Computer Science

CSC380: Principles of Data Science

Wrap-up 1

Kyoungseok Jang

Outline

- Data Science Ethics and Fairness
- Course Review
- Additional Resources
- Final Exam Overview

Outline

- Data Science Ethics and Fairness
- Course Review
- Additional Resources
- Final Exam Overview

Data Science Ethics

The movement to hold AI accountable gains more steam

First-in-US NYC law requires algorithms used in hiring to be "audited" for bias.

KHARI JOHNSON, WIRED.COM – 12/5/2021, 6:10 AM

ars technica

[Article Link]

As Data Science / AI / ML become more standard,
we need to address fairness and ethics…

State of Michigan's mistake led to man filing bankruptcy

Paul Egan Detroit Free Press

The secret bias hidden in mortgage-approval algorithms

By EMMANUEL MARTINEZ and LAUREN KIRCHNER/The Markup August 25, 2021

Senators Question Regulators About Tenant Screening Oversight

Facebook's race-blind
around hate speech
e expense of Black
v documents show

- Washington Post

ExamSoft's remote bar exam sparks privacy and facial recognition concerns

- Venture Beat

Data Science Ethics

- NYC adopted law requiring audits of algorithms used in hiring
- White house proposes an AI bill of rights to disclose when AI makes decisions with societal impact
- EU lawmakers require inspection of AI deemed high-risk
- Analysis of automated hiring software found to be biased to appearance, software program used to create resume, accent, or whether applicants have a bookshelf in the background
- Photo ID software works well for white men—black women, not so much

Data Science Ethics

Aspects of ethics include…

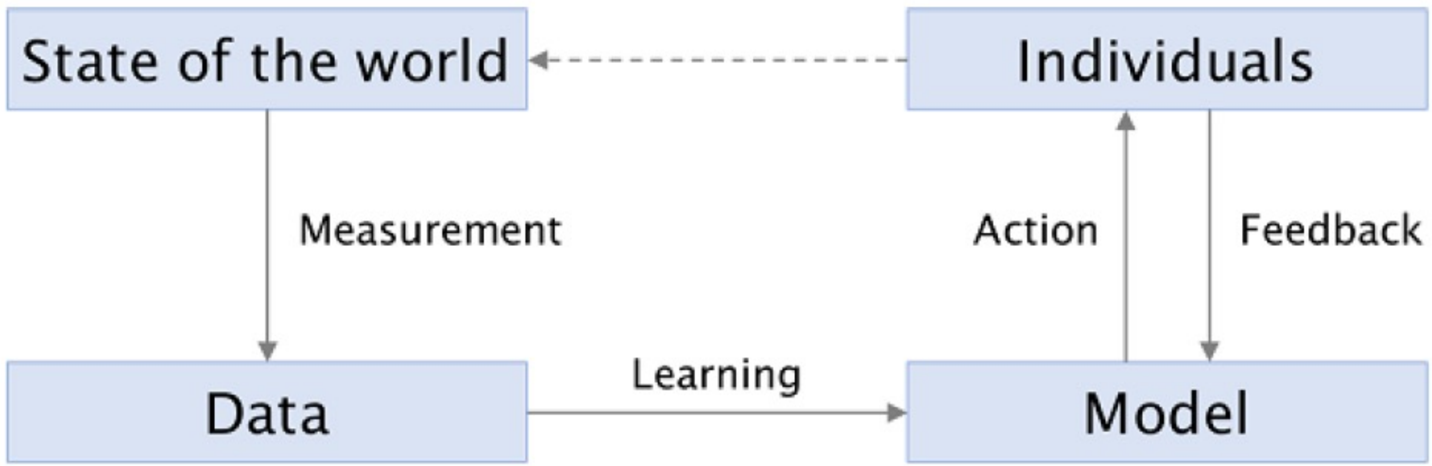
Security Who has access to the data?

Privacy Can data be used to identify individuals?

Fairness Are predictions biased across groups?

Transparency Do users know what they are consenting to? Are model decisions interpretable?

Impacts of Data Science



It is rare for data science to *not* involve people in some way

- Of the top 30 recent Kaggle competitions in 2021, 14 involve making decisions that directly affect people
- An additional 5 have obvious indirect affect on people
- Only 9 had no obvious impact on people

[Source: Brocas et al. “Fairness and ML”]

Data Science Fairness

Fairness issues can arise from biases in the data…

- Are there observable biases in the data?
- Can we correct for them?

- Differences in the distributions of training / test data?
- Can we detect these differences and avoid / correct them?

Training data reflect disparities, distortions, and biases from the real world and measurement process…

For each model a data scientist should ask… Does learning the model preserve, mitigate, or exacerbate these disparities?

Example Machine translation “She is a doctor” reverse translates to “He is a doctor” in many languages due to data biases.

Data Science Ethics

A real-live example of dataset bias…

https://translate.google.com/

Exhibits gender bias in many languages…

…largely the result of using highly-parameterized neural networks with inadequate training data

Assessing Gender Bias in Machine Translation – A Case Study with Google Translate

Marcelo Prates MORPRATES@INF.UFRGS.BR
Pedro Avelar PEDRO.AVELAR@INF.UFRGS.BR
Luis C. Lamb LAMB@INF.UFRGS.BR
Federal University of Rio Grande do Sul

Abstract

Recently there has been a growing concern in academia, industrial research labs and the mainstream commercial media about the phenomenon dubbed as *machine bias*, where trained statistical models – unbeknownst to their creators – grow to reflect controversial societal asymmetries, such as gender or racial bias. A significant number of Artificial Intelligence tools have recently been suggested to be harmfully biased towards some minority, with reports of racist criminal behavior predictors, Apple's Iphone X failing to differentiate between two distinct Asian people and the now infamous case of Google photos' mistakenly classifying black people as gorillas. Although a systematic study of such biases can be difficult, we believe that automated translation tools can be exploited through gender neutral languages to yield a window into the phenomenon of gender bias in AI.

