Which machine learning algori should I use?

By Hui Li on Subconscious Musings | April 12, 2017

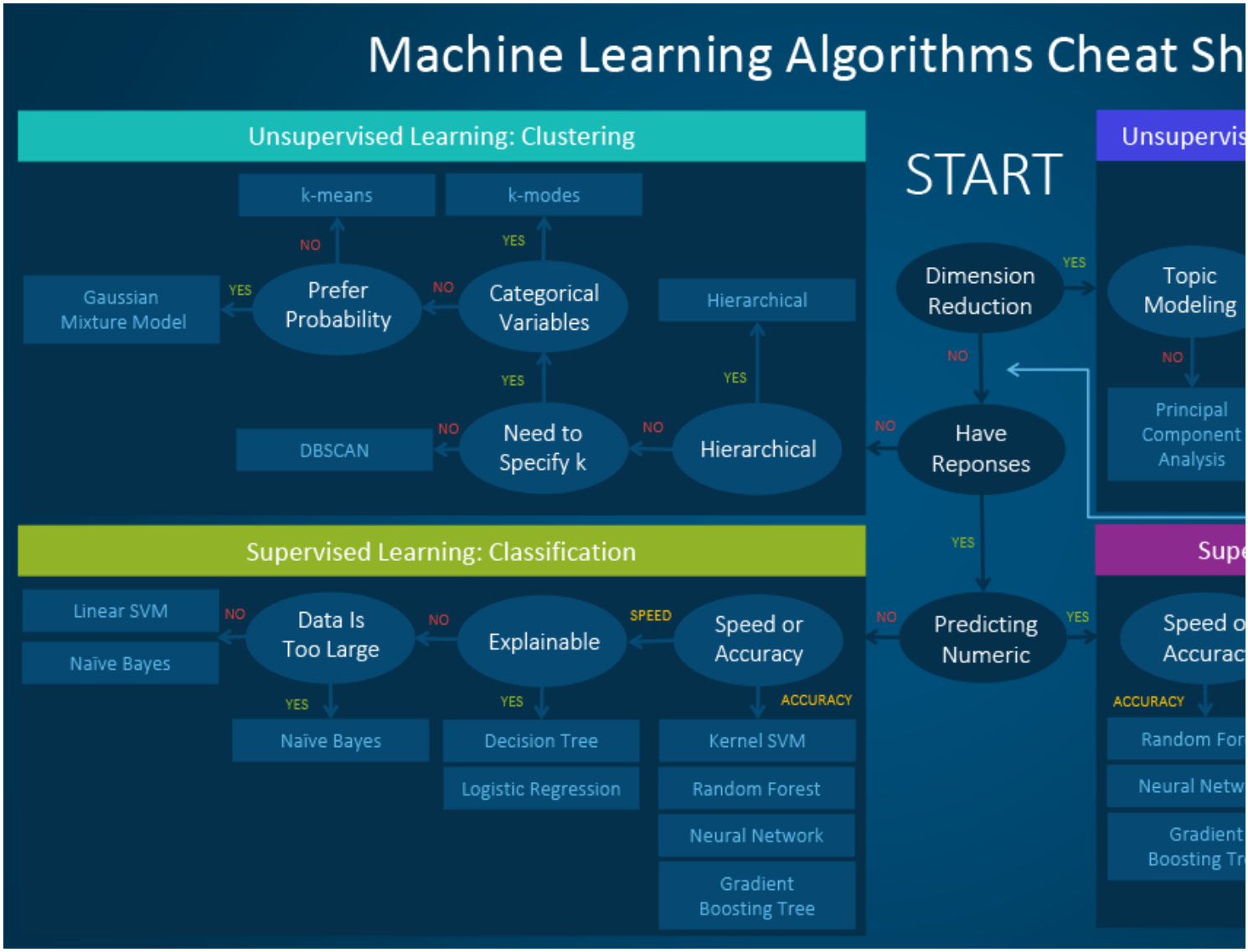
This resource is designed primarily for beginner to intermediate data scientists or analysts who are applying machine learning algorithms to address the problems of their interest.

A typical question asked by a beginner, when facing a wide variety of machine learning algorithms, The answer to the question varies depending on many factors, including:

- The size, quality, and nature of data.
- The available computational time.
- The urgency of the task.
- What you want to do with the data.

