DS 4400

Machine Learning and Data Mining I

Alina Oprea
Associate Professor, CCIS
Northeastern University

January 10 2019

Class Outline

- Introduction – 1 week
  - Probability and linear algebra review
- Supervised learning - 7 weeks
  - Linear regression
  - Classification (logistic regression, LDA, kNN, decision trees, random forest, SVM, Naïve Bayes)
  - Model selection, regularization, cross validation
- Neural networks and deep learning – 2 weeks
  - Back-propagation, gradient descent
  - NN architectures (feed-forward, convolutional, recurrent)
- Unsupervised learning – 1-2 weeks
  - Dimensionality reduction (PCA)
  - Clustering (k-means, hierarchical)
- Adversarial ML – 1 lecture
  - Security of ML at testing and training time

Schedule and Resources

- Instructors
  - Alina Oprea
  - TA: Ewen Wang
- Schedule
  - Tue 11:45am – 1:25pm, Thu 2:50-4:30pm
  - Shillman Hall 210
  - Office hours:
    - Alina: Thu 4:30 – 6:00 pm (ISEC 625)
    - Ewen: Monday 5:30-6:30pm (ISEC 605)
- Online resources
  - Slides will be posted after each lecture
  - Use Piazza for questions, Gradescope for homework and project submission

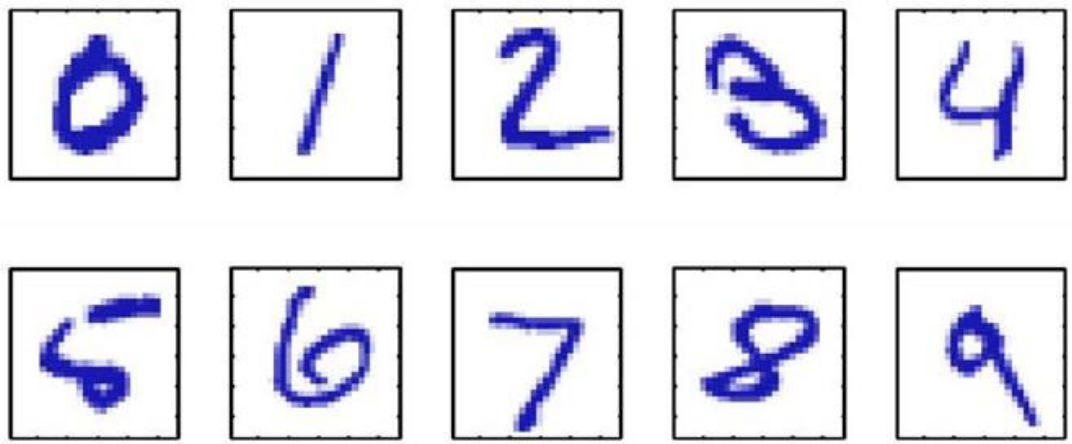
Grading

- Assignments – 25%
  - 4-5 assignments and programming exercises based on studied material in class
- Final project – 35%
  - Select your own project based on public dataset
  - Submit short project proposal and milestone
  - Presentation at end of class (10 min) and report
- Exam – 35%
  - One exam about 3/4 in the class
  - Tentative end of March
- Class participation – 5%
  - Participate in class discussion and on Piazza

Outline

- Supervised learning
    - Classification
    - Regression
- Unsupervised learning
    - Clustering
- Bias-Variance Tradeoff
- Occam’s Razor
- Probability review

Example 1
Handwritten digit recognition



Images are 28 x 28 pixels

Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

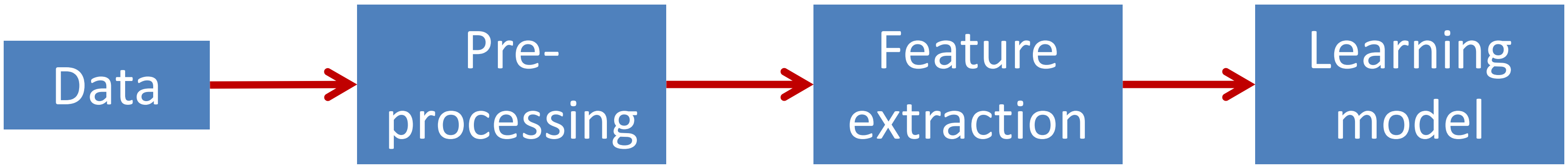
$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

MNIST dataset: Predict the digit

Multi-class classifier

Supervised Learning: Classification

Training



Labeled

$x^{(i)}, y^{(i)} \in \{0,1\}$

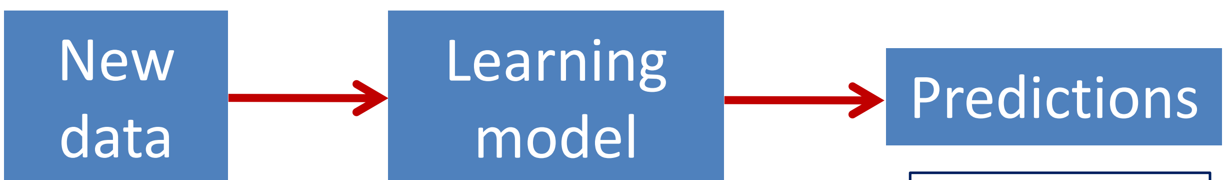
Normalization

Feature Selection

Classification

$f(x)$

Testing



$y' = f(x') \in \{0,1\}$

Unlabeled

x'

$f(x)$

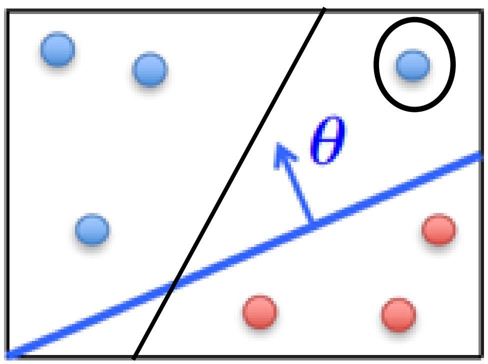
Positive
Negative

Classification

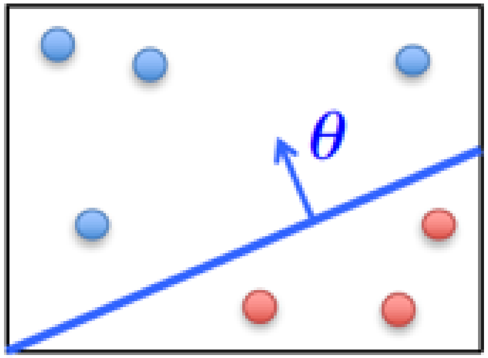
Classification

- Training data
  - $x^{(i)} = [x_1^{(i)}, \dots x_d^{(i)}]$: vector of image pixels
  - Size $d = 28\text{x}28 = 784$
  - $y^{(i)}$: image label (in {0,1})
- Models (hypothesis)
  - Example: Linear model
    - $f(x) = wx + b$
  - Classify 1 if $f(x) > \mathrm{T}$; 0 otherwise
- Classification algorithm
  - Training: Learn model parameters w, b to minimize error (number of training examples for which model gives wrong label)
  - Output: "optimal" model
- Testing
  - Apply learned model to new data and generate prediction

