1 with 0 indicating that the series is completely predictable and 1 indicating the series is completely unpredictable.

I looked at some entropy measures like Kolmogorov Complexity and Shannon entropy, but neither seem to fit my requirement. In Kolmogorov complexity, the statistic value changes depending on the length of the time series (as in “1 0 1 0 1” and “1 0 1 0” have different complexities, so its not possible to compare predictability of two time series with differing number of observations). In Shannon entropy, the order of observations didn’t seem to matter.

Any pointers on what would be a good statistic for my requirement?

Tags: [time-series](#) ([Prev Q](#))

User: [rajesh](#) 

[Answer](#)  by [aleksandr-blekh](#) 

Since you have looked at Kolmogorov-Smirnov and Shannon entropy measures, I would like to suggest some other hopefully relevant options. First of all, you could take a look at the so-called [approximate entropy ApEn](#) . Other potential statistics include *block entropy*, *T-complexity (T-entropy)* as well as *Tsallis entropy*:

<http://members.noa.gr/anastasi/papers/B29.pdf> 

In addition to the above-mentioned potential measures, I would like to suggest to have a look at available statistics in *Bayesian inference-based* model of *stochastic volatility* in time series, implemented in R package *stochvol*: <http://cran.r-project.org/web/packages/stochvol>  (see detailed [vignette](#) http://simpsonm.public.iastate.edu/BlogPosts/btcvol/KastnerFruhwirthSchnatterASISstoch. A **comprehensive example** of using stochastic volatility model approach and *stochvol* package can be found in the excellent blog post “[Exactly how volatile is bitcoin?](#)”  by Matt Simpson.

Tags: [time-series](#) ([Prev Q](#))

Scalability

Questions

[Q: Data Science Tools Using Scala](#)

Tags: [scalability](#) ([Prev Q](#))

I know that Spark is fully integrated with Scala. It's use case is specifically for large data sets. Which other tools have good Scala support? Is Scala best suited for larger data sets? Or is it also suited for smaller data sets?

Tags: [scalability](#) ([Prev Q](#))

User: [sheldonkreger](#) 

[Answer](#)  by [thegrimmscientist](#) 

Re: size of data

The short answer

Scala works for both small and large data, but its creation and development is motivated by needing something scalable. [Scala is an acronym for “Scalable Language”](#) .

The long answer

Scala is a [functional programming language](#)  that runs on the [jvm](#) languages written on the jvm  and plenty of other [functional programming languages](#) 

[This talk](#)  give a good overview of the motivation behind Scala.

Re: other tools that have good Scala support:

As you mentioned, Spark (distributable batch processing better at iterative algorithms than its counterpart) is a big one. With Spark comes its libraries [Mllib](#)  for machine learning and [GraphX](#)  for graphs. As mentioned by Erik Allik and Tris Nefzger, [Akka](#)  and [Factorie](#)  exist. There is also [Play](#) .

Generally, I can't tell if there is a specific use case you're digging for (if so, make that a part of your question), or just want a survey of big data tools and happen to know Scala a bit and want to start there.

[Answer](#) by [brandon-loudermilk](#)

ScalaNLP is a suite of machine learning and numerical computing libraries with support for common natural language processing tasks. <http://www.scala-nlp.org/>

[Answer](#) by [tris-nefzger](#)

From listening to presentations by Martin Odersky, the creator of Scala, it is especially well suited for building highly scalable systems by leveraging functional programming constructs in conjunction with object orientation and flexible syntax. It is also useful for development of small systems and rapid prototyping because it takes less lines of code than some other languages and it has an interactive mode for fast feedback. One notable Scala framework is Akka which uses the actor model of concurrent computation. Many of Odersky's presentations are on YouTube and there is a list of tools implemented with Scala on wiki.scala-lang.org.

An implicit point is that tools and frameworks written in Scala inherently have Scala integration and usually a Scala API. Then other APIs may be added to support other languages beginning with Java since Scala is already integrated and in fact critically depends on Java. If a tool or framework is not written in Scala, it is unlikely that it offers any support for Scala. That is why in answer to your question I have pointed towards tools and frameworks written in Scala and Spark is one example. However, Scala currently has a minor share of the market but its adoption rate is growing and the high growth rate of Spark will enhance that. The reason I use Scala is because Spark's API for Scala is richer than the Java and Python APIs.

The main reasons I prefer Scala generally is because it is much more expressive than Java because it allows and facilitates the use of functions as objects and values while retaining object oriented modularity, which enables development of complex and correct programs with far less code than Java which I had preferred because of widespread use, clarity and excellent documentation.

