CST 463: Advanced Machine Learning

Glenn Bruns
CSUMB

image credit: Martin Pyka

Learning Outcomes

After this lecture, you should be able to:

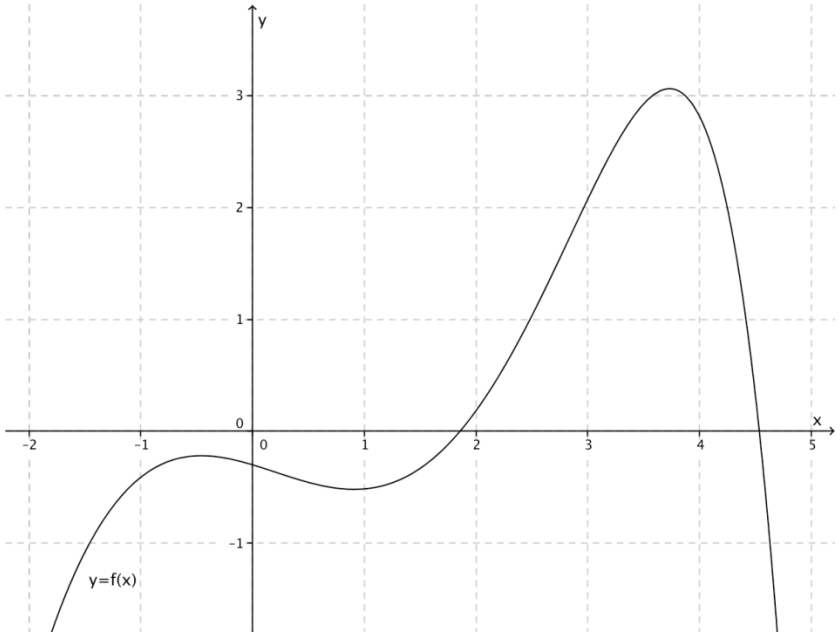
- explain the topics we'll cover in class at a high level
- explain why we are covering these topics

Python and Machine Learning

- R and Python are the biggest languages for data science
- Advantages of Python:
  - speed
  - widely-used, general-purpose language
- Toolset we'll use:
  - NumPy, SciPy, Scikit-Learn, Pandas (ML libraries)
  - Matplotlib, Seaborn (plotting)
  - Spyder (IDE, similar to R Studio)
  - Jupyter notebooks (similar to R Markdown)
  - Anaconda (Python distribution for ML; environments)

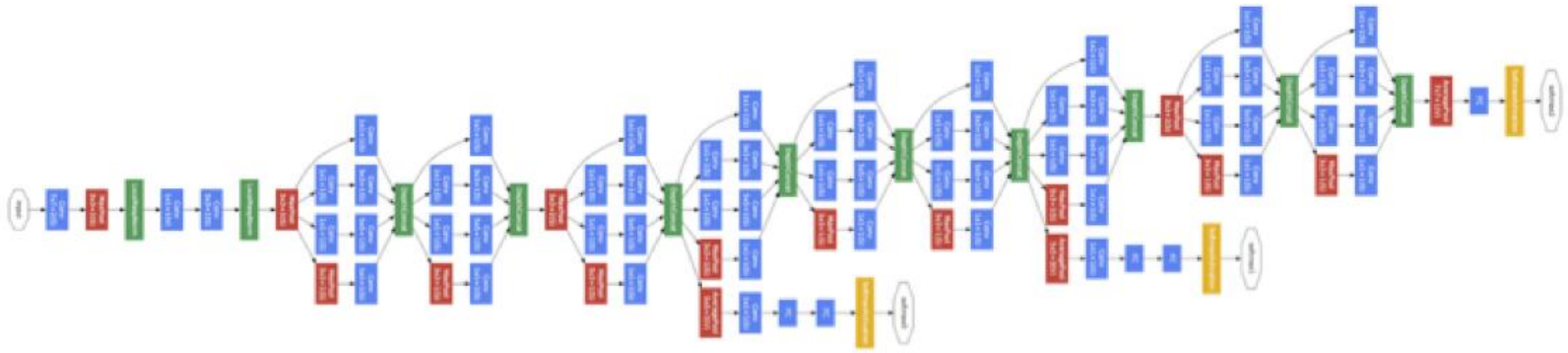
Optimization

- ML as an optimization problem
- versus ML as a probability problem
- Idea:
  - define model
  - define some concept of error
  - optimize model params to minimize error
- Optimize algorithmically