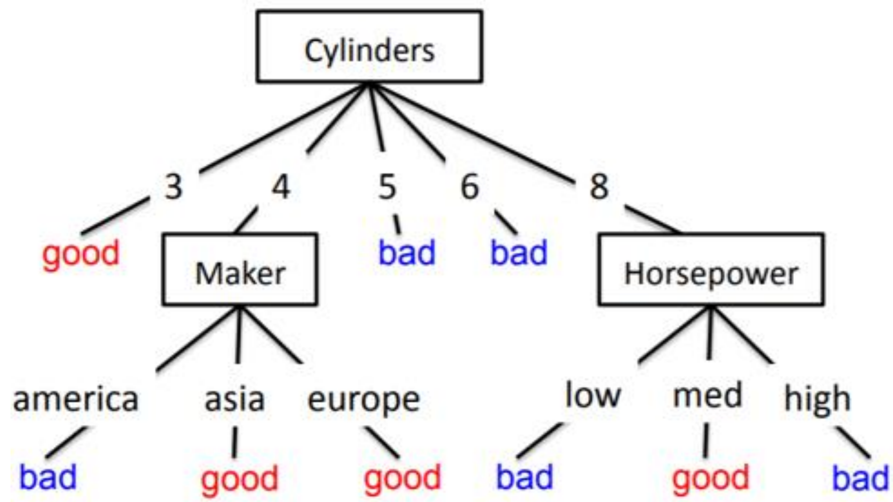
Error



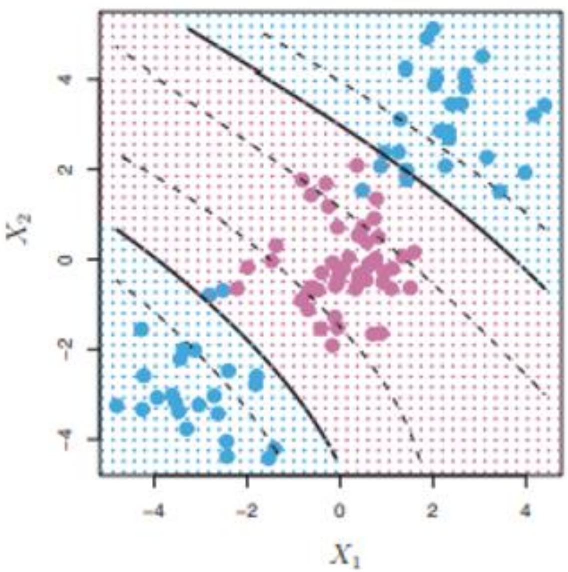
Example Classifiers



Linear classifiers: logistic regression, SVM, LDA

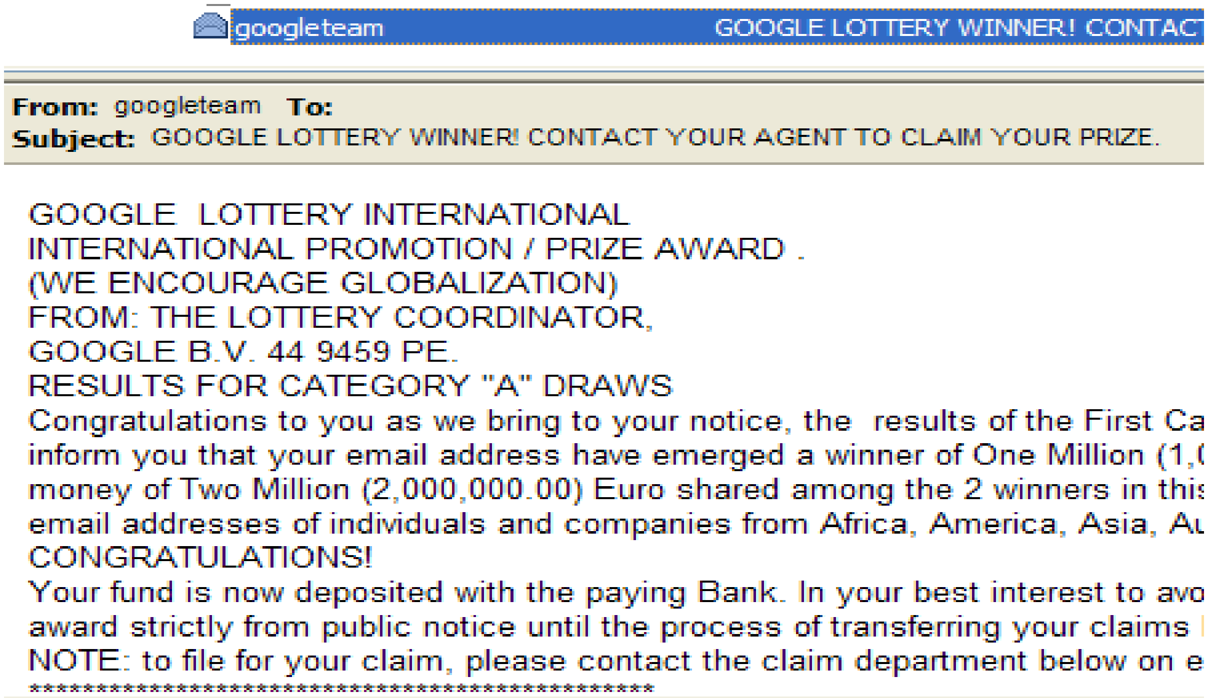


Decision trees



SVM polynomial kernel

Real-world example: Spam email



googleteam GOOGLE LOTTERY WINNER! CONTACT

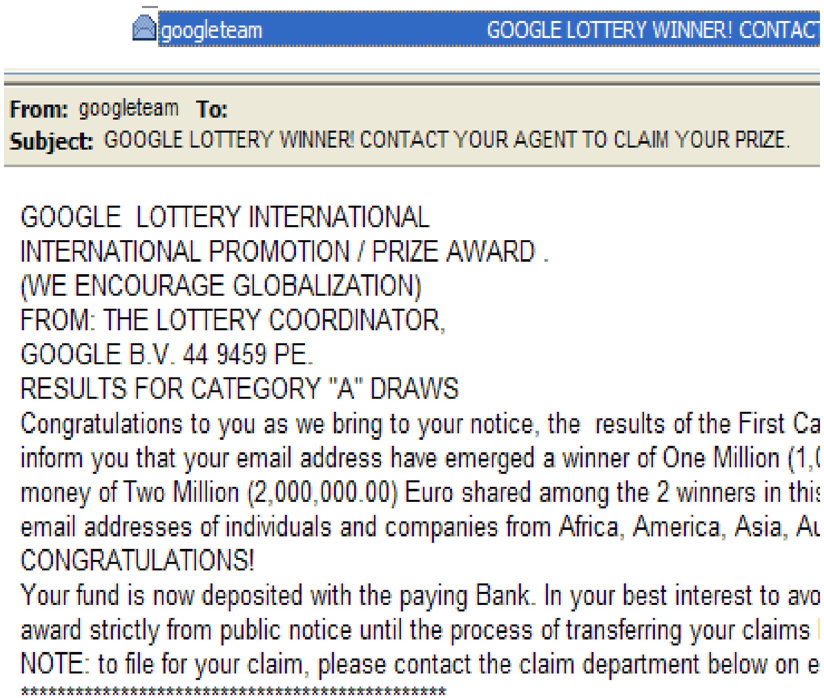
From: googleteam **To:**
Subject: GOOGLE LOTTERY WINNER! CONTACT YOUR AGENT TO CLAIM YOUR PRIZE.