Tags: [scalability](#) ([Prev Q](#))

Beginner

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Some good beginner resources for getting started in Data Science are:

1. [Data Science wiki of Quora](#) 

Some good beginner resources for getting started in Machine Learning are:

1. [Professor Andrew Ng's course on Machine Learning](#) 
-

Questions

[Q: How to self-learn data science?](#)

Tags: [beginner](#) ([Prev Q](#))

I am a self-taught web developer and am interested in teaching myself data science, but I'm unsure of how to begin. In particular, I'm wondering:

1. What fields are there within data science? (e.g., Artificial Intelligence, machine learning, data analysis, etc.)
2. Are there online classes people can recommend?
3. Are there projects available out there that I can practice on (e.g., open datasets).
4. Are there certifications I can apply for or complete?

Tags: [beginner](#) ([Prev Q](#))

User: [martin](#) 

[Answer](#)  by [kyle.](#) 

Welcome to the site, Martin! That's a pretty broad question, so you're probably going to get a variety of answers. Here's my take.

1. *Data Science* is an interdisciplinary field generally thought to combine classical statistics, machine learning, and computer science (again, this depends on who you ask, but other might include business intelligence here, and possible information visualization or knowledge discovery as well; for example, [the wikipedia article on data science](#)). A good data scientist is also skilled at picking up on the domain-specific characteristics of the domain in which they working, as well. For example, a data scientist working on analytics for hospital records is much more effective if they have a background in Biomedical Informatics.
2. There are many options here, depending on the type of analytics you're interested in. [Andrew Ng's coursera course is the first resource mentioned by most](#), and rightly so. If you're interested in machine learning, that's a great starting place. If you want an in-depth exploration of the mathematics involved, [Tibshirani's The Elements of Statistical Learning](#) is excellent, but fairly advanced text. There are many online courses available on coursera in addition to Ng's, but you should select them with a mind for the type of analytics you want to focus on, and/or the domain in which you plan on working.
3. [Kaggle](#). Start with kaggle, if you want to dive in on some real-world analytics problems. Depending on your level of expertise, it might be good to start of simpler, though. [Project Euler](#) is a great resource for one-off practice problems that I still use as warm-up work.
4. Again, this probably depends on the domain you wish to work in. However, I know Coursera offers a data science certificate, if you complete a series of data science-related courses. This is probably a good place to start.

Good luck! If you have any other specific questions, feel free to ask me in the comments, and I'll do my best to help!

Tags: [beginner](#) ([Prev Q](#))

Data Cleaning

[Skip to questions,](#)

Wiki by user [dawny33](#) 

Data cleaning is a preliminary step to statistical analysis in which the data-set is edited to correct errors and to put it into a form suitable for processing by statistical software. Exploratory data analysis techniques are often used to identify problems.

Questions

Q: How can I transform names in a confidential data set to make it anonymous, but preserve some of the characteristics of the names? 

Tags: [data-cleaning](#) ([Prev Q](#)) ([Next Q](#))

Motivation

I work with datasets that contain personally identifiable information (PII) and sometimes need to share part of a dataset with third parties, in a way that doesn't expose PII and subject my employer to liability. Our usual approach here is to withhold data entirely, or in some cases to reduce its resolution; e.g., replacing an exact street address with the corresponding county or census tract.

This means that certain types of analysis and processing must be done in-house, even when a third party has resources and expertise more suited to the task. Since the source data is not disclosed, the way we go about this analysis and processing lacks transparency. As a result, any third party's ability to perform QA/QC, adjust parameters or make refinements may be very limited.

Anonymizing Confidential Data

One task involves identifying individuals by their names, in user-submitted data, while taking into account errors and inconsistencies. A private individual might be recorded in one place as "Dave" and in another as "David," commercial entities can have many different abbreviations, and there are always some typos. I've developed scripts based on a number of criteria that determine when two records with non-identical names represent the same individual, and assign them a common ID.

At this point we can make the dataset anonymous by withholding the names and replacing them with this personal ID number. But this means the recipient has almost no information about e.g. the strength of the match. We would prefer to be able to pass along as much information as possible without divulging identity.

What Doesn't Work

For instance, it would be great to be able to encrypt strings while preserving edit distance. This way, third parties could do some of their own QA/QC, or choose to do further processing on their own, without ever accessing (or being able to potentially reverse-engineer) PII. Perhaps we match strings in-house with edit distance ≤ 2 , and the recipient wants to look at the implications of tightening that tolerance to edit distance ≤ 1 .

But the only method I am familiar with that does this is [ROT13](#)  (more generally, any [shift cipher](#) ), which hardly even counts as encryption; it's like writing the names upside down and saying, "Promise you won't flip the paper over?"

Another **bad** solution would be to abbreviate everything. “Ellen Roberts” becomes “ER” and so forth. This is a poor solution because in some cases the initials, in association with public data, will reveal a person’s identity, and in other cases it’s too ambiguous; “Benjamin Othello Ames” and “Bank of America” will have the same initials, but their names are otherwise dissimilar. So it doesn’t do either of the things we want.

