An Introduction to Machine Learning and the R Programming Language

Vlad-Ovidiu Lupu

What is Machine Learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Tom Mitchell)

Who does it and why?

- Google - NLP, spam detection, ad placement, translation
- Facebook - NLP, spam detection, newsfeed and post grouping
- Twitter - NLP, sentiment analysis
- Microsoft - NLP, malware detection
- Baidu - NLP, search, image recognition
- Amazon and retailers - recommender systems

Other industries:

- Medicine
- Banking
- Sports
- Energy

What does Machine Learning do?

Some general tasks:

- regression
- classification
- clustering

General purposes:

- prediction (maximize accuracy)
- inference (maximize interpretability)

Evaluation criteria may vary:

- minimizing mean error vs reducing the number of extreme results
- accuracy vs reducing the number of false positives/negatives

What is R?

- R is better described not as a programming language, but as an environment for statistical computing which has a programming language
- R was created by Ross Ihaka and Robert Gentleman at the University of Auckland in 1993
- interpreted language
- supports both procedural and functional programming
- vectors and matrices are first class citizens, many functions are vectorized
- open source

The Good Parts

Pros:

- large community, many learning resources
- package management is integrated in the core language
- vectors and matrices are core language features
- great IDE (RStudio)
- very good visualizations, more alternatives(lattice, ggplot2, ggvis, ggmaps)
- easy data manipulation(dplyr, tidyr, lubridate)

The Bad Parts

Cons:

- very slow, no industry ready tools (servers, web frameworks)
- poor concurrency
- the documentation assumes you know statistics
- R core is stable, which is good for backwards compatibility, bad because language design mistakes accumulate(The R Inferno)
- Everything must be in memory

R vs Python

Pros:

- vectors and matrices are core language features, unlike NumPy
- RStudio is a much better IDE than the Python alternatives
- less package interdependencies
- easier data manipulation, faster prototyping

Cons:

- Python is a much better production language, going from prototype to product is feasible
- R is slower than Python
- easier to find general purpose programming libraries (feature engineering)
- for a machine learning pipeline, scikit-learn is more mature than caret

What is data science?

Data Science refers to the whole process of deriving actionable insight from data.

“80% of the work is data wrangling”, study presented at the Strata conference

“It is an absolute myth that you can send an algorithm over raw data and have insights pop up”, Jeffrey Heer, co-founder of Trifacta

“Data wrangling is a huge - and surprisingly so - part of the job”, Monica Rogati, VP at Jawbone

“Data Science is 20% statistics and 80% engineering”, Sandy Ryza, Cloudera

Data Science tasks

Actual machine learning tasks:

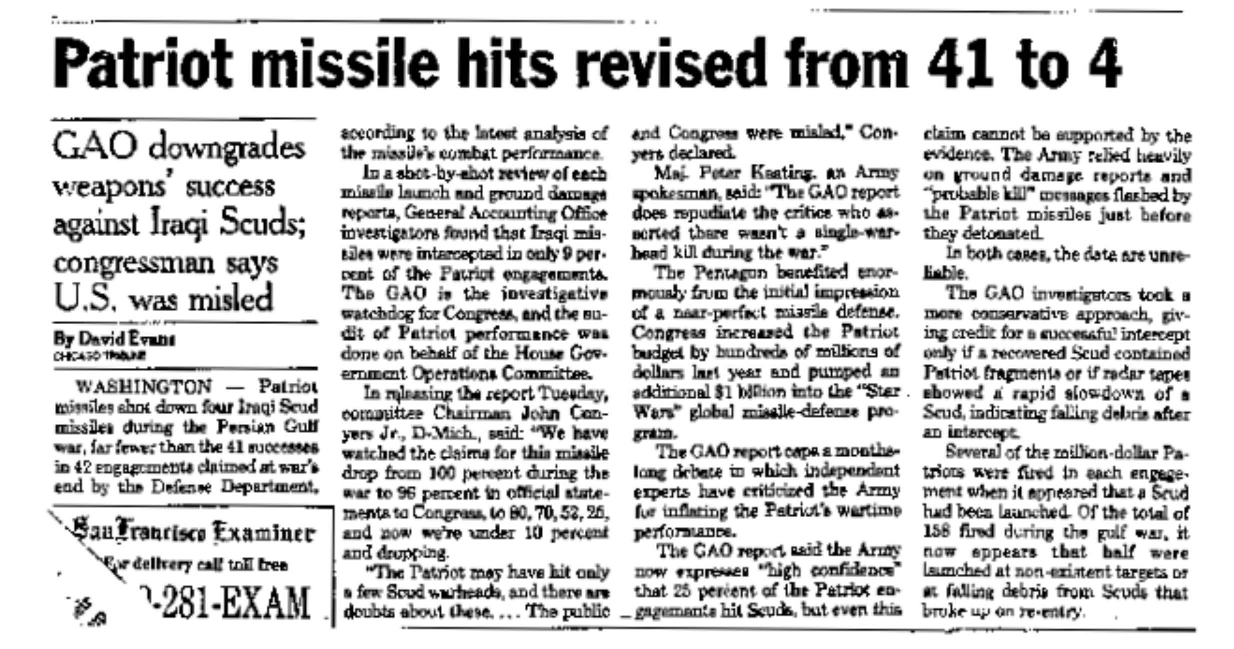
- what question do you want to answer
- what data do you need to answer it?
- what data can you obtain?
- can you answer the initial question using the data you have?
- exploratory data analysis
- data cleaning and processing
- feature engineering
- selecting a model that answers the question in the best way
- training the model; parameter selection
- evaluating the model

Machine Learning pitfalls

- Machine Learning is only a small part of the Data Science workflow
- A principle that is valid for all steps is “garbage in garbage out”
- A mistake at any step can invalidate the whole process
- Many Machine Learning courses and books focus only on the methods themselves, not on the context in which they are applied
- The next slides will present some common errors that can influence the results of applying Machine Learning methods

Misleading Data

George H.W. Bush(1991): "Forty-two Scuds engaged, 41 intercepted. Thank God for the Patriot missile”



Patriot missile hits revised from 41 to 4

GAO downgrades weapons' success against Iraqi Scuds; congressman says U.S. was misled

By David Evans
CHICAGO TRIBUNE

WASHINGTON — Patriot missiles shot down four Iraqi Scud missiles during the Persian Gulf war, far fewer than the 41 successes in 42 engagements claimed at war's end by the Defense Department, according to the latest analysis of the missile's combat performance.

In a shot-by-shot review of each missile launch and ground damage reports, General Accounting Office investigators found that Iraqi missiles were intercepted in only 9 percent of the Patriot engagements. The GAO is the investigative watchdog for Congress, and the audit of Patriot performance was done on behalf of the House Government Operations Committee.

In releasing the report Tuesday, committee Chairman John Conyers Jr., D-Mich., said: "We have watched the claims for this missile drop from 100 percent during the war to 96 percent in official statements to Congress, to 80, 70, 52, 25, and now we're under 10 percent and dropping.

"The Patriot may have hit only a few Scud warheads, and there are doubts about these. . . . The public and Congress were misled," Conyers declared.

Maj. Peter Keating, an Army spokesman, said: "The GAO report does repudiate the critics who asserted there wasn't a single-warhead kill during the war."

The Pentagon benefited enormously from the initial impression of a near-perfect missile defense. Congress increased the Patriot budget by hundreds of millions of dollars last year and pumped an additional $1 billion into the "Star Wars" global missile-defense program.

The GAO report caps a months-long debate in which independent experts have criticized the Army for inflating the Patriot's wartime performance.

The GAO report said the Army now expresses "high confidence" that 25 percent of the Patriot engagements hit Scuds, but even this claim cannot be supported by the evidence. The Army relied heavily on ground damage reports and "probable kill" messages flashed by the Patriot missiles just before they detonated.