Even an experienced data scientist cannot tell which algorithm will perform the best before trying d advocating a one and done approach, but we do hope to provide some guidance on which algorith clear factors.

The machine learning algorithm cheat sheet



The **machine learning algorithm cheat sheet** helps you to choose from a variety of machine lear appropriate algorithm for your specific problems. This article walks you through the process of how

Since the cheat sheet is designed for beginner data scientists and analysts, we will make some sin about the algorithms.

The algorithms recommended here result from compiled feedback and tips from several data scien experts and developers. There are several issues on which we have not reached an agreement an highlight the commonality and reconcile the difference.

Additional algorithms will be added in later as our library grows to encompass a more complete set

How to use the cheat sheet

Read the path and algorithm labels on the chart as "If *<path label>* then use *<algorithm>*." For exa

- If you want to perform dimension reduction then use principal component analysis.
- If you need a numeric prediction quickly, use decision trees or logistic regression.
- If you need a hierarchical result, use hierarchical clustering.

Sometimes more than one branch will apply, and other times none of them will be a perfect match. paths are intended to be rule-of-thumb recommendations, so some of the recommendations are nc talked with said that the only sure way to find the very best algorithm is to try all of them.

Types of machine learning algorithms

This section provides an overview of the most popular types of machine learning. If you're familiar move on to discussing specific algorithms, you can skip this section and go to "When to use specifi

Supervised learning

Supervised learning algorithms make predictions based on a set of examples. For example, histori the future prices. With supervised learning, you have an input variable that consists of labeled train variable. You use an algorithm to analyze the training data to learn the function that maps the input function maps new, unknown examples by generalizing from the training data to anticipate results i

- **Classification:** When the data are being used to predict a categorical variable, supervised le This is the case when assigning a label or indicator, either dog or cat to an image. When ther binary classification. When there are more than two categories, the problems are called multi
- **Regression:** When predicting continuous values, the problems become a regression problen
- **Forecasting:** This is the process of making predictions about the future based on the past a commonly used to analyze trends. A common example might be estimation of the next year s current year and previous years.

Semi-supervised learning

The challenge with supervised learning is that labeling data can be expensive and time consuming unlabeled examples to enhance supervised learning. Because the machine is not fully supervised i is semi-supervised. With semi-supervised learning, you use unlabeled examples with a small amou learning accuracy.

Unsupervised learning

When performing unsupervised learning, the machine is presented with totally unlabeled data. It is patterns that underlies the data, such as a clustering structure, a low-dimensional manifold, or a sp

- **Clustering:** Grouping a set of data examples so that examples in one group (or one cluster) some criteria) than those in other groups. This is often used to segment the whole dataset in performed in each group to help users to find intrinsic patterns.
- **Dimension reduction:** Reducing the number of variables under consideration. In many appl high dimensional features and some features are redundant or irrelevant to the task. Reducin the true, latent relationship.

Reinforcement learning

Reinforcement learning analyzes and optimizes the behavior of an agent based on the feedback fr different scenarios to discover which actions yield the greatest reward, rather than being told which and delayed reward distinguishes reinforcement learning from other techniques.

Considerations when choosing an algorithm

When choosing an algorithm, always take these aspects into account: accuracy, training time and e accuracy first, while beginners tend to focus on algorithms they know best.