An inelegant alternative is to introduce additional fields to track certain attributes of the name, e.g.:

[Skip code block](#)

Row	ID	Name	WordChars	Origin
1	17	"AMELIA BEDELIA"	(6, 7)	Eng
2	18	"CHRISTOPH BAUER"	(9, 5)	Ger
3	18	"C J BAUER"	(1, 1, 5)	Ger
4	19	"FRANZ HELLER"	(5, 6)	Ger

I call this “inelegant” because it requires anticipating which qualities might be interesting and it’s relatively coarse. If the names are removed, there’s not much you can reasonably conclude about the strength of the match between rows 2 & 3, or about the distance between rows 2 & 4 (i.e., how close they are to matching).

Conclusion

The goal is to transform strings in such a way that as many useful qualities of the original string are preserved as possible while obscuring the original string. Decryption should be impossible, or so impractical as to be effectively impossible, no matter the size of the data set. In particular, a method that preserves the edit distance between arbitrary strings would be very useful.

I’ve found a couple papers that might be relevant, but they’re a bit over my head:

- [Privacy Preserving String Comparisons Based on Levenshtein Distance](#) 
- [An Empirical Comparison of Approaches to Approximate String Matching in Private Record Linkage](#) 

Tags: [data-cleaning](#) ([Prev Q](#)) ([Next Q](#))

User: [air](#) 

[Answer](#)  by [air](#) 

One of the references I mentioned in the OP led me to a potential solution that seems quite powerful, described in “Privacy-preserving record linkage using Bloom filters” ([doi:10.1186/1472-6947-9-41](#) ):

A new protocol for privacy-preserving record linkage with encrypted identifiers allowing for errors in identifiers has been developed. The protocol is based on Bloom filters on q-grams of identifiers.

The article goes into detail about the method, which I will summarize here to the best of my ability.

A Bloom filter is a fixed-length series of bits storing the results of a fixed set of independent hash functions, each computed on the same input value. The output of each hash function should be an index value from among the possible indexes in the filter; i.e., if you have a 0-indexed series of 10 bits, hash functions should return (or be mapped to) values from 0 to 9.

The filter starts with each bit set to 0. After hashing the input value with each function from the set of hash functions, each bit corresponding to an index value returned by any hash function is set to 1. If the same index is returned by more than one hash function, the bit at that index is only set once. You could consider the Bloom filter to be a superposition of the set of hashes onto the fixed range of bits.

The protocol described in the above-linked article divides strings into n-grams, which are in this case sets of characters. As an example, "hello" might yield the following set of 2-grams:

```
["_h", "he", "el", "l1", "lo", "o_"]
```

Padding the front and back with spaces seems to be generally optional when constructing n-grams; the examples given in the paper that proposes this method use such padding.

Each n-gram can be hashed to produce a Bloom filter, and this set of Bloom filters can be superimposed on itself (bitwise OR operation) to produce the Bloom filter for the string.

If the filter contains many more bits than there are hash functions or n-grams, arbitrary strings are relatively unlikely to produce exactly the same filter. However, the more n-grams two strings have in common, the more bits their filters will ultimately share. You can then compare any two filters A, B by means of their Dice coefficient:

$$D_{A, B} = 2h / (a + b)$$

Where h is the number of bits that are set to 1 in both filters, a is the number of bits set to 1 in *only* filter A, and b is the number of bits set to 1 in *only* filter B. If the strings are exactly the same, the Dice coefficient will be 1; the more they differ, the closer the coefficient will be to 0.

Because the hash functions are mapping an indeterminate number of unique inputs to a small number of possible bit indexes, different inputs may produce the same filter, so the coefficient indicates only a *probability* that the strings are the same or similar. The number of different hash functions and the number of bits in the filter are important parameters for determining the likelihood of false positives - pairs of inputs that are much less similar than the Dice coefficient produced by this method predicts.

I found [this tutorial](#) to be very helpful for understanding the Bloom filter.

There is some flexibility in the implementation of this method; see also [this 2010 paper](#) (also linked at the end of the question) for some indications of how performant it is in relation to other methods, and with various parameters.

[Answer](#) by [emre](#)

If feasible I would link related records (e.g., Dave, David, etc.) and replace them with a sequence number (1,2,3, etc.) or a [salted hash of the string](#) that is used to represent all related records (e.g., David instead of Dave).

I assume that third parties need not have any idea what the real name is, otherwise you might as well give it to them.

edit: You need to define and justify what kind of operations the third party needs to be able to do. For example, what is wrong with using initials followed by a number (e.g., BOA-1, BOA-2, etc.) to disambiguate Bank of America from Benjamin Othello Ames? If that's too revealing, you could bin some of the letters or names; e.g., [A-E] -> 1, [F-J] -> 2, etc. so BOA would become 1OA, or ["Bank", "Barry", "Bruce", etc.] -> 1 so Bank of America is again 1OA.

For more information see [k-anonymity](#).

[Answer](#) by [dave-challis](#)

One option (depending on your dataset size) is to just provide edit distances (or other measures of similarity you're using) as an additional dataset.