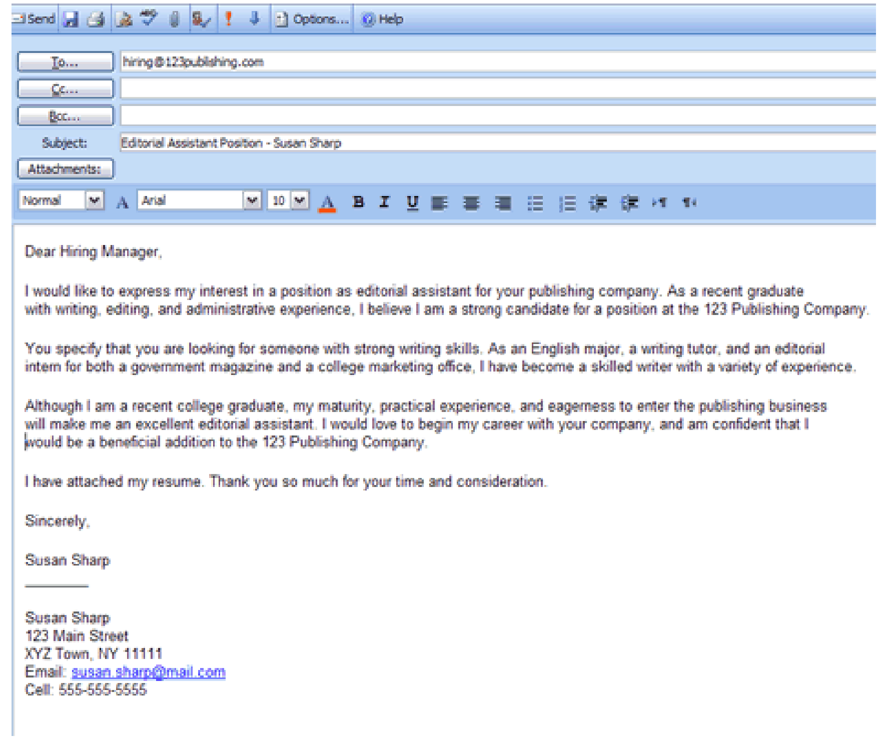
GOOGLE LOTTERY INTERNATIONAL
INTERNATIONAL PROMOTION / PRIZE AWARD .
(WE ENCOURAGE GLOBALIZATION)
FROM: THE LOTTERY COORDINATOR,
GOOGLE B.V. 44 9459 PE.
RESULTS FOR CATEGORY "A" DRAWS
Congratulations to you as we bring to your notice, the results of the First Ca
inform you that your email address have emerged a winner of One Million (1,
money of Two Million (2,000,000.00) Euro shared among the 2 winners in thi
email addresses of individuals and companies from Africa, America, Asia, A
CONGRATULATIONS!
Your fund is now deposited with the paying Bank. In your best interest to avo
award strictly from public notice until the process of transferring your claims
NOTE: to file for your claim, please contact the claim department below on e
****

SPAM email

- Unsolicited
- Advertisement
- Sent to a large number of people

Classifying spam email





Content-related features

- Use of certain words
- Word frequencies
- Language
- Sentence

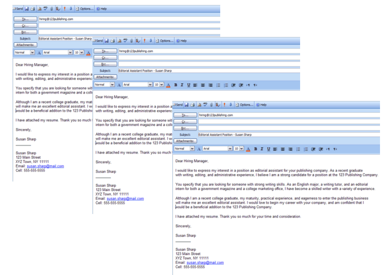
Structural features

- Sender IP address
- IP blacklist
- DNS information
- Email server
- URL links (non-matching)

Classifying spam email

SPAM

REGULAR



Feature extraction
- Content
- Structural

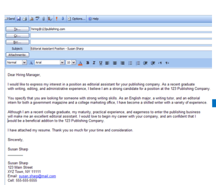
Numerical

Labeled data
- SPAM
- REGULAR

Training

Classifier
- Logistic regression
- Decision tree
- SVM

New email



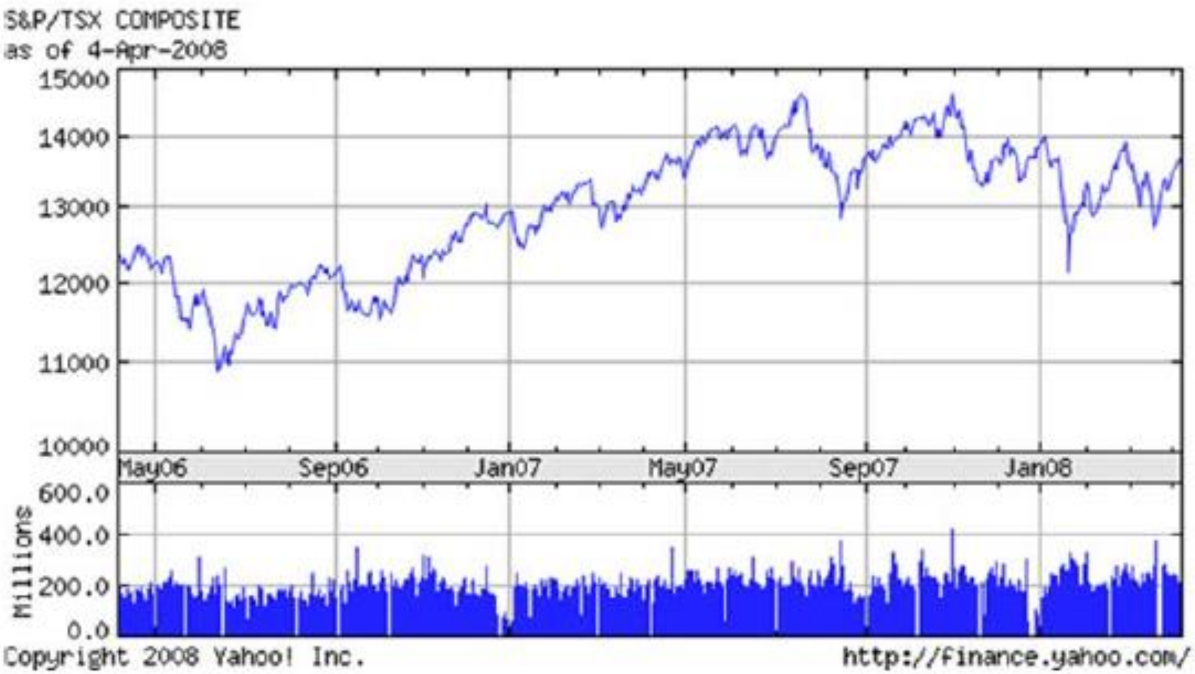
Model

SPAM → Filter

REGULAR → Allow

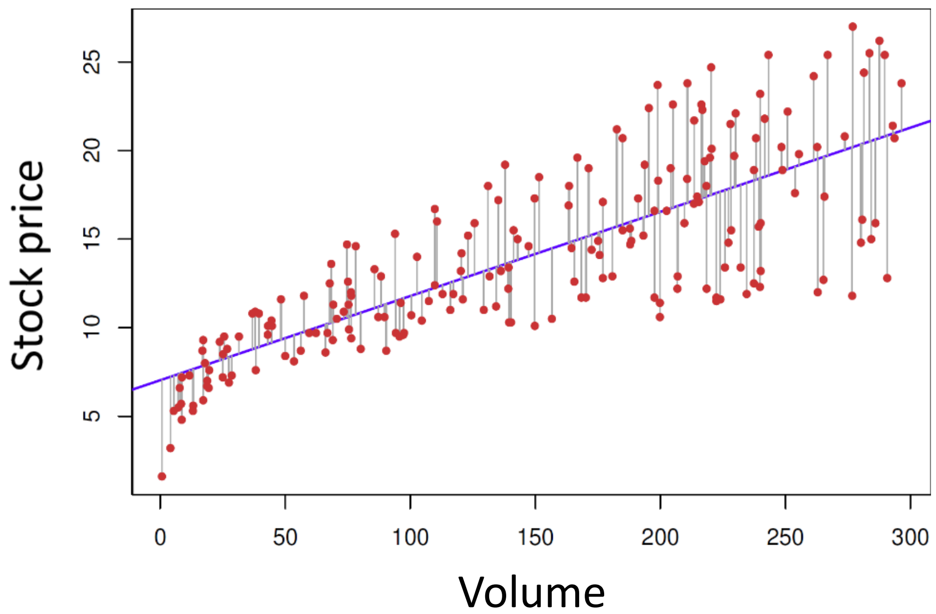
Testing

Example 2
Stock market prediction



- Task is to predict stock price at future date
- This is a regression task, as the output is continuous

