Statistics, Data Science, and Machine Learning: A Very Brief Introduction

Professor Edgar Dobriban[1]

Department of Statistics, The Wharton School, University of Pennsylvania

August 11, 2018

[1]e-mail: dobriban@wharton.upenn.edu

Overview

Overview

The Age of Data

- We live in the *Age of Data*
- An enormous variety of digital data is being generated and recorded every day

Figure: A datacenter

The Age of Data

- Examples:

The Age of Data

- Examples:
  - Internet:

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce:

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science:

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science: Bio-medical data: health records, genetics, pharmaceuticals; Astronomy; Physics;

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science: Bio-medical data: health records, genetics, pharmaceuticals; Astronomy; Physics;
  - Government:

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science: Bio-medical data: health records, genetics, pharmaceuticals; Astronomy; Physics;
  - Government: census, demographics, economics, income, polls

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science: Bio-medical data: health records, genetics, pharmaceuticals; Astronomy; Physics;
  - Government: census, demographics, economics, income, polls
  - Computer vision:

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science: Bio-medical data: health records, genetics, pharmaceuticals; Astronomy; Physics;
  - Government: census, demographics, economics, income, polls
  - Computer vision: Images, movies

The Age of Data

- Examples:
  - Internet: Webpages, links, files, user activity, ads
  - Commerce: Product prices and quantities, transactions, descriptions, reviews, user history
  - Science: Bio-medical data: health records, genetics, pharmaceuticals; Astronomy; Physics;
  - Government: census, demographics, economics, income, polls
  - Computer vision: Images, movies
  - Many others...

Thoughts on the Data Age

- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)

Thoughts on the Data Age

- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)
- Some of the data is *big*, but small data is still valuable.

Thoughts on the Data Age

- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)
- Some of the data is *big*, but small data is still valuable.
- Many successful businesses are highly data-centric

Thoughts on the Data Age

- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)
- Some of the data is *big*, but small data is still valuable.
- Many successful businesses are highly data-centric
  - e-commerce (Amazon, Alibaba)

Thoughts on the Data Age

- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)
- Some of the data is *big*, but small data is still valuable.
- Many successful businesses are highly data-centric
  - e-commerce (Amazon, Alibaba)
  - advertising and social (Facebook, WeChat)

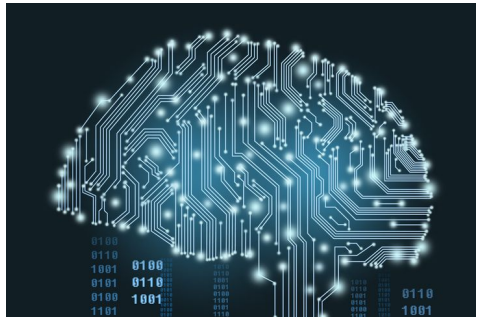
Thoughts on the Data Age

- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)
- Some of the data is *big*, but small data is still valuable.
- Many successful businesses are highly data-centric
  - e-commerce (Amazon, Alibaba)
  - advertising and social (Facebook, WeChat)
  - financial (hedge funds, credit agencies)

Thoughts on the Data Age

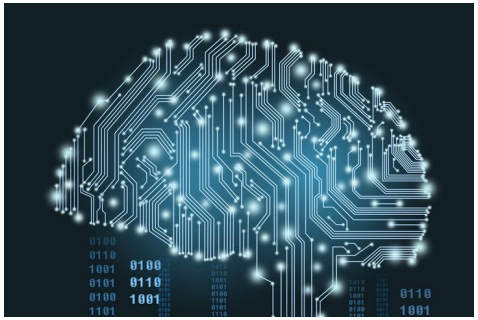
- Data has been around for a long time, but it is now increasingly common (easy to record, store, share, analyze, present)
- Some of the data is *big*, but small data is still valuable.
- Many successful businesses are highly data-centric
  - e-commerce (Amazon, Alibaba)
  - advertising and social (Facebook, WeChat)
  - financial (hedge funds, credit agencies)
  - transportation (Uber, Didi)

The Age of Data (ctd)



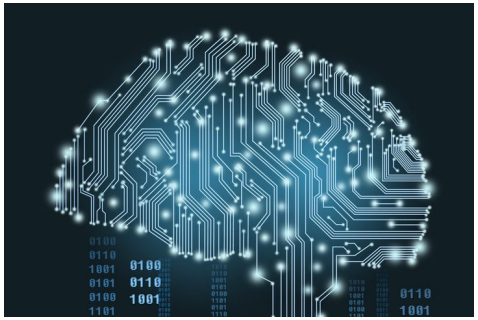
- However, there are surely great opportunities left to use data better

The Age of Data (ctd)



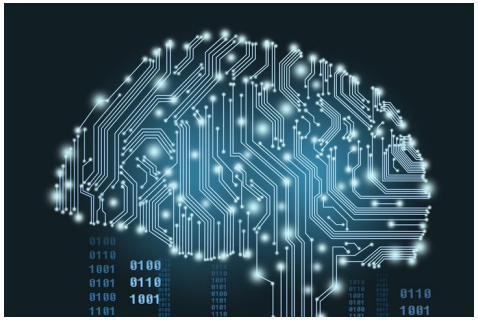
- However, there are surely great opportunities left to use data better
  - online education (eg personalized degrees)

The Age of Data (ctd)



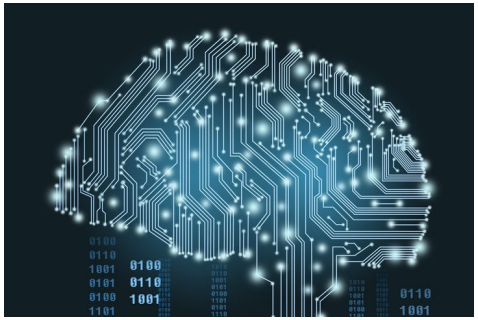
- However, there are surely great opportunities left to use data better
  - online education (eg personalized degrees)
  - augmented intelligence (eg ...)