E.g.:

1. Generate a set of unique names in the dataset
2. For each name, calculate edit distance to each other name
3. Generate an ID or irreversible hash for each name
4. Replace names in the original dataset with this ID
5. Provide matrix of edit distances between ID numbers as new dataset

Though there's still a lot that could be done to deanonymise the data from these even.

E.g. if "Tim" is known to be the most popular name for a boy, frequency counting of IDs that closely match the known percentage of Tims across the population might give that away. From there you could then look for names with an edit distance of 1, and conclude that those IDs might refer to "Tom" or "Jim" (when combined with other info).

Tags: [data-cleaning](#) ([Prev Q](#)) ([Next Q](#))

[Q: Dealing with training set of questionable quality](#)

Tags: [data-cleaning](#) ([Prev Q](#))

Most of the material I have read in the past usually assumes that the training set is flawless. However that doesn't seem to be the case here with what I am given. The data that is meant to send into the training set is often questionable (I don't even know how to start separating good from bad ones). I have spent much more time trying to pre-process the files in order to increase the validity of the data, than actually building the prototype.

I have only dumped a part of the data that is claimed to be better. There are still data from

other sources to complement the missing part in the first set of training data. Given the amount of data, it is practically impossible for me to spend too much time on them.

So the question is, how do people deal with really horribly, inconsistent data in real life (assuming I am working with addresses, where people often write the wrong postcode due to bad city planning or lack of clear instruction, misspell the name of neighbourhood, plus data got “autocorrected” by excel such that “1-2” becomes “1-Feb”, etc. etc.)?

Specifically, how do I ensure the quality of the train classifier given the training set (and cross validation set) has questionable quality?

Tags: [data-cleaning](#) ([Prev Q](#))

User: [jeffrey04](#) 

[Answer](#)  by [dawny33](#) 

Welcome to the real world of data science. Here, the data sets are not as clean as you thought while doing those courses/tutorials online. Those are super polished and refined. But, the real world data is not so.

The step where one does the cleaning and scrubbing is called the data pre-processing step. So, some nice data cleaning techniques, in addition to @jknappen’s excellent answer are:

1. **Elimination of zero variance columns/predictors:** These columns are not important, and they cause the model and the fit to crash and leak errors. So, eliminating them would make complete sense.
2. **Correlated Predictors:** Reducing the level of correlation between the predictors would be a very nice step in the pre-processing process.
3. **Scaling:** You must be knowing why scaling is important during pre-process.
4. **Predictor Transformations**

A [nice reference from Kaggle forums](#)  where the pre-processing and cleaning of data sets is discussed.

[Answer](#)  by [jknappen](#) 

You can use techniques of semi-supervised learning where you have a small clean training set and some dirty data. You than extend your data base by judging how good the other data are and incorporate the “best” data points into your training set.

Tags: [data-cleaning](#) ([Prev Q](#))

Aws

Questions

[Q: Instances vs. cores when using EC2](#)

Tags: [aws](#) ([Prev Q](#))

Working on what could often be called “medium data” projects, I’ve been able to parallelize my code (mostly for modeling and prediction in Python) on a single system across anywhere from 4 to 32 cores. Now I’m looking at scaling up to clusters on EC2 (probably with StarCluster/IPython, but open to other suggestions as well), and have been puzzled by how to reconcile distributing work across cores on an instance vs. instances on a cluster.

Is it even practical to parallelize across instances as well as across cores on each instance? If so, can anyone give a quick rundown of the pros + cons of running many instances with few cores each vs. a few instances with many cores? Is there a rule of thumb for choosing the right ratio of instances to cores per instance?

Bandwidth and RAM are non-trivial concerns in my projects, but it’s easy to spot when those are the bottlenecks and readjust. It’s much harder, I’d imagine, to benchmark the right mix of cores to instances without repeated testing, and my projects vary too much for any single test to apply to all circumstances. Thanks in advance, and if I’ve just failed to google this one properly, feel free to point me to the right answer somewhere else!

Tags: [aws](#) ([Prev Q](#))

User: [therriault](#) 

[Answer](#)  by [alex-i](#) 

When using IPython, you very nearly don’t have to worry about it (at the expense of some loss of efficiency/greater communication overhead). The parallel IPython plugin in StarCluster will by default start one engine per physical core on each node (I believe this is configurable but not sure where). You just run whatever you want across all engines by using the DirectView api (map_sync, apply_sync, ...) or the %px magic commands. If you are already using IPython in parallel on one machine, using it on a cluster is no different.

Addressing some of your specific questions:

“how to reconcile distributing work across cores on an instance vs. instances on a cluster”
- You get one engine per core (at least); work is automatically distributed across all cores and across all instances.

“Is it even practical to parallelize across instances as well as across cores on each instance?” - Yes :) If the code you are running is embarrassingly parallel (exact same algo on multiple data sets) then you can mostly ignore where a particular engine is running. If

the core requires a lot of communication between engines, then of course you need to structure it so that engines primarily communicate with other engines on the same physical machine; but that kind of problem is not ideally suited for IPython, I think.