Regression



Linear regression
1 dimension

- Suppose we are given a training set of N observations

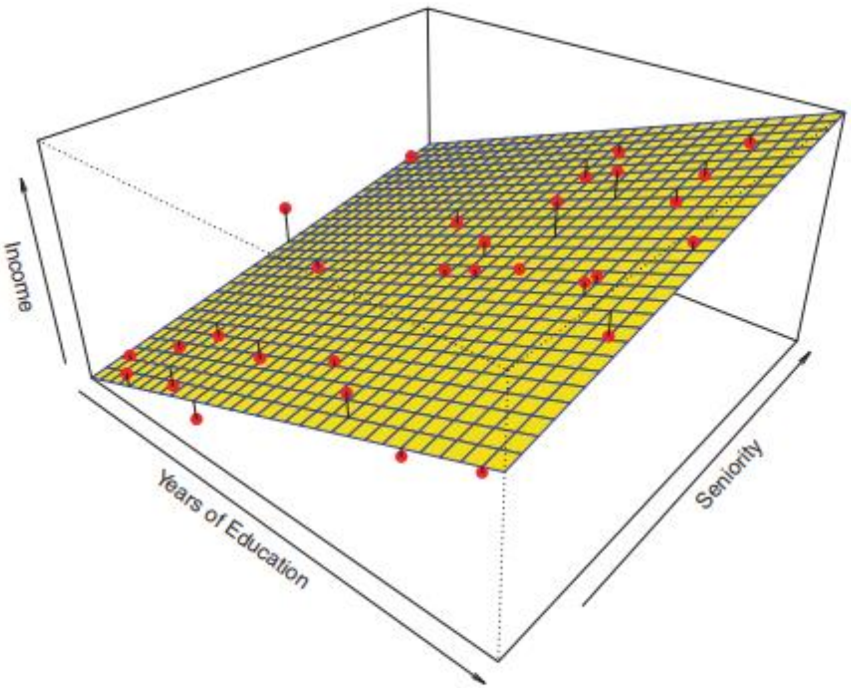
$$x^{(1)}, \ldots, x^{(N)} \quad \text{and} \quad y^{(1)}, \ldots, y^{(N)} \in R$$ Numerical

- Regression problem is to estimate y(x) from this data

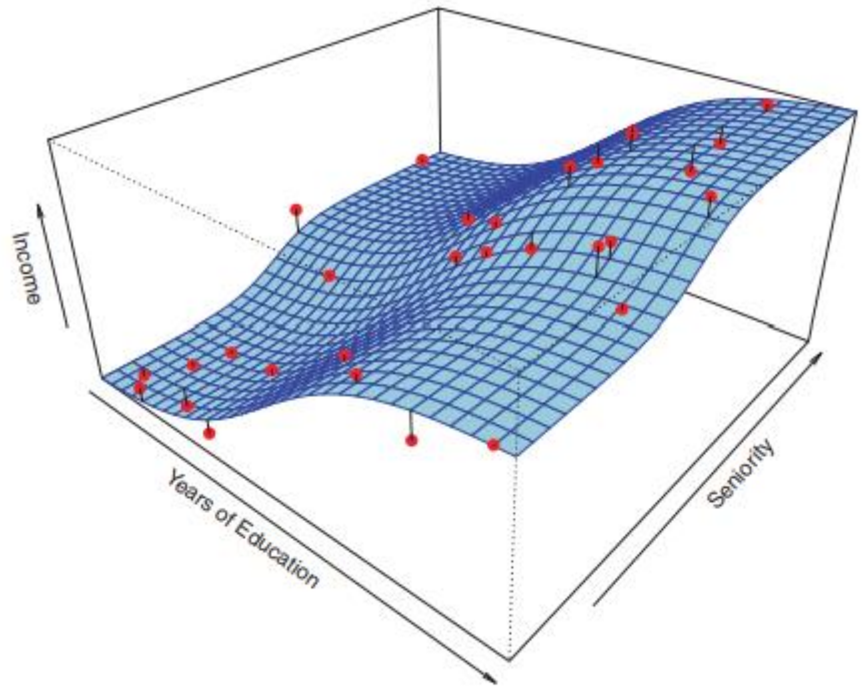
$$x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})$$ d predictors (features)

$$y^{(i)}$$ response variable

Income Prediction



Linear Regression



Non-Linear Regression
Polynomial/Spline Regression

Supervised Learning: Regression

Training

Data → Pre-processing → Feature extraction → Learning model

Labeled

$x^{(i)}, y^{(i)} \in R$

Normalization

Feature Selection

Regression

$f(x)$

Testing

New data → Learning model → Predictions

Unlabeled

x'