Deep Learning

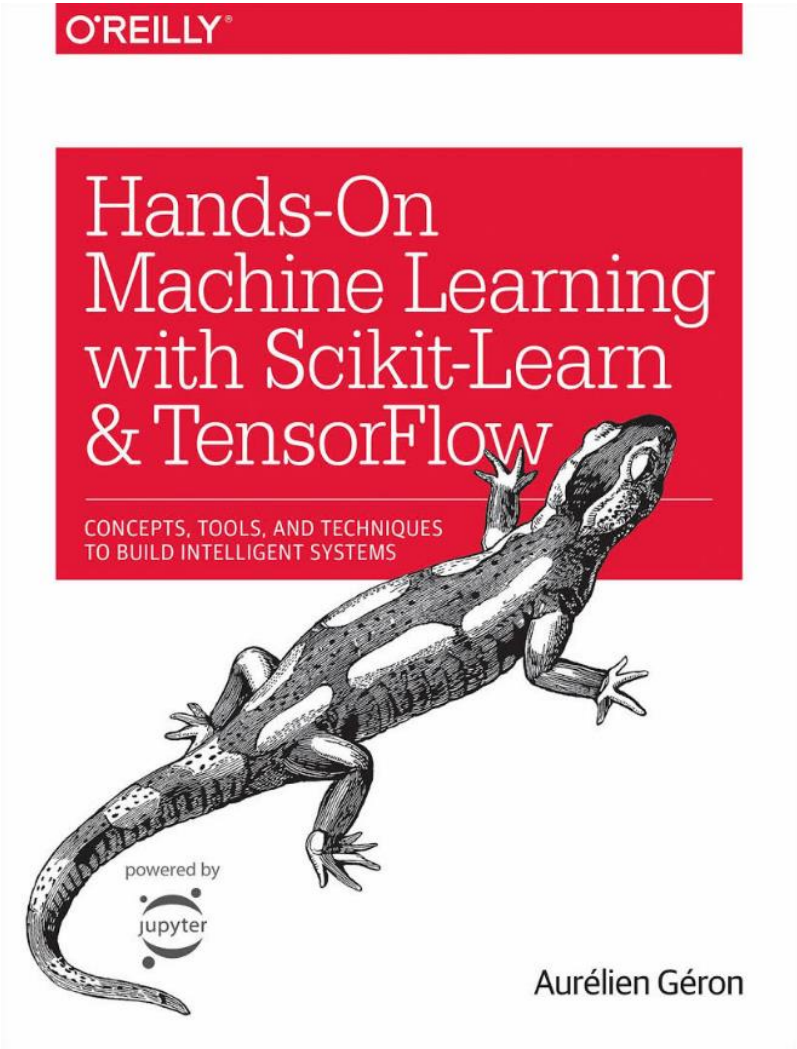
- Neural nets with many hidden layers
- Revolution in machine learning
- Can handle large and complex tasks:
  - image processing – Google Images
  - speech recognition - Siri
  - game playing - AlphaGo



Google's GoogLeNet

Our text

- Excellent book!
- Covers all course topics (except time series)
- Very up to date (published March 2017)



End-to-End machine learning

- Get started with Python
- What we did in Intro to Data Science, but with Python
- A couple of new ideas:
  - language support for “pipelines”
  - standardized API

Linear algebra

- Linear Algebra is widely used in machine learning
- we'll cover the basics
- we'll get more comfortable with Python, too