"If so, can anyone give a quick rundown of the pros + cons of running many instances with few cores each vs. a few instances with many cores? Is there a rule of thumb for choosing the right ratio of instances to cores per instance?" - Use the largest c3 instances for compute-bound, and the smallest for memory-bandwidth-bound problems (or small enough that the problem almost stops being memory-bandwidth-bound); for message-passing-bound problems, also use the largest instances but try to partition the problem so that each partition runs on one physical machine and most message passing is within the same partition. Problems which run significantly slower on N quadruple c3 than on $2N$ double c3 are rare (an artificial example may be running multiple simple filters on a large number of images, where you go through all images for each filter rather than all filters for the same image). Using largest instances is a good rule of thumb.

[Answer](#)  by [sean-owen](#) 

A general rule of thumb is to not distribute until you have to. It's usually more efficient to have N servers of a certain capacity than $2N$ servers of half that capacity. More of the data access will be local, and therefore fast in memory versus slow across the network.

At a certain point, scaling up one machine becomes uneconomical because the cost of additional resource scales more than linearly. However this point is still amazingly high.

On Amazon in particular though, the economics of each instance type can vary a lot if you are using spot market instances. The default pricing more or less means that the same amount of resource costs about the same regardless of the instance type, that can vary a lot; large instances can be cheaper than small ones, or N small instances can be much cheaper than one large machine with equivalent resources.

One massive consideration here is that the computation paradigm can change quite a lot when you move from one machine to multiple machines. The tradeoffs that the communication overhead induce may force you to, for example, adopt a data-parallel paradigm to scale. That means a different choice of tools and algorithm. For example, SGD looks quite different in-memory and in Python than on MapReduce. So you would have to consider this before parallelizing.

You may choose to distribute work across a cluster, even if a single node and non-distributed paradigms work for you, for reliability. If a single node fails, you lose all of the computation; a distributed computation can potentially recover and complete just the part of the computation that was lost.

[Answer](#)  by [damienfrancois](#) 

All things considered equal (cost, CPU perf, etc.) you could choose the smallest instance that can hold all of my dataset in memory and scale out. That way

- you make sure not to induce unnecessary latencies due to network communications, and

- you tend to maximize the overall available memory bandwidth for your processes.

Assuming you are running some sort of [cross-validation scheme](#) to optimize some [meta parameter](#) of your model, assign each core a value to test and choose as many instances as needed to cover all the parameter space in as few rounds as you see fit.

If your data does not fit in the memory of one system, of course you'll need to distribute across instances. Then it is a matter of balancing memory latency (better with many instances) with network latency (better with fewer instances) but given the nature of EC2 I'd bet you'll often prefer to work with few fat instances.

Tags: [aws](#) ([Prev Q](#))

Graphs

Questions

[Q: Facebook's Huge Database](#)

Tags: [graphs](#) ([Prev Q](#))

I assume that each person on Facebook is represented as a node (of a Graph) in Facebook, and relationship/friendship between each person(node) is represented as an edge between the involved nodes.

Given that there are millions of people on Facebook, how is the Graph stored?

Tags: [graphs](#) ([Prev Q](#))

User: [rohitjv](#) 

[Answer](#)  by [indico](#) 

Strange as it sounds, graphs and graph databases are typically implemented as [linked lists](#). As alluded to [here](#), even the most popular/performant graph database out there (neo4j), is secretly using something akin to a doubly-linked list.

Representing a graph this way has a number of significant benefits, but also a few drawbacks. Firstly, representing a graph this way means that you can do edge-based insertions in near-constant time. Secondly, this means that traversing the graph can happen extremely quickly, if we're only looking to either step up or down a linked list.

The biggest drawback of this though comes from something sometimes called The Justin Bieber Effect, where nodes with a large number of connections tend to be extremely slow to evaluate. Imagine having to traverse a million semi-redundant links every time someone linked to Justin Bieber.

I know that the awesome folks over at Neo4j are working on the second problem, but I'm not sure how they're going about it, or how much success they've had.

[Answer](#)  by [laurik](#) 

Having worked with Facebook data a bit (harvested from Facebook users) we stored it just as a pair of values: USER_ID, FRIEND_USER_ID.

But I guess your question is a bit deeper? You can store it in different ways, depending on your research question. One interesting option is triads for example -

http://mypersonality.org/wiki/doku.php?id=list_of_variables_available#triads 

Tags: [graphs](#) ([Prev Q](#))

Cross Validation

Questions

[Q: Cross-validation: K-fold vs Repeated random sub-sampling](#)

Tags: [cross-validation](#) ([Prev Q](#))

I wonder which type of model cross-validation to choose for classification problem: K-fold or random sub-sampling (bootstrap sampling)?