In both cases, the date are unreliable.

The GAO investigators took a more conservative approach, giving credit for a successful intercept only if a recovered Scud contained Patriot fragments or if radar tapes showed a rapid slowdown of a Scud, indicating falling debris after an intercept.

Several of the million-dollar Patriots were fired in each engagement when it appeared that a Scud had been launched. Of the total of 158 fired during the gulf war, it now appears that half were launched at non-existent targets or at falling debris from Scuds that broke up on re-entry.

Correlation vs Causation

Quote and figure from an article in the New England Journal of Medicine(2012):

“chocolate consumption could hypothetically improve cognitive function not only in individuals but also in whole populations”

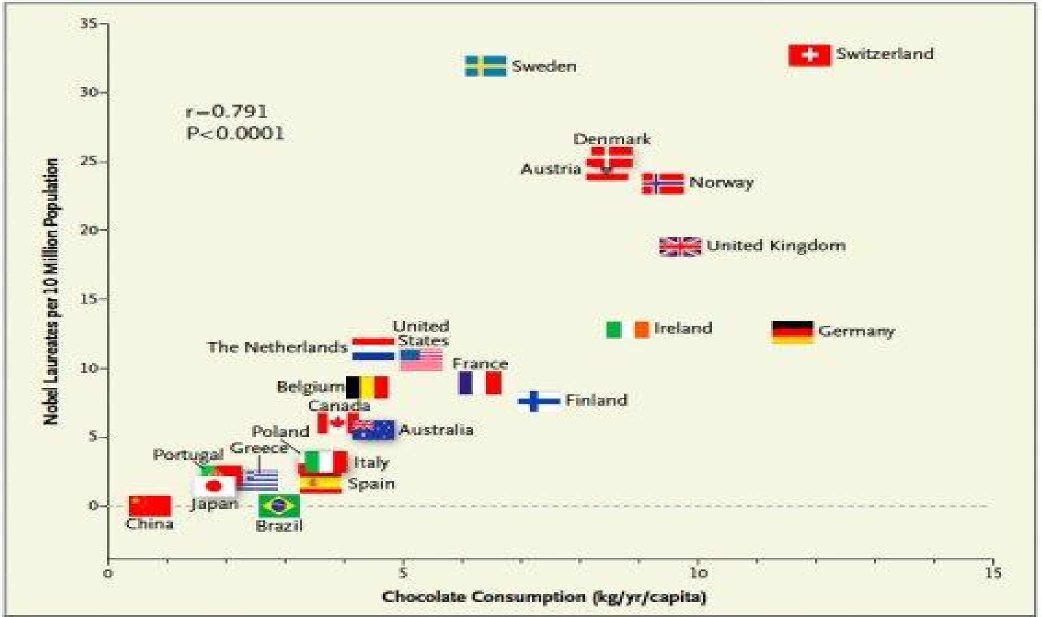
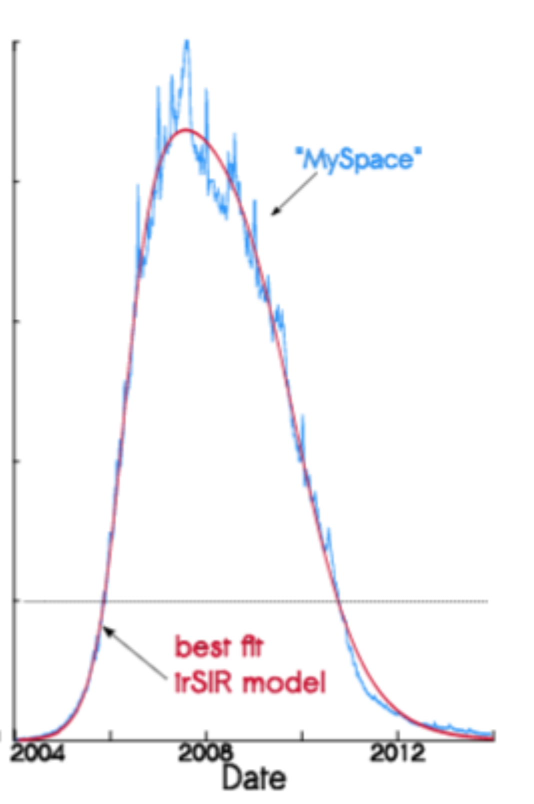
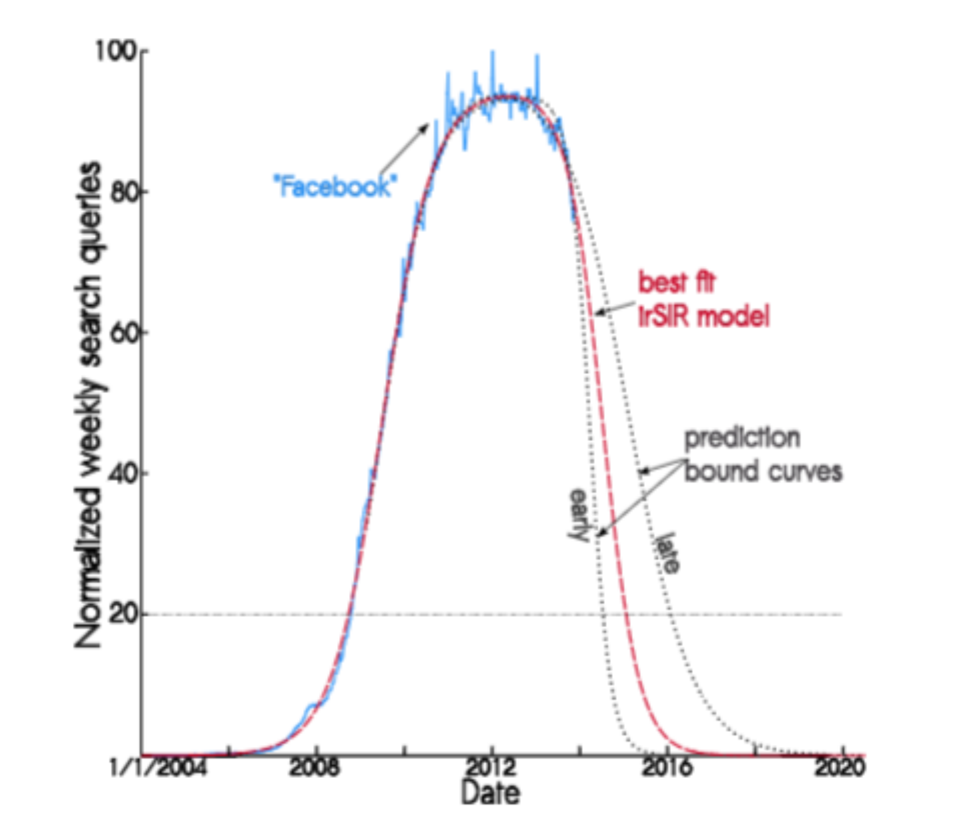