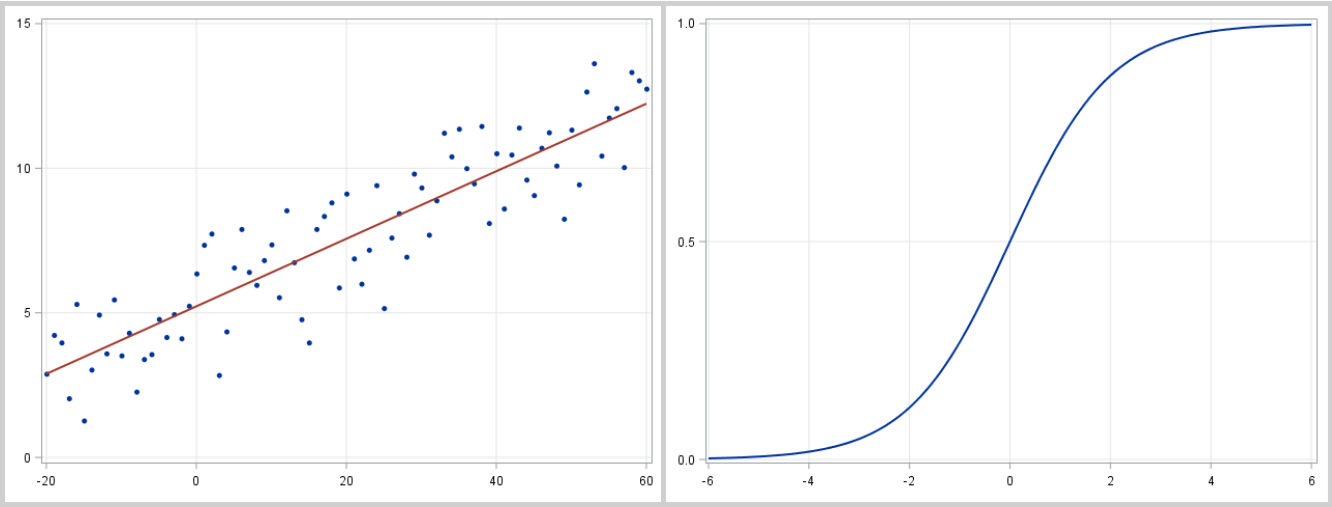
When presented with a dataset, the first thing to consider is how to obtain results, no matter what t Beginners tend to choose algorithms that are easy to implement and can obtain results quickly. Thi the first step in the process. Once you obtain some results and become familiar with the data, you sophisticated algorithms to strengthen your understanding of the data, hence further improving the

Even in this stage, the best algorithms might not be the methods that have achieved the highest re
usually requires careful tuning and extensive training to obtain its best achievable performance.

When to use specific algorithms

Looking more closely at individual algorithms can help you understand what they provide and how
provide more details and give additional tips for when to use specific algorithms, in alignment with

Linear regression and Logistic regression

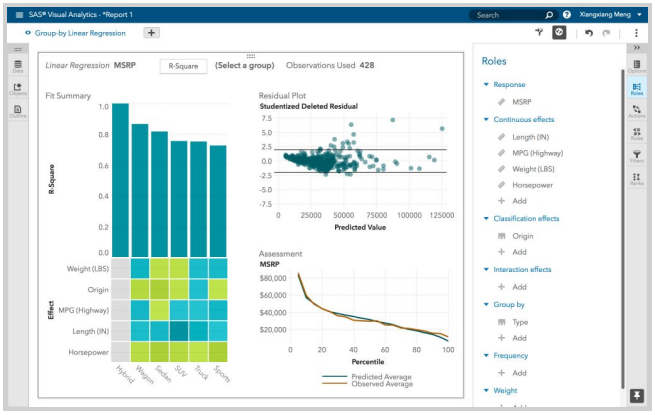


Linear regression *Logistic regression*

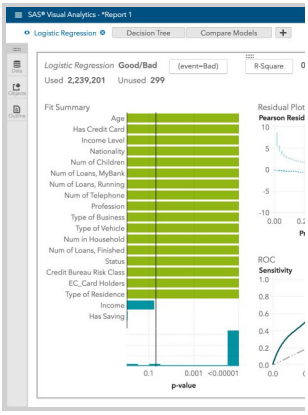
Linear regression is an approach for modeling the relationship between a continuous dependent va
predictors X. The relationship between y and X can be linearly modeled as $y = \beta^T X + \epsilon$ Given th
the parameter vector β can be learnt.

If the dependent variable is not continuous but categorical, linear regression can be transformed to
link function. Logistic regression is a simple, fast yet powerful classification algorithm. Here we disc
dependent variable y only takes binary values $\{y_i \in (-1, 1)\}_{i=1}^{N}$ (it which can be easily extended
problems).

In logistic regression we use a different hypothesis class to try to predict the probability that a giver
versus the probability that it belongs to the "-1" class. Specifically, we will try to learn a function of t
and $p(y_i = -1|x_i) = 1 - \sigma(\beta^T x_i)$. Here $\sigma(x) = \frac{1}{1+exp(-x)}$ is a sigmoid function. Given the trainir
parameter vector β can be learnt by maximizing the log-likelihood of β given the data set.