My best guess is to use 2/3 of the data set (which is ~1000 items) for training and 1/3 for validation.

In this case K-fold gives only three iterations(folds), which is not enough to see stable average error.

On the other hand I don't like random sub-sampling feature: that some items won't be ever selected for training/validation, and some will be used more than once.

Classification algorithms used: random forest & logistic regression.

Tags: [cross-validation](#) ([Prev Q](#))

User: [ihars](#) 

[Answer](#)  by [bogatron](#) 

If you have an adequate number of samples and want to use all the data, then k-fold cross-validation is the way to go. Having ~1,500 seems like a lot but whether it is adequate for k-fold cross-validation also depends on the dimensionality of the data (number of attributes and number of attribute values). For example, if each observation has 100 attributes, then 1,500 observations is low.

Another potential downside to k-fold cross-validation is the possibility of a single, extreme outlier skewing the results. For example, if you have one extreme outlier that can heavily bias your classifier, then in a 10-fold cross-validation, 9 of the 10 partitions will be affected (though for random forests, I don't think you would have that problem).

Random subsampling (e.g., bootstrap sampling) is preferable when you are either undersampled or when you have the situation above, where you don't want each observation to appear in k-1 folds.

[Answer](#)  by [rhand](#) 

I guess you say that you want to use 3-fold cross-validation because you know something about your data (that using k=10 would cause overfitting? I'm curious to your reasoning). I am not sure that you know this, if not then you can simply use a larger k.

If you still think that you cannot use standard k-fold cross-validation, then you could

modify the algorithm a bit: say that you split the data into 30 folds and each time use 20 for training and 10 for evaluation (and then shift up one fold and use the first and the last 9 as evaluation and the rest as training). This means that you're able to use all your data.

When I use k-fold cross-validation I usually run the process multiple times with a different randomisation to make sure that I have sufficient data, if you don't you will see different performances depending on the randomisation. In such cases I would suggest sampling. The trick then is to do it often enough.

Tags: [cross-validation](#) ([Prev Q](#))

Apache Spark

[Skip to questions,](#)

Wiki by user [dawny33](#) 

From <http://spark.apache.org/>:

Apache Spark is an open source cluster computing system that aims to make data analytics fast — both fast to run and fast to write.

To run programs faster, Spark offers a general execution model that can optimize arbitrary operator graphs, and supports in-memory computing, which lets it query data faster than disk-based engines like [hadoop](#).

Spark is not tied to the two-stage [mapreduce](#) paradigm, and promises performance up to 100 times faster than Hadoop MapReduce

Spark provides primitives for in-memory cluster computing that allows user programs to load data into a cluster's memory and query it repeatedly, making it well suited to machine learning algorithms. To make programming faster, Spark provides clean, concise APIs in [scala](#), [java](#) and [python](#). You can also use Spark interactively from the [scala](#) and [python](#) shells to rapidly query big datasets.

Spark runs on [yarn](#), [mesos](#), standalone, or in the cloud. It can access diverse data sources including [hdfs](#), [cassandra](#), [hbase](#), and [amazon-s3](#).

Recommended reference sources:

[Spark Documentation](#) 

[Learning Spark](#) - Lightning-Fast Big Data Analysis

[AMP Camp 5](#) (Berkeley, CA, November 20-21, 2014)

[AMP Camp 4](#) (Strata Santa Clara, Feb 2014) — focus on BlinkDB, MLlib, GraphX, Tachyon

[AMP Camp 3](#) (Berkeley, CA, Aug 2013)

[AMP Camp 2](#) (Strata Santa Clara, Feb 2013)

[AMP Camp 1](#) (Berkeley, CA, Aug 2012)

Questions

[Q: Local Development for Apache Spark](#)

Tags: [apache-spark](#) ([Prev Q](#)) ([Next Q](#))

I'm wondering how other developers are setting up their local environments for working on Spark projects. Do you configure a 'local' cluster using a tool like Vagrant? Or, is it most common to SSH into a cloud environment, such as a cluster on AWS? Perhaps there are many tasks where a single-node cluster is adequate, and can be run locally more easily.

Tags: [apache-spark](#) ([Prev Q](#)) ([Next Q](#))

User: [sheldonkreger](#) 

[Answer](#)  by [j.a.gartner](#) 

Spark is intended to be pointed at large distributed data sets, so as you suggest, the most typical use cases will involve connecting to some sort of Cloud system like AWS.

In fact, if the data set you aim to analyze can fit on your local system, you'll usually find that you can analyze it just as simply using pure python. If you're trying to leverage a series of local VMs, you're going to run out of memory pretty quickly and jobs will either fail or grind to a halt.

With that said, a local instance of spark is very useful for the purpose of development.

One way that I've found that works is if I have a directory in HDFS with many files, I'll pull over a single file, develop locally, then port my spark script to my cloud system for execution. If you're using AWS, this is really helpful for avoiding big fees while you're developing.