The Age of Data (ctd)



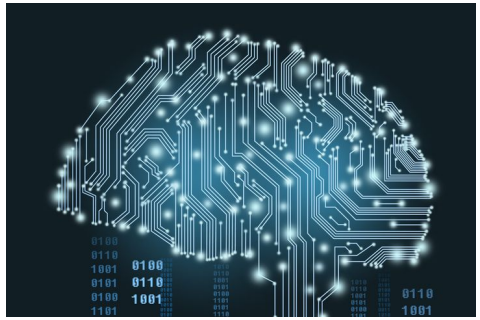
- However, there are surely great opportunities left to use data better
  - online education (eg personalized degrees)
  - augmented intelligence (eg ...)
  - health tracking and recommendation
  - ...

The Age of Data (ctd)



- However, there are surely great opportunities left to use data better
  - online education (eg personalized degrees)
  - augmented intelligence (eg ...)
  - health tracking and recommendation
  - ...
- Many great businesses will be started to solve these problems

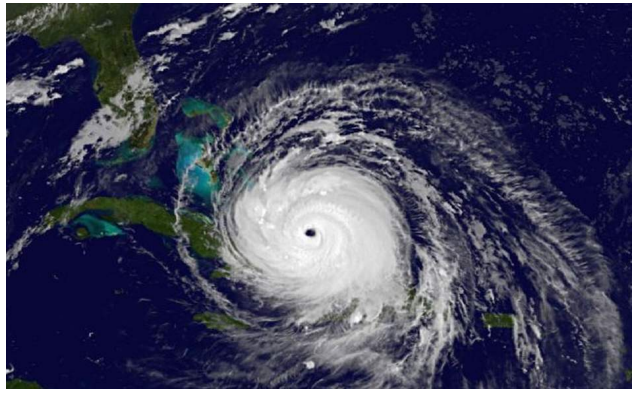
The Age of Data (ctd)



- However, there are surely great opportunities left to use data better
  - online education (eg personalized degrees)
  - augmented intelligence (eg ...)
  - health tracking and recommendation
  - ...
- Many great businesses will be started to solve these problems
- They will use tools from statistics, data science, and machine learning
- Good news: same tools can solve many different problems

Overview

Example: Hurricane Frances[2]

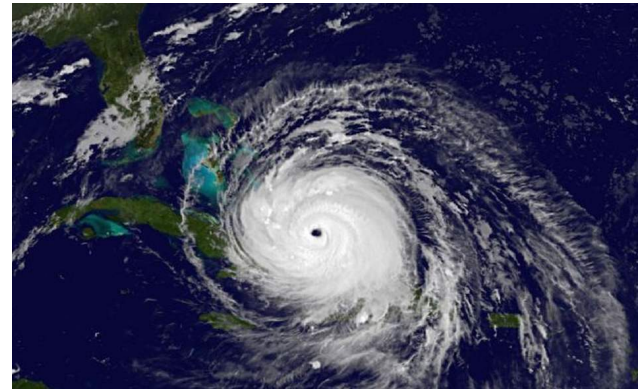


- New York Times story from 2004:

  Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Arkansas, executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven methods: predictive technology.

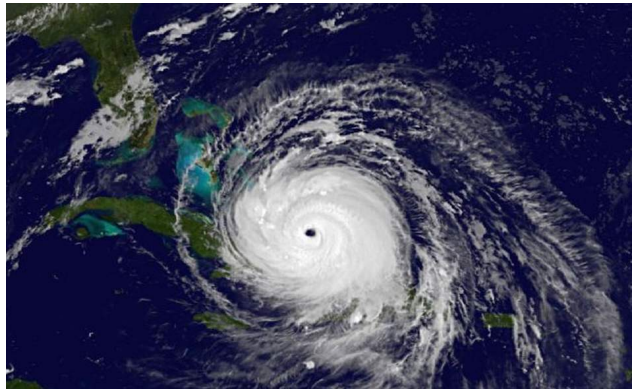
[2]Many examples borrowed from Provost & Fawcett - Data Science for Business. Rights belong to the owners.

Example: Hurricane Frances (ctd)



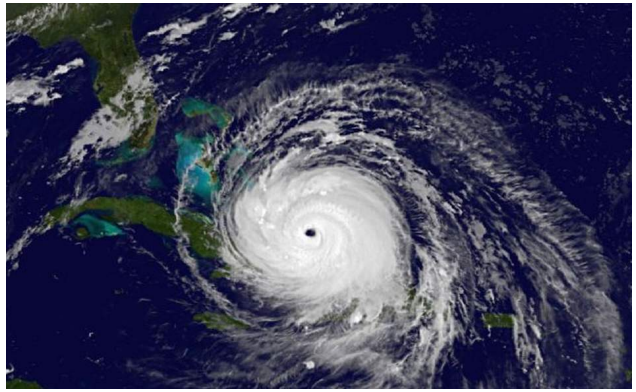
A week ahead of the storm's landfall, Linda Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could "start predicting what is going to happen, instead of waiting for it to happen," as she put it. (Hays, 2004)

Example: Hurricane Frances (ctd)



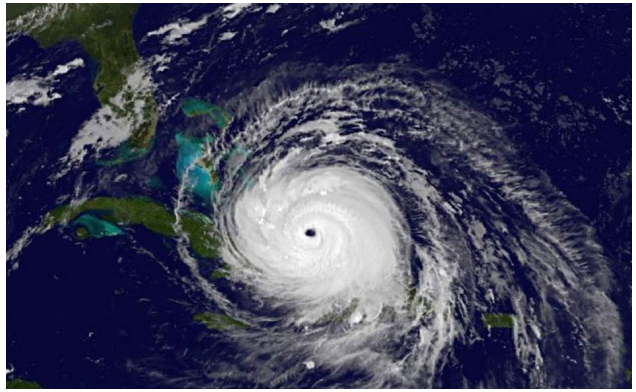
- What are they trying to do? Why?
- What do they expect to find?