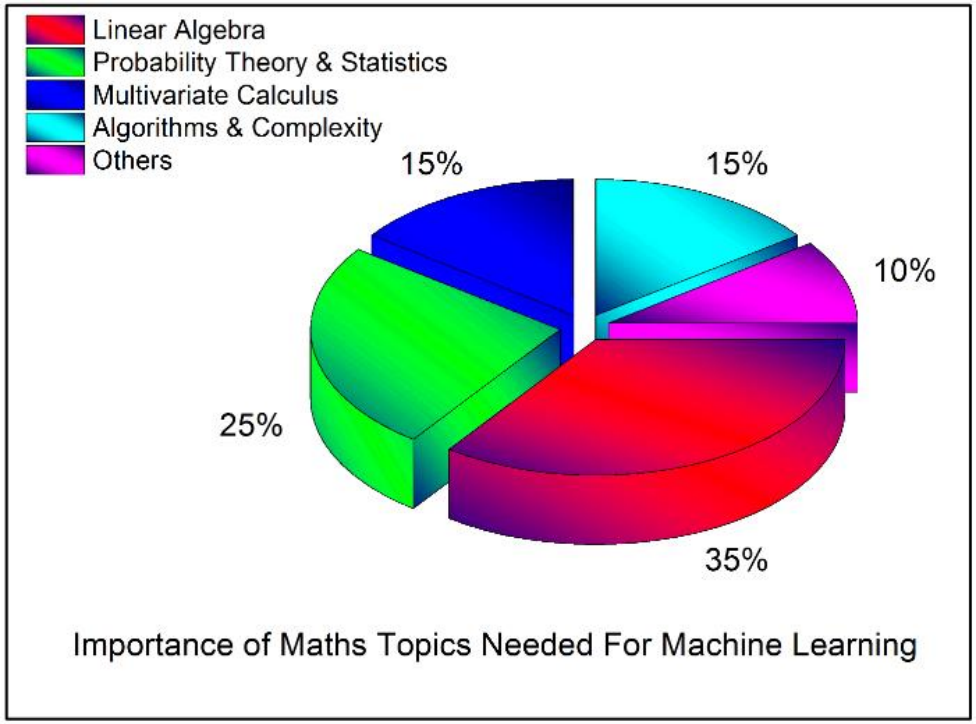
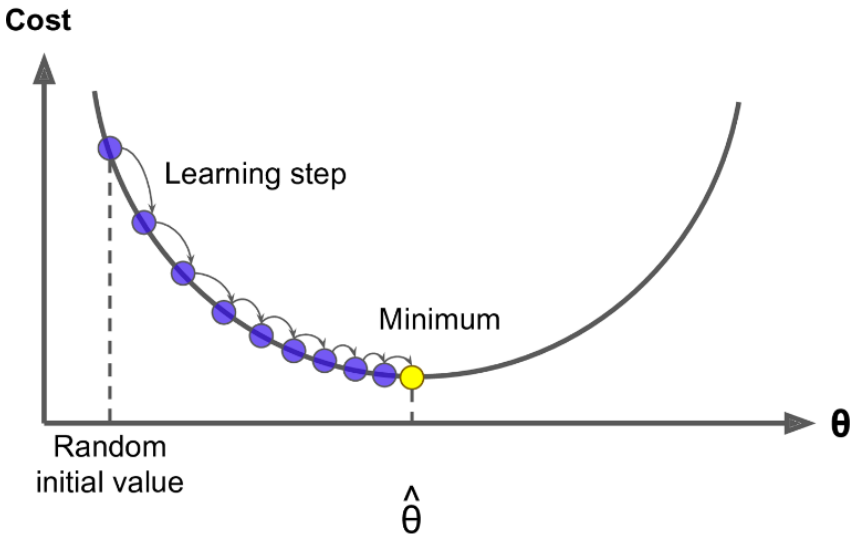


figure source: http://datascience.ibm.com/blog/the-mathematics-of-machine-learning/