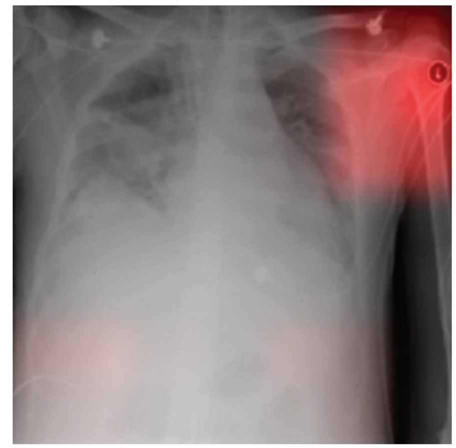
In this paper, we start with a comprehensive list of job positions from the U.S. Bureau of Labor Statistics (BLS) and used it in order to build sentences in constructions like "He/She is an Engineer" (where "Engineer" is replaced by the job position of interest) in 12 different gender neutral languages such as Hungarian, Chinese, Yoruba, and several others. We translate these sentences into English using the Google Translate API, and collect statistics about the frequency of female, male and gender-neutral pronouns in the

Data Science Ethics

- Short-cut learning

- E.g.) X-ray scan
  - Trained a classifier, but turns out it works well on some hospital, and works poorly on other hospital
  - Turns out, the deep neural network classifier learned other things than symptoms! (detecting hospital-specific metal token, posture of the X-ray picture)

- Shows why interpretability is important



Shortcut Learning in Deep Neural Networks

Robert Geirhos[1,2,*,§], Jörn-Henrik Jacobsen[3,*], Claudio Michaelis[1,2,*], Richard Zemel[†,3], Wieland Brendel[†,1], Matthias Bethge[†,1] & Felix A. Wichmann[†,1]

[1] University of Tübingen, Germany
[2] International Max Planck Research School for Intelligent Systems, Germany
[3] University of Toronto, Vector Institute, Canada
[*] Joint first / [†] joint senior authors
[§] To whom correspondence should be addressed: robert.geirhos@wichmannlab.org

Data Science Fairness

Example We are building a system to screen mortgage applications. Suppose we collect training data from two demographic groups: 85% White and 15% Black

- Predictive accuracy on the held-out validation set is 95%
- Only 5% error
- Should we sign off on the system as good?

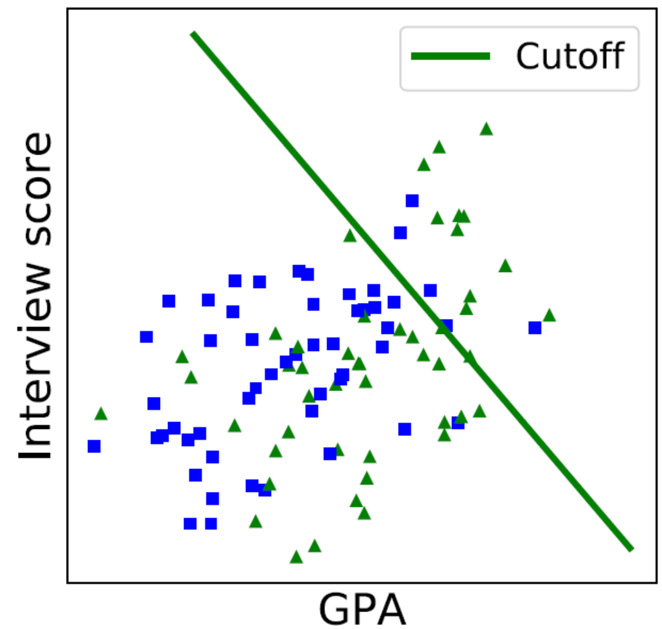
With 5% total error can have up to 66% error on the underrepresented group. Need to report error by each group and aim for 95% accuracy in any group.

Data Science Fairness

Example You are building a system for college admissions based on GPA and interview score (obviously a toy example)

- Fit a least squares regression model
- Model does not account for two demographic groups (blue / green)
- Does this make it fair? (fairness-as-blindness)

Admission rate much lower for blue cohort

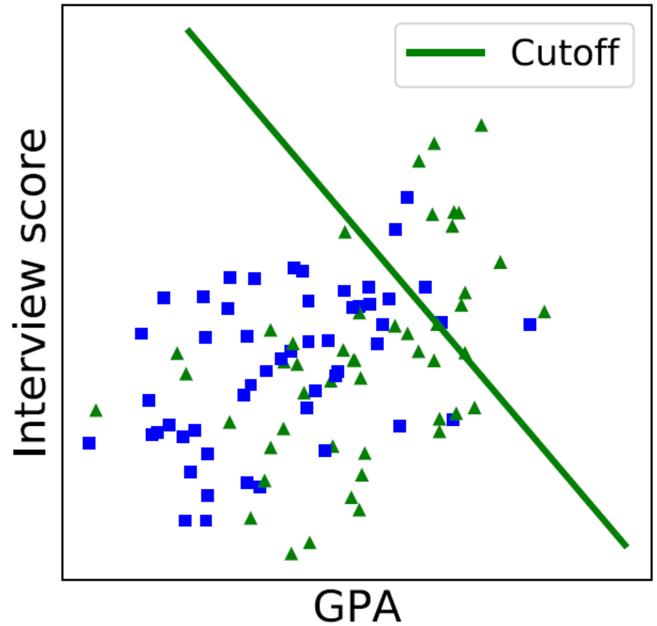


[Source: Brocas et al. “Fairness and ML”]

Data Science Fairness

How to address this behavior?

