CSE/STAT 416

Ethics, Explainable ML
Course Review

Vinitra Swamy
University of Washington
Aug 17, 2020

Announcements

- Homework 7 grades have been released
- Congratulations to Team Otterhog! + creative team names
  - weak learner
  - machine teachers
  - indecisive tree classifiers
  - beff jezos
  - deep mind 2.0
  - confused matrix
  - ham in spam
- Final Exam is on Wednesday on Gradescope
  - Have scratch paper handy
  - Compile your notes / study materials
  - Work fast (if you're getting stuck on a problem, skip it, and come back to it later)

Fairness

ML Ethics

Assessing Accuracy

Always dig in and ask critical questions of your accuracy.

- Is there a **class imbalance**?
- How does it compare to a baseline approach?
  - Random guessing
  - Majority class
  - …
- Most important: **What does my application need?**
  - What's good enough for user experience?
  - What is the impact of a mistake we make?

Which is Worse?

What's worse, a false negative or a false positive?

- It entirely depends on your application!

Detecting Spam

False Negative: Annoying

False Positive: Email lost

Medical Diagnosis

False Negative: Disease not treated

False Positive: Wasteful treatment

In almost every case, how treat errors depends on your context.

When ML goes wrong...

COMPAS, the algorithm used for recidivism prediction produces much higher false positive rate for African American people than Caucasian people. (Larson et al. ProPublica, 2016)

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

XING, a job platform, was found to rank less qualified male candidates higher than more qualified female candidates (Lahoti et al. 2018)

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

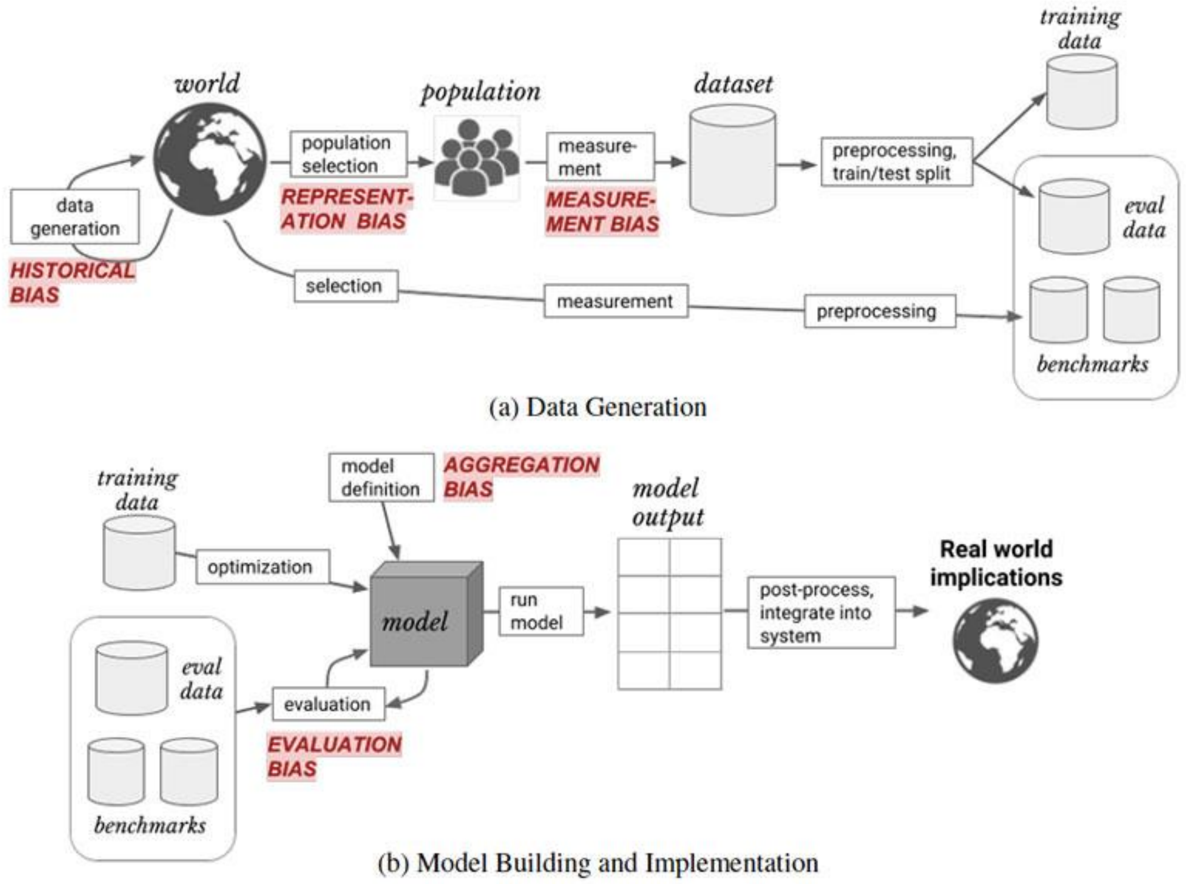
TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

5 sources of bias in ML

- **Historical bias** arises when the world, as it is, is biased
  - “CEO” Image Search
- **Representation bias** occurs when some groups of the population are underrepresented in the training data
  - ImageNet: 45% US, 1% China
- **Measurement bias** arises when there are granularity issues with measuring features of interest.
  - GPA -> Student Success
- **Aggregation bias** occurs when a one-size-fits-all model is used for groups that have different conditional distributions
  - Hb1ac level for diabetes across different populations
- **Evaluation bias** arises when evaluation and/or benchmark datasets are not representative of the target population
  - Facial Recognition datasets

A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING,
BY HARINI SURESH AND JOHN V. GUTTAG, 2019

Bias in the ML pipeline



Solutions for mitigating bias need to be tailored to the source of the bias.

Fairness

- Can we define fairness?
  - Legal precedent
    - Civil Rights Act of 1964, Title VII, Equal Employment Opportunities, 1964.
  - Barocas and Selbst, 2016
    - **disparate treatment**: decisions based on sensitive attributes
    - **disparate impact**: outcomes disproportionately hurt or benefit people with certain sensitive attribute values

Fairness Frameworks

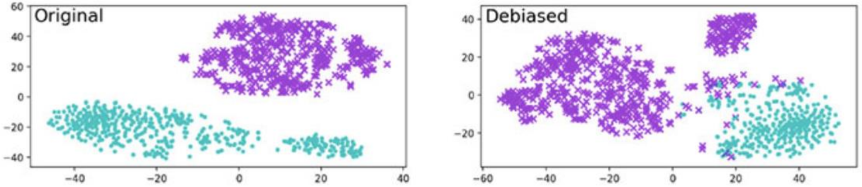
- No consensus on the mathematical definition of fairness
- Many papers have attempted to add frameworks
  - unawareness
  - demographic parity
  - equalized odds
  - predictive rate parity
  - individual fairness
  - counterfactual fairness
- Tradeoff between accuracy and fairness

Research directions in de-biasing ML

- Debiasing ML when you don't have access to private data
  - ["race", "gender", "age"] are protected attributes
  - Let's take advantage of the societal biases in names.

[WHAT'S IN A NAME? REDUCING BIAS IN BIOS WITHOUT ACCESS TO PROTECTED ATTRIBUTES](), BY ALEXEY ROMANOV ET. AL

- Bias in word embeddings exist
  - Debiasing methods by zero-ing projection onto the gender direction conceals the bias instead of removing it



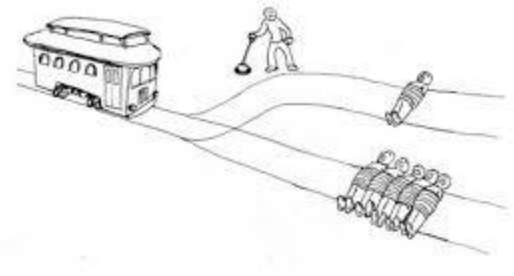
(b) Clustering for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

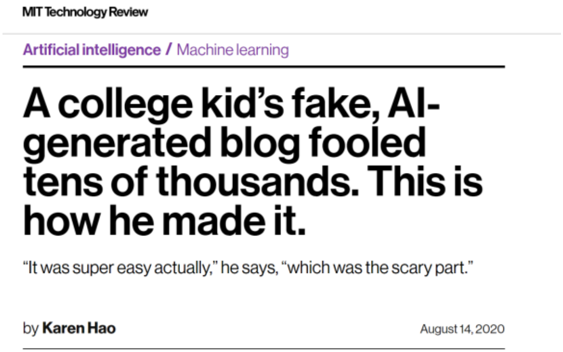
LIPSTICK ON A PIG: DEBIASING METHODS COVER UP SYSTEMATIC GENDER BIASES IN WORD EMBEDDINGS BUT DO NOT REMOVE THEM, BY HILA GONEN AND YOAV GOLDBERG

Other interesting ethics thoughts

- What effect does deep learning research have on the planet?
  - “training one model on GPU, with tuning and experimentation, results in CO_2 emissions that are comparable to the **two-year carbon footprint** of an average American”
  - “cloud computing costs for developing a state-of-the-art NLP model may account for **$103–350K**”

ENERGY AND POLICY CONSIDERATIONS FOR DEEP LEARNING IN NLP, BY EMMA STRUBELL, ANANYA GANESH, ANDREW MCCALLUM





MIT Technology Review

Artificial intelligence / Machine learning

A college kid’s fake, AI-generated blog fooled tens of thousands. This is how he made it.

“It was super easy actually,” he says, “which was the scary part.”

by Karen Hao

August 14, 2020

Poll Everywhere

Think

pollev.com/cse416

- Select the text sample that was generated by GPT-2.

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctiv
These four-horned, silver-white unicorns were previously u

Now, after almost two centuries, the mystery of what spark
is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the Univer
several companions, were exploring the Andes Mountains whe
valley, with no other animals or humans. Pérez noticed tha
appeared to be a natural fountain, surrounded by two peaks
silver snow.

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

For today’s homework assignment, please describe the reasor
Civil War.