Figure 1. Correlation between Countries’ Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

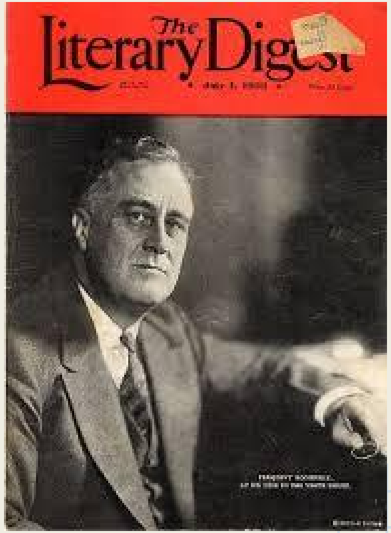
Bad Domain Modeling





Incorrect Sampling

- The Literary Digest was one of the most well known magazines in the USA
- In 1936 they set out to make the most precise election prediction and they used the biggest sample the world has ever seen(10 million)
- They predicted that Roosevelt would lose the election with 43%, instead he won with 61%
- The magazine never recovered from the reputation loss and failed soon after
- The sample suffered from many biases, like over-representation of certain groups and non-response bias



Data Dredging

- Data dredging (or p-Hacking) is the statisticians’ version of overfitting
- Many statistical methods produce a measure of significance called the p-value (the probability that you can get results at least as extreme as the ones you got given that some hypothesis is true)
- The main idea: Measure many things about few subjects and you are almost guaranteed to find statistically significant results
- Examples: most social sciences studies with ~30 subjects
- An xkcd comic on the subject: https://xkcd.com/882/
- In some parts of academia, due to publication bias this is most of the time not an individual mistake, but an emergent collective behaviour

Other things to consider

- multicorrelation
- statistical power
- independence

Other Examples:

- the decision to allow turning right at red lights is the result of an underpowered study, which said that there would be no difference in the number of accidents, when there actually are 100% more
- underpowered studies with <30 subjects also make data dredging very efficient

Recommended free e-book: http://www.statisticsdonewrong.com/

The Curse of Dimensionality

- Let’s say we have a dataset consisting of variables with values uniformly distributed between 0 and 1
- We want to eliminate the extreme values from all variables(<=0.05 and >=0.95)
- For one dimension, the proportion of remaining data would be 0.9
- For 2 dimensions we would be left with 0.81 of the data
- For 50 dimensions, the proportion of data left is 0.005
- In high dimensional spaces almost all points are at the edges
- As the number of dimensions grows, you need exponentially more data to build good models

Summary so far

- machine learning algorithms are not magic
- they do not understand phenomena, but rely on correlations instead
- the previous examples were funny, but it can happen to anyone working in an unfamiliar domain
- the larger the number of predictors, the more likely it is to find one correlated with the variable you want to predict
- more funny correlations: http://www.tylervigen.com/spurious-correlations
- The curse of dimensionality and the need for more data to train complex models is one of the reasons why Big Data is such a buzzword

Intuition on the bias-variance tradeoff

Example from “An Introduction to Statistical Learning with Applications in R” (James, Witten, Hastie, Tibshirani)

The book is free and can be downloaded from:

http://www-bcf.usc.edu/~gareth/ISL/

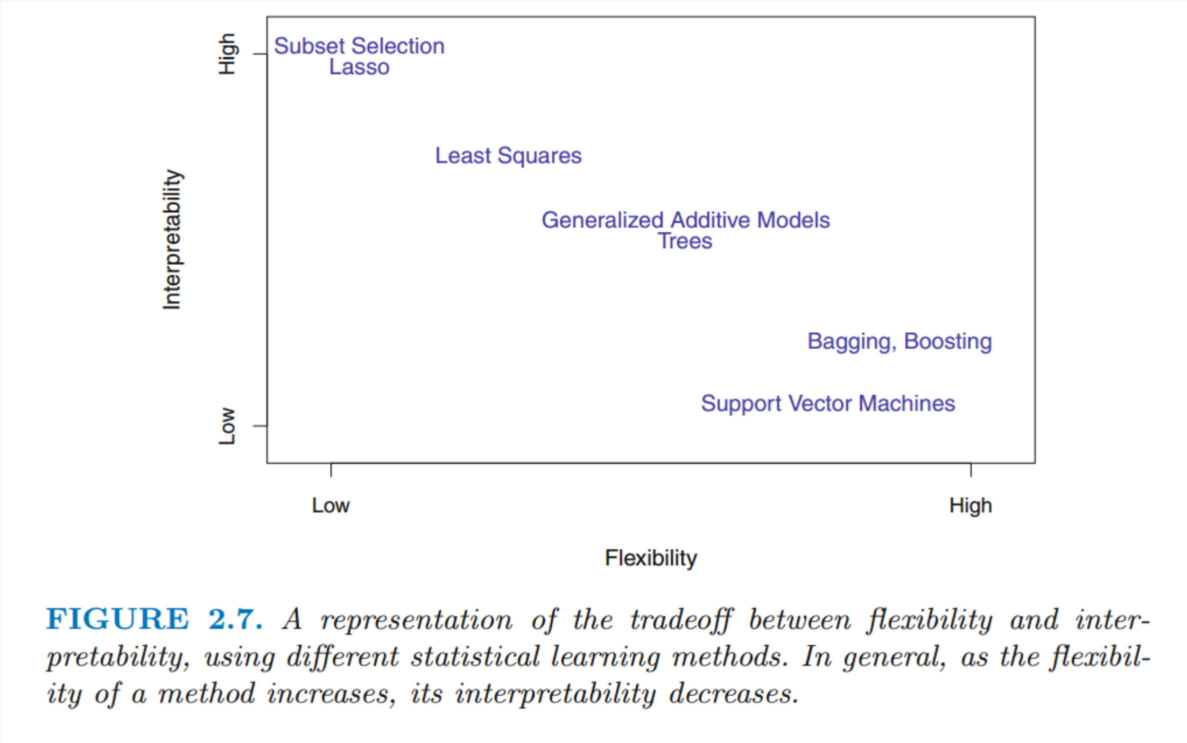


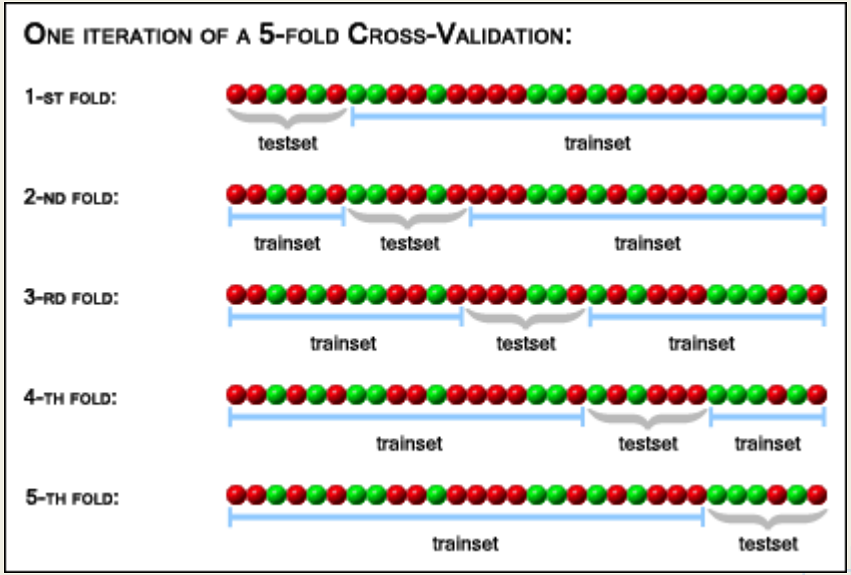
FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Overfitting

- By far the most common examples of overfitting are applying a very complex model on too few data points or on a noisy dataset
- Overfitting can be much subtler, as seen on the “Data dredging” slide
- One of the most common forms of overfitting that is not recognized as such is picking a model or model parameters based on the test set results
- A very good list of overfitting methods: http://hunch.net/?p=22

Cross-validation

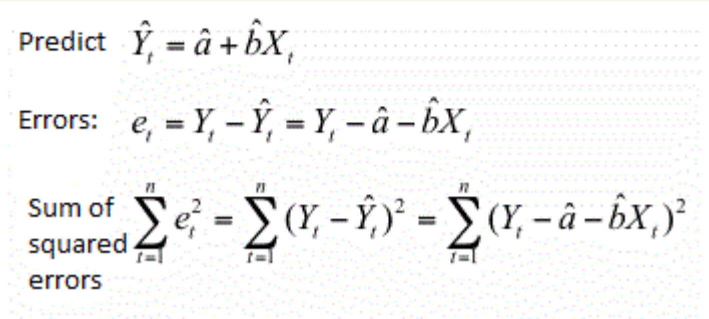
- one of the most widely used model validation methods
- helps to prevent overfitting
- allows choosing model parameters using the training set



Linear regression

2D case: Fitting a line through a set of points on a plane; can be generalized to an arbitrary number of dimensions.