Group By Linear Regression



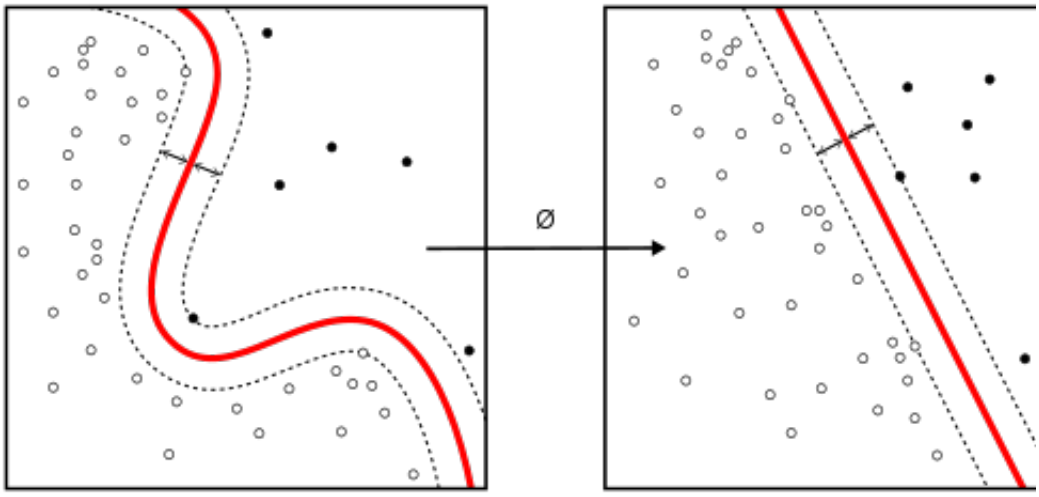
Logistic Regression i

Linear SVM and kernel SVM

Kernel tricks are used to map a non-linearly separable functions into a higher dimension linearly se machine (SVM) training algorithm finds the classifier represented by the normal vector w and bias hyperplane (boundary) separates different classes by as wide a margin as possible. The problem c constrained optimization problem:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & ||w|| \\ \text{subject to} \quad & y_i(w^T X_i - b) \geq 1, \ \ i = 1, \ldots, n. \end{aligned}$$

A support vector machine (SVM) training algorithm finds the classifier represented by the normal ve This hyperplane (boundary) separates different classes by as wide a margin as possible. The prob constrained optimization problem:

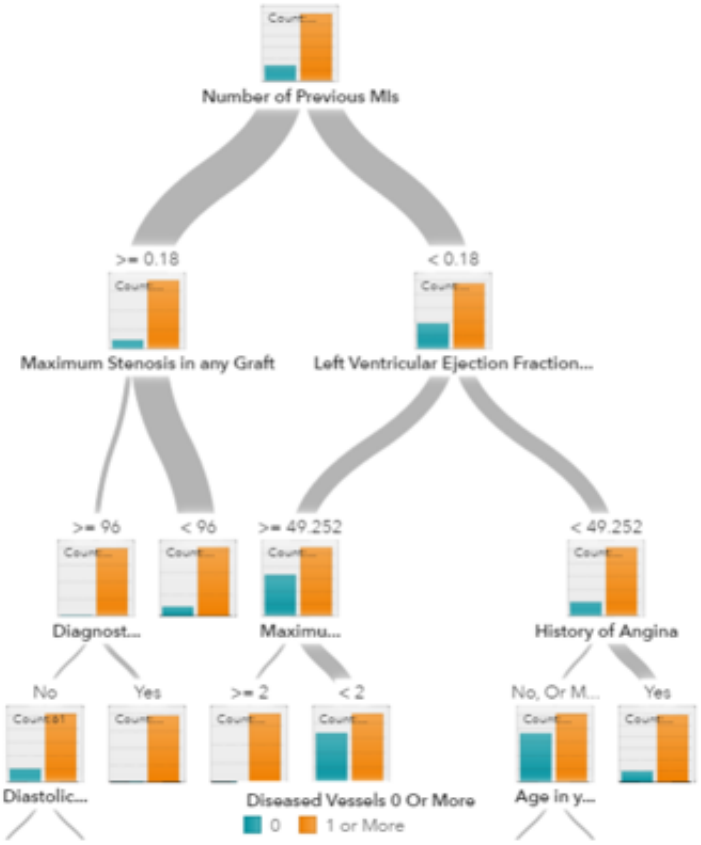


Kernel tricks are used to map a non-linearly separable functions into a hig dimension linearly separable function.

When the classes are not linearly separable, a kernel trick can be used to map a non-linearly sepa dimension linearly separable space.