Example: Hurricane Frances (ctd)



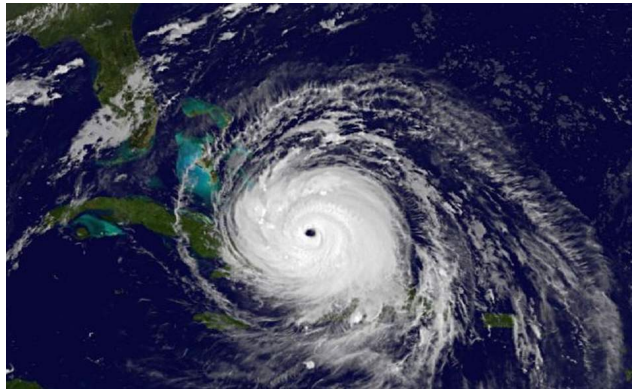
- What are they trying to do? Why?
- What do they expect to find?
  - Forecasting and prediction: Sale of some supplies will increase.

Example: Hurricane Frances (ctd)



- What are they trying to do? Why?
- What do they expect to find?
  - Forecasting and prediction: Sale of some supplies will increase.
  - How much? When? (non-obvious but crucial)

Example: Hurricane Frances (ctd)



- What are they trying to do? Why?
- What do they expect to find?
  - Forecasting and prediction: Sale of some supplies will increase.
  - How much? When? (non-obvious but crucial)
- How to do it?

Overview

Table of Contents

Prediction

- Can we make informed guesses about what is likely to happen?
  - e.g., forecast demand, earnings
  - ad click prediction: will a user click on an ad?
- Physics, chemistry: study highly structured problems, find simple laws

$$F = ma$$

- Real-life phenomena are complex, unlikely to be governed by simple laws
- Instead, try to find approximate model (not "correct", but "useful")

$$Y \approx f(X)$$

- This is possibly the most important current topic in stats/ML/data science, so we will focus on this

Prediction: general formulation

- Also known as the "Supervised learning" problem.
- Each datapoint has two parts:
  - outcome (to predict): e.g., did the user click on the ad?
  - features: e.g., user characteristics (age, gender, preferences), ad characteristics (product, company, topic)
- Goal: want to predict outcome of new example, aka *test data*
- Using a rule learned from old examples and outcomes, aka *training data*

Prediction: toy "ad click prediction" example

- Example data:

Datapoint	Age	Gender	Product	Clicked?
1	28	Male	Dress	No
2	23	Female	Skirt	Yes
3	53	Female	Skirt	No
4	31	Male	Coat	Yes
5	20	Female	Gloves	?

- First 4: training data. last: test data.
- What would you predict?
- Why?

Prediction: notation

- Organize data as follows:
  - n training samples total ($n = 4$ above)
  - Numbered $i = 1, 2, \ldots, n$
  - Data encoded numerically: e.g.,
    1. gender: male $= 0$, female $= 1$.
    2. products: dress $= 0$, skirt $= 1$, etc
  - Features of each user arranged in a vector x_i, e.g.,
    - (28, male, dress) $\rightarrow x_1 = (28, 0, 0)$
  - Outcome is a number y_i, e.g., $0 =$ no click, $1 =$ click
- Summary: We have training data (x_i, y_i), where x_i is a feature vector, and y_i is the outcome of interest.

Prediction: notation

- Want to construct a predictor f, such that
  - Given the features of a new sample x_t (test data)
  - We predict $\hat{y}_t = f(x_t)$, as a good guess for the unobserved outcome y_t
- How to construct f?

Prediction: summary of setup

- Given training data (x_i, y_i), $i = 1, \ldots, n$
- Want to construct a predictor f, to predict on future test samples: $\hat{y}_t = f(x_t)$

Table of Contents

How to do prediction and solve supervised learning problems?

- "Choose a class of predictors and find the best". In more detail:
- Choose a set of potential predictors f
  - e.g., Nearest neighbors, Linear models, Neural networks...
- Find the best one.
  - Optimize test error.

How to choose potential predictors?

- What aspects of the data will I use to construct predictors?
- e.g., In ad-click example: predict based on similar users
- e.g., In image classification:
  - extract higher-level features.
  - Animal vs plant: is there an eye in the picture?
- Choose a collection of potential predictors $f_1, f_2, \ldots$
  - look at 1,2,...? most similar users
  - Or: typically, infinitely many predictors f_θ, where θ is a vector $\theta \in \mathbb{R}^p$ of real numbers

How to find the best one?

- How to choose from the potential predictors f_θ?
- Choose the "best", or "best fit" one.
- How to measure quality? Introduce loss function $L(y, \hat{y})$.
  - Suppose the truth is y, and I predict $\hat{y}$
  - How much does that cost me?
  - e.g., if $y \in \{0, 1\}$, can use 0/1 loss

$$L(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases}$$

  - If I predict correctly, no cost
  - If I predict incorrectly, fixed cost

How to find the best one? ctd

- Choose "best" f_θ, with smallest loss function
- Loss on one training example (x_i, y_i) is

$$L(y_i, \hat{y}_i) = L(y_i, f_\theta(x_i))$$

- Total loss (known as risk) on dataset is

$$R(\theta) = L(y_1, f_\theta(x_1)) + \ldots + L(y_n, f_\theta(x_n)) = \sum_{i=1}^{n} L(y_i, f_\theta(x_i))$$

- Find predictor f_θ such that risk $R(\theta)$ is as small as possible

Overview

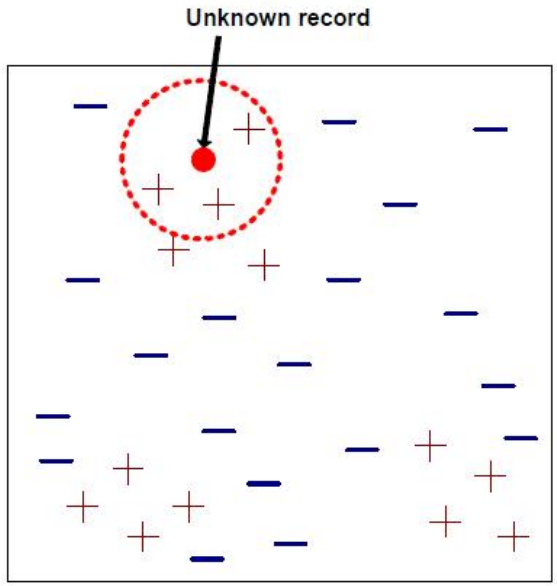
How to choose potential predictors?