- GPA correlates with group—omit it as a predictor?
  - Would dramatically impact accuracy
- Pick separate cutoffs (fit separate model) for each group
  - No longer blind to demographics
  - What is the goal for picking cutoffs? Same admission rates?
- Could optimize for diversity among selected candidates
  - Measuring similarity is non-trivial



[Source: Brocas et al. “Fairness and ML”]

Classification Fairness Criteria

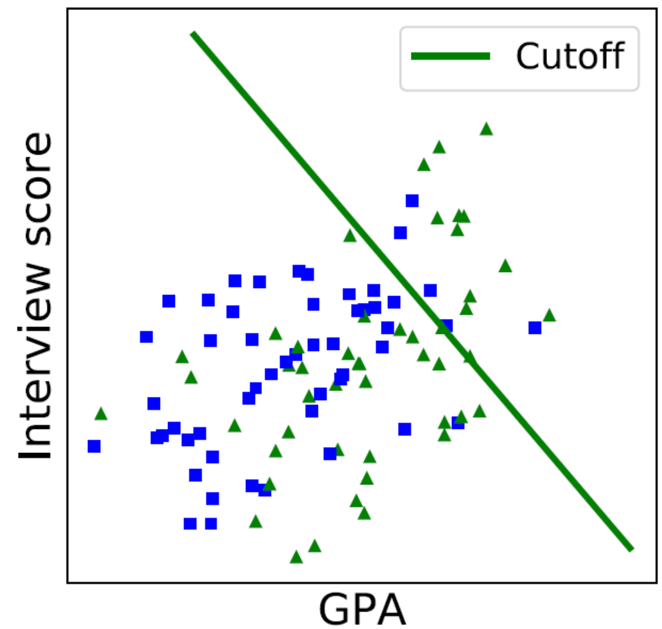
Let *A* be a sensitive attribute, target variable *Y*, and classifier prediction *R.*

Example In our admissions case,

A : Demographic group

R : Prediction of admission

Y : Actual acceptance outcome



Classification Fairness Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Independence The prediction and attribute are independent

Example The probability of predicting admission doesn't differ across demographic groups,

$$P(R \mid A = a) = P(R \mid A = b)$$

Demographic parity, statistical parity, group fairness, disparate impact

Classification Fairness Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Separation Score and attribute are conditionally independent, given the classifier decision

Example There is no relationship between prediction and attribute within accepted / non-accepted groups,

$$P(R \mid Y = 1, A = a) = P(R \mid Y = 1, A = b)$$

$$P(R \mid Y = 0, A = a) = P(R \mid Y = 0, A = b)$$

Classification Fairness Criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Sufficiency Outcome and attribute are independent given the model prediction

Example There is no relationship between whether someone is admitted and their demographic group within predictions

$$P(Y \mid R = 1, A = a) = P(Y \mid R = 1, A = b)$$

$$P(Y \mid R = 0, A = a) = P(Y \mid R = 0, A = b)$$

Data Science Fairness

In short… there is a lot to say on ethics and fairness… and much can be quantified rigorously…

FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

https://fairmlbook.org/

Outline

- Data Science Ethics and Fairness
- Course Review
- Additional Resources
- Final Exam Overview

Probability and Statistics

- We've learned many definitions from this part.
- I want you to remember at least these concepts.
  - How to calculate (joint) probability and expectation
  - How to calculate conditional probability and expectation
  - Understand independence and how to prove/disprove it
- I will submit 9 problems for this part. They are from:
  - Last year's final
  - Our midterm
  - A few new problems

- The following problems are candidates, and I will remove or split some of them to ensure that you can solve all the problems within the given time.
  - I prepared too many problems as subproblems.

- The problem in a completely different form that we did not discuss during today's lecture will not appear.
  - But slight variations (numbers, questions) are possible.

- I recommend you discuss these problems on Piazza.

Problem 1(a)

- modification of the following problem (our midterm)

1. Suppose we have two independent random variables $X \sim \text{Uniform}(\{1,2\})$ and $Y \sim \text{Uniform}(\{1,2,3\})$.
(a) (7 points) Compute the probability mass function of random variable $Z = X \cdot Y$.
(b) (3 points) Compute $\mathbb{E}[Z]$.

- Possible difference
  - X and Y might follow different distributions. (Y~Unif({0,2})? X~Ber(0.7)?)
  - Z might be different (Z=X+Y, Z=X-Y, Z=X^2 + Y^2…)

Problem 2

- modification of the following problem (last final 3)

- Suppose we throw a biased six-sided die twice in a row. We know the distribution of this die – let X be the outcome of this die, then P(X=1)=0.2, P(X=2)=0.1, P(X=3)=0.3, P(X=4)=0.2, P(X=5)=0.1, P(X=6)=0.1. Let S be the sum of both throws. Calculate P(S=5).

- Possible difference
  - Distribution of X might be different (fair die?)
  - Maybe I can ask for a different probability (S=5 and the first throw is even? S=7?)

Problem 3

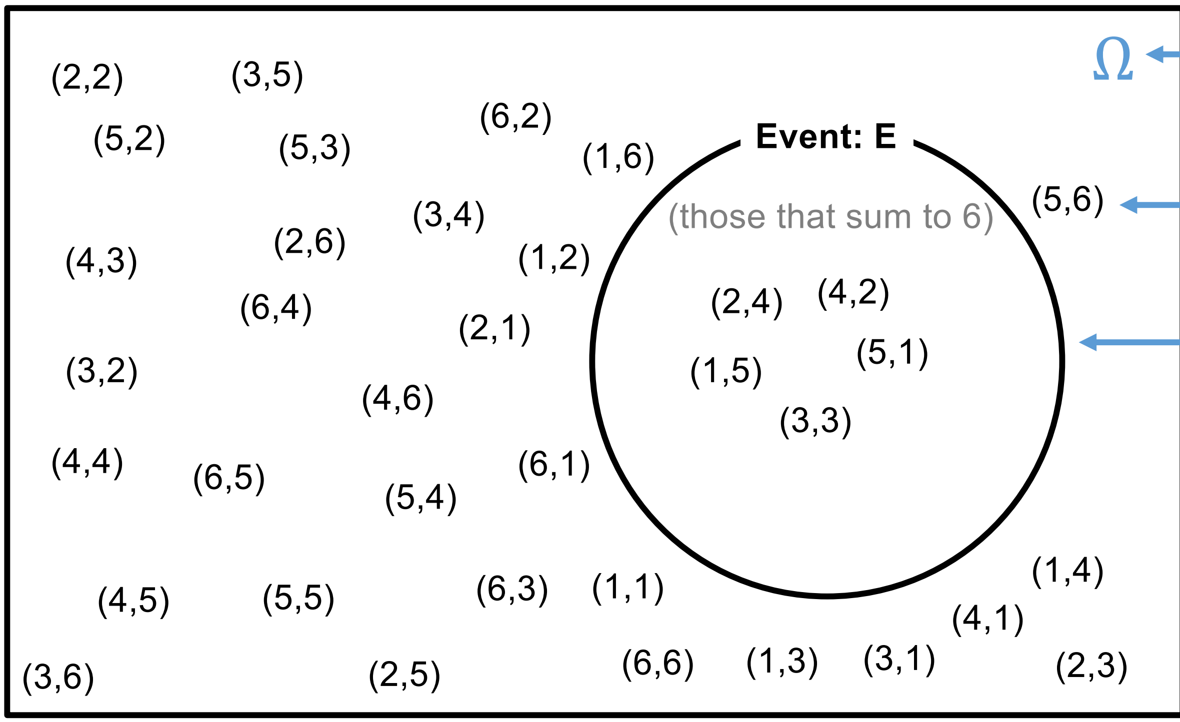
- Modification of the following problem:
- Suppose that we have three events A, B, and C, and we know the following facts:
  - P(A)=0.5, P(B)=0.2, P(C)=0.1
  - Events A and B are independent.
  - Event C is disjoint from A and B.