By Donny Ferguson

It is easy to identify why the Civil War hap
many books and so much television and films
it has something to do with race or economic
agreement that it was essentially a war of s
a century of slavery. But that’s not what mo
the Civil War. Many people think of the war
wrong about that, but it’s kind of misleadin
conflict between states’ rights and federali
idea that has inspired so many people to go
and Antietam and Gettysberg and Petersburg a
of the republic--a notion of limited governm

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred Lines’ on the front and ‘Fashion Police’ on the back.

Scroll down for video

Poll Everywhere

Pair

pollev.com/cse416

- Select the text sample that was generated by GPT-2.

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

"The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation," said Tom Hicks, the U.S. Energy Secretary, in a statement. "Our top priority is to secure the theft and ensure it doesn't happen again."

For today's homework assignment, please describe the reasons for the US Civil War.

By Donny Ferguson

It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion. There is much agreement that it was essentially a war of slavery on behalf of capitalism, about a century of slavery. But that's not what most people think of when they think of the Civil War. Many people think of the war as about states rights. You're not wrong about that, but it's kind of misleading to say that the Civil War was a conflict between states' rights and federalism. So let's try again. What's the idea that has inspired so many people to go onto the battlefields of Gettysburg and Antietam and Gettysberg and Petersburg and Fredericksburg? The American idea of the republic--a notion of limited government--is a great part of the history.

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label 'Blurred Lines' on the front and 'Fashion Police' on the back.

Scroll down for video

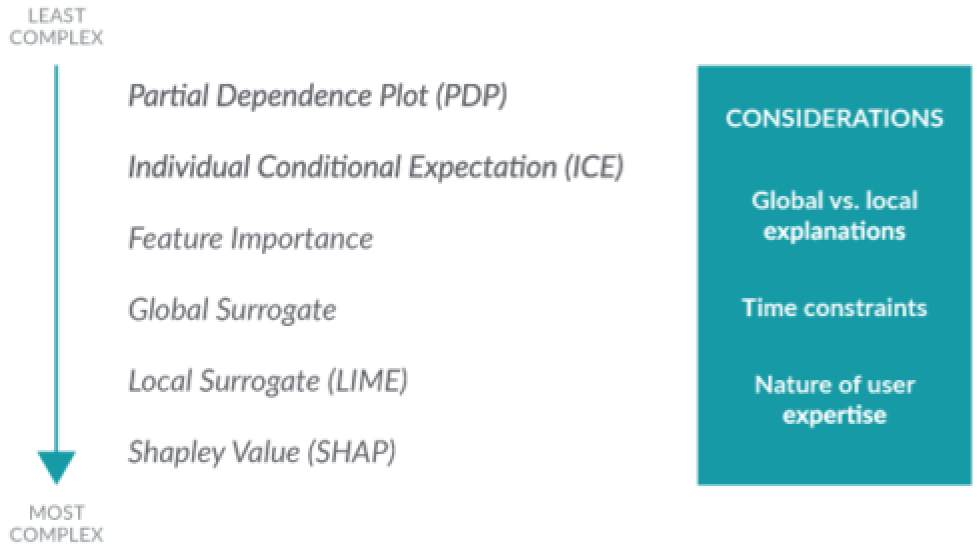
Explanations

Interpretable ML

Doge, our vacuum cleaner, got stuck. As an explanation for the accident, Doge told us that it needs to be on an even surface.

Interpretability

- What is an explanation?
- Interpretable Models
- Model-Agnostic Methods
- Example-Based Explanations
- Future of Interpretability



Explanations

- **What Is an explanation?**
  - An explanation is the answer to a why-question (Miller 2017)
    - Why did the treatment not work on the patient?
    - Why was my loan rejected?
    - Why have we not been contacted by alien life yet?

- **What is a “good” explanation?**
  - Humanities research can help us out!
  - Good explanations are contrastive (Lipton 1990)
  - Good explanations are selected
  - Good explanations are context-driven
  - Good explanations focus on the abnormal
  - Good explanations are general and probable

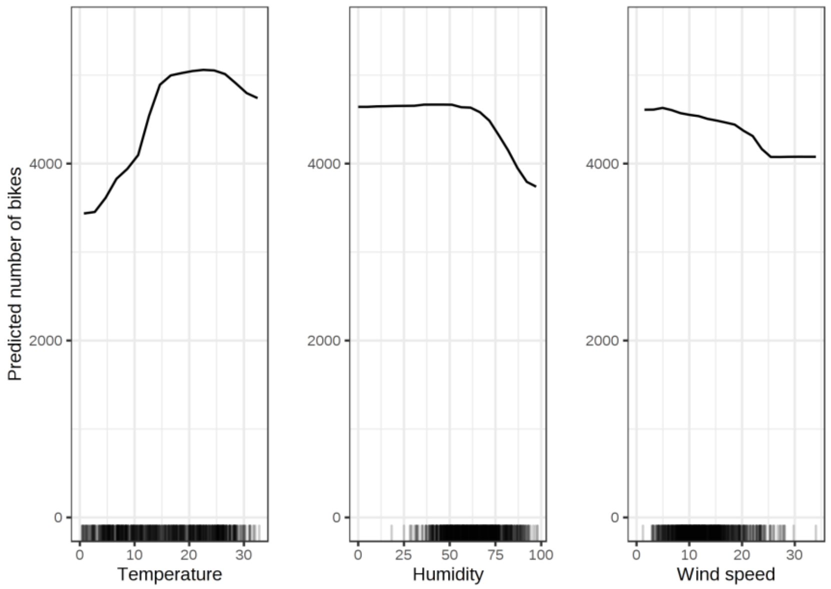
Interpretable Models

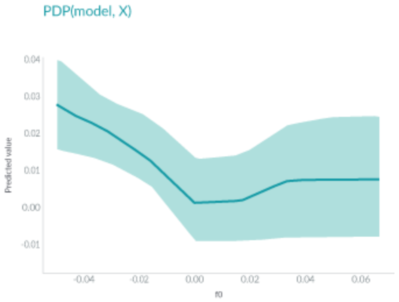
- Linear Regression
- Logistic regression
- Decision Trees
- Naïve Bayes
- K Nearest Neighbors

(GLM, RuleFit, Decision Rules)

Model-Agnostic Methods: PDP

- The partial dependence plot (PDP plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model (J. H. Friedman 2001).

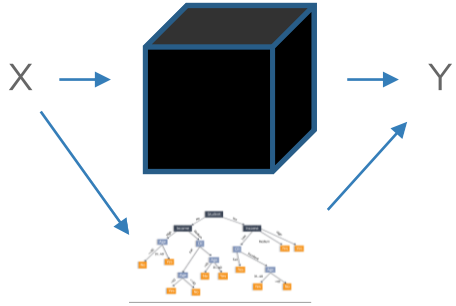




pro: easy to implement
con: avg. line hides heterogenous effects

Model-Agnostic Methods: Surrogate Models

- Global Surrogate Model
  - An interpretable model is trained to approximate the prediction of a black box model.

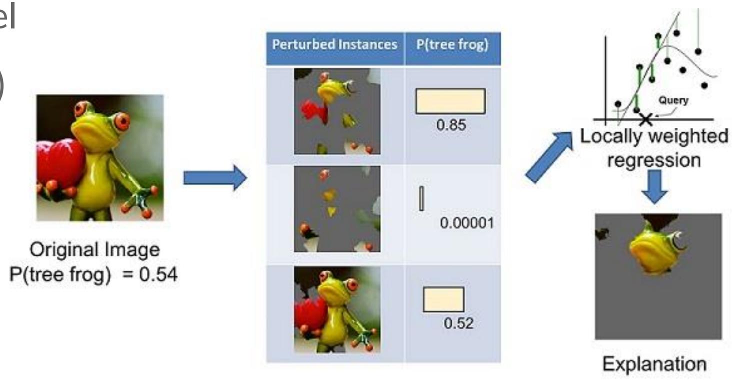


pro: flexible, intuitive
con: draw conclusions about model, not about data -- can differ by subset of dataset

- Local Surrogate Model

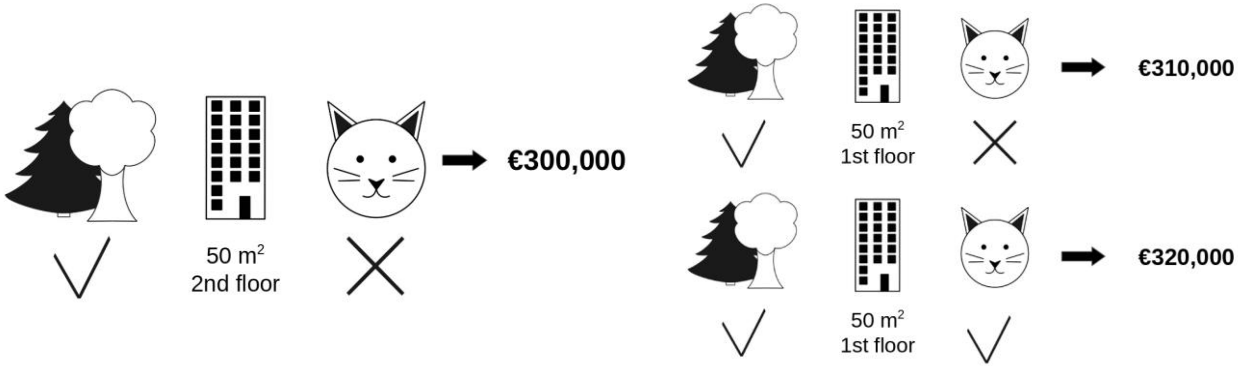
LIME (Riberio et. al, 2016)

pro: human-centric explanations
con: instability, hard to define “neighborhood”



