Machine Learning in Real Life

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS424/524, SML302

Lecture 24

Data are everywhere.

In this course, we learned how to pair data with methods to answer questions.

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read or write.
–H.G. Wells

Data are complex, and there is a lot of it.

Every day, three times per second, we produce the equivalent of the amount of data that the Library of Congress has in its entire print collection, right? But most of it is like cat videos on YouTube or 13-year-olds exchanging text messages about the next Twilight movie. –Nate Silver

This doesn't make answering questions easier.

The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. – John Tukey

Interacting with data

We have begun to learn about fields of study—statistics, machine learning, data science, data mining—that address the problem of data analysis.

Goal today:

- review what we have learned,
- apply what we have learned to data scenarios,
- highlight some recurring themes,
- send you off as confident data scientists and machine learnists.

The purpose of this class is to convince you that you now have the tools to *interact with data in real life*.

What have we learned?

Course was divided into four parts, each representing one data analysis:

- *Classification*: e.g., sentiment analysis
- *Prediction*: e.g., predicting children's outcomes
- *Clustering*: e.g., finding groups of similar documents
- *Dimension reduction*: e.g., Netflix ratings

Data Analysis

The goal is to turn data into information, and information into insight. –Carly Fiorina

Today I want to talk about the general process of data analysis.

1. Understand the complexities of and patterns in your data
2. Explore your data by appropriately modeling those complexities and patterns
3. Use what you learned to cluster, predict, classify, find anomalies, answer questions.

Assumptions in data analysis

All models are wrong, but some are useful. –George Box

- Probability models capture the generative process for data.
- Probability models encode assumptions about data.
- Assumptions are necessary for a model to be tractable, but they must be included in the interpretation of the results.

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
–George Box

What are the probability models we have learned about?

Distributions

- Gaussian, multivariate Gaussian
- Bernoulli, binomial, multinomial
- Poisson
- Dirichlet

Classification

- Naive Bayes classification
- Support vector machines
- K-nearest neighbors
- Decision trees; random forests

What are the probability models we have learned about?

Prediction

- Linear regression
- Regularized linear regression
- Sparse linear regression
- Generalized linear models
- Gaussian processes

What are the probability models we have learned about?

Clustering

- K-means clustering
- Mixture models
- Hidden Markov models
- Dirichlet process mixture models

Dimension reduction

- Principal component analysis
- Factor analysis
- Latent Dirichlet allocation

Theme one: design features, clean data

- data do not go from collection directly to statistical analysis
- the success of the analysis depends heavily on
  - what features are extracted and used from each sample
  - how biases, outliers, and missing data are handled
- essential to rely on how the data were collected and the natural phenomena measured

Theme one: design features, clean data

For example: Spam filtering

- data goes from email dumps to a bag-of-words matrix representation
- vocabulary is chosen carefully
    - stopwords removed
    - non-dictionary words removed
    - words indicative of the data collection process removed
    - words stemmed
    - infrequent words removed
- then, ensure that every email has sufficient “words in the bag”

Theme two: Models generalize easily to data

We often began with the Gaussian model. We never ended there...

- Naive Bayes classification to general probabilistic classification
- Linear regression to generalized linear models
- Mixtures of Gaussians to mixtures of anything
- Hidden Markov models for any emission distribution
- Factor analysis to latent Dirichlet allocation

You can generalize now

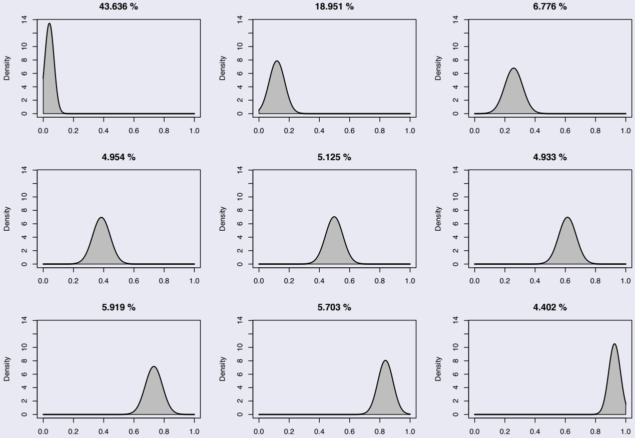
- *Across models*: for a given model, think about what data problems you can address by changing the data generating distribution