Training models

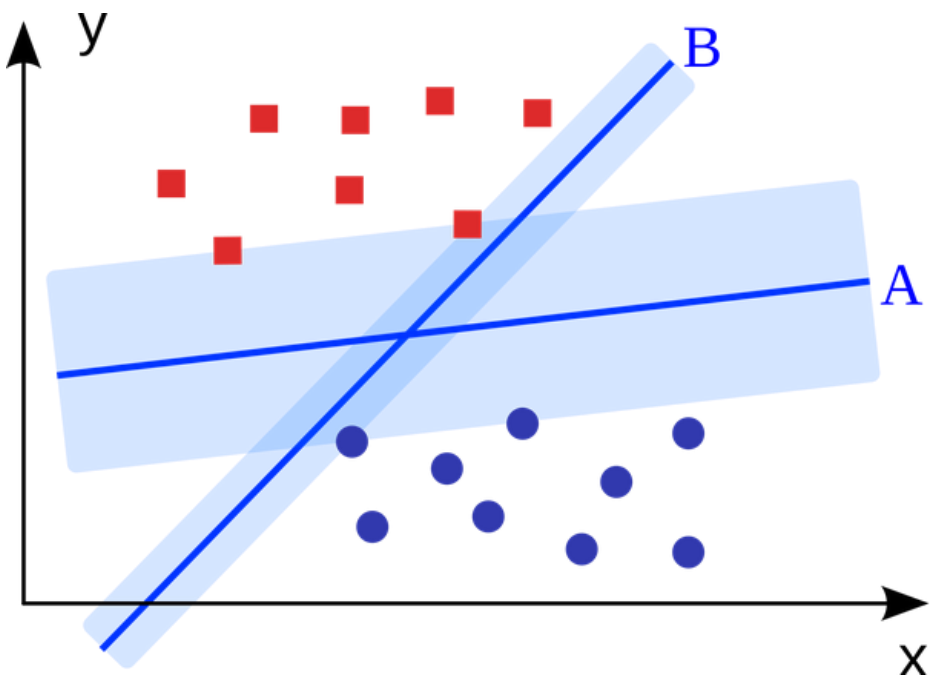
- How to solve the optimization problems in ML?
- Gradient descent, stochastic gradient descent, and more
- Example: linear regression
  - can solve with "closed form" equation
  - can solve with "iterative" optimization, like gradient descent
- regularization and early stopping



source: Géron text

Support Vector Machines

- Try to find the widest "street" that separates classes
- Related to logistic regression
- Clever math reduces cost of extra features

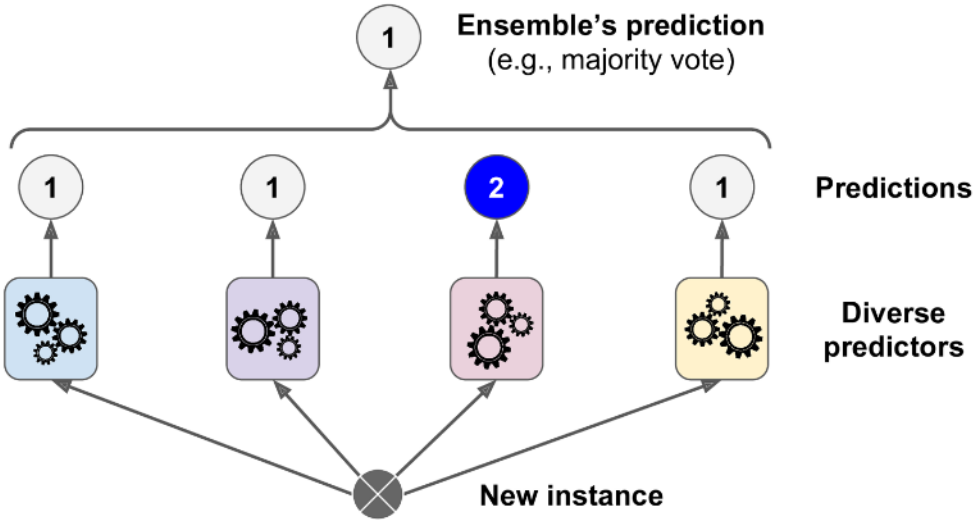


source: Quazi Marufur Rahman (Quora)

Ensemble learning

Examples:

- Voting: get majority class from diverse classifiers
- Bagging: train multiple classifiers on random samples of the training set



A majority-voting classifier

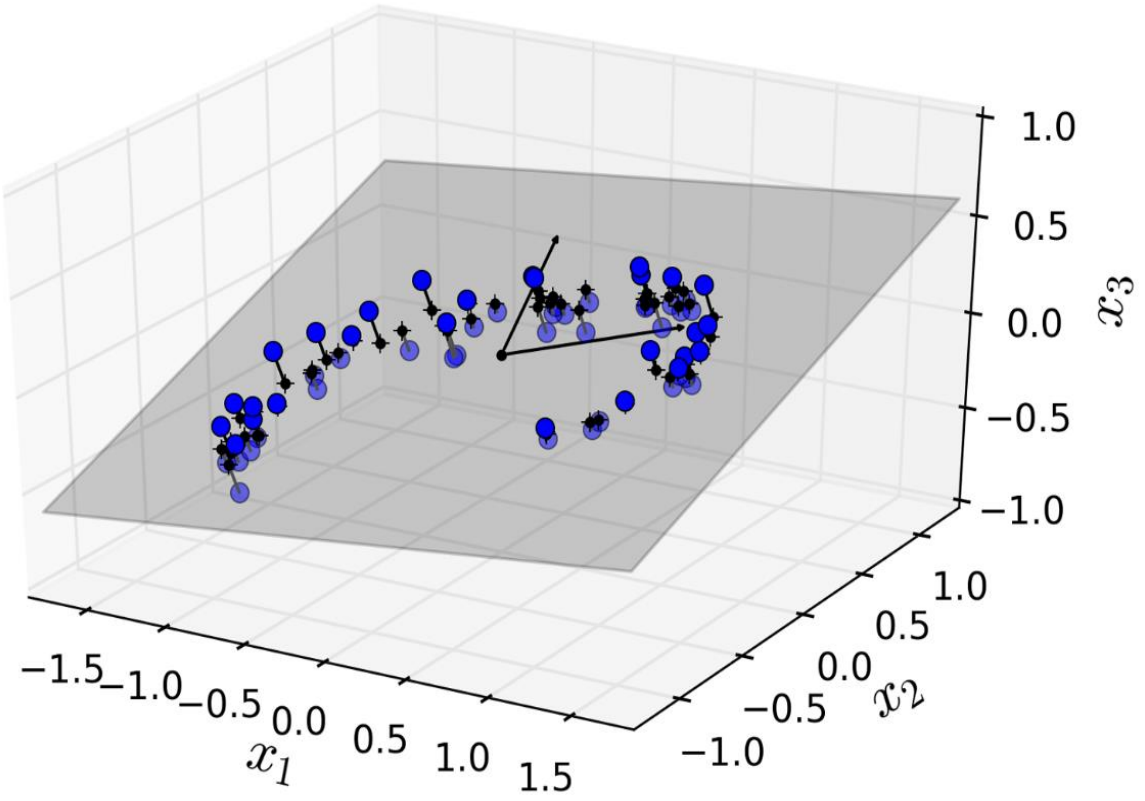
source: Géron text

Dimensionality reduction

Some problems involve thousands of features.

High-dimensional spaces behave strangely.