Model-Agnostic Methods: Shapley values

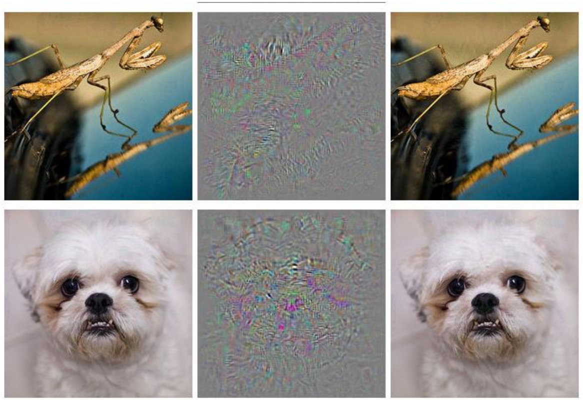
- Shapley value draws from game theory.
  - Each feature value of the instance is a “player” in a game.
  - The contribution of each player is measured by adding/removing the player from all subsets of players.
  - The Shapley Value for one player is the weighted sum of all its contributions.
- If you add up the Shapley Values of all the features, plus the base value, which is the prediction average, you will get the exact prediction value.



pro: good explanations, fairly distributed
con: lot of computing time, no prediction model

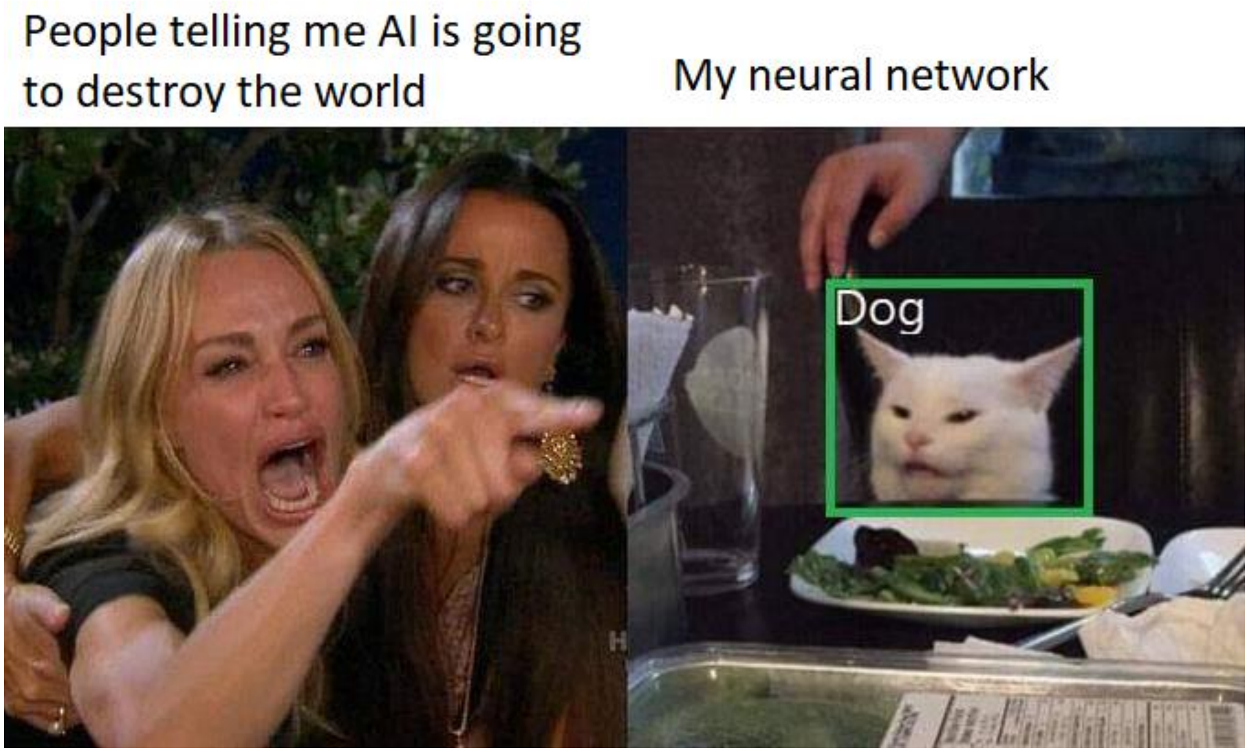
Example-based explanations

- Adversarial examples
  - small, intentional feature perturbations that cause a machine learning model to make a false prediction



- Counterfactual examples
  - "If X had not occurred, Y would not have occurred“
  - describes the **smallest change to the feature values** that changes the prediction to a predefined output