Now what is the probability of $P(A \cup B \cup C)$?

- Possible difference:
  - number of events, relationship, numbers…

Random Events and Probability

What is the probability of having two outcomes of fair dices sum to 6?



Outcome space consists of all possible outcomes

Each **outcome** $\omega \in \Omega$.

Event is a collection of outcomes and a subset of the outcome space
$E \subseteq \Omega$

Note: outcome = sample

Random Events and Probability

But, what is probability, really?

(e.g., can explain the probability of seeing an event when throwing two dice)

Mathematicians have found a set of conditions that 'makes sense'.

- Probability is a map P. ⇒ i.e., takes in an event, spits out a real value
- P must map events to a real value in interval [0,1].
- P is a (valid) **probability distribution** if it satisfies the following **axioms of probability**,
  1. For any event E, $\mathrm{P}(\mathrm{E}) \geq 0$
  2. $P(\Omega) = 1$
  3. For any *finite* or *countably infinite* sequence of disjoint events $E_1, E_2, E_3, \ldots$

$$P\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} P(E_i)$$

disjoint: intersection is empty

Random Events and Probability

- Many properties follows (i.e., can be proved mathematically)

$$
\begin{aligned}
\mathbb{P}(\emptyset) &= 0 \\
A \subset B &\implies \mathbb{P}(A) \leq \mathbb{P}(B) \\
0 \leq \mathbb{P}(A) &\leq 1 \\
\mathbb{P}(A^c) &= 1 - \mathbb{P}(A) \\
A \bigcap B = \emptyset &\implies \mathbb{P}\left(A \bigcup B\right) = \mathbb{P}(A) + \mathbb{P}(B).
\end{aligned}
$$

E.g., throw a die. A= getting 1, B=getting an odd number

E.g., A= getting 1, B=getting 3 or 5

Random Events and Probability

Law of total probability: Let A be an event. For any events $B_1, B_2, \ldots$ that partitions Ω, we have

$$P(A) = \sum_i P(A \cap B_i)$$

Example Roll two fair dice. Let X be the outcome of the first die. Let Y be the sum of both dice. What is the probability that both dice sum to 6 (i.e., Y=6)?

$$p(Y=6) = \sum_{x=1}^{6} p(Y=6, X=x)$$

$$\mathrm{P}(A,B) := P(A \cap B)$$

$$= p(Y=6, X=1) + p(Y=6, X=2) + \ldots + p(Y=6, X=6)$$

$$= p(X' = 5, X = 1) + p(X' = 4, X = 2) + \cdots + p(X' = 0, X = 6)$$

(X': the outcome of the second die)

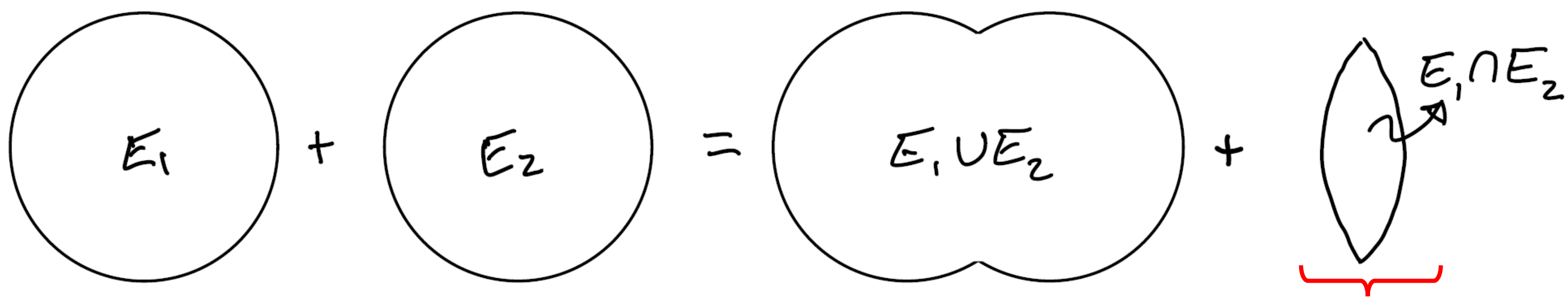
$$= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + 0 = \frac{5}{36}$$

Random Events and Probability

Lemma: (inclusion-exclusion rule) *For <u>any</u> two events* E_1 *and* E_2,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Graphical Proof:



Subtract from both sides

Problem 4

- Modification of the following problem

6. Suppose we have two random variables $X \sim \text{Bernoulli}(0.5)$ and $Y = 0.5X + 0.5$.
(a) (2 points) Compute $\mathbf{E}[X]$ and $\mathbf{E}[Y]$.
(b) (4 points) Compute $\text{Var}(X)$ and $\text{Var}(Y)$
(c) (4 points) Compute $\mathbf{E}[XY]$ and $\text{Cov}(X, Y)$.