- Why do we need to choose a set of potential predictors? Can't we consider *all* predictors?
- No, because of overfitting. There are infinitely many predictors that fit the train data perfectly.
- Choosing prediction methods = deciding what structure of the data we want to exploit.
  - Domain-specific. No "universal best"
  - E.g., Nearest neighbors, Linear models, Neural networks

Table of Contents

Nearest neighbors

- Predict the same as for most similar training example seen
- Or, average of k most similar training data samples



Nearest neighbors

- Formally: let $d(x, x')$ be a measure of distance on the features
  - e.g., usual "Euclidean" distance ($x[i]$ is i-th coordinate of x)

$$d(x, x') = \|x - x'\|_2 = \sqrt{(x[1] - x'[1])^2 + \ldots + (x[p] - x'[p])^2}$$

  - e.g., "Hamming distance", where recall $L()$ is the binary loss

$$d_H(x, x') = L(x[1], x'[1]) + \ldots + L(x[p], x'[p])$$

    how many differences are there between x, x'?

Nearest neighbors (ctd): 1-nearest neighbor

- For a test datapoint x_t, let x_j be the "nearest neighbor", minimizing distance, solving

$$\min_{i \in \{1,\ldots,n\}} d(x_t, x_i)$$

- Predict $\hat{y}_t = y_j$, outcome associated with x_j

1 - Nearest neighbor example

- Example: ad prediction (see before). Define

$$d(x, x') = |x[1] - x'[1]| + 10L(x[2], x[2]') + 10L(x[3], x[3]')$$

- So: absolute value for age; binary loss for gender, product
- What do we predict for x_5?

k-Nearest neighbors

- Similar, but take majority vote of the k nearest neighbors

Table of Contents

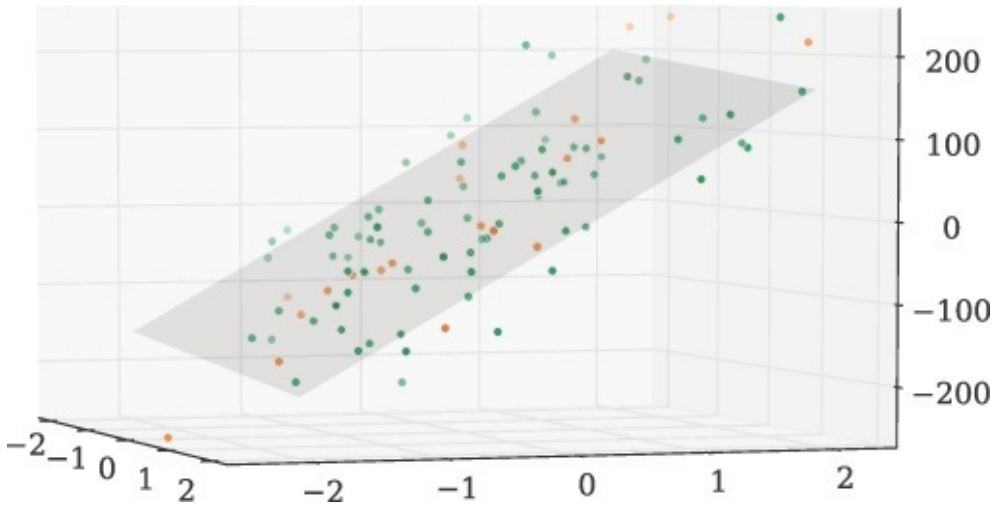
Linear models

- Model the outcome y as a linear function of the features:

$$y \approx x[1]\theta_1 + x[2]\theta_2 + \ldots + x[p]\theta_p$$

- The coefficients θ_i are the unknown parameters of interest.
- This makes sense when outcome is continous. When outcome is binary, model its probability. (logistic regression)

Linear models



$$y \approx x[1]\theta_1 + x[2]\theta_2$$

for all datapoints x_i, so

$$y_i \approx x_i[1]\theta_1 + x_i[2]\theta_2$$

Why linear models?

- Widely used and successful
- Linear function $x \to ax + b$ is the simplest nontrivial function.
- May not have data to estimate more complicated function
- Interpretable: $\theta_1 =$ effect of one extra unit of $x[1]$, given all other explanatory variables

$$y \approx x[1]\theta_1 + x[2]\theta_2 + \ldots + x[p]\theta_p$$

Change $x[1]$ to $x_{new}[1] = x[1] + 1$, get

$$y_{new} \approx y + \theta_1$$

Why linear models?

- Easy to estimate/fit: only need to solve linear equations
- "Statistical inference" (assessing uncertainty) easy
- Flexible: many other methods use it. for instance polynomial regression

$$y \approx x[1]\theta_1$$

$$\rightarrow$$

$$y \approx x[1]\theta_1 + x[1]^2\theta_2 + x[1]^3\theta_3$$

Table of Contents

Neural networks and deep learning

- Recently popular and very successful methods, with long history
- Used to process images, sounds, text etc, at many major companies

Neural networks and deep learning (NNs)

- A neural network is simply a multilayer nonlinear regression. aka "multilayer perceptron"
- Input data vector x: $x = a^{[0]}$.. called "activation"
- Construct hidden layers: activations $a^{[1]}, a^{[2]}, \ldots, a^{[L]}$