- *Across data*: when you encounter new data, imagine how models might approximate the structure

- *Generative models:* We can adapt many models to many kinds of observations—continuous, discrete, categorical, multivariate, ordinal—by modifying the generating distribution

Theme two: generalize

For example: Estimate an empirical distribution on $[0, 1]$



[From Lock & Dunson 2013]

Theme three: Regularization and feature selection

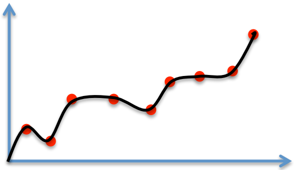
We saw that models sometimes *overfit* data, which means that they fit sample noise rather than population signal.

Regularization, including sparsity and feature selection, helps:

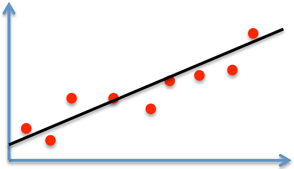
- moderate bias/variance trade-off
- gives sparse solutions, for interpretability and tractabiity
- Bayesian models regularize via prior distributions;
  classical statistics use penalty terms.

Regularization: intuition

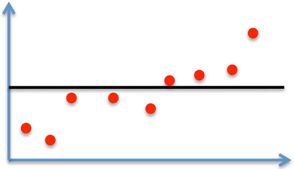
Regularization can be interpreted as *adding bias* to the parameter estimates and reducing variance.



Large variance



Some bias, variance



Large bias

Regularization, and sparse regularization in particular, helps to avoid modeling noise and facilitates interpretation.

Theme four: Hierarchical generative models

Generative models:

- encode probabilistic assumptions and conditional independences;
- mathematically characterize the factorization of a joint probability;
- encourage generalization and modularity;
- enable prior domain-specific knowledge;
- share statistical support across samples and observations;
- connect assumptions to algorithms.

Theme five: Understanding results of the application

For the theory-practice iteration to work, the scientist must be, as it were, mentally ambidextrous; fascinated equally on the one hand by possible meanings, theories, and tentative models to be induced from data and the practical reality of the real world, and on the other with the factual implications deducible from tentative theories, models and hypotheses.
–George Box

- Consider what questions the data can shed light on, and what questions they cannot;
- Understand the models sufficiently to interpret the models parameters fitted to those data in the context of the phenomenon being analyzed.

Theme six: data science is open-ended

In scientific subjects, the natural remedy for dogmatism has been found in research. By temperament and training, the research worker is the antithesis of the pundit. What he is actively and constantly aware of is his ignorance, not his knowledge; the insufficiency of his concepts, of the terms and phrases in which he tries to excogitate his problems: not their final and exhaustive sufficiency. He is, therefore, usually only a good teacher for the few who wish to use their mind as a workshop, rather than a warehouse. – R.A. Fisher

- All data science problems can be approached in different ways
- your analysis can always be improved; this improvement may come from adapting many aspects of the analysis
- know when to move on to next challenge ("good enough" analysis).

Leaving the nest: Data analysis in real life

A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product. –Hillary Mason

You encounter a huge data set you want to analyze. What do you do?

1. Define the problem: frame a specific question or hypothesis
2. Dig into the data: explore structure and existing methods
3. Clean the data: remove outliers, impute missing data, define features
4. Design a model: start with the simplest model, rely on previous work
5. Fit model: estimate parameters, start simple and adapt
6. Evaluate results: sift carefully through false positive, false negatives
7. Revise: extend model, add more data or supervision, adapt methods
8. Present analysis: write, publish, blog, present your work