- Possible change:
  - Maybe change X and Y to something else (possible scenario: I can introduce additional Z~Bernoulli(0.5) and say Y=X+Z)
  - Maybe remove some unnecessary sub-problems.

Problem 5

- Modification of the following problem

3. Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and $X_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$ are three independent random variables. Let $\hat{\mu} = \frac{1}{3}\sum_{i=1}^{3} X_i$. What distribution does $\hat{\mu}$ follow? Fully specify the parameters of the distribution. Show your work and reasoning.

- Possible change:
  - Number of random variables
  - Maybe I can use explicit numbers instead of μ_i and σ_i

Problem 7(b)

- Modification of the following problem (last year final)

7. Suppose we have placed two advertisements next to each other in a website. A user can either click both, click one of them, or not click at all. Let $A \in \{1, 0\}$ and $B \in \{1, 0\}$ be the random variables indicating whether each ad is clicked (1) or not(0). They follow the following joint probability table.

	$B = 1$	$B = 0$
$A = 1$	$1/8$	$3/8$
$A = 0$	$3/8$	$1/8$

(i) Compute E[B|A=1]

(ii) Is A and B independent or not? Justify your answer formally.

Possible changes: numbers, questions (like, compute E[A|B=1]), but I will ask (ii) definitely.

Independence

- Informally, given two events A and B, they are **<u>independent</u>** if the probability of A is not affected by whether B is true or false (and vice versa)
  - E.g., A = “die1=1” and B=“die2=1” are independent.

    ⟹ we know that the probability of die1 being 1 would not be changed just because die2=1.

- Mathematically, this can be written as $P(A|B) = P(A)$ or $P(B|A) = P(B)$.

- E.g., A = “die1=6” and B=“two dice sum to 6” are **<u>not independent</u>**.

  ∵ intuitively, when B is true, A can never happen! So, P(A|B)=0 but P(A) = 1/6.

- E.g., A = “die1=1” and B=“two dice sum to 6” are **<u>not independent</u>**.

  ∵ P(A) = 1/6 = 0.166... . However, P(A|B) = 1/5 = 0.2

More examples

- Q: A = “die1=1” and B=“two dice sum to 5”. Independent?

No

$\because$ P(A) = 1/6 , P(A|B) = 1/4 = .25

- Q: A = “die1=even” and B=“two dice sum to 5”. Independent?

Yes

$\because$ P(A) = 1/2 , P(A|B) = 2/4 = 1/2

Independence

[Def] Two events A and B are **independent** if

$$P(A, B) = P(A)P(B)$$

"joint probability is product of two marginal probabilities"

$A \perp B$ means A and B are independent

=> note: symmetric!

(skipping the following..)

Also, a set of events $\{A_i \in \mathcal{F}\}_{i=1}^{n}$ (n can be ∞) are **mutually independent** if

for every J ⊆ $\{1, \ldots, n\}$, we have $P\left(\cap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$

(∃ a notion of 'pairwise' independence, but not much useful, so we omit it here)

Independence

- Ex) recall two fair dice
  - We took it for granted that P((1,1)) is 1/36.
  - But why is it true, really?
  - To be rigorous,

$$P(die1 = 1, die2 = 1) = P(die1 = 1)P(die2 = 1) = \frac{1}{6} \cdot \frac{1}{6}$$

  due to independence.

or, …= P(die1=1 | die2=1) * P(die2=1) = P(die1=1) * P(die2=1)

- E.g., two biased coin **C1** and **C2**. Suppose P(C1=H) = 0.3 and P(C2=H) = 0.4. Compute the probability of P(C1=H,C2=T).

0.3·0.6 = 0.18

quiz candidate

Independence

Definition *Two random variables* X *and* Y *are independent given if and only if*

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

for all values x and y, and we say $X \perp Y$.

- From now on, we will just write it down as $p(X, Y) = p(X)p(Y)$
- Property: X and Y are independent if and only if $p(X) = p(X|Y)$ (or $p(Y) = p(Y|X)$)

➢ N RVs are independent if

$$p(X_1, \dots, X_N) = \prod_{i=1}^{N} p(X_i)$$

(Again, for all the possible values $x_1, \dots, x_N$)