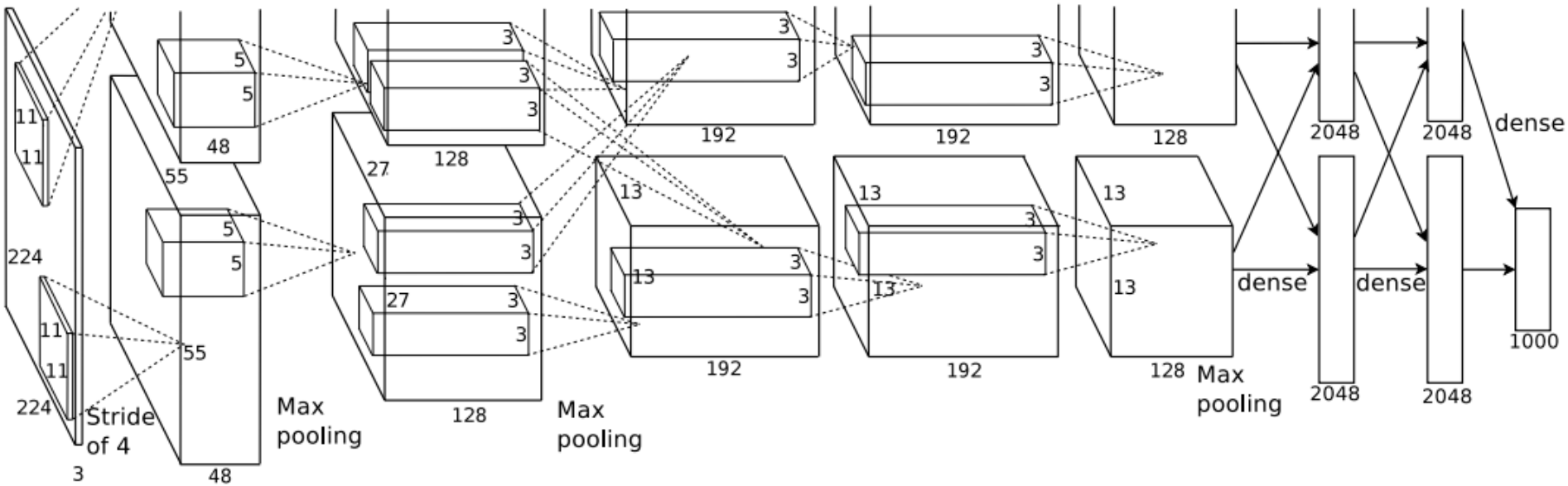


Figure: "AlexNet." Krizhevsky et al (2012, NIPS).

Intuition for neural networks

- NNs are nonlinear functions.
- Two steps at each layer:
  1. Linear combinations of the input features. (like linear model)
  2. Activation: when there is a sufficient amount of that feature, propagate to the next layer.

Neural networks

- At every layer, a linear filter is applied to the activations of the previous layer. computes "pre-activation"

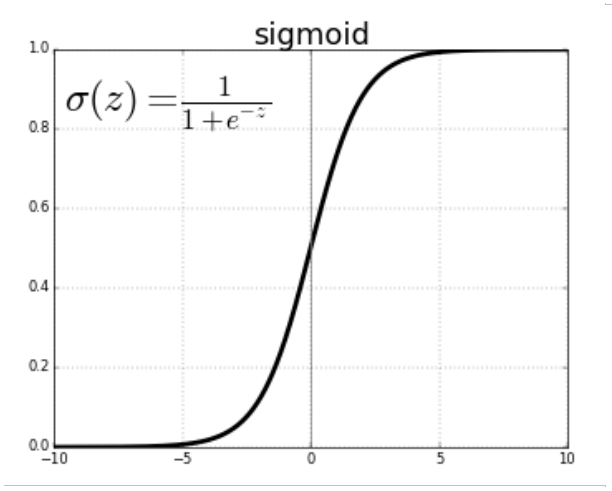
$$z^{[i]} = W^{[i]} a^{[i-1]} + b^{[i]}.$$

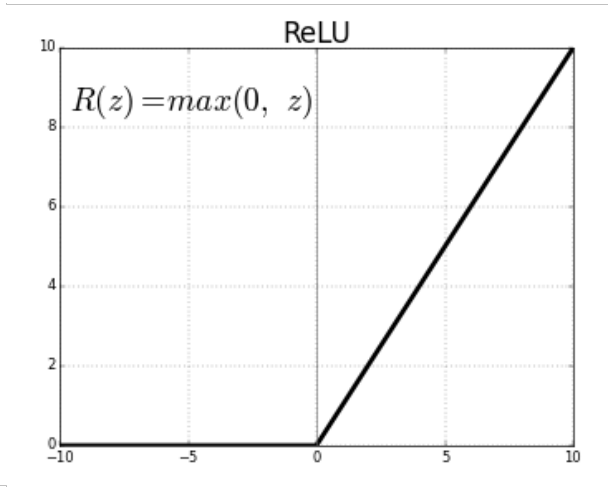
- Parameters: weight matrices $W^{[i]}$, bias vectors $b^{[i]}$
- Next, a nonlinear activation is applied to the coordinates of z-s:

$$a^{[i]} = \sigma(z^{[i]}).$$

Neural networks: activations

- Popular examples:
  1. rectified linear unit (ReLU) $\sigma(x) = \max(x, 0)$.
  2. sigmoid $\sigma(x) = 1/(1 + \exp(-x))$.