Brain Break


People telling me AI is going to destroy the world
My neural network
Dog

CSE/STAT 416

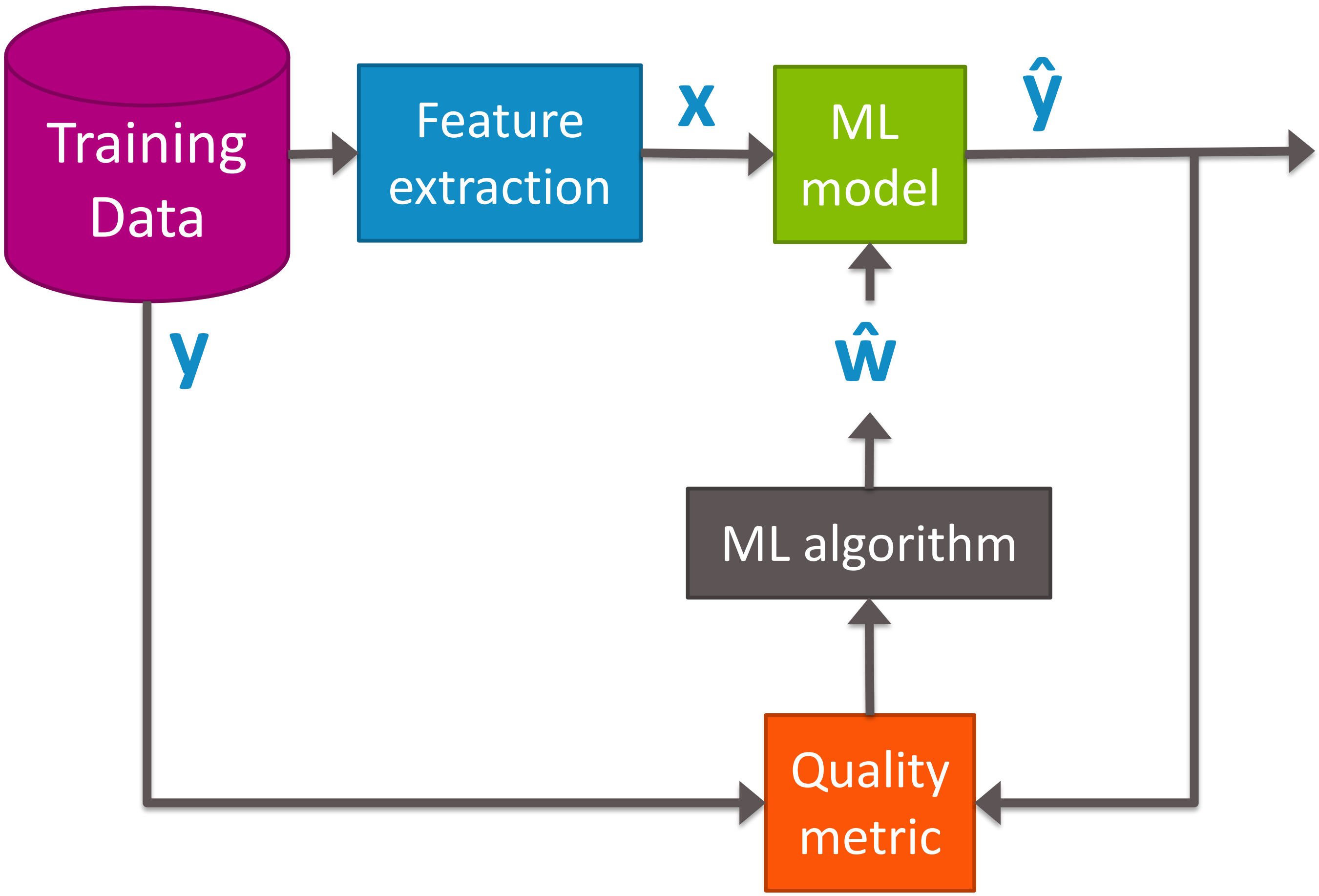
Course Wrap Up

Vinitra Swamy
University of Washington
Aug 17, 2020

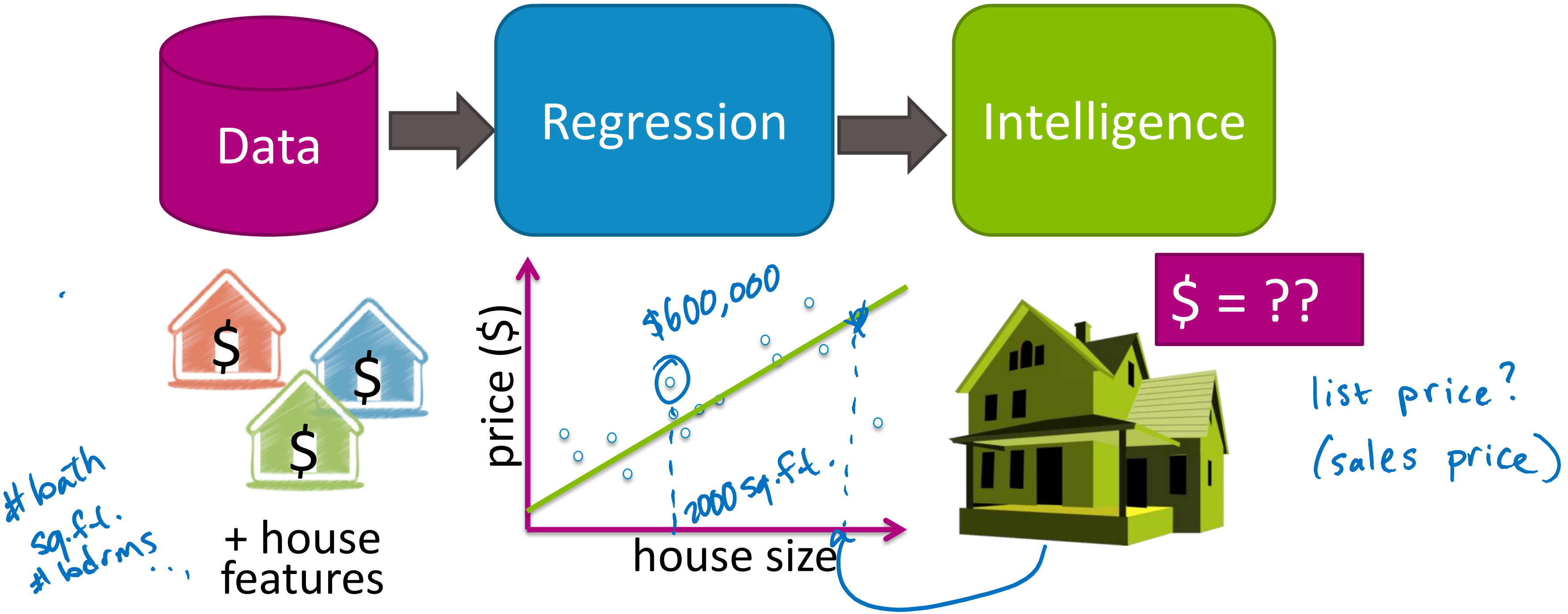
Slides borrowed from Emily Fox

One Slide

- Regression
- Overfitting
- Training, test, generalization error
- Bias-Variance tradeoff
- Ridge/LASSO regularization
- Cross validation
- Gradient Descent
- Classification
- Logistic Regression
- Decision trees
- Boosting
- Precision and recall
- Nearest-neighbor retrieval, regression, and classification
- Kernel regression
- Locality Sensitive Hashing
- Dimensionality Reduction (PCA)
- K-means clustering
- Hierarchical clustering
- Supervised vs. unsupervised learning
- Recommender systems
- Matrix Factorization (NMF)
- Coordinate Descent
- Neural Networks
- Convolutional Neural Networks
- Transfer learning


Training Data
Feature extraction
x
ML model
ŷ
y
ŵ
ML algorithm
Quality metric

Case Study 1:
Predicting house prices



Regression

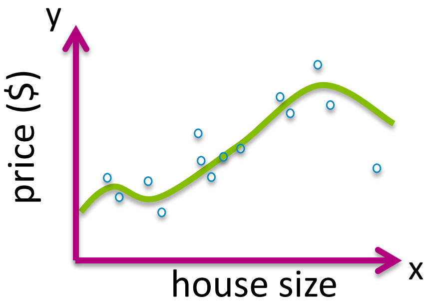
Case study: Predicting house prices

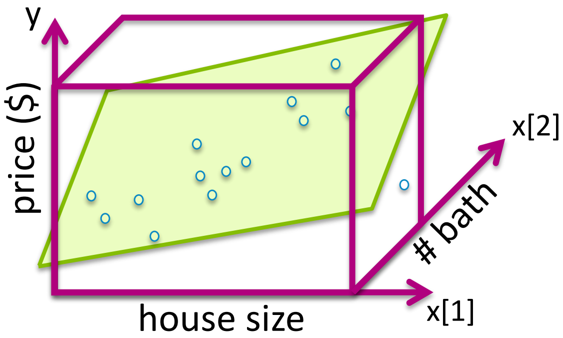
Models

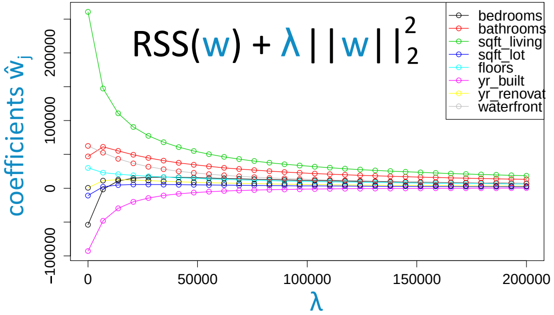
- Linear regression
- Regularization:
  Ridge (L2), Lasso (L1)

Including many features:

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- …







Regression

Case study: Predicting house prices

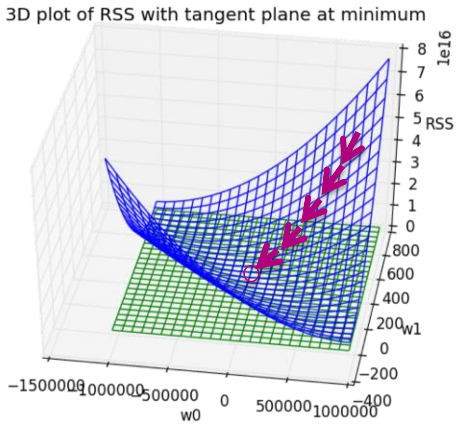
Algorithms

- Gradient descent

$$\text{RSS}(w_0,w_1) = (\$_{\text{house }1}-[w_0+w_1\text{sq.ft.}_{\text{house }1}])^2 + (\$_{\text{house }2}-[w_0+w_1\text{sq.ft.}_{\text{house }2}])^2 + (\$_{\text{house }3}-[w_0+w_1\text{sq.ft.}_{\text{house }3}])^2 + \ldots \text{[include all houses]}$$

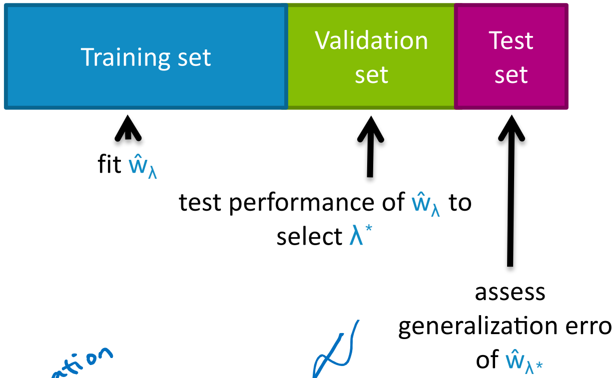
$$\downarrow$$

$$\hat{w}$$



Regression

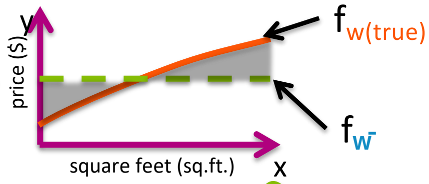
Case study: Predicting house prices

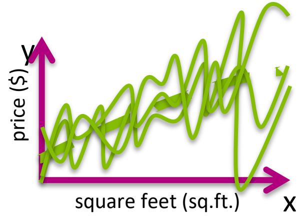


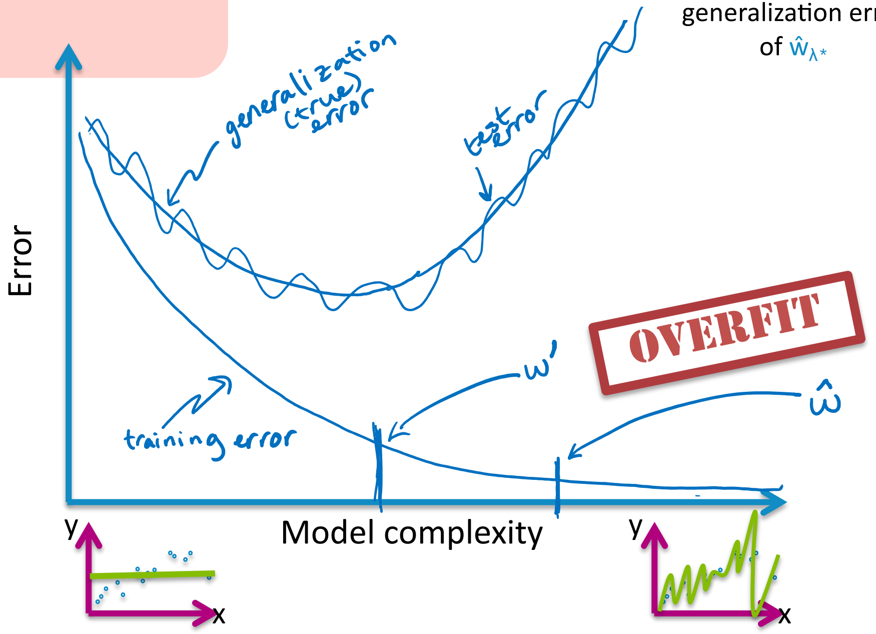
Concepts

- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

1. Noise
2. Bias
3. Variance







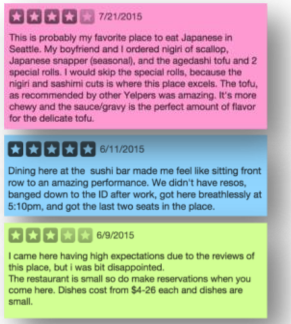
Case Study 2:
Sentiment analysis

Sushi was awesome,
the food was awesome,
but the service was awful.

stars / +/-

text

All reviews:


7/21/2015

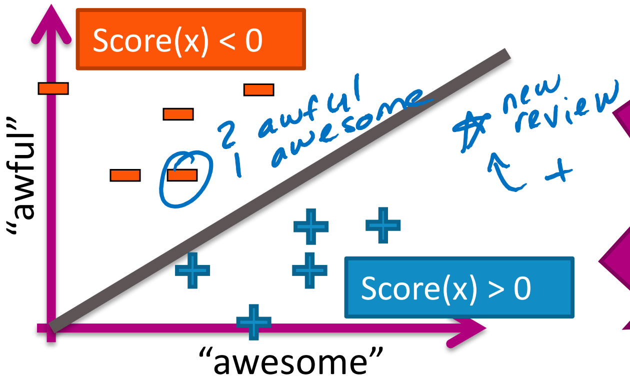
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

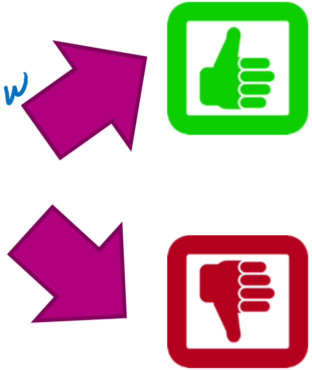
6/11/2015

Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have resos, banged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

6/9/2015

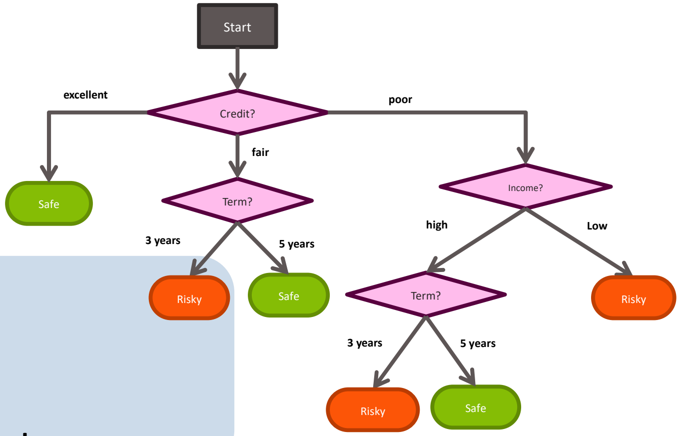
I came here having high expectations due to the reviews of this place, but i was bit disappointed.
The restaurant is small so do make reservations when you come here. Dishes cost from $4-26 each and dishes are small.