Tags: [apache-spark](#) ([Prev Q](#)) ([Next Q](#))

[Q: Scan-based operations Apache Spark](#)

Tags: [apache-spark](#) ([Prev Q](#))

Looking at the first paper on RDDs/Apache Spark, I found a statement saying that "RDDs degrade gracefully when there is not enough memory to store them, as long as they are only being used in scan-based operations"

What are scan-based operations in the context of RDDs and which of the [Transformations in Spark](#)  are scan-based operations

Tags: [apache-spark](#) ([Prev Q](#))

User: [mb_ce](#) 

[Answer](#)  by [eliasah](#) 

Scan based operations are basically all the operations that require evaluating the predicate on an RDD.

In other terms, each time you create an RDD or a DataFrame in which you need to compute a *predicate* like performing a filter, map on a case class, per example, or even explain method will be considered as a scan based operation.

To be more clear, let's review the definition of a predicate.

A predicate or a functional predicate is a logical symbol that may be applied to an object term to produce another object term.

Functional predicates are also sometimes called *mappings*, but that term can have other meanings as well.

Example :

```
// scan based transformation
rdd.filter(!_.contains("#")) // here the predicate is !_.contains("#")

// another scan based transformation
rdd.filter(myfunc) // myfunc is a boolean function

// a third also trivial scan based transformation followed by a non scan based one.
rdd.map(myfunc2)
    .reduce(myfunc3)
```

If you want to understand how spark internals work, I suggest that you watch the [presentation](#) made by Databricks about the topics

Tags: [apache-spark](#) ([Prev Q](#))

Categorical Data

[Skip to questions,](#)

Wiki by user [dawny33](#) 

For analysis, categorical values are considered as abstract entities without any mathematical structure such as an order or a topology, regardless of how they are coded and stored. For more, see [Wikipedia](#) 

Questions

[Q: How can I dynamically distinguish between categorical data and numerical data?](#)

Tags: [categorical-data](#) ([Prev Q](#))

I know someone who is working on a project that involves ingesting files of data without regard to the columns or data types. The task is to take a file with any number of columns and various data types and output summary statistics on the numerical data.

However, he is unsure of how to go about dynamically assigning data types for certain number-based data. For example:

```
CITY
Albuquerque
Boston
Chicago
```

This is obviously not numerical data and will be stored as text. However,

```
ZIP
80221
60653
25525
```

are not clearly marked as categorical. His software would assign the ZIP code as numerical and output summary statistics for it, which does not make sense for that sort of data.

A couple ideas we had were:

1. If a column is all integers, label it as categorical. This clearly wouldn't work, but it was an idea.
2. If a column has fewer than n unique values and is numeric, label it categorical. This might be closer, but there could still be issues with numerical data falling through.
3. Maintain a list of common numeric data that should actually be categorical and compare the column headers to this list for matches. For example, anything with "ZIP" in it would be categorical.

My gut tells me that there is no way to accurately assign numeric data as categorical or numerical, but was hoping for a suggestion. Any insight you have is greatly appreciated.

Tags: [categorical-data](#) ([Prev Q](#))

User: [poisson-fish](#) 

[Answer](#)  by [jncraton](#) 

I'm not aware of a foolproof way to do this. Here's one idea off the top of my head:

1. Treat values as categorical by default.

2. Check for various attributes of the data that would imply it is actually continuous. Weight these attributes based on how likely they are to correlate with continuous data. Here are some possible examples:
 - Values are integers: +.7
 - Values are floats: +.8
 - Values are normally distributed: +.3
 - Values contain a relatively small number of unique values: +.3
 - Values aren't all the same number of characters: +.1
 - Values don't contain leading zeros: +.1
 3. Treat any columns that sum to greater than 1 as being numerical. Adjust the factors and weights based on testing against different data sets to suit your needs. You could even build and train a separate machine learning algorithm just to do this.
-

[Answer](#) by [bernardo-aflalo](#)

If you have, for example, number of children of a family (which could range, for example, between 0 and 5), is it a categorical or numerical variable? Actually it depends on your problem and how you intend to solve it. In this sense, you can do the following:

- Compute the number of unique values of that column
- Divide this number by the total number of rows
- If this ratio is below some threshold (for example, 20%), you consider it categorical.

In case of discrete values, one additional test could be: use a regression model to estimate some of the parameters and check if the estimated values are contained in the original set of values. If this is not true, you are probably dealing with categorical data (as it is the case of ZIP).

It worked relatively well for me in the past...

Tags: [categorical-data](#) ([Prev Q](#))

Hierarchical Data Format

Questions

[Q: Hierarchical Data Format. What are the advantages compared to alternative formats?](#)

Tags: [hierarchical-data-format](#) ([Prev Q](#))

What are the main benefits from storing data in HDF? And what are the main data science tasks where HDF is really suitable and useful?