Neural networks

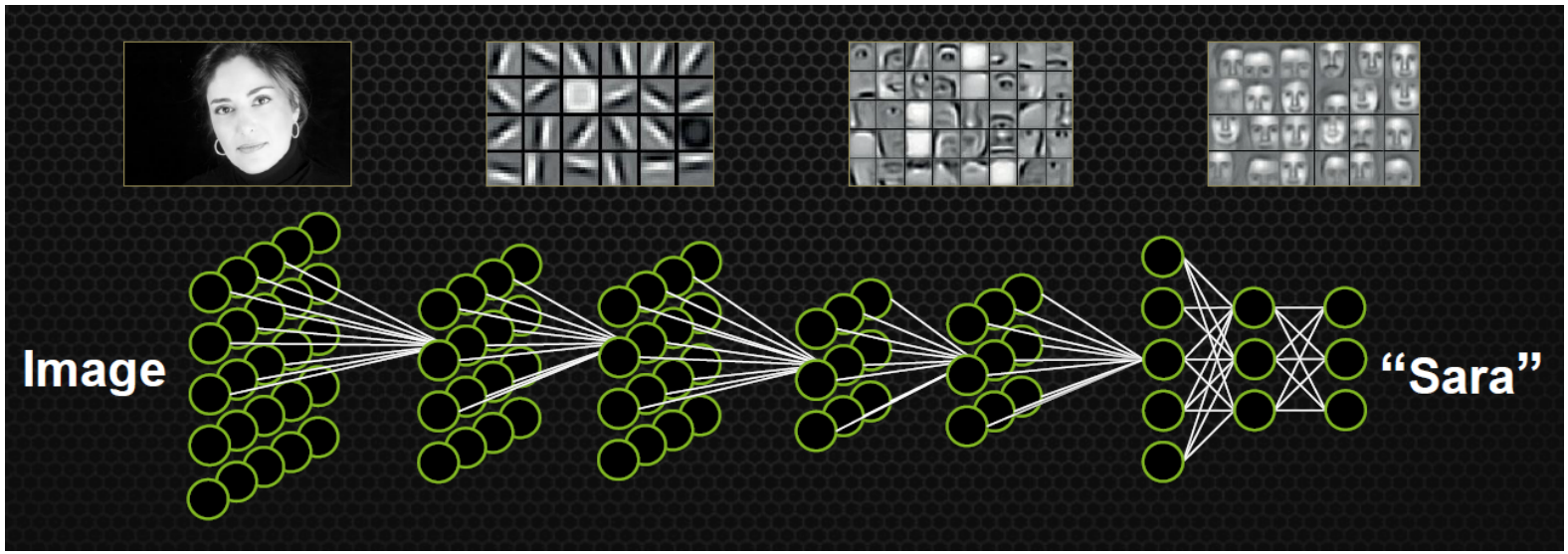
- Output layer:

$$\hat{y} = a^{[L]} = \sigma(W^{[L]} a^{[L-1]} + b^{[L]}).$$

- For binary classification, the final layer is like a logistic regression.
- Find best weights and biases.

Intuition for neural networks

- NNs learn features: e.g., in image classification,
  1. lowest level features are often edge detectors,
  2. higher-level features look for combinations of edges in the right position, such as eyes in a face.



Thoughts on neural networks

- Natural when there is a lot of "hierarchical structure" (images, text, ..)
- But also widely used in other contexts
- Requires "big data", and massive computation
- Many important aspects not discussed here

Overview

Summary

- Data is everywhere
- Statistics, ML, and data science are reusable tools
- Many prediction methods are available

Summary

- There is a lot that we have not covered.
  - Collecting data, Visualization, Evaluating methods, Workflow, Data management, Security, Software ...
- Entire courses, and even degrees cover this
- I encourage you to be curious and learn more
- I hope that this will help your careers. Thanks!

Questions?

References

- Provost, Fawcett: Data Science for Business, O'Reilly, 2013
- Hastie, Friedman, Tibshirani: The Elements of Statistical Learning, Springer, 2009
- Bertsekas, Tsitsiklis: Probability, 2008

Overview

Supplement

- The material in this section is a bonus and will not be covered during the class. It is only included for completeness and for interested students.

Is data science worth it?

- Data science can be costly: infrastructure, personnel...
- Is the investment worth it?

Is data science worth it?

- Data science can be costly: infrastructure, personnel...
- Is the investment worth it?
- Yes. Several academic studies have shown this. e.g., Bynjolfsson, Hitt, & Kim, 2011:

  One standard deviation higher of "data-driven decision-making" $\rightarrow$ a 4%6% increase in productivity.

Is data science worth it?

- Data science can be costly: infrastructure, personnel...
- Is the investment worth it?
- Yes. Several academic studies have shown this. e.g., Bynjolfsson, Hitt, & Kim, 2011:

  One standard deviation higher of "data-driven decision-making" $\rightarrow$ a 4%6% increase in productivity.

- It is worth learning about data science/statistics/machine learning even if that is not your profession, to be able to interact productively with the analytics teams.
- Statistics, data science, data mining, machine learning, predictive analytics, business inteliigence, etc, are all the *same*.

Stats/ML/Data Science problems

- Recall "same tools can solve many different problems". There are only a small number of truly different problems.

Stats/ML/Data Science problems

- Recall "same tools can solve many different problems". There are only a small number of truly different problems.
  - Description: summarization

Stats/ML/Data Science problems

- Recall "same tools can solve many different problems". There are only a small number of truly different problems.
  - Description: summarization
  - Prediction: classification, regression, forecasting

Stats/ML/Data Science problems

- Recall "same tools can solve many different problems". There are only a small number of truly different problems.
  - Description: summarization
  - Prediction: classification, regression, forecasting
  - Detection: testing, inference

Stats/ML/Data Science problems

- Recall "same tools can solve many different problems". There are only a small number of truly different problems.
  - Description: summarization
  - Prediction: classification, regression, forecasting
  - Detection: testing, inference
  - Clustering: matching, grouping

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative
  - ... There is an obvious tradeoff

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative
  - ... There is an obvious tradeoff
- Tools: "descriptive statistics"

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative
  - ... There is an obvious tradeoff
- Tools: "descriptive statistics"
  - Average (Mean). Median. "centrality"

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative
  - ... There is an obvious tradeoff
- Tools: "descriptive statistics"
  - Average (Mean). Median. "centrality"
  - Standard deviation. Interquartile range. "spread"

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative
  - ... There is an obvious tradeoff
- Tools: "descriptive statistics"
  - Average (Mean). Median. "centrality"
  - Standard deviation. Interquartile range. "spread"
  - Histogram. "distribution"

Description

- How to summarize and describe observations in a dataset? (e.g., last year's business performance)
- What do we want?
  - Simple
  - Informative
  - ... There is an obvious tradeoff
- Tools: "descriptive statistics"
  - Average (Mean). Median. "centrality"
  - Standard deviation. Interquartile range. "spread"
  - Histogram. "distribution"
- These are elementary, so we will not talk much about them.