Moments of Random Variables

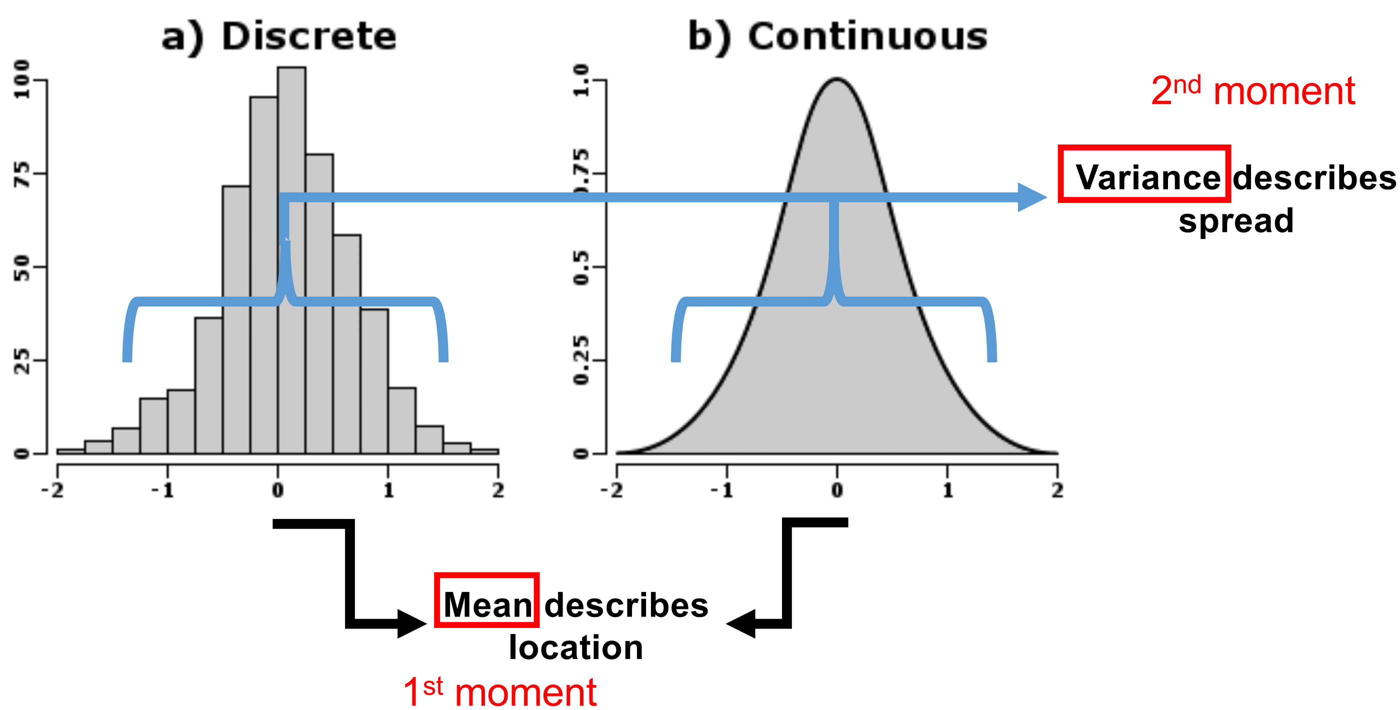
(informal introduction)

Properties of a RV are characterized by its distribution / PMF / PDF
But there are “summary” numbers capturing important characteristics
This is called “**moments**”.

Moment ordinal	Moment			Cumulant	
	Raw	Central	Standardized	Raw	Normalized
1	Mean	0	0	Mean	N/A
2	–	Variance	1	Variance	1
3	–	–	Skewness	–	Skewness
4	–	–	(Non-excess or historical) kurtosis		Excess kurtosis

(Wikipedia)

Moments of Random Variables



Moments characterize properties of the distribution "shape"

Mean = Expectation = Expected Value

Definition *The <u>expectation</u> of a discrete RV* X*, denoted by* $\mathbf{E}[X]$*, is:*

(with PMF) $$\mathbf{E}[X] = \sum_x x \cdot p(X = x)$$

Summation over all values in domain of X

- **<u>Effectively, a weighted average</u>**: *each outcome weighted by probability of occurring*

Some people call it **<u>average</u>** rather than **<u>mean</u>**, but I wouldn't.

$\Rightarrow$ average is a particular 'operator': $\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} x$

$\Rightarrow$ in data science, average is something about the data, not the distribution behind the data

Expected Value

Example Let X be the sum of two fair dice, compute $\mathbf{E}[X]$:

	count	prob.
2: (1,1)	1	1/36
3: (1,2), (2,1)	2	2/36
…	..	…
6: (1,5), (2,4), (3,3), (4,2), (5,1)	5	5/36
7: (1,6), (2,5), (3,4), (4,3), (5,2), (6,1)	6	6/36
8: (2,6), (3,5), (4,4), (5,3), (6,2)	5	5/36
…	…	…
12: (6,6)	1	1/36

Expectation: $2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} +$ ... $+ 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} +$ ... $+ 12 \cdot \frac{1}{36}$ =7

Expected Value

Theorem (Linearity of Expectations) *For any finite collection of RVs $X_1, \cdots, X_n$ with finite expectations,*

$$\mathbf{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbf{E}[X_i]$$

E.g. for two RVs X and Y
$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$

you do not need an independence!

Example Throw two fair dice. What is the expected sum? Let X and Y be the outcome of the first and second die, respectively. Then,

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y] = 3.5 + 3.5 = 7$$

Expected Value

Theorem For any random variable X and constant c,

$$\mathbf{E}[cX] = c\mathbf{E}[X]$$

Example Let X and Y be the outcome of two fair dice, then:

$$\begin{aligned} \mathbf{E}[2(X+Y)] &= \mathbf{E}[2X] + \mathbf{E}[2Y] \\ &= 2\mathbf{E}[X] + 2\mathbf{E}[Y] \\ &= 2 \cdot 3.5 + 2 \cdot 3.5 = 14 \end{aligned}$$

Caveat: c has to be a constant, not a random variable!
E.g., X: outcome of a fair die, c: outcome of another fair die

Expected Value

Definition *The <u>conditional expectation</u> of a discrete RV X, given Y is:*

cf. $\mathbf{E}[X] = \sum_x x \cdot p(X = x)$

$$\mathbf{E}[X \mid Y = y] = \sum_x x\, p(X = x \mid Y = y)$$

Example Roll two fair dice. X_1: first die outcome, Y: sum of two dice

quiz candidate

$$\mathbf{E}[X_1 \mid Y = 5] = \sum_{x=1}^{4} x\, p(X_1 = x \mid Y = 5)$$

$$= \sum_{x=1}^{4} x \frac{p(X_1 = x, Y = 5)}{p(Y = 5)} = \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2}$$

Conditional expectation follows properties of expectation (linearity, etc.)

Expected Value

Example: Two fair dice.

Y=outcome of die 1
X=sum of two dice

X|Y=1 ~ U{2,3,4,5,6,7} E[X|Y=1] = 4.5 $P(Y=1)=\frac{1}{6}$

X|Y=2 ~ U{3,4,5,6,7,8} E[X|Y=2] = 5.5 $P(Y=2)=\frac{1}{6}$

E[X|Y=3] = 6.5

… E[X|Y=4] = 7.5 …

E[X|Y=5] = 8.5

X|Y=6 ~ U{7,8,9,10,11,12} E[X|Y=6] = 9.5 $P(Y=6)=\frac{1}{6}$

$\mathbf{E}_X[X|Y]$ is a random variable:
$\mathbf{E}_X[X|Y] \sim U\{4.5, 5.5, 6.5, 7.5, 8.5, 9.5\}$

Expectation is 7
=> $\mathrm{E}_Y\left[\mathbf{E}_X[X|Y]\right] = 7$
=> coincides with $\mathbf{E}[X]$ we computed before!