Tags: [hierarchical-data-format](#) ([Prev Q](#))

User: [ihars](#) 

[Answer](#)  by [alex-i](#) 

Perhaps a good way to paraphrase the question is, what are the advantages compared to alternative formats?

The main alternatives are, I think: a database, text files, or another packed/binary format.

The database options to consider are probably a columnar store or NoSQL, or for small self-contained datasets SQLite. The main advantage of the database is the ability to work with data much larger than memory, to have random or indexed access, and to add/append/modify data quickly. The main *dis*advantage is that it is much slower than HDF, for problems in which the entire dataset needs to be read in and processed. Another disadvantage is that, with the exception of embedded-style databases like SQLite, a database is a system (requiring administration, setup, maintenance, etc) rather than a simple self-contained data store.

The text file format options are XML/JSON/CSV. They are cross-platform/language/toolkit, and are a good archival format due to the ability to be self-describing (or obvious :). If uncompressed, they are huge (10x-100x HDF), but if compressed, they can be fairly space-efficient (compressed XML is about the same as HDF). The main disadvantage here is again speed: parsing text is much, much slower than HDF.

The other binary formats (npy/npz numpy files, blz blaze files, protocol buffers, Avro, ...) have very similar properties to HDF, except they are less widely supported (may be limited to just one platform: numpy) and may have specific other limitations. They typically do not offer a compelling advantage.

HDF is a good complement to databases, it may make sense to run a query to produce a roughly memory-sized dataset and then cache it in HDF if the same data would be used more than once. If you have a dataset which is fixed, and usually processed as a whole, storing it as a collection of appropriately sized HDF files is not a bad option. If you have a

dataset which is updated often, staging some of it as HDF files periodically might still be helpful.

To summarize, HDF is a good format for data which is read (or written) typically as a whole; it is the lingua franca or common/preferred interchange format for many applications due to wide support and compatibility, decent as an archival format, and very fast.

P.S. To give this some practical context, my most recent experience comparing HDF to alternatives, a certain small (much less than memory-sized) dataset took 2 seconds to read as HDF (and most of this is probably overhead from Pandas); ~1 minute to read from JSON; and 1 *hour* to write to database. Certainly the database write could be sped up, but you'd better have a good DBA! This is how it works out of the box.

[Answer](#)  by [steve-kallestad](#) 

One benefit is wide support - C, Java, Perl, Python, and R all have HDF5 bindings.

Another benefit is speed. I haven't ever seen it benchmarked, but HDF is supposed to be faster than SQL databases.

I understand that it is very good when used with both large sets of scientific data and time series data - network monitoring, usage tracking, etc.

I don't believe there is a size limitation for HDF files (although OS limits would still apply).

Tags: [hierarchical-data-format](#) ([Prev Q](#))

Xgboost

Questions

[Q: xgboost: give more importance to recent samples](#)

Tags: [xgboost](#) ([Prev Q](#))

Is there a way to add more importance to points which are more recent when analyzing data with xgboost?

Tags: [xgboost](#) ([Prev Q](#))

User: [kilojoules](#) 

Answer  by [mark-heiler](#) 

You could try building multiple xgboost models, with some of them being limited to more recent data, then weighting those results together. Another idea would be to make a customized evaluation metric that penalizes recent points more heavily which would give them more importance.

Tags: [xgboost](#) ([Prev Q](#))

Sequence

Questions

[Q: Classification of DNA Sequences](#)

Tags: [sequence](#)

I have a [database](#)  of 3190 instances of DNA consisting of 60 sequential DNA nucleotide positions classified according to 3 types: EI, IE, Other.

I want to formulate a supervised classifier.

My present approach is to formulate a 2nd order Markov Transition Matrix for each instance and apply the resulting data to a Neural Network.

How best to approach this classification problem, given that the Sequence of the data should be relevant? Is there a better approach than the one I came up with?

Tags: [sequence](#)

User: [akellyirl](#) 

[Answer](#)  by [nitesh](#) 

One way would be to create 20 features (each feature representing a codon). In this way, you would have a dataset with 3190 instances and 20 categorical features. There is no need to treat the sequence as a Markov chain.

Once the dataset has been featurized as suggested above, any supervised classifier can work well. I would suggest using a gradient boosting machine as it might be better suited to handle categorical features.

Tags: [sequence](#)

Copyright

This book was compiled from [datascience](#), using the latest data dump available at [archive.org](#). A selection of the best questions about Data Science, as voted by the site's users have been selected for inclusion in this book.

I am not affiliated or endorsed by StackExchange Inc, although I do have a [Stack Exchange](#) user profile there under the username [George Duckett](#).

All questions and content contained within this book are licensed under [cc-wiki](#) with [attribution required](#) as per Stack Exchange Inc's requirements.

If you have or know of copyrighted content included in this book and want it to be removed please let me know at georgeduckett@hotmail.com.

The cover image was altered from it's original source, found below.

[YouTube](#) / [CC BY 3.0](#)