How to find the best one? problems and solutions

- For instance, can try to find "best"

$$\min_{\theta} R(\theta)$$

- The minimization can be hard/impossible. $\rightarrow$ Replace with easier problem
- We can overfit. $\rightarrow$ Regularize
- How to do it fast? $\rightarrow$ Whole field of "optimization"

Linear regression: 1D, no probability

- Observe (x_i, y_i), $i = 1, \ldots, n$.
- Model $y \approx \theta_0 + \theta_1 x$
- θ_0, θ_1 unknown.
- Goal: find good estimates $\hat{\theta}_0, \hat{\theta}_1$. How?

Linear regression: 1D, no probability ctd

In particular, given some estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the resulting parameters, the value y_i corresponding to x_i, as predicted by the model, is

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i.$$

Generally, $\hat{y}_i$ will be different from the given value y_i, and the corresponding difference

$$\tilde{y}_i = y_i - \hat{y}_i,$$

is called the ith **residual**. A choice of estimates that results in small residuals is considered to provide a good fit to the data. With this motivation, the linear regression approach chooses the parameter estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ that minimize the sum of the squared residuals.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2,$$

over all θ_1 and θ_2;

Linear regression: 1D, no probability ctd

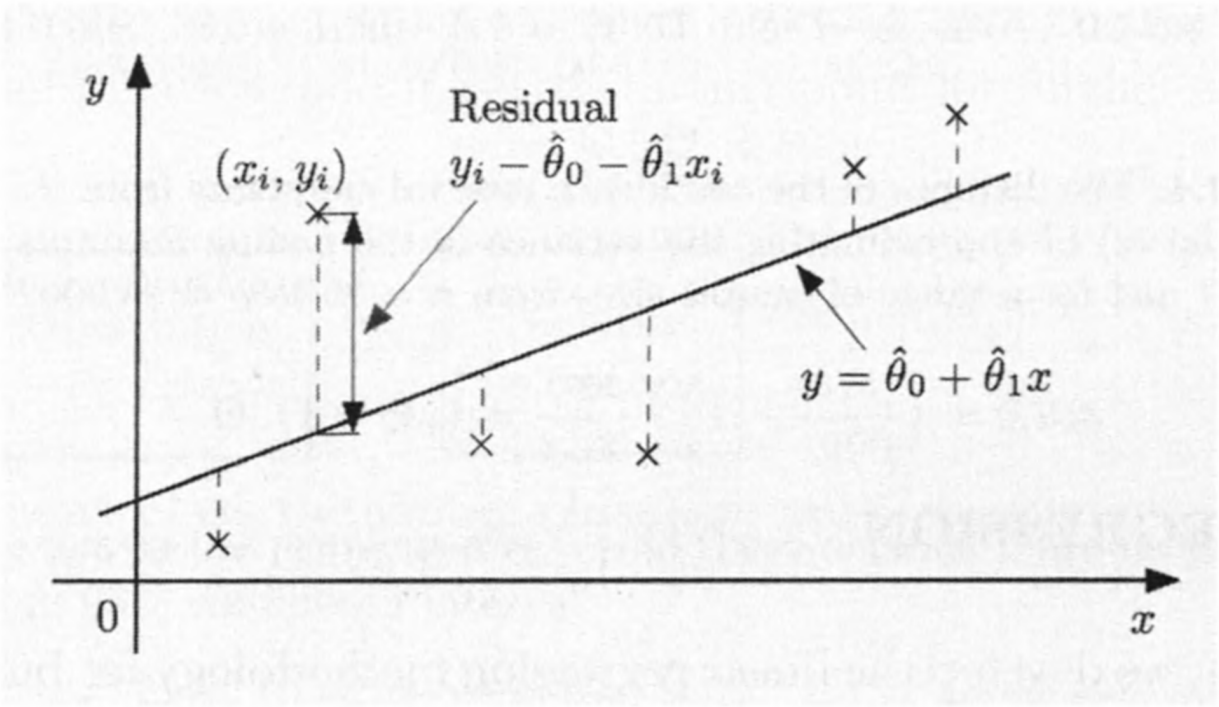


Figure 9.5: Illustration of a set of data pairs (x_i, y_i), and a linear model $y = \hat{\theta}_0 + \hat{\theta}_1 x$, obtained by minimizing over θ_0, θ_1 the sum of the squares of the residuals $y_i - \theta_0 - \theta_1 x_i$.

Linear regression: 1D, no probability ctd

After some algebra, we find:

Given n data pairs (x_i, y_i), the estimates that minimize the sum of the squared residuals are given by

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \qquad \hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x},$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Example

Example 9.9. The leaning tower of Pisa continuously tilts over time. Measurements between years 1975 and 1987 of the "lean" of a fixed point on the tower (the distance in meters of the actual position of the point, and its position if the tower were straight) have produced the following table.