When most dependent variables are numeric, logistic regression and SVM should be the first try fo easy to implement, their parameters easy to tune, and the performances are also pretty good. So t beginners.

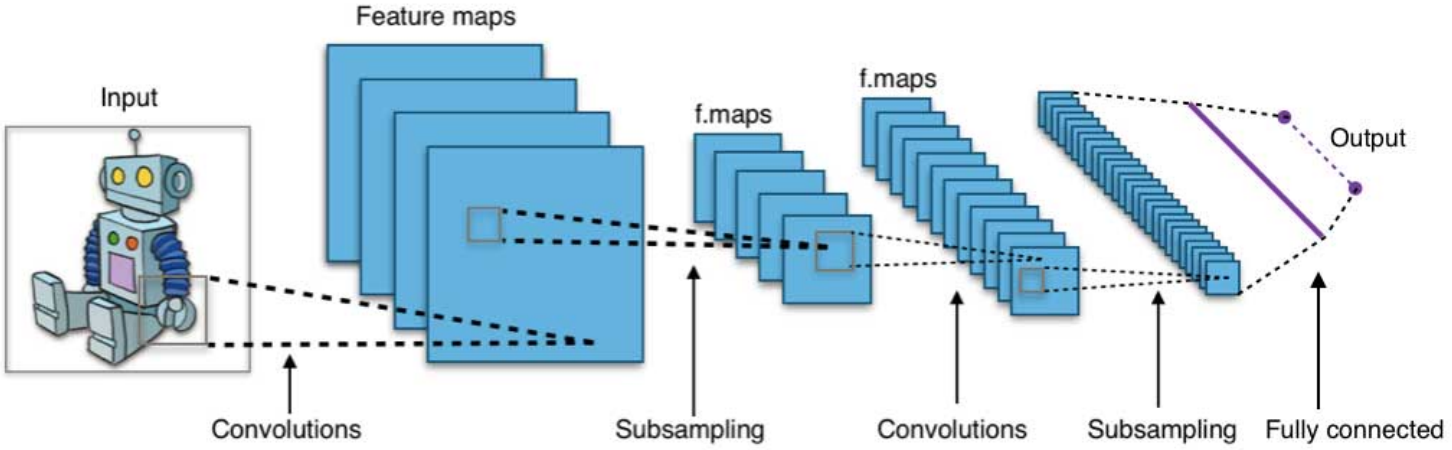
Trees and ensemble trees



A decision tree for prediction model

Decision trees, random forest and gradient boosting are all algorithms based on decision trees. Th trees, but they all do the same thing – subdivide the feature space into regions with mostly the sam to understand and implement. However, they tend to over fit data when we exhaust the branches a Random Forrest and gradient boosting are two popular ways to use tree algorithms to achieve goo the over-fitting problem.

Neural networks and deep learning

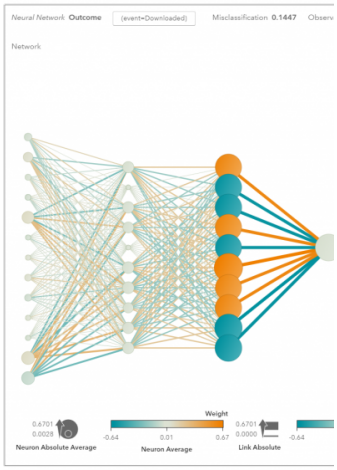


A convolution neural network architecture (image source: wikipedia creative commons)

Neural networks flourished in the mid-1980s due to their parallel and distributed processing ability. impeded by the ineffectiveness of the back-propagation training algorithm that is widely used to op networks. Support vector machines (SVM) and other simpler models, which can be easily trained b problems, gradually replaced neural networks in machine learning.

In recent years, new and improved training techniques such as unsupervised pre-training and laye a resurgence of interest in neural networks. Increasingly powerful computational capabilities, such (GPU) and massively parallel processing (MPP), have also spurred the revived adoption of neural in neural networks has given rise to the invention of models with thousands of layers.