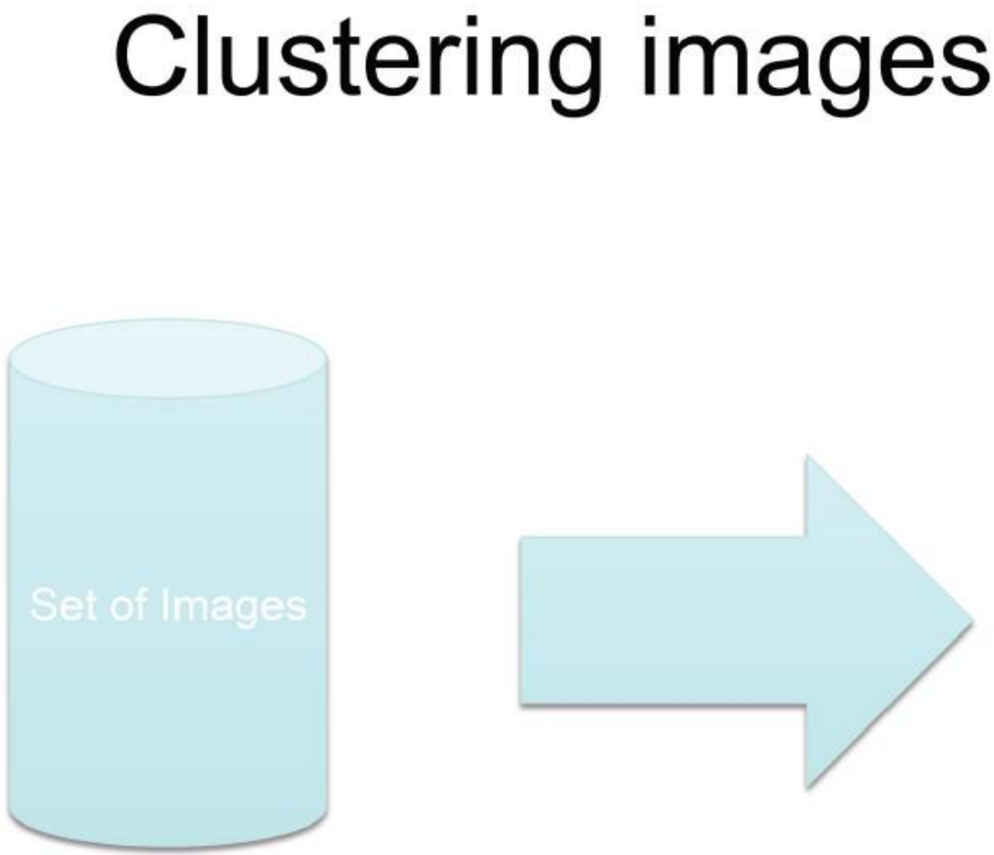
$f(x)$

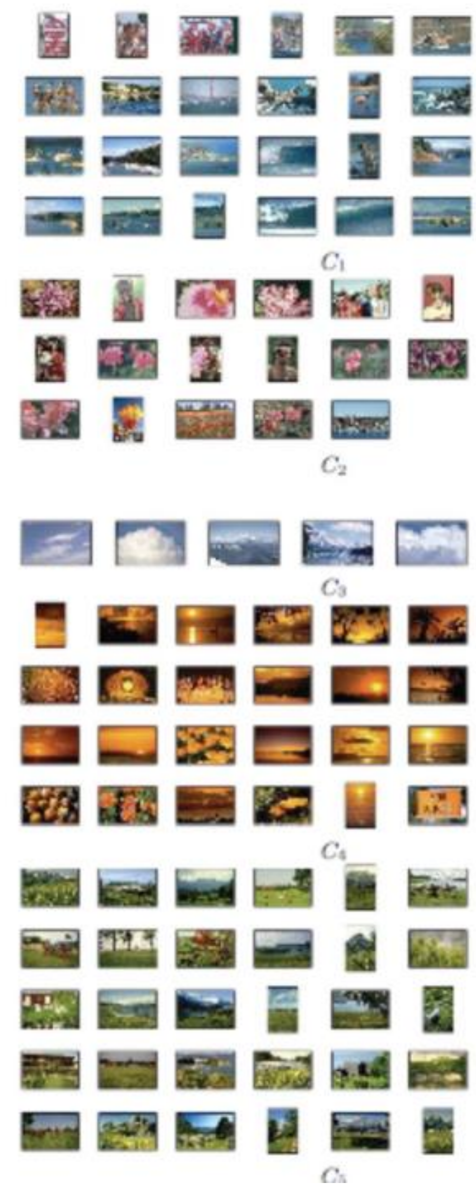
$y' = f(x') \in R$

Response variable

Regression

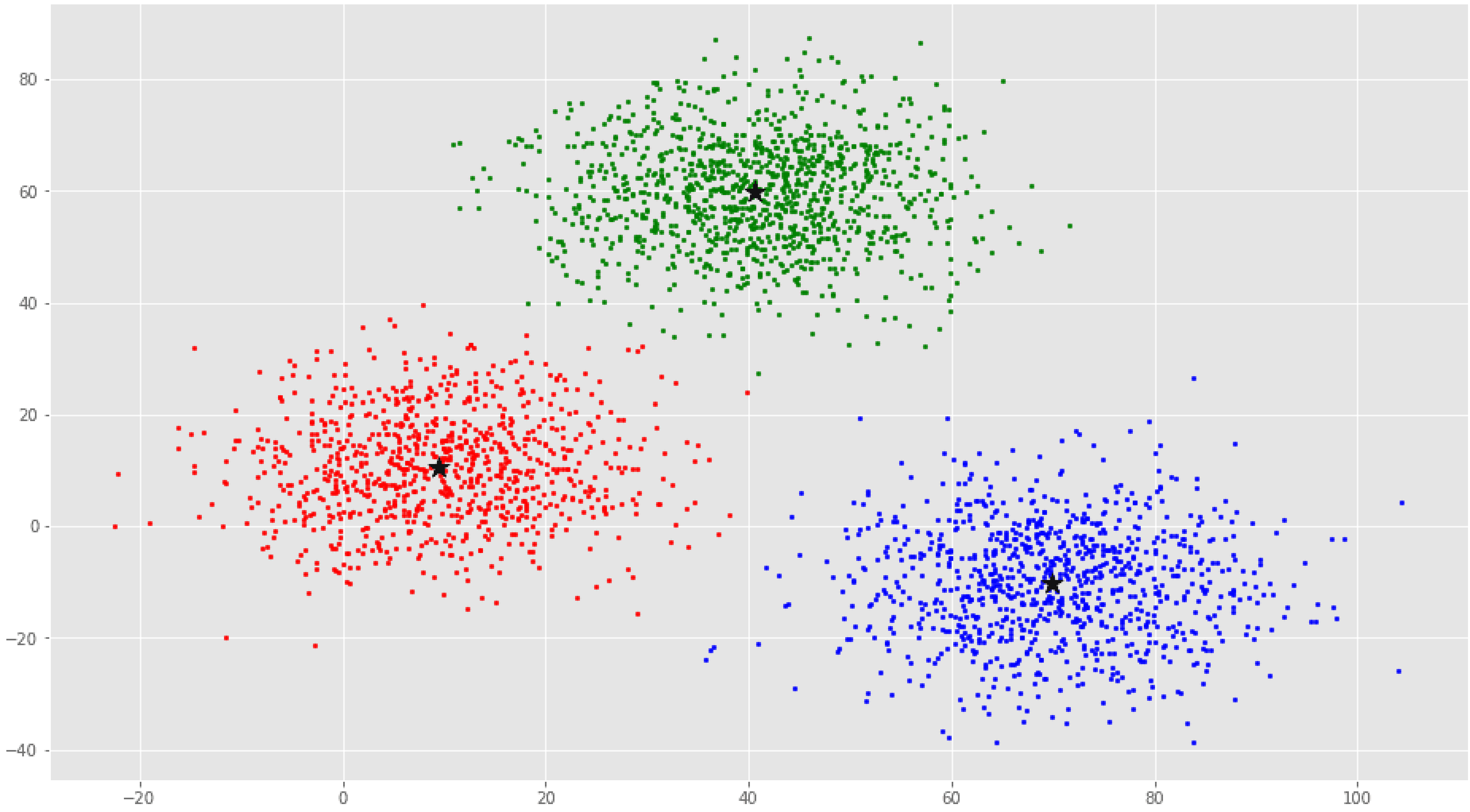
Example 3: image search





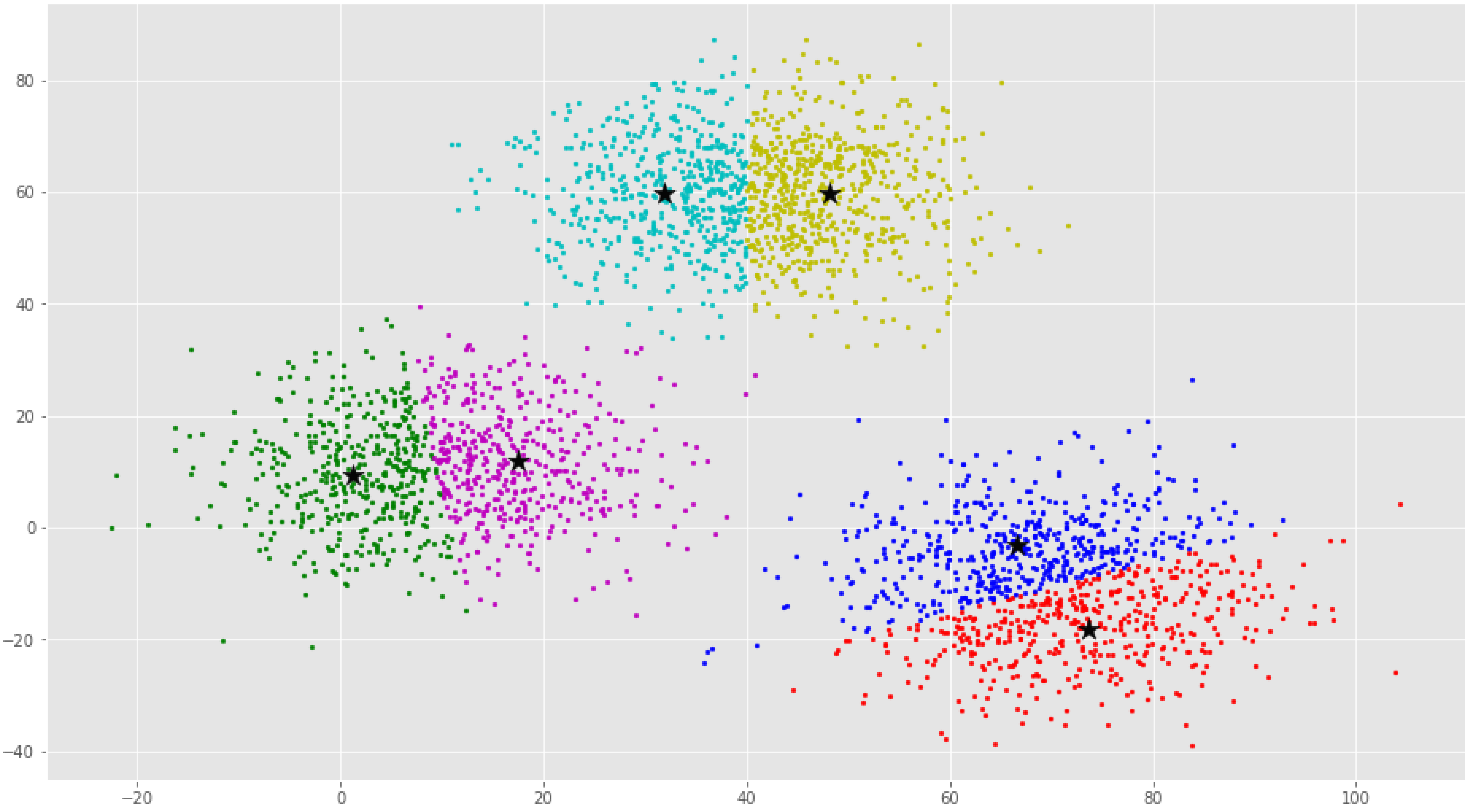
Find similar images to a target one

K-means Clustering



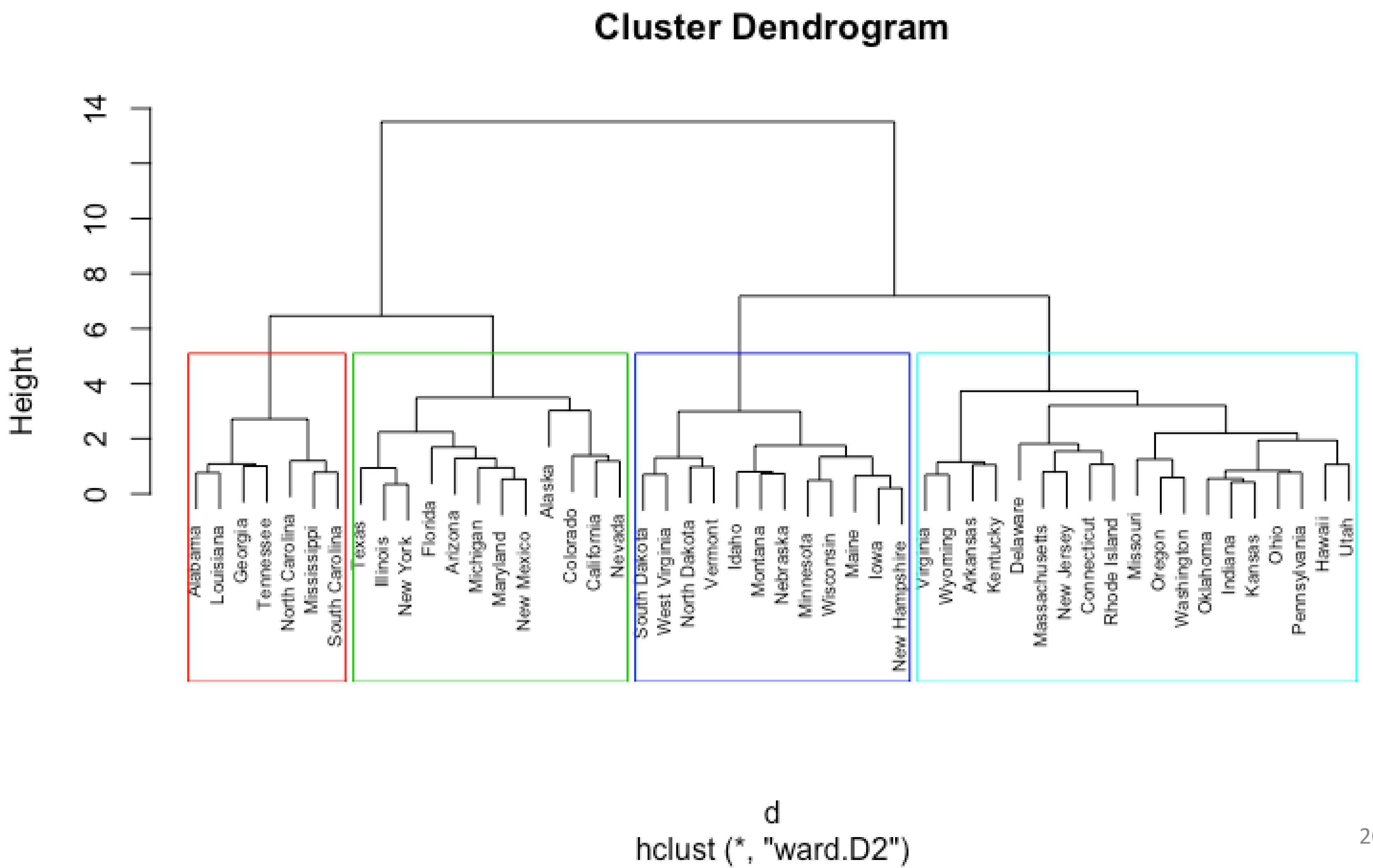
K=3

K-means Clustering



K=6

Hierarchical Clustering



Unsupervised Learning

- Clustering
  - Group similar data points into clusters
  - Example: k-means, hierarchical clustering
- Dimensionality reduction
  - Project the data to lower dimensional space
  - Example: PCA (Principal Component Analysis)
- Feature learning
  - Find feature representations
  - Example: Autoencoders

Supervised Learning Tasks

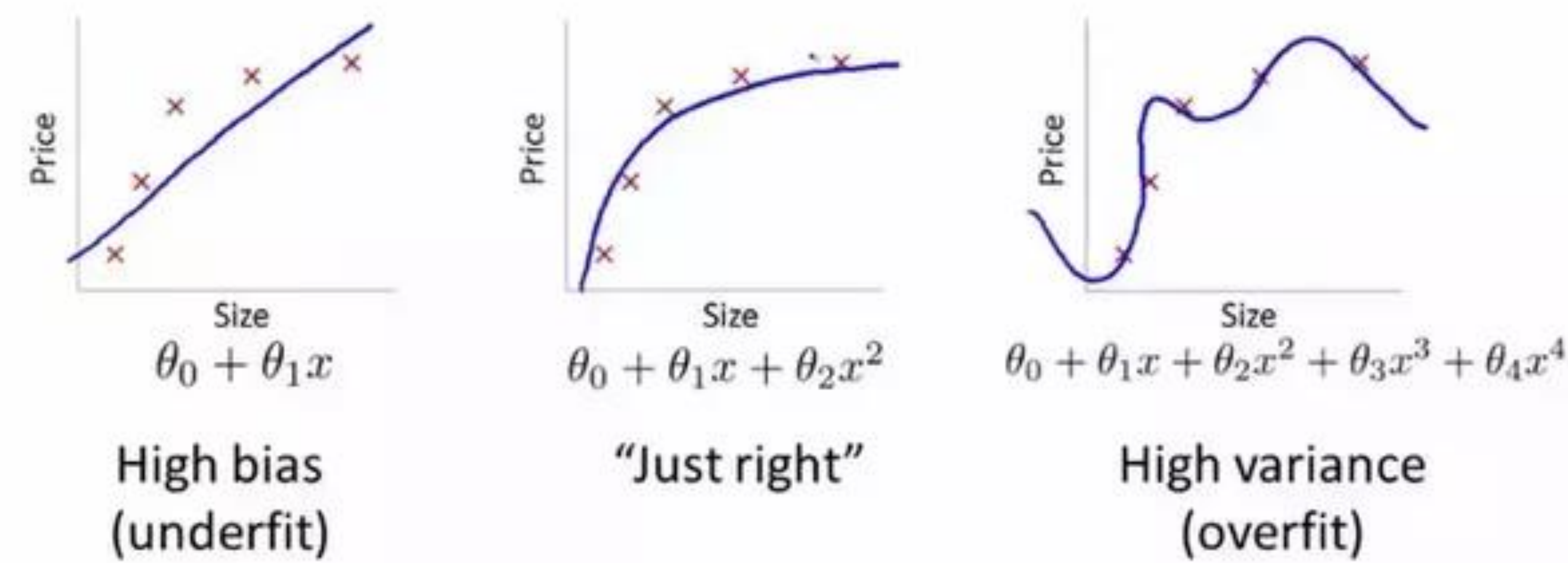
- Classification
  - Learn to predict class (discrete)
  - Minimize classification error 1/N $\sum_{i=1}^{N}[y^{(i)} \neq f(x^{(i)})]$
- Regression
  - Learn to predict response variable (numerical)
  - Minimize MSE (Mean Square Error)
  - $1/N \sum_{i=1}^{N}\left[y^{(i)} - f\left(x^{(i)}\right)\right]^2$
- Both classification and regression
  - Training and testing phase
  - "Optimal" model is learned in training and applied in testing

Learning Challenges

- Goal
  - Classify well new testing data
  - Model generalizes well to new testing data
- Variance
  - Amount by which model would change if we estimated it using a different training data set
  - More complex models result in higher variance
- Bias
  - Error introduced by approximating a real-life problem by a much simpler model
  - E.g., assume linear model (linear regression), then error is high
  - More complex models result in lower bias