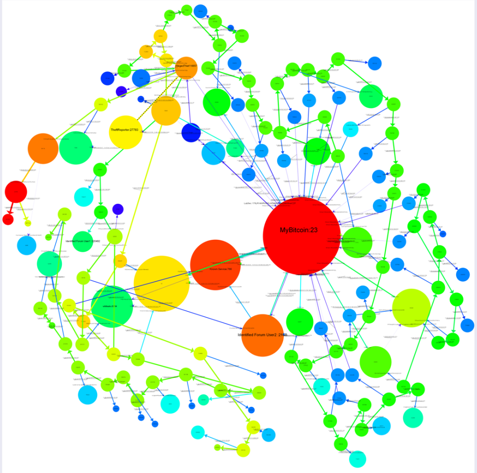
Define the problem

How can I search more efficiently for shoes?

[From Kristin Grauman]

Define the problem

What Bitcoin addresses are involved in illicit trade? *from [Reid & Harrigan 2012]*



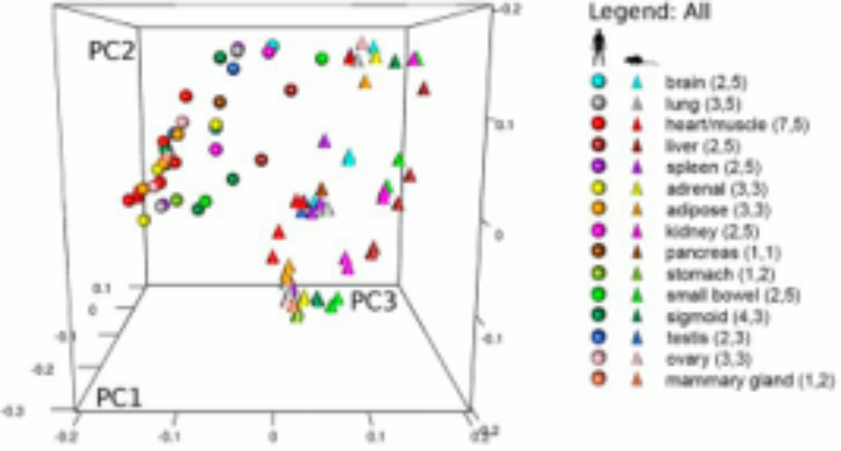
Define the problem

- Did you specify the type of data analytic question (e.g., exploration, association, causality) before touching the data?
- Did you define the metric for success before beginning?
- Did you understand the context for the question and the scientific or business application?
- Did you record the experimental design?
- Did you consider whether the question could be answered with the available data?

[From Jeff Leek]

Dig into the data

Is gene expression data more similar within species or within tissues?



[From Lin et al. 2014]

Dig into the data

Tell me about the data.

- Are the data Gaussian? Poisson? Overdispersed?
- Is the number of features greater than the number of samples?
- Plot many statistics of the data! Use unsupervised learning methods + intuition.
- Are there outliers? Is the structure different without those samples?

Clean the data

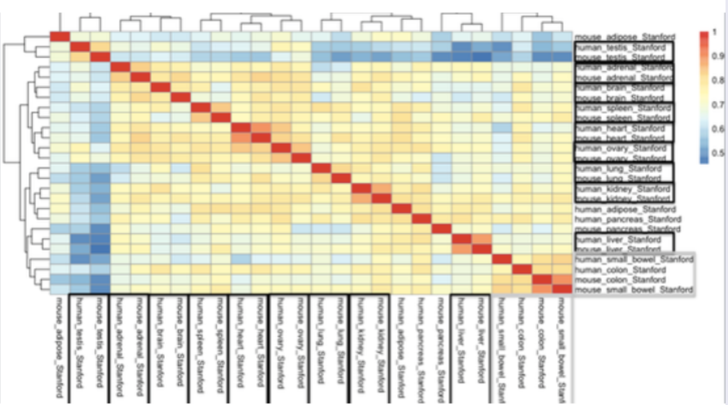
To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of. –R.A. Fisher

Consider collection process

- Are we trying to find differences between two sets of data that are collected differently? (spam/not spam)
- Are the data missing at random? (movie ratings)
- Are my residual errors Gaussian? (gene expression data)
- Is the data collection process designed to allow me to make specific conclusions from the data?
- **Process of going from raw to clean data must be transparent and reproducible!**

Clean the data

Control for batch effects.