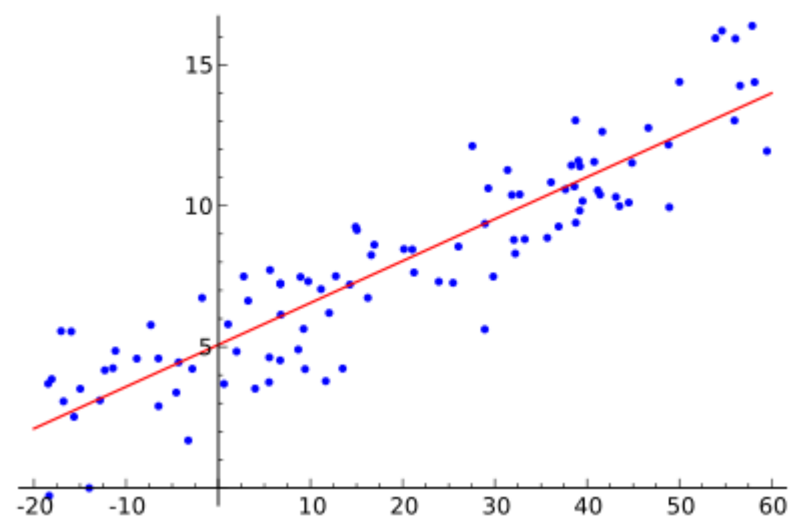
We can think of an infinite number of lines. Which is the best line?



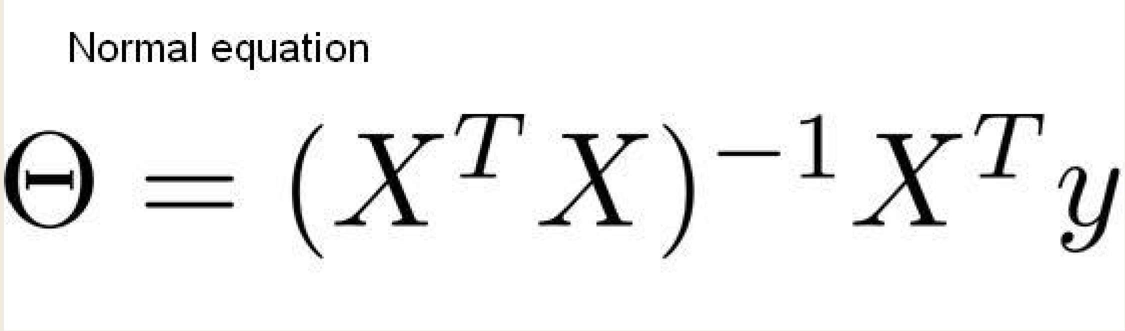
Predict $\hat{Y}_t = \hat{a} + \hat{b}X_t$

Errors: $e_t = Y_t - \hat{Y}_t = Y_t - \hat{a} - \hat{b}X_t$

Sum of squared errors $\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{a} - \hat{b}X_t)^2$



Is the Theory Necessary?