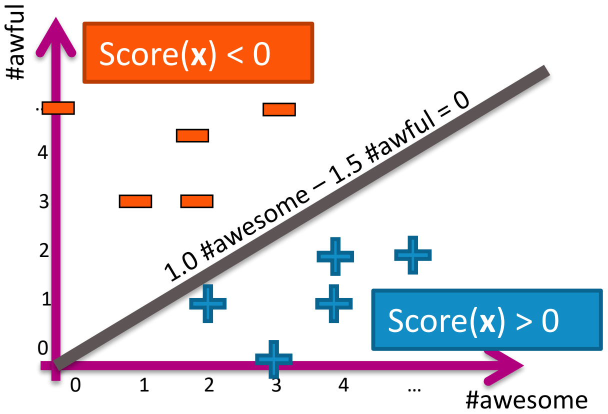
Classification

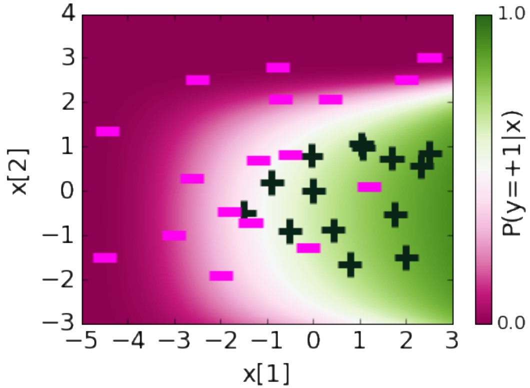
Case study: Analyzing sentiment



Models

- Linear classifiers (logistic regression)
- Multiclass classifiers
- Decision trees
- Boosted decision trees and random forests



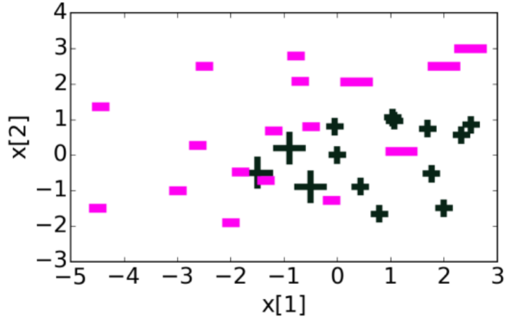


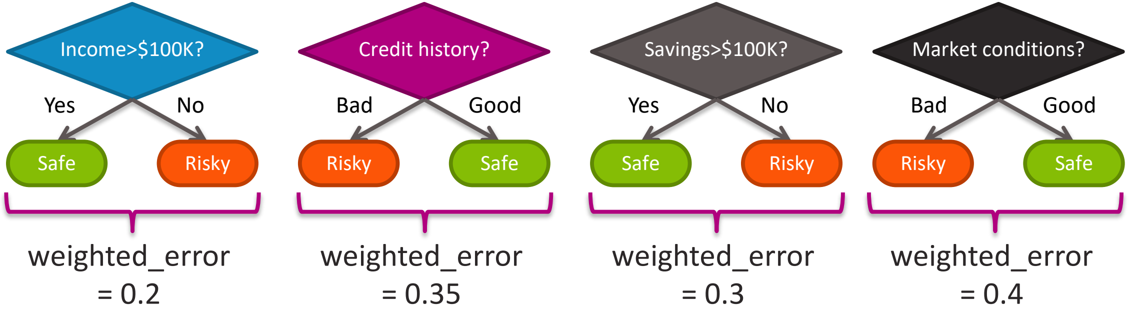
Classification

Case study: Analyzing sentiment

Algorithms

- Boosting
- Learning from weighted data



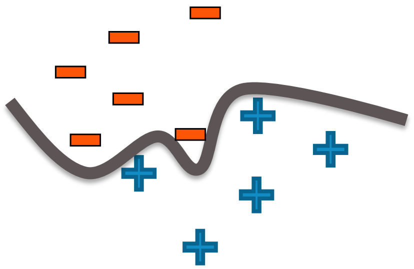


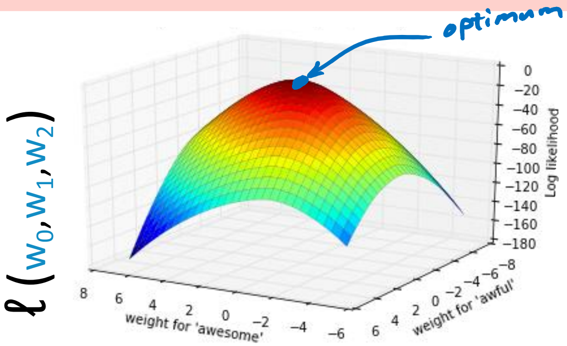
Classification

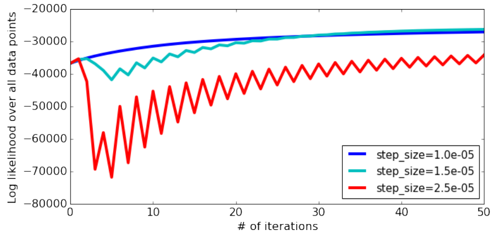
Case study: Analyzing sentiment

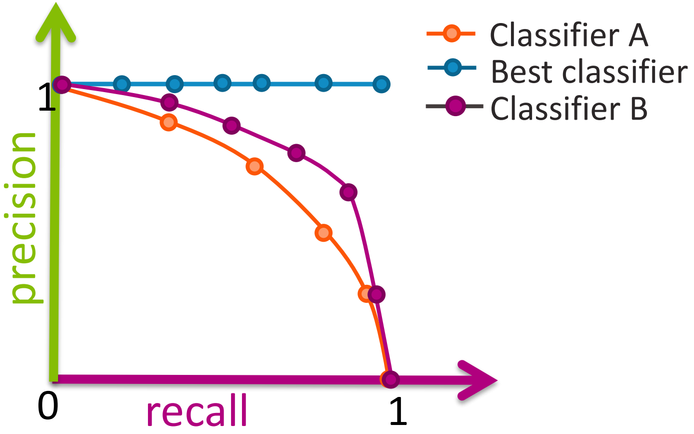
Concepts

- Decision boundaries, maximum likelihood estimation, ensemble methods, random forests
- Precision and recall









Case Study 3:
Document retrieval

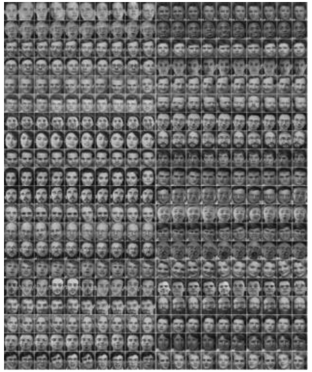
Case Study 3+:

Document structuring for retrieval

Case Study 3++:
Dimensionality reduction

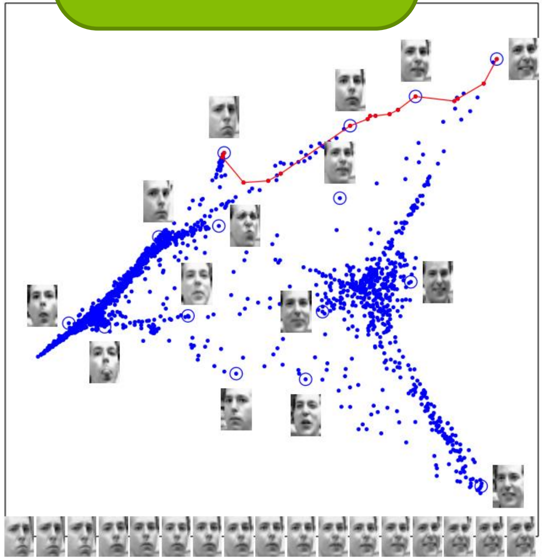
Data → PCA → Intelligence

simple



Images with thousands or millions of pixels

Can we give each image a coordinate, such that similar images are near each other?