Independence and Moments

Theorem: *If* $X \perp Y$ *then* $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$ *.*

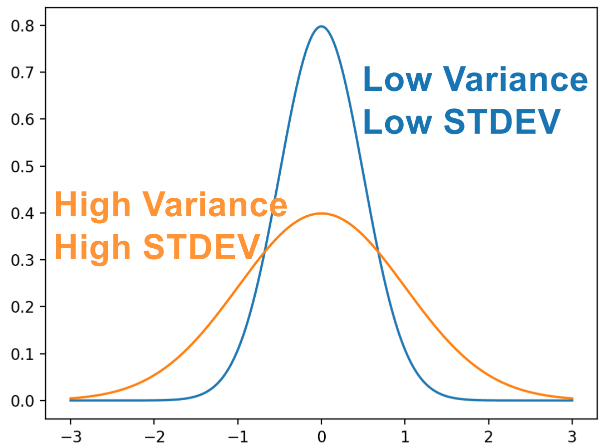
Comparison: $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ regardless of independence!

Variance

Definition The variance of a RV X is defined as,

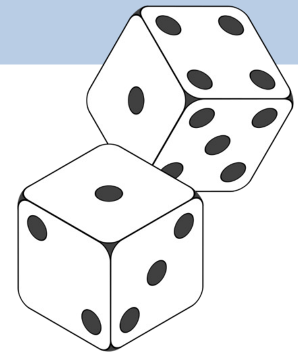
$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$$

The standard deviation (STDEV) is $\sigma[X] = \sqrt{\mathbf{Var}[X]}$.



- *Describes the "spread" of a distribution*
- *Describes uncertainty of outcome*
- *STDEV is in original units (more intuitive), variance is in units^2*
- *Variance is more mathematically useful than STDEV*

Variance



Example Let X be the result of a fair six-sided die.

The variance is then,

$$\mathrm{Var}(X) = \sum_{i=1}^{6} \frac{1}{6}\left(i - \frac{7}{2}\right)^2$$

$$= \frac{1}{6}\left((-5/2)^2 + (-3/2)^2 + (-1/2)^2 + (1/2)^2 + (3/2)^2 + (5/2)^2\right)$$

$$= \frac{35}{12} \approx 2.92.$$

The STDEV is $\sqrt{\mathrm{Var}(X)} \approx 1.71$, which suggests we should expect outcomes to vary around the mean of 3.5 by $\pm$ 1.71

Variance

Lemma An equivalent form of variance is:

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

Proof

$$\mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] \quad \textbf{(Expand it)}$$

$$= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 \quad \textbf{(Linearity of expectations)}$$

$$= \mathbf{E}[X^2] - 2\mathbf{E}[X]^2 + \mathbf{E}[X]^2 \quad \textbf{(Algebra)}$$

$$= \mathbf{E}[X^2] - \mathbf{E}[X]^2 \quad \textbf{(Algebra)}$$

Variance

- If c is a constant, $Var[cX] = c^2Var[X]$

- Important that c has to be a constant here!

Independence and Moments

Recall that for any two RVs X and Y variance is not a linear function,

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}(X,Y)$$

If X and Y are independent then they have zero covariance,

$$\mathbf{Cov}(X,Y) = 0$$

Thus, $$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$$

And, for a collection of independent RVs $X_1, X_2, \ldots, X_N$ we have,

$$\mathbf{Var}(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} \mathbf{Var}(X_i)$$

Q: Is variance is a linear operator under independence?

A: No! $Var(cX) \neq c\,Var(X)$ for a constant c. Rather, $Var(cX) = c^2 Var(X)$.

Covariance

Definition *The covariance of two RVs* X *and* Y *is defined as,*

$$\mathbf{Cov}(X,Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

Question *What is Cov(X,X)?*

Answer *Cov(X,X) = Var(X)*

Covariance

- A shortcut to compute covariance.
- $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$

$$= E\big[XY - X \cdot E[Y] - Y \cdot E[X] + E[X]E[Y]\big]$$

$$= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]$$

$$= E[XY] - E[X]E[Y]$$

- Safety check: $Cov(X, X) = E[XX] - E[X]E[X] = Var(X)$

Problem 6

- Modification of the following problem:
- You tossed a fair coin three times. Let X be the number of heads you observed after three tosses.
  - 1) What is the probability that you observe exactly one head?
  - 2) Compute $E[X^2]$.
  - 3) Given that you observed the result of the first coin was tail, what is the conditional expectation of X?

- Possible difference:
  - Number of tosses, question details, but for 1) I will ask probability, 2) I will ask expectation(or variance), 3) I will ask conditional expectation.
  - Maybe I can remove one sub-problem from here…

Problem 7(a)

- Modification of the following problem (last year final)

7. Suppose we have placed two advertisements next to each other in a website. A user can either click both, click one of them, or not click at all. Let $A \in \{1, 0\}$ and $B \in \{1, 0\}$ be the random variables indicating whether each ad is clicked (1) or not(0). They follow the following joint probability table.

	$B=1$	$B=0$
$A=1$	$1/8$	$3/8$
$A=0$	$3/8$	$1/8$

(i) Compute E[B|A=1]

(ii) Is A and B independent or not? Justify your answer formally.

Problem 8

- Modification of the following problem:(our midterm, last final 9)

8. A spam email detector has the following property:

- Given a spam email, it will raise an alarm with probability 0.9.
- Given a non-spam email, it will raise an alarm with probability 0.01.

In addition, suppose that 1 out of 10 emails is a spam email. Given an email, we use $S \in \{\text{True}, \text{False}\}$ to denote whether it is a spam, and use $R \in \{\text{True}, \text{False}\}$ to denote whether the spam detector raises an alarm.
(a) (5 points) Compute the joint probability distribution of (S, R).
(b) (5 points) Conditioned on that the spam detector raises an alarm, what is the probability that the email is a spam?