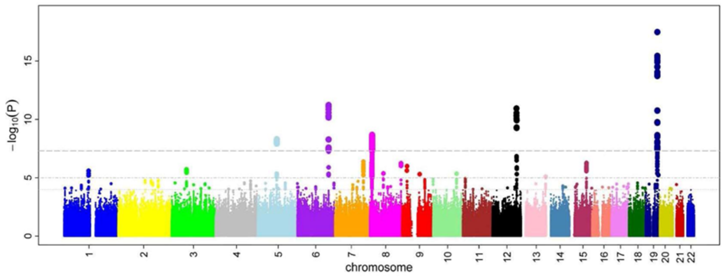


[From Yoav Gilad]

Design a model

...models leave out a lot of phenomena, but they are good enough for the task at hand. –Peter Norvig

Genome-wide association studies



[From wikipedia]

How we select the proper model to use for a data set?

A mechanistic model has the following advantages:

1. It contributes to our scientific understanding of the phenomenon under study.
2. It usually provides a better basis for extrapolation (at least to conditions worthy of further experimental investigation if not through the entire range of all input variables).
3. It tends to be parsimonious (i.e., frugal) in the use of parameters and to provide better estimates of the response

–George Box

1. Interpretable
2. Low dimensional representation of data
3. Simple; not overparameterized

How is my analytic answer quantified probabilistically?

You have made assumptions and fit a model.

With a fitted model in hand, solve your problem with a probabilistic computation that uses the model and some input.

Solve your problem:

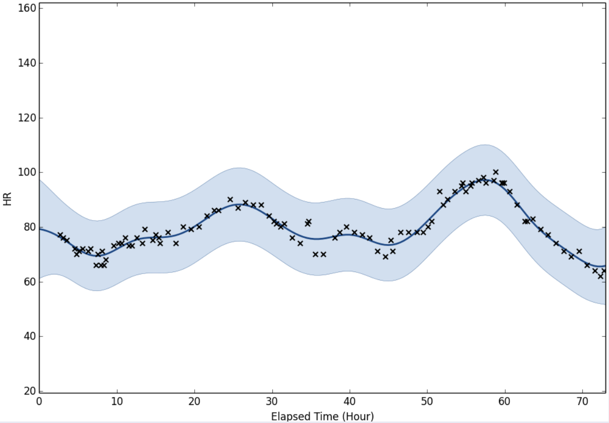
- Cast what you are trying to do in terms of a predictive quantity
- Cast that quantity as a probabilistic computation

What does an analysis solution look like?

- Classification: the maximum probability class
- Speech recognition: the maximum probability sequence
- Target tracking: The expected location at the next time point
- Collaborative filtering: The expected ratings of unrated movies
- Exploration: how often are two samples in the same group when clustering into K groups?

Fit a model

Fit the model parameters to data



Trade-off accuracy, speed, model complexity, and reproducibility.

Learning about data

The problem of how to discover patterns in data—also called "learning"—can be formulated as an optimization of an appropriate objective function.

We focused on *maximum likelihood estimation*.

- The model is indexed by parameters.
- The observed data have a probability under each setting.
- Find the parameters that make the observed data most likely.
- Use exact methods, gradient-based methods, or approximate methods

Alternative objective functions

Other objectives are related to likelihoods under a model.

- Expectation-maximization: iteratively optimizing auxiliary function
- PCA: maximizing variance and minimizing reconstruction error
- Traditional regression: minimizing residual sum of squares

Analysis to scale to data

In modern machine learning we need to think about scale.

- Parallelize computation: take advantage of clusters of computers
- Stochastic optimization: repeatedly subsample the data
- Online methods: update parameters as data roll in
- Approximate methods: variational approaches

These methods take advantage of structure in objective function.