Normal equation

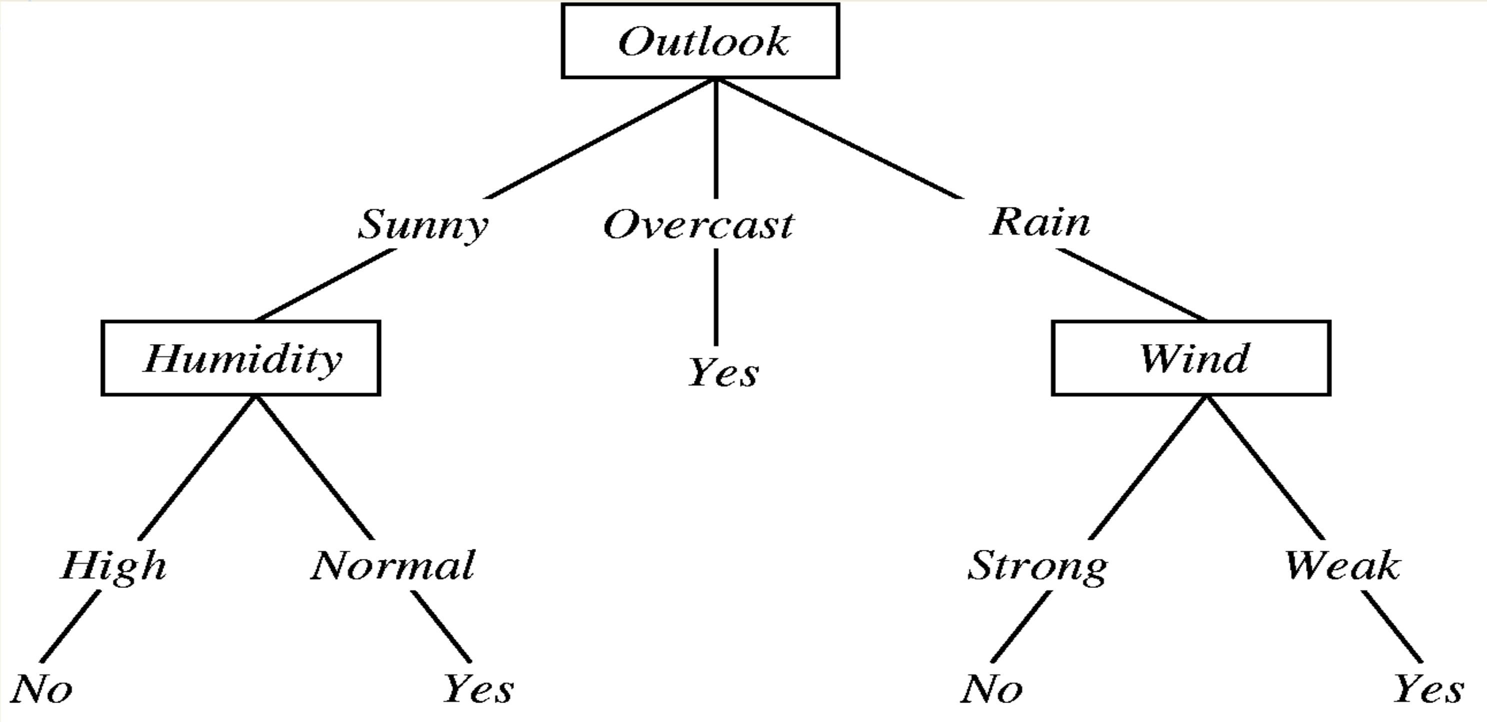
$$\Theta = (X^T X)^{-1} X^T y$$

From the point of view of a software engineer, does knowing this formula help you in any way?

Alternative Solving Method

- The normal equation is important because it is used by most linear regression libraries
- It doesn’t work for big data
- An alternative is gradient descent
- There are many methods for gradient descent like L-BFGS, SGD, AdaGrad

Decision trees



The bootstrap

- An effective statistical estimation technique
- It is useful when you have a small but representative sample from the whole population
- It involves generating new samples by sampling with replacement from the original sample

Bagging

- Bagging = bootstrap aggregation
- Bagging was originally used with decision trees trained on bootstrap samples and then aggregated

Random Forest

- The decision trees produced by the bagging procedure were still too similar
- Individual trees are only allowed to choose one variable out of a random subset to make a split
- A good value for the size of the random subset is sqrt(nr. of variables), but it is usually chosen with cross-validation
- It is the best performing algorithm on Kaggle(at least until now)
- Other studies show that especially for classification problems, it is really hard to beat:

Do we Need Hundreds of Classifiers to Solve Real World problems?

The End

Questions?