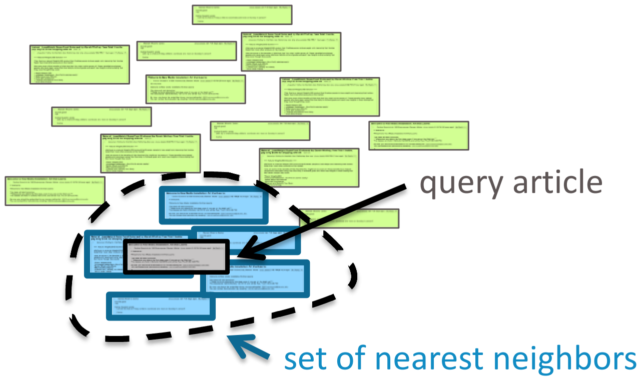
[Saul & Roweis '03]

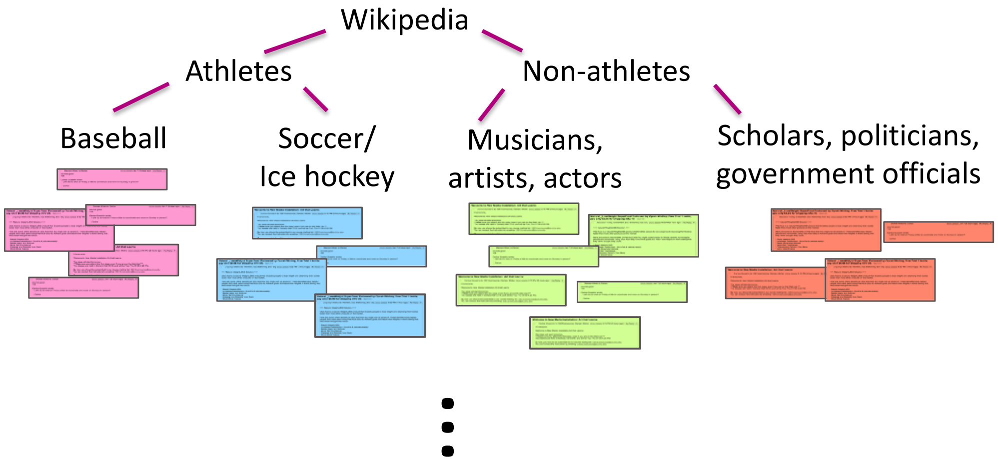
Clustering & Retrieval

Case study: Finding documents

Models

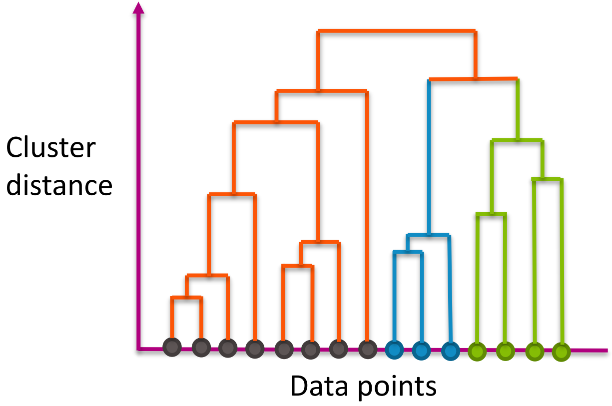
- Nearest neighbors
- Clustering
- Hierarchical clustering





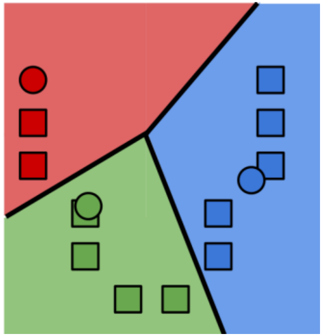
Clustering & Retrieval

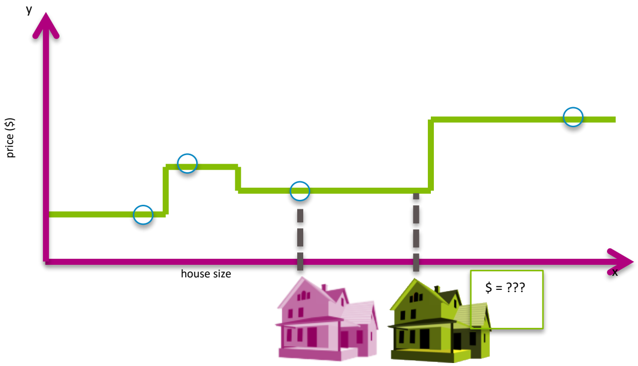
Case study: Finding documents

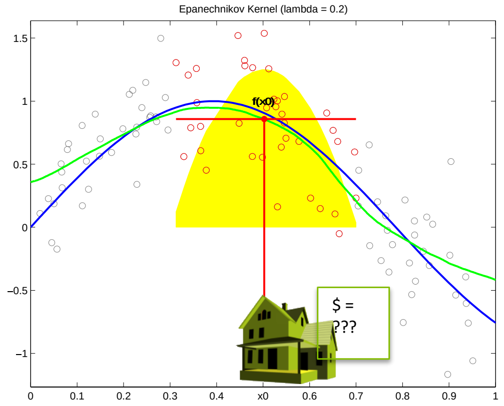


Algorithms

- k-means
- Locality-sensitive hashing (LSH)
- NN regression and classification
- Kernel regression
- Agglomerative and divisive clustering
- PCA

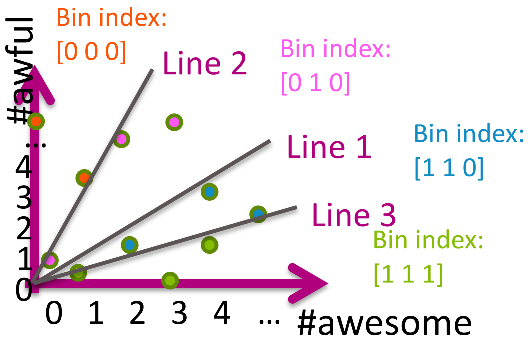






Clustering & Retrieval

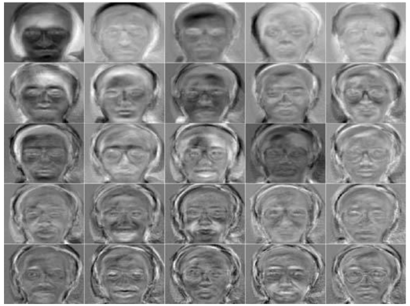
Case study: Finding documents



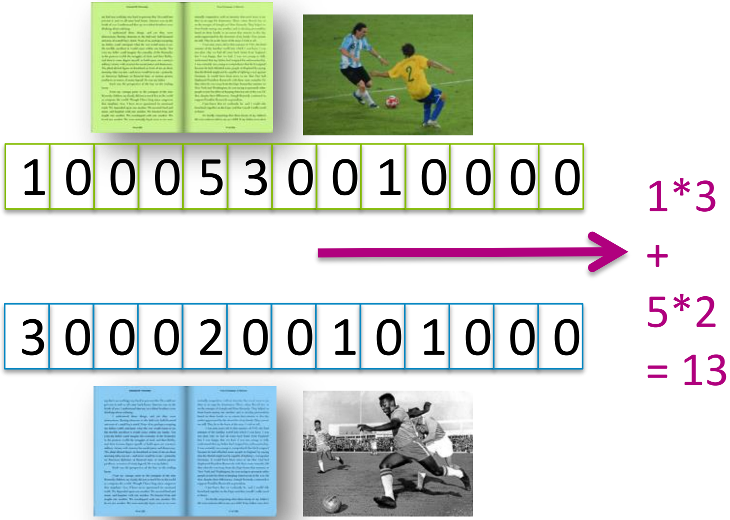
Concepts

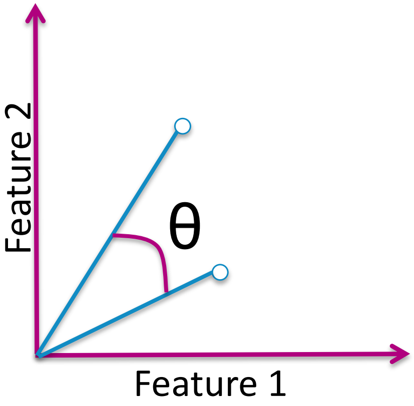
- Distance metrics, kernels, approximation algorithms, dimensionality reduction

Principal components:

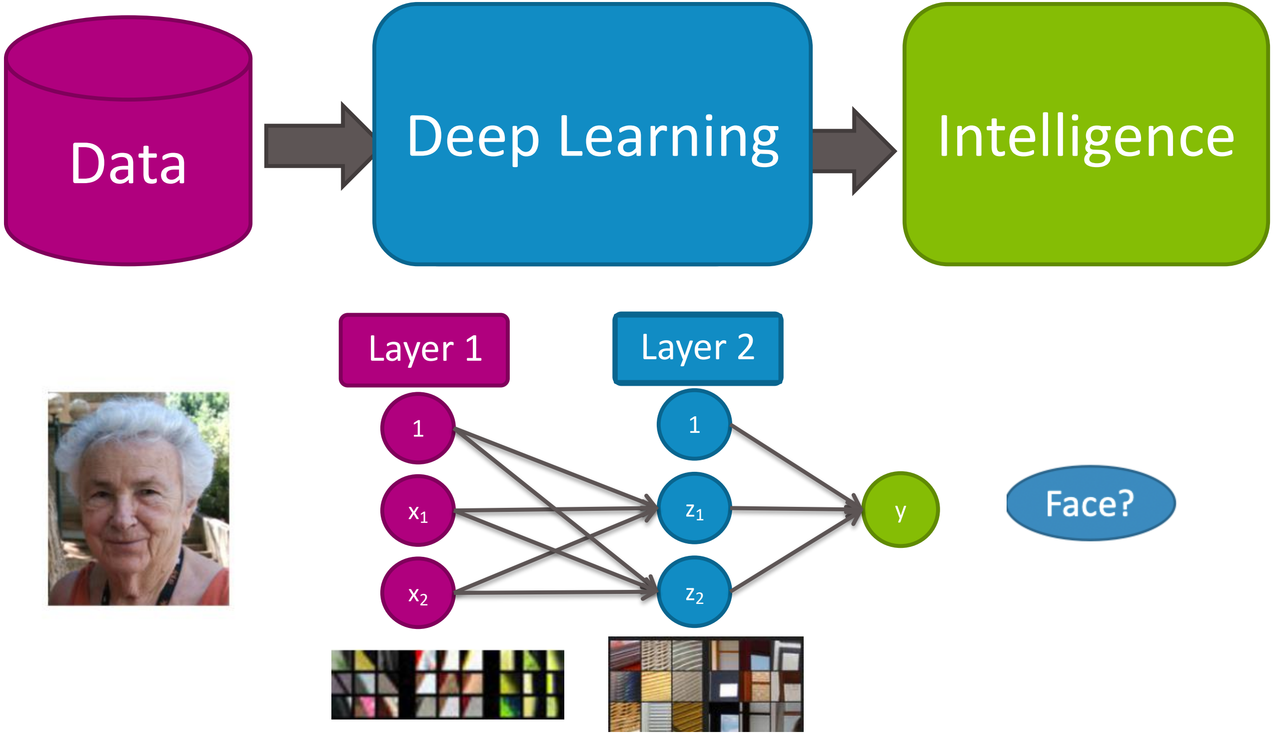


Reconstructing:



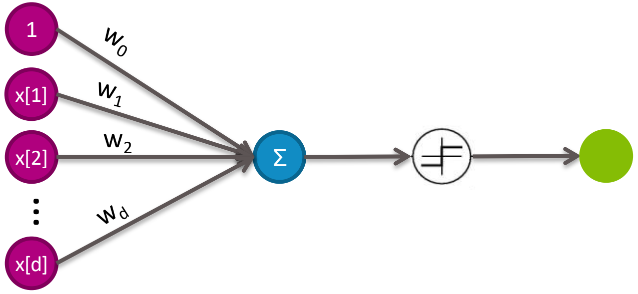


Case Study 4:
Image classification



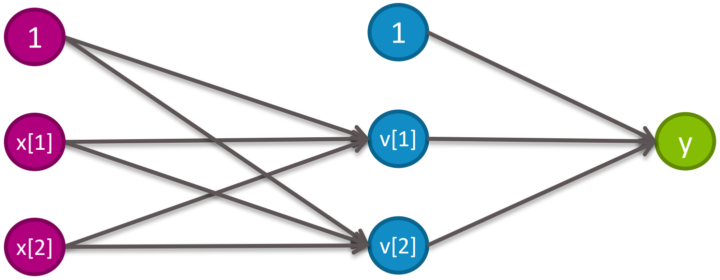
Deep Learning

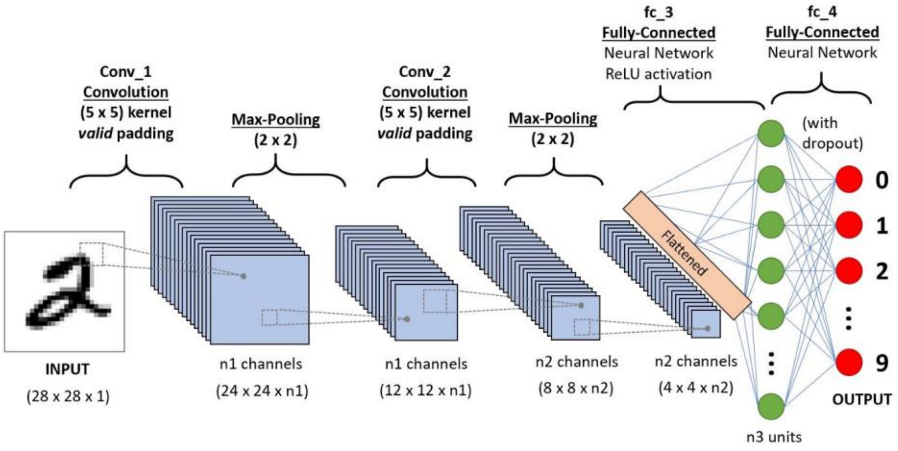
Case study: Image classification



Models

- Perceptron
- General neural network
- Convolutional neural network



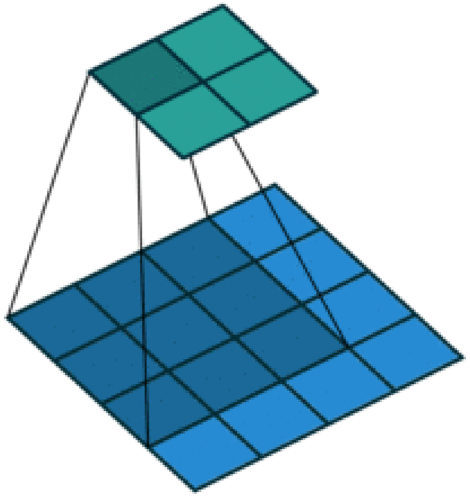


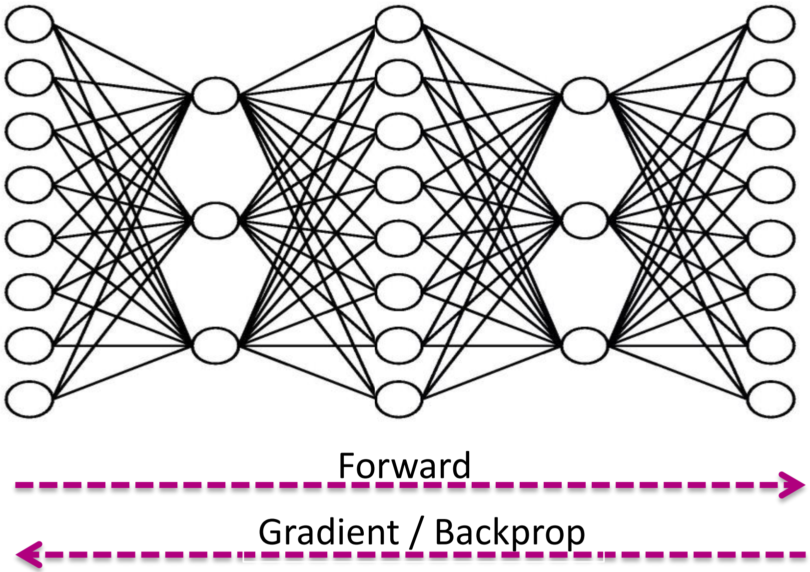
Deep Learning

Case study: Image classification

Algorithms

- Convolutions
- Backpropagation (high level only)



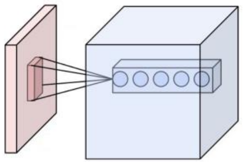


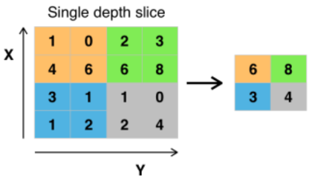
Deep Learning

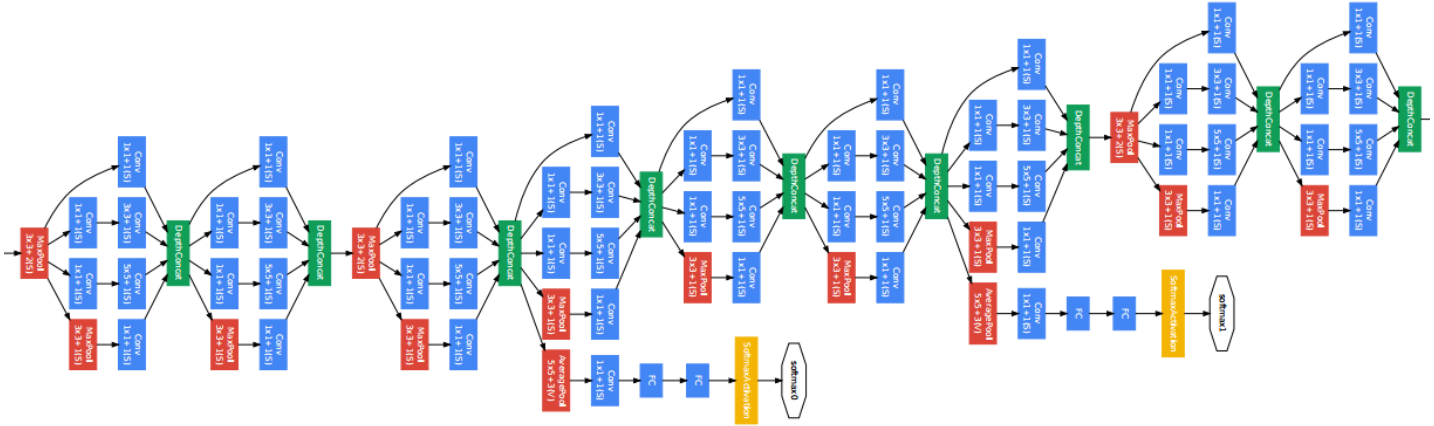
Case study: Image classification

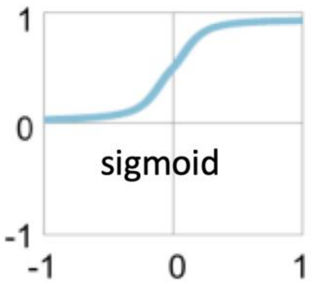
Concepts

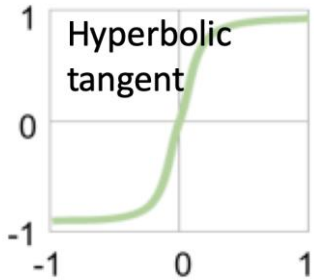
- Activation functions, hidden layers, architecture choices

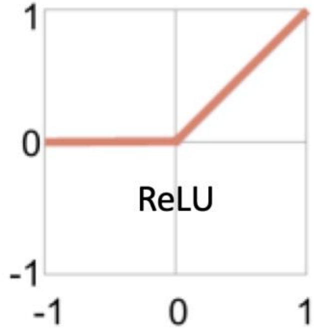








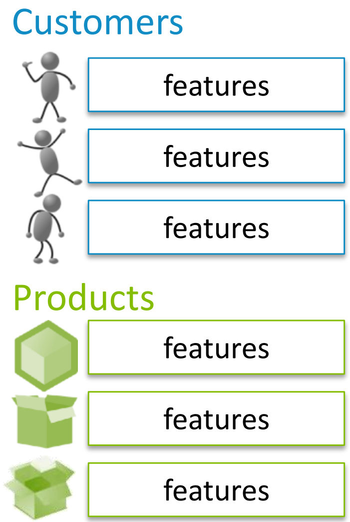




Case Study 5:
Product recommendation

Your past purchases:

+ purchase histories of all customers



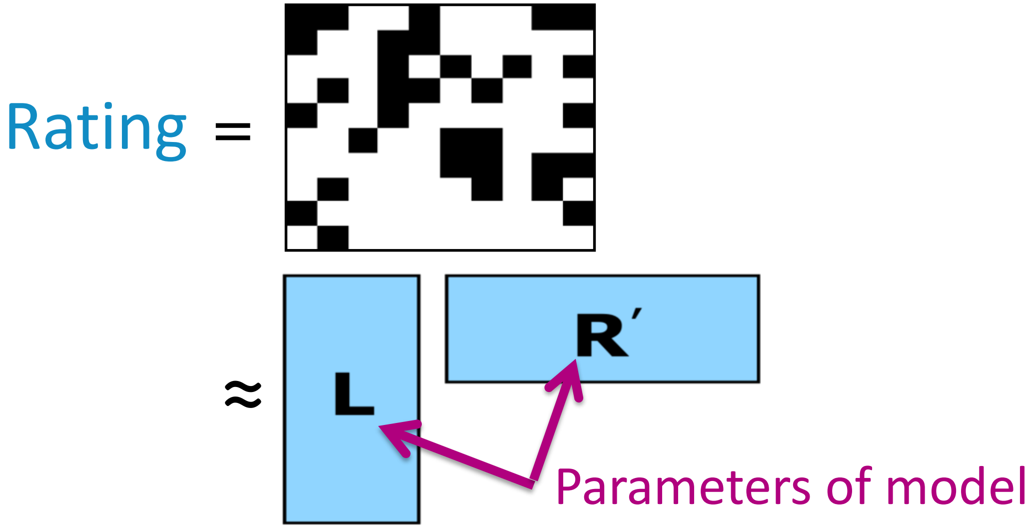
Recommended items:

Recommender Systems & Matrix Factorization

Case study: Recommending Products

Models

- Collaborative filtering
- Matrix factorization

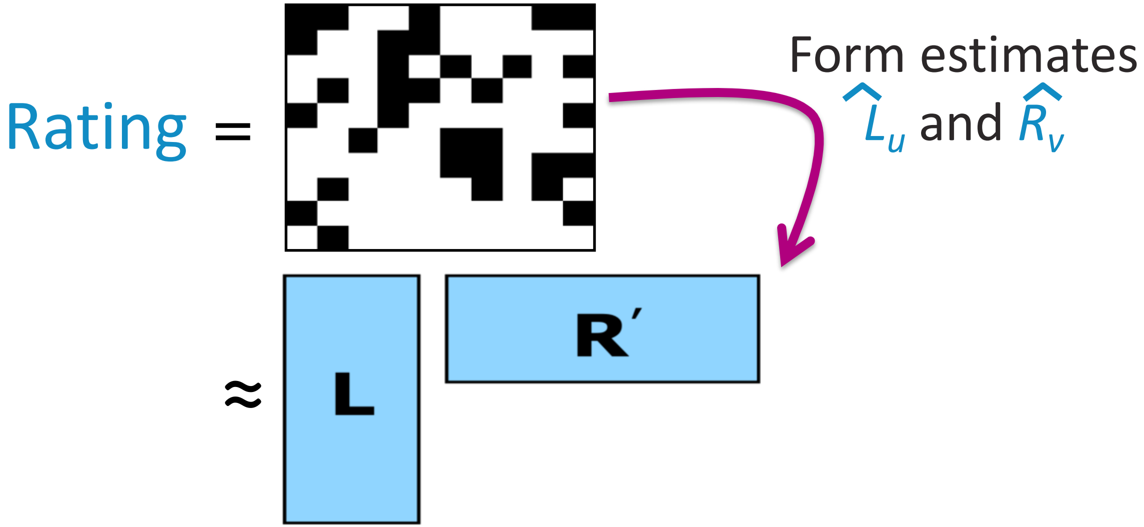


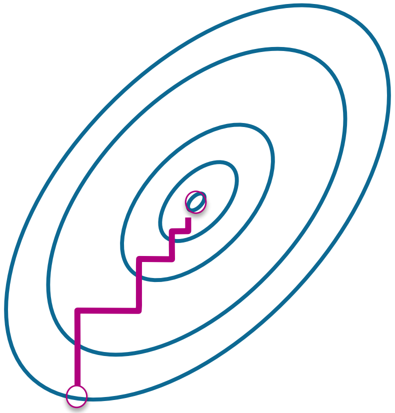
Recommender Systems & Matrix Factorization

Case study: Recommending Products

Algorithms

- Coordinate descent



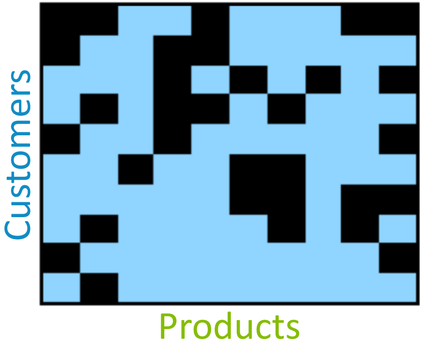


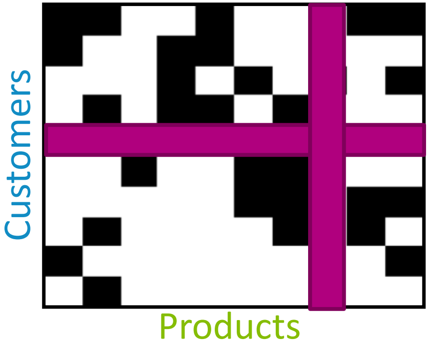
Recommender Systems & Matrix Factorization

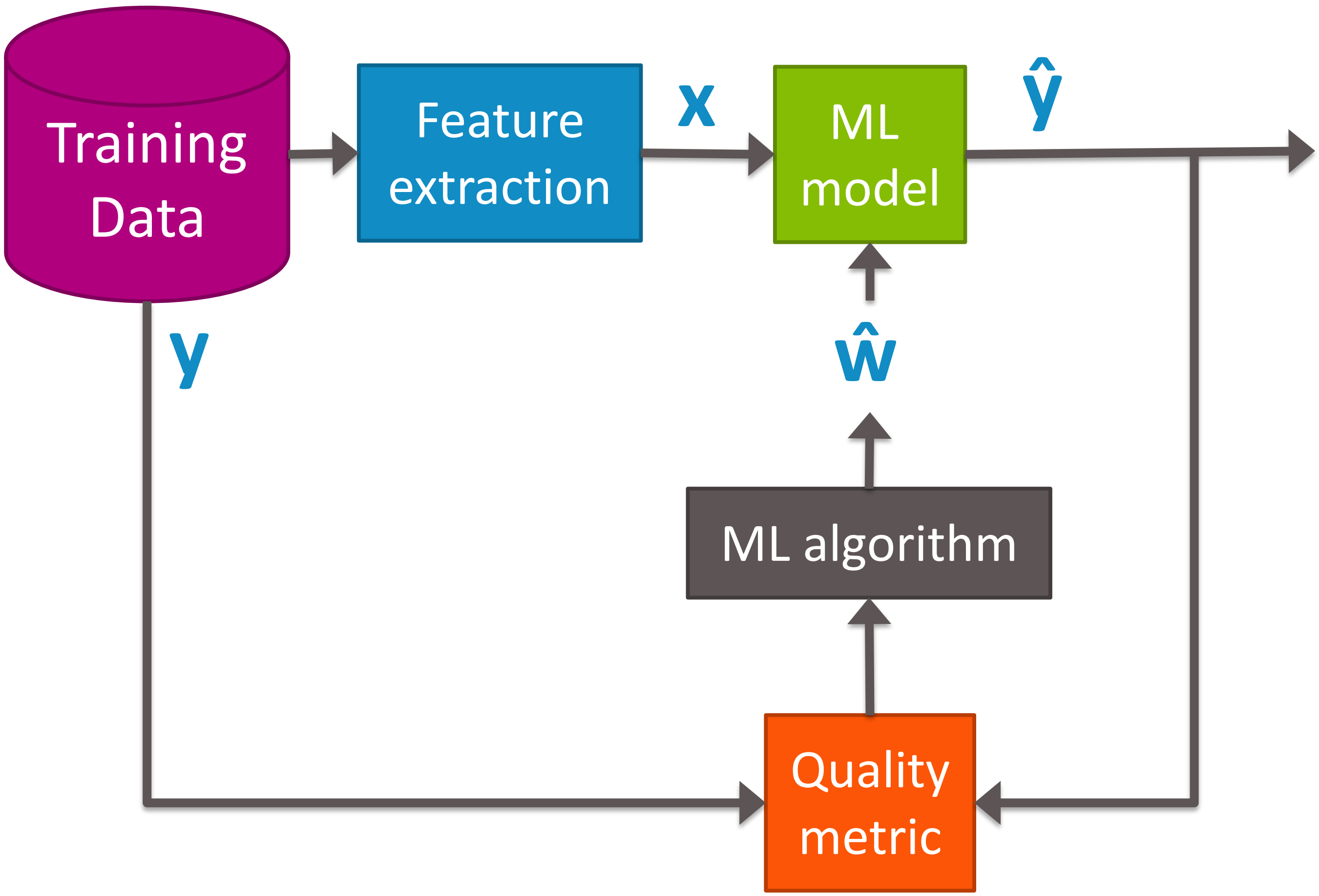
Case study: Recommending Products

Concepts

- Matrix completion, cold-start problem






Training Data
Feature extraction
x
ML model
ŷ
y
ŵ
ML algorithm
Quality metric

Big Picture

Improving the performance at some task through experience!

- Before you start any learning task, remember fundamental questions that will impact how you go about solving it

What is the learning problem?

From what experience?

What model?

What loss function are you optimizing?

Are there any guarantees?

With what optimization algorithm?

How will you evaluate the model?

What consequences does your model have?

Congrats on finishing CSE/STAT 416!
Thanks for the hard work!