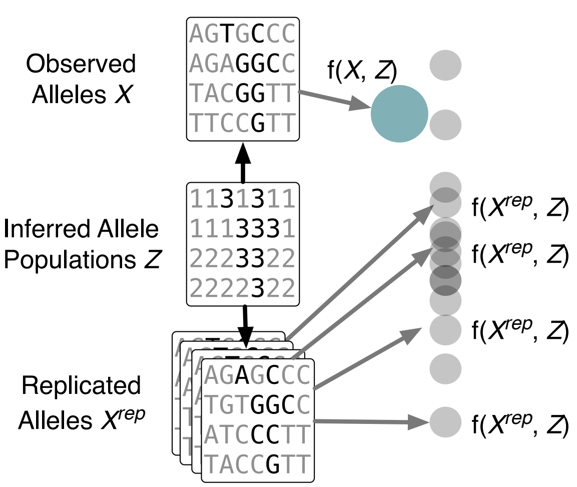
Evaluate results

We have considered a number of ways to validate results

- Confidence intervals: compute variance of the estimated parameters
- Bootstrapping: resample data; refit model
- Hold out data: fit model to subset of data; look at generalization
- Replicate results: Is there another study whose results can be compared?

Evaluate results

Posterior predictive checks *from [Mimno et al. 2015]*



AGTGCCC
AGAGGCC
TACGGTT
TTCCGTT

1131311
1113331
2223322
2222322

AGAGCCC
TGTGGCC
ATCCCTT
TACCGTT

Look at your results carefully

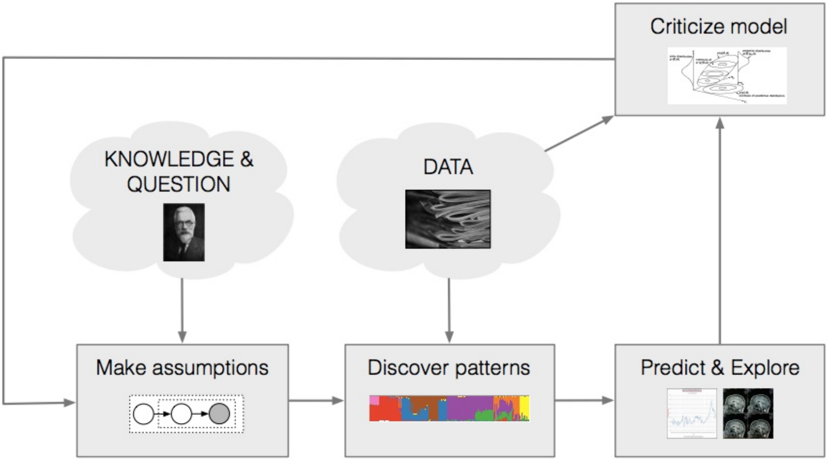
Politicians use statistics like drunkards use lampposts: not for illumination, but for support.
–Hans Kuhn

Error analysis is important

- Where did my method go wrong?
- Are there patterns the residual error?
- Do these patterns suggest specific model extensions?

Note the iterative nature of this process—now we go back to add new features or to redefine a model.

Cycle of data analysis *[Blei 2016]*



Criticize model

KNOWLEDGE & QUESTION

DATA

Make assumptions

Discover patterns

Predict & Explore

Evaluating results

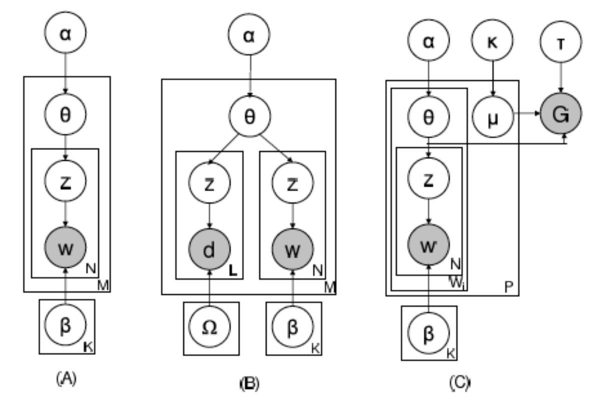
Generalization error, or performance on held-out samples, is important to predict how well model will perform on future data.