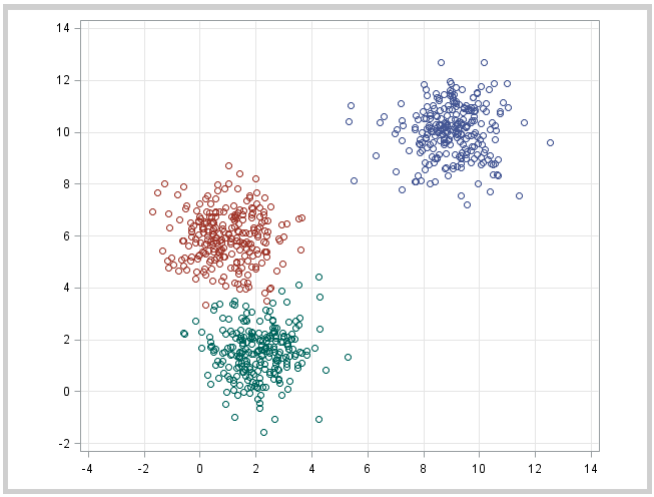
In other words, shallow neural networks have evolved into deep learning neural networks. Deep neural networks have been very successful for supervised learning. When used for speech and image recognition, deep learning performs as well as, or even better than, humans. Applied to unsupervised learning tasks, such as feature extraction, deep learning also extracts features from raw images or speech with much less human intervention.



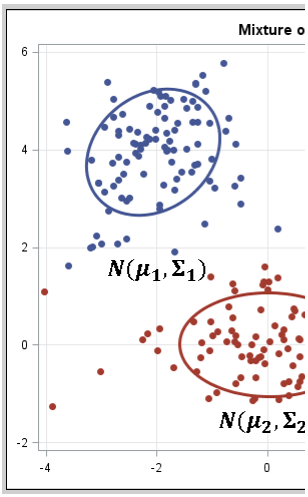
A neural netwo

A neural network consists of three parts: input layer, hidden layers and output layer. The training samples define the input and output layers. When the output layer is a categorical variable, then the neural network is a way to address classification problems. When the output layer is a continuous variable, then the network can be used to do regression. When the output layer is the s network can be used to extract intrinsic features. The number of hidden layers defines the model c

k-means/k-modes, GMM (Gaussian mixture model) clustering



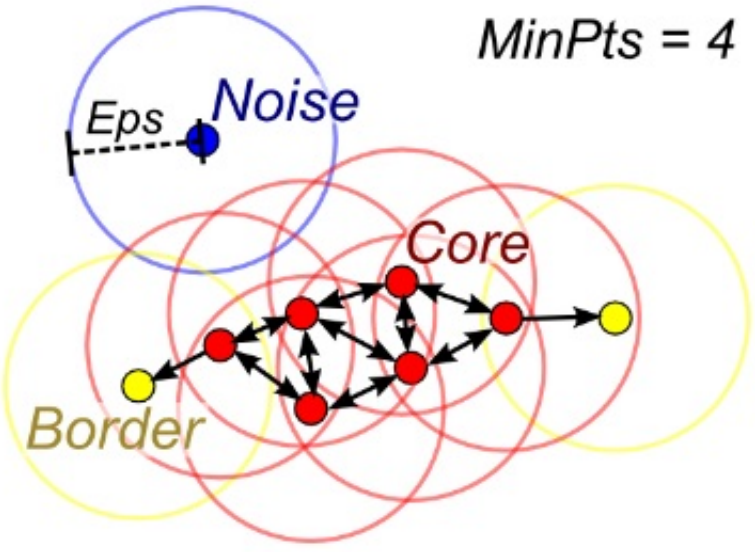
K Means Clustering



Gaussian M

Kmeans/k-modes, GMM clustering aims to partition n observations into k clusters. K-means define to be and only to be associated to one cluster. GMM, however define a soft assignment for each sa probability to be associated with each cluster. Both algorithms are simple and fast enough for clust k is given.

DBSCAN

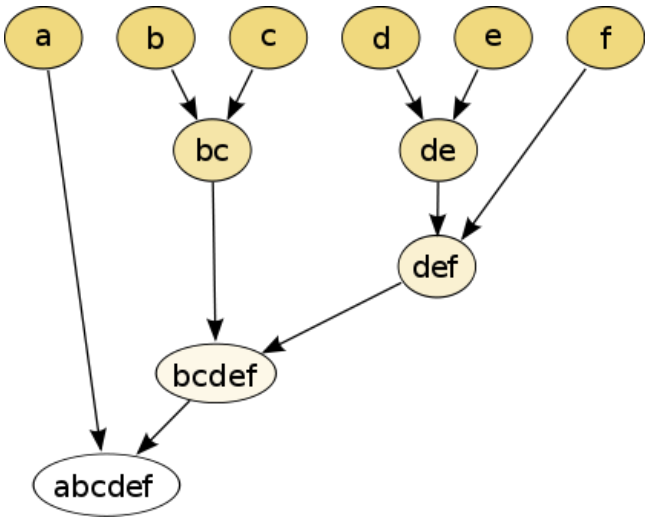


A DBSCAN illustration (image source: Wikipedia)

When the number of clusters k is not given, DBSCAN (density-based spatial clustering) can be use

density diffusion.