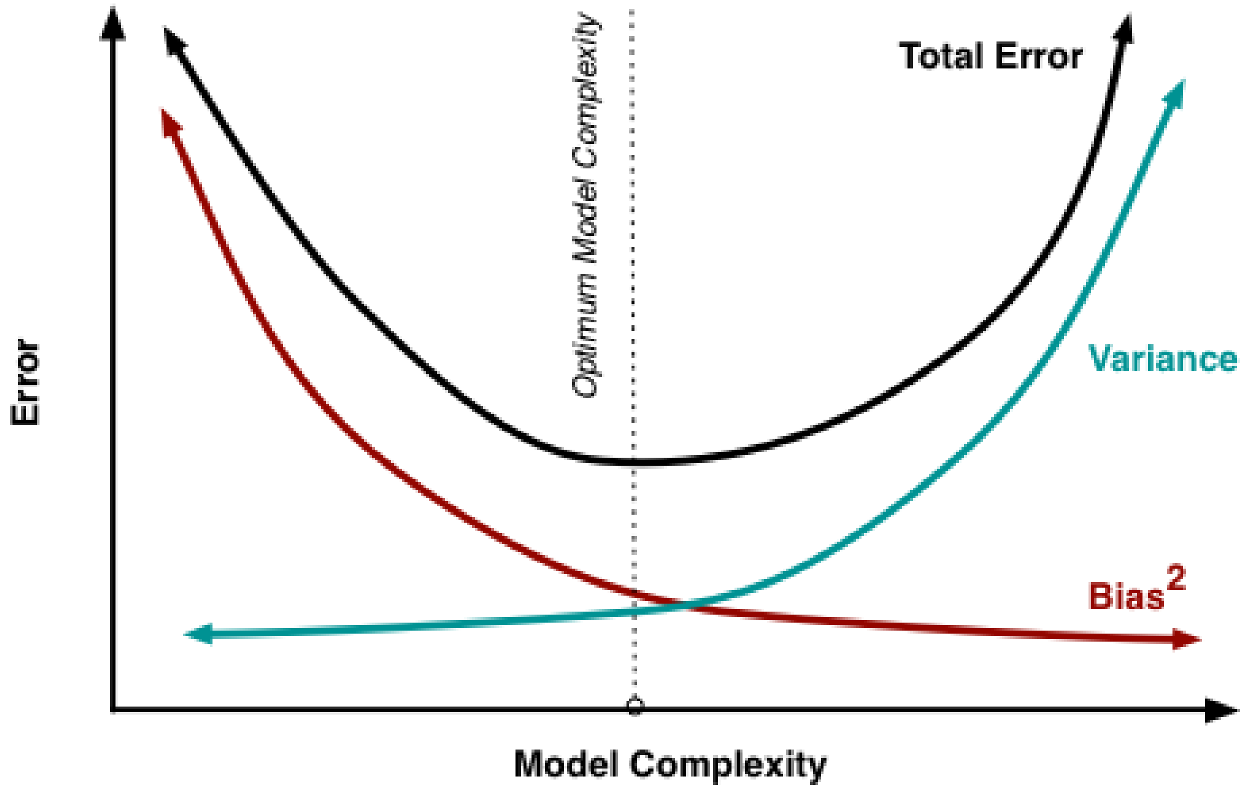
Bias-Variance tradeoff

Example: Regression



Bias-Variance Tradeoff

Generalizes well on new data



Model underfits the data

Model overfits the data

Occam's Razor

- William of Occam: Monk living in the 14th century
- Principle of parsimony:

"One should not increase, beyond what is necessary, the number of entities required to explain anything"

- When many solutions are available for a given problem, we should select the simplest one

Select the simplest machine learning model that gets reasonable accuracy for the task at hand

Recap

- ML is a subset of AI designing learning algorithms
- Learning tasks are *supervised* (e.g., classification and regression) or *unsupervised* (e.g., clustering)
  - Supervised learning uses labeled training data
- Learning the “best” model is challenging
  - Design algorithm to minimize the error
  - Bias-Variance tradeoff
  - Need to generalize on new, unseen test data
  - Occam’s razor (prefer simplest model with good performance)

Probability review

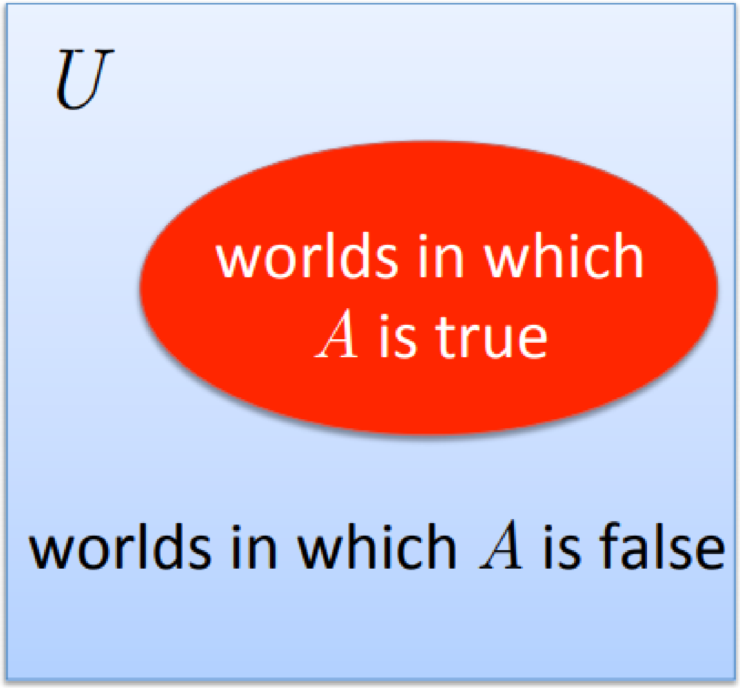
Discrete Random Variables

- Let A denote a random variable
  - A represents an event that can take on certain values
  - Each value has an associated probability
- Examples of binary random variables:
  - A = I have a headache
  - A = Sally will be the US president in 2020
- $\mathrm{P}(A)$ is "the fraction of possible worlds in which A is true"

Visualizing A

- Universe U is the event space of all possible worlds
  - Its area is 1
  - $\mathrm{P}(U)$ = 1
- $\mathrm{P}(A)$ = area of red oval
- Therefore:

$$P(A) + P(\neg A) = 1$$
$$P(\neg A) = 1 - P(A)$$



Axioms of Probability

Kolmogorov showed that three simple axioms lead to the rules of probability theory

- de Finetti, Cox, and Carnap have also provided compelling arguments for these axioms

1. All probabilities are between 0 and 1:

$$0 \leq \mathrm{P}(A) \leq 1$$

2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0:

$$\mathrm{P}(\mathrm{true}) = 1 \,; \quad \mathrm{P}(\mathrm{false}) = 0$$

3. The probability of a disjunction is given by:

$$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$$

Interpreting the Axioms

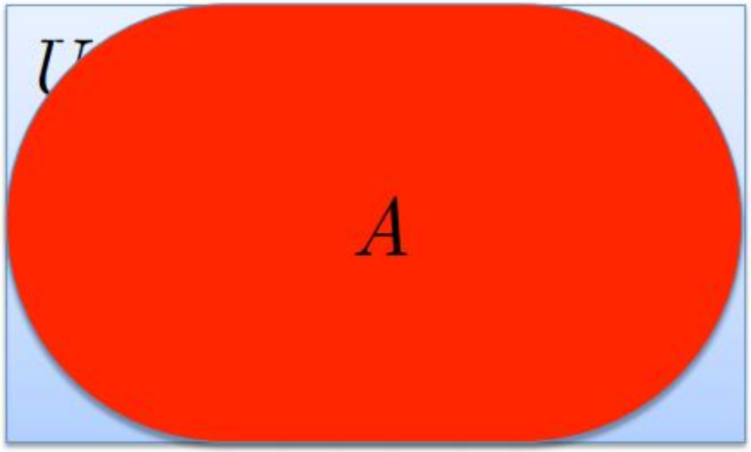
- $0 \leq \mathrm{P}(A) \leq 1$
- $\mathrm{P}(\text{true}) = 1$
- $\mathrm{P}(\text{false}) = 0$
- $\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