Similarly, cross-validation is essential for setting hyperparameters.

Other evaluations should be used: human studies, money earned, enriched peripheral data

Revise analysis

What structure in the data should I model explicitly?



[From Liu et al. 2009]

Revise analysis: Feature selection

- What features are uninformative (feature selection)?
- What features can I add or elaborate on?
- Is there other available data I can pull in to help with my task?
- What do I know about this phenomenon I am studying?
- If ground truth is lacking, how can I creatively validate my analysis conclusions?

Set your analysis free

We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover all the tracks, to not worry about the blind alleys or to describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work. –Richard Feynman

- Data analysis is hard; manuscripts make it look easy and direct.
- Publications—Methods sections in particular—should read like a recipe book for someone wanting to replicate precisely your results.
- Publish code, both models and data cleaning code, to make replication possible

Set your analysis free

How to present your data analysis?

- Did you describe the question of interest?
- Did you describe the data set, experimental design, and question you are answering?
- Did you specify the type of data analytic question you are answering?
- Did you specify in clear notation the exact model you are fitting?
- Did you explain on the scale of interest what each estimate and measure of uncertainty means?
- Did you report a measure of uncertainty for each estimate on the scientific scale?
- Does each figure communicate an important piece of information or address a question of interest?
- Did you create a script that reproduces all your analyses?

[From Jeff Leek]

Choose your analytic goals bravely

- Can smartphone behavior be used to identify victims of domestic violence?
- Can pedometer data be used for early diagnosis of Parkinson's disease?
- Can online course material be tailored for people with dyslexia?
- Can video analysis be used to translate sign language to text?
- Can we predict when an elderly adult living at home alone has fallen?
- Can we understand what drives at-risk children who beat the odds?
- Can we predict suicide risk from social network post patterns?
- Can we automatically flag social media posts with online bullying?

Change people's minds with your analyses

- How are social institutions – school, financial systems, housing, childcare, criminal justice, social media – unfairly treating Black people?
- What patterns are expected? What patterns are unexpected, but lead to insights that impact policy?
  - Black and white Covid-19 patients at the UPenn Hospitals have similar outcomes.
  - But, Black patients are on average five years younger than white patients.
  - COMPAS recidivism scores are much higher for Black parolees than white.
  - But, Black parolees are more likely to have second offense be a traffic stop than white.
  - Prosecutors strike Black potential jurors at much higher rates than white potential jurors when defendant is Black.
  - But, only a small proportion of prosecutors make up the bulk of these biased strikes.

Continue to educate yourself as a data scientist

A significant constraint on realizing value from Big Data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from Big Data.
–McKinsey Report

Learn

Take classes

- Foundations of probabilistic models, theory, NLP, computer vision
- Ethics of AI: Ruha Benjamin, Arvind Narayanan, Janet Vertesi
- Courses in CS, EE, ORFE, MoBio, Political Science, others
- Coursera courses, video lectures, Talking Machines podcasts

Join the ml-stat-talks mailing list

Go to graduate school, or don't go to graduate school

Read

Journal and conference proceedings:

- ESL, PRML, MLAPA
- Conference proceedings: ICML, NIPS, AI-STATS, EMNLP, CVPR, ...
- Journals: JMLR, JASA, AAS, AS, BA, JCGS, ...
- Blogs: `simplystatistics.org`, Andrew Gelman, Lior Pachter, Ryan P Adams, ...

Participate in reading groups: CSML reading group

Collaborate

- Work with classmates, colleagues, scientists, engineers, government, startups, companies, non-profits, the internet, and everyone else with exciting data.
- Grill them on the data: biases, missing data, existing approaches.
- Ask hard questions of the data. Find unexpected patterns. Change people's minds.
- Present your work: publish, write, blog, tweet data and solutions.
- Accept criticism gracefully and admit mistakes. Continue to learn.
- Let me know what you have done! I love hearing ML success stories.

Do

Dig into data and solve important problems.