Hierarchical clustering



Hierarchical partitions can be visualized using a tree structure (a dendrogram). It does not need the and the partitions can be viewed at different levels of granularities (i.e., can refine/coarsen clusters

PCA, SVD and LDA

We generally do not want to feed a large number of features directly into a machine learning algori irrelevant or the “intrinsic” dimensionality may be smaller than the number of features. Principal cor value decomposition (SVD), and latent Dirichlet allocation (*LDA*) all can be used to perform dimens

PCA is an unsupervised clustering method which maps the original data space into a lower dimens much information as possible. The PCA basically finds a subspace that most preserves the data va by the dominant eigenvectors of the data’s covariance matrix.

The SVD is related to PCA in the sense that SVD of the centered data matrix (features versus sam singular vectors that define the same subspace as found by PCA. However, SVD is a more versatil things that PCA may not do. For example, the SVD of a user-versus-movie matrix is able to extract profiles which can be used in a recommendation system. In addition, SVD is also widely used as a latent semantic analysis, in natural language processing (NLP).

A related technique in NLP is latent Dirichlet allocation (LDA). LDA is probabilistic topic model and topics in a similar way as a Gaussian mixture model (GMM) decomposes continuous data into Gau the GMM, an LDA models discrete data (words in documents) and it constrains that the topics are Dirichlet distribution.

Conclusions

This is the work flow which is easy to follow. The takeaway messages when trying to solve a new p

- Define the problem. What problems do you want to solve?
- Start simple. Be familiar with the data and the baseline results.
- Then try something more complicated.

SAS Visual Data Mining and Machine Learning provides a good platform for beginners to learn ma learning methods to their problems. Sign up for a free trial today!

Tags machine learning algorithms | machine learning | data science basics | data scie | regression

Share

ABOUT AUTHOR

Hui Li

Principal Staff Scientist, Data Science

Dr. Hui Li is a Principal Staff Scientist of Data Science Technologies at SAS. Her curren Cognitive Computing and SAS recommendation systems in SAS Viya. She received he

degree in Electrical and Computer Engineering from Duke University. Before joining SA as a research scientist and at Signal Innovation Group, Inc. as a research engineer. He machine learning for big, heterogeneous data, collaborative filtering recommendations, reinforcement learning.

8 COMMENTS

Daymond Ling on April 12, 2017 7:58 pm

Thank you for the cheat-sheet, it provides a nice taxonomy for people to understand the rela use it in my machine learning class to help students round out their world view.

Hui Li on April 17, 2017 9:54 am

Thank Daymond.

Let us know if you have any questions when teaching the students using the information.

Hector Alvaro Rojas on April 21, 2017 11:12 am

This is a great cheat-sheet to understand and remember the relationship between the most u I have not seen something similar like this published online yet.

I think it could be nice to incorporate the "cost" variable, the principal's reasons why each se examples of applications for each one. I know that this suggestion means a lot of work and s Anyway, it could be a nice new project to be done, don't you think so?

Congratulations for the work already done anyway!

Hui Li on April 24, 2017 11:29 am

Thanks, Hector. Incorporating the "cost" variable is a pretty wider area in machine lea

considered as a subfield of reinforcement learning -- based on the cost (reward), the a
he/she wants to take. I considered this problem for a while and haven't found a good e
I have a time, I will write a blog specifically for the reinforcement learning.

charles on April 24, 2017 9:54 am

An excellent blog. Thank you

Hui Li on April 24, 2017 11:30 am

Thank you.

Don Maclean on April 25, 2017 11:34 am

Excellent summary but I think the target audience is a few steps beyond "beginner". I showe
study machine learning, and they were overwhelmed.

Anastassia Dr Lauterbach on April 26, 2017 6:31 am

Great blog, thank you. I will use it when talking to non tech companies about starting doing M