Year	1975	1976	1977	1978	1979	1980	1981
Lean	2.9642	2.9644	2.9656	2.9667	2.9673	2.9688	2.9696
Year	1982	1983	1984	1985	1986	1987	
Lean	2.9698	2.9713	2.9717	2.9725	2.9742	2.9757	

Example ctd

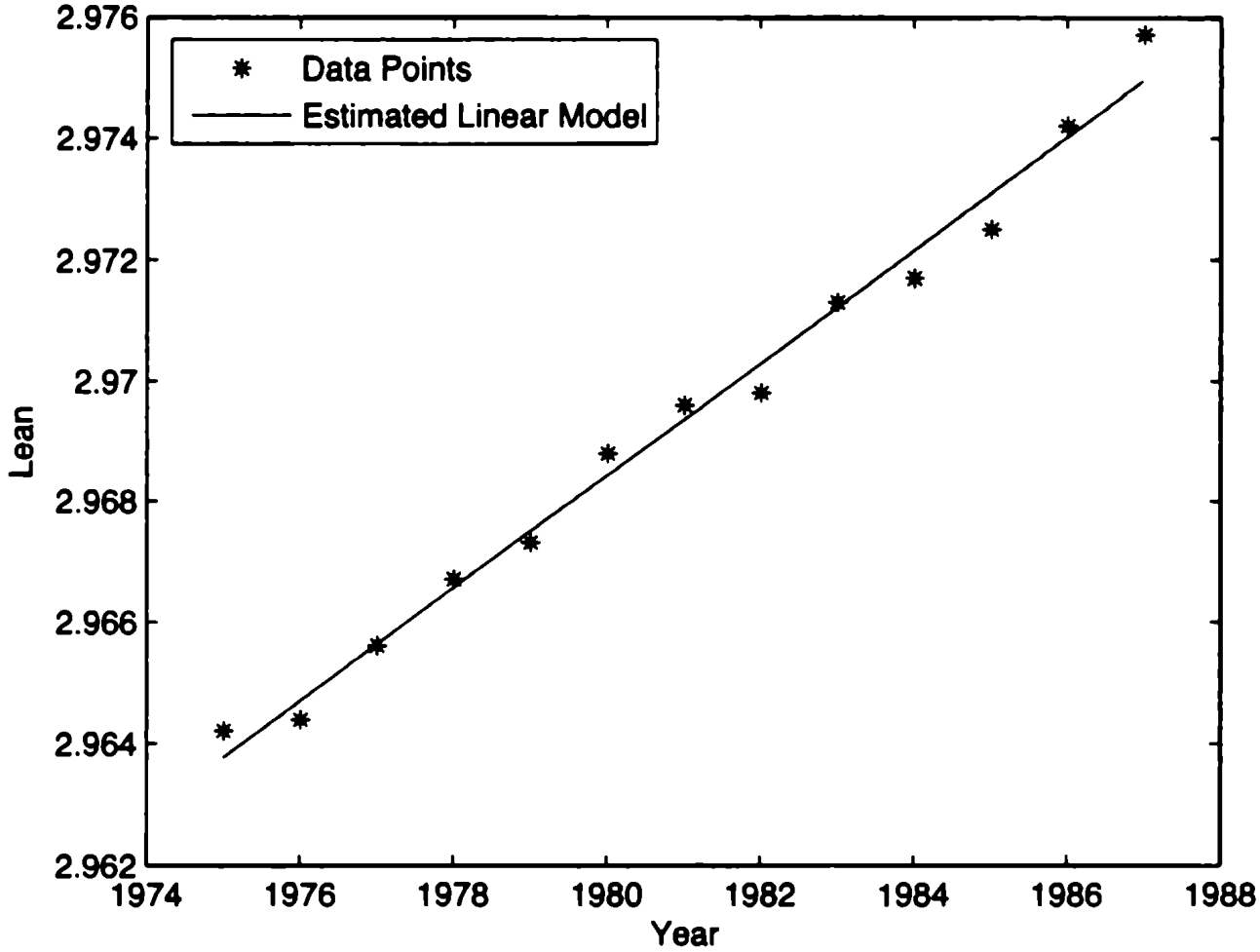


Figure 9.6: The data and the estimated linear model for the lean of the tower of Pisa (Example 9.9).

Linear regression with multiple covariates

- Observe (x_i, y_i), $i = 1, \ldots, n$. Now $x_i \in \mathbb{R}^p$ are vectors of p covariates for the i-th sample.
- Model $y \approx \theta^\top x$. Now $\theta \in \mathbb{R}^p$ is a vector of p unknown parameters.
- Goal: a find good estimate of θ.

Matrix description

- Arrange y_i into a $n \times 1$ vector Y.
- Arrange x_i into the rows of a $n \times p$ matrix X.
  - So the x_{ij} entry of X is the j-th feature for the i-th sample (e.g., height of i-th patient)
  - The j-th feature is the j-th column of X
- Previous model becomes $Y \approx X\theta$.
  - The i-th entry of this approximate equality is $y_i \approx \theta^\top x_i$
  - Check that Y is $n \times 1$, X is $n \times p$, β is $p \times 1$ so this makes sense in terms of matrix multiplication
  - $Y \approx X_1\theta_1 + \ldots + X_p\theta_p$ if X_i are now *columns* of X

Minimizing MSE

- Ordinary least squares (OLS): Find θ to minimize

$$\sum_{i=1}^{n}(y_i - \theta^\top x_i)^2 = \|Y - X\theta\|^2,$$

where $\|x\|^2 = x^\top x = \sum_{i=1}^{n} x_i^2$ is the squared length, or Euclidean norm of a vector

Minimizing MSE

- Can check $\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2x^\top y$, so need to minimize

$$S(\theta) := \|Y - X\theta\|^2 = \|Y\|^2 + \|X\theta\|^2 - 2Y^\top X\theta$$

- Multivariate calculus: Gradients of mv functions
  - $\nabla_\theta b^\top \theta = b$
  - $\nabla_\theta \theta^\top M\theta = 2M\theta$
- Recognize $\|X\theta\|^2 = \theta^\top X^\top X\theta$, (because $(X\theta)^\top = \theta^\top X^\top$)
- So, $\nabla_\theta S(\theta) = 2(X^\top X\theta - X^\top Y)$

OLS

- At minimizer, the gradient is zero. Thus, OLS estimator solves

$$X^\top X\theta - X^\top Y = 0,$$

so

$$\hat{\theta}_{OLS} = (X^\top X)^{-1} X^\top Y.$$

- Intuition: $(X^\top X)^{-1} X^\top$ "undoes" the effect of X in $Y = X\theta + \varepsilon$ to recover θ