The area of A can't get any smaller than 0

A zero area would mean no world could ever have A true

Interpreting the Axioms

- $0 \leq P(A) \leq 1$
- $P(\text{true}) = 1$
- $P(\text{false}) = 0$
- $P(A \lor B) = P(A) + P(B) - P(A \land B)$

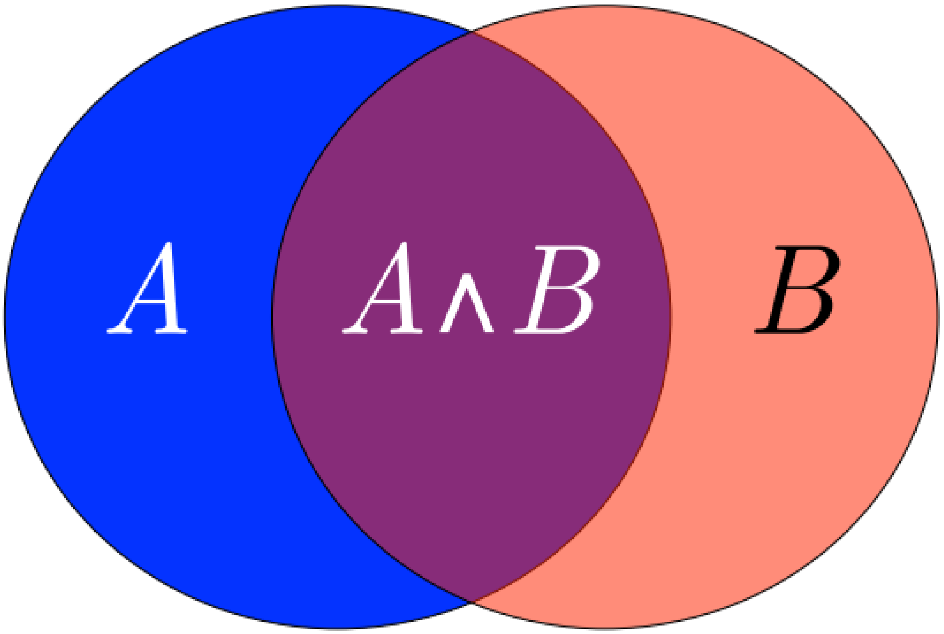


The area of A can't get any bigger than 1

An area of 1 would mean A is true in all possible worlds

Interpreting the Axioms

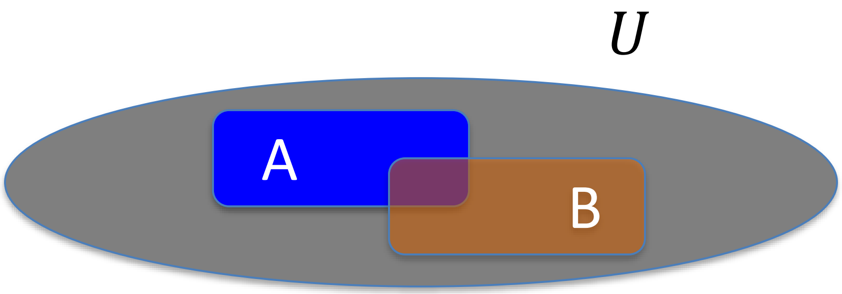
- $0 \leq P(A) \leq 1$
- $P(\mathrm{true}) = 1$
- $P(\mathrm{false}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



The union bound

- For events A and B

$$P[\,A \cup B\,] \le P[A] + P[B]$$



Axiom: $P[\,A \cup B\,] = P[A] + P[B] - P[A \cap B]$

If $A \cap B = \Phi$, then $P[\,A \cup B\,] = P[A] + P[B]$

Example:

$A_1 = \{$ all x in $\{0,1\}^n$ s.t $lsb_2(x)=11$ $\}$; $A_2 = \{$ all x in $\{0,1\}^n$ s.t. $msb_2(x)=11$ $\}$

$P[\,lsb_2(x)=11$ or $msb_2(x)=11\,] = P[A_1 \cup A_2] \le ¼+¼ = ½$

Negation Theorem

$0 \leq \mathrm{P}(A) \leq 1$

$\mathrm{P}(\text{true}) = 1; \quad \mathrm{P}(\text{false}) = 0$

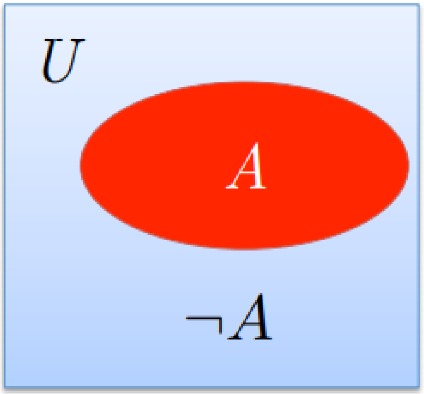
$\mathrm{P}(A \vee B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \wedge B)$

From these we can prove:

$$P(\neg A) = 1 - P(A)$$

Proof: Let $B = \neg A$. Then, we have

$$\begin{aligned} P(A \vee B) &= P(A) + P(B) - P(A \wedge B) \\ P(A \vee \neg A) &= P(A) + P(\neg A) - P(A \wedge \neg A) \\ P(\text{true}) &= P(A) + P(\neg A) - P(\text{false}) \\ 1 &= P(A) + P(\neg A) - 0 \\ P(\neg A) &= 1 - P(A) \qquad \square \end{aligned}$$



Random Variables (Discrete)

Def: a random variable X is a function $X : U \longrightarrow V$
Def: A discrete random variable takes a finite number of values: $|V|$ is finite

Example: X is modeling a coin toss with output 1 (heads) or 0 (tail)
$\Pr[X=1] = p$, $\Pr[X=0] = 1-p$

Bernoulli Random Variable

We write $X \longleftarrow U$ to denote a **uniform random variable (discrete)** over U

for all $u \in U$: $\Pr[\, X = u \,] = 1/|U|$

Example: If $p=1/2$; then X is a uniform coin toss

Probability Mass Function (PMF): $p(u) = \Pr[X = u]$

Example

1. X is the number of heads in a sequence of n coin tosses

What is the probability $\mathrm{P}[X = k]$?

$$\mathrm{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Random Variable

2. X is the sum of two fair dice

What is the probability $\mathrm{P}[X = k]$ for $k \in \{2, \ldots, 12\}$?

P[X=2]=1/36; P[X=3]=2/36; P[X=4]= 3/36

For what k is $\mathrm{P}[X = k]$ highest?

Expectation and variance

Expectation for discrete random variable X

$$E[X] = \sum_{v} v Pr[X = v]$$

Properties

- $E[aX] = a\,E[X]$
- Linearity: $E[X + Y] = E[X] + E[Y]$

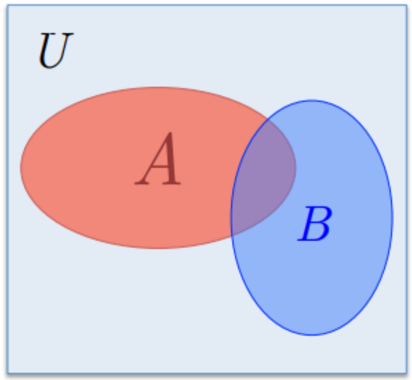
Variance

$$Var[X] \triangleq E[(X - E(X))^2]$$

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2, \end{aligned}$$

Conditional Probability

- $\mathrm{P}(A \mid B)$ = Fraction of worlds in which B is true that also have A true



What if we already know that B is true?

That knowledge changes the probability of A

- Because we know we're in a world where B is true

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

<u>Def</u>: Events A and B are **independent** if and only if

$$\Pr[\, A \cap B \,] = \Pr[A] \cdot \Pr[B]$$

If A and B are independent

$$\Pr[A|B] = \frac{\Pr[A \cap \; B]}{\Pr[B]} = \frac{\Pr[A]\Pr[B]}{\Pr[B]} = \Pr[\mathrm{A}]$$

Acknowledgements

- Slides made using resources from:
  - Andrew Ng
  - Eric Eaton
  - David Sontag
- Thanks!