Principal Component Analysis (PCA) is one way to reduce # of dimensions



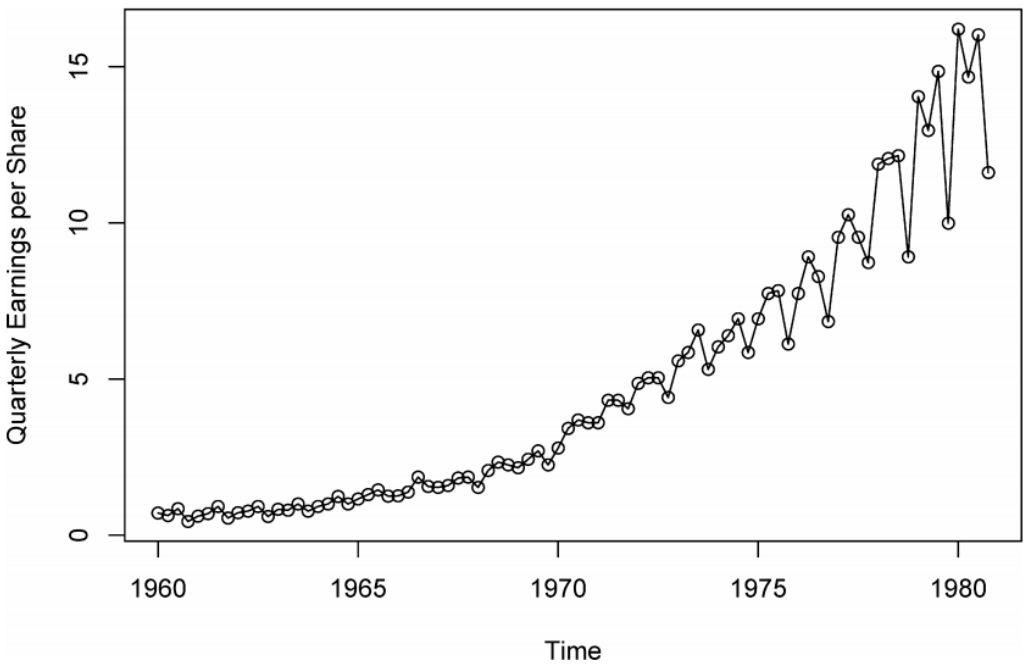
source: Géron text

Time series data

Time series data is data collected at regular intervals.

Many ML problems use such data.

Stock market data, unemployment data, birthrates, flu outbreaks, brain-wave data, global temps, …



Johnson & Johnson quarterly earnings, 1960 to 1980

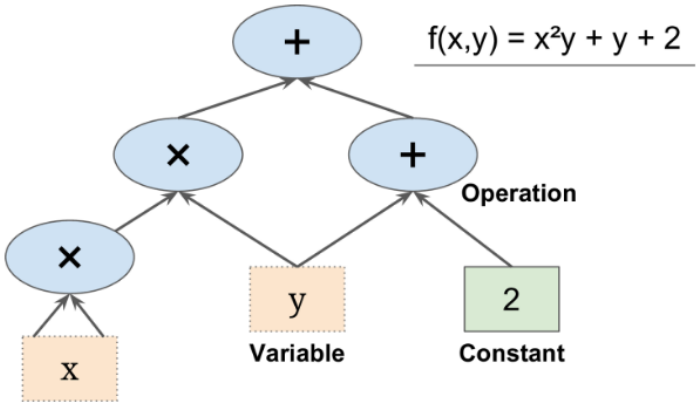
source: Shumway and Stoffer, Time Series Analysis and Application, Springer.

TensorFlow and neural nets

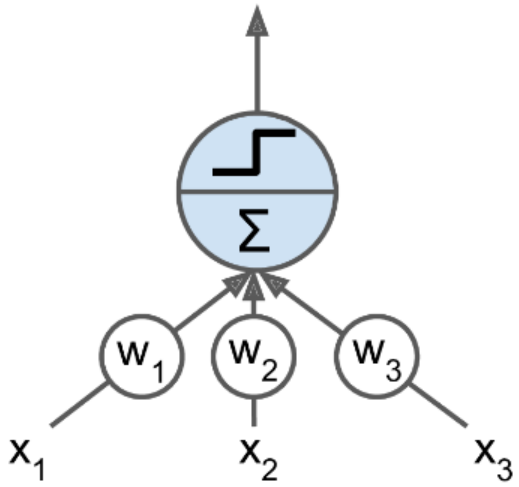
TensorFlow is a library for "computation graphs".

Neural nets are inspired by neurons in the brain.

They're used for image classification, speech recognition, and also simple ML problems.



a computation graph



an artificial neuron

Summary

Intro to Data Science:

- R
- probability view of machine learning

This class:

- Python
- optimization view of machine learning