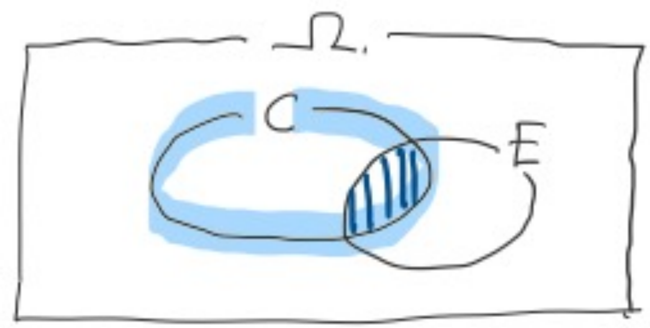
- I will only change numbers.

Problem 9

- Modification of the following problem:
- A box contains two coins. One fair coin and one two-headed coin. You picked a coin at random and tossed it.
  - You observed it lands heads up. What is the probability that the coin you chose was a fair coin?
  - You observed it lands heads up. What is the expected number of heads inside the box? (note: the fair coin has only one head, and the two-headed coin has two heads)
- Possible changes: number of coins, bias, question slightly…

Conditional Probability

- Two fair dice example:
  - Suppose I roll two dice secretly and tell you that one of the dice is 2. C
  - **<u>In this situation</u>**, find the probability of two dice summing to 6. E
- Turns out, such a probability can be computed by $\frac{P(E \cap C)}{P(C)}$
- It's like "zooming in" to the condition.
- This happens a lot in practice, so let's give it a notation:



$$P(E|C) := \frac{P(E \cap C)}{P(C)}$$

Say: probability of "E given C", "E conditioned on C"

"it's the ratio"

Conditional Probability

Chain rule

- $P(A \cap B) = P(A|B)P(B)$ ←just a rearrangement of definition
- $P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$
- $P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) \prod_{i=2}^{n} P(E_i | \cap_{j=1}^{i-1} E_j)$ valid for any ordering!

Law of total probability: If $A \in \mathcal{F}$ and $\{B_i \in \mathcal{F}\}_i$ partitions Ω, then

$$P(A) = \sum_i P(A, B_i) = \sum_i P(B_i)P(A|B_i)$$

$$= \sum_i P(A)P(B_i|A)$$

(trivially true by definition)

Shortcut:
P(A,B) := P(A ∩ B)

Conditional Probability

[W:Ex.6.9] The Public Health Department gives us the following information:

- A test for the disease yields a positive result 90% of the time when the disease is present

  P(test=+ | disease=Y) = 0.9

- A test for the disease yields a positive result 1% of the time when the disease is not present

  P(test=+ | disease=N) = 0.01

- One person in 1,000 has the disease.

  P(disease=Y) = 0.001

<u>**Q**</u>: What is the probability that a person with positive test has the disease?

P(disease=Y | test=+)

Conditional Probability

What we know:

P(test=+ | D=Y) = 0.9 $\quad$ P(test=- | D=Y) = 0.1

P(test=+ | D=N) = 0.01 $\quad\Rightarrow\quad$ P(test=- | D=N) = 0.99

P(D=Y) = 0.001 $\quad$ P(D=N) = 0.999

Question: P(D=Y | test=+)

$$= \frac{P(D = Y\ , test = +)}{P(test = +)}$$

$$P(test = +) = P(test = +, D = Y) + P(test = +, D = N)$$

$$P(test = +, D = Y) = P(test = +|\ D = Y)P(D = Y)$$

$$P(test = +, D = N) = P(test = +|\ D = N)P(D = N)$$

The answer is 0.0826…

CAVEAT: P(test=+ | D=Y) = 0.9 ≠ P(D=Y | test=+)

Also: P(D=Y) = 0.001 vs P(D=Y | test=+)

Expected Value

Definition *The <u>conditional expectation</u> of a discrete RV X, given Y is:*

cf. $\mathbf{E}[X] = \sum_x x \cdot p(X = x)$

$$\mathbf{E}[X \mid Y = y] = \sum_x x\, p(X = x \mid Y = y)$$

Example Roll two fair dice. X_1: first die outcome, Y: sum of two dice

quiz candidate

$$\mathbf{E}[X_1 \mid Y = 5] = \sum_{x=1}^{4} x\, p(X_1 = x \mid Y = 5)$$

$$= \sum_{x=1}^{4} x \frac{p(X_1 = x, Y = 5)}{p(Y = 5)} = \sum_{x=1}^{4} x \frac{1/36}{4/36} = \frac{5}{2}$$

Conditional expectation follows properties of expectation (linearity, etc.)

Problem 10

- I will change the question to some other example.

10. Suppose we would like to do an exit poll for a presidential election where we ask people coming out of poll stations who they have voted for. Due to resource constraints, we cannot ask every single person. Which sampling method is appropriate?

(1) Simple random sampling.

(2) Systematic sampling.

(3) Stratified sampling.

(4) Cluster sampling.

Problem Candidates

- These are my current (format) candidates for the first half, but I will talk with our TAs and will
  - 1) Remove some of the problems or sub-problems
  - 2) Split one problem into two
  - 3) Substitute some problems with candidates on the next slides.

Candidate 1

- This is Problem 1

1. Suppose we have two independent random variables $X \sim \text{Uniform}(\{1, 2\})$ and $Y \sim \text{Uniform}(\{1, 2, 3\})$.
(a) (7 points) Compute the probability mass function of random variable $Z = X \cdot Y$.
(b) (3 points) Compute $\mathbb{E}[Z]$.

- Adding some continuing problem…?

2. Continuing the previous problem,
(a) (5 points) Compute $\mathbb{P}(X = 2 \mid Z = 2)$.
(b) (5 points) Compute $\mathbb{E}\left[X \mid Z = 2\right]$.

Candidate 2

- Last year final

2. Suppose we throw a fair six-sided die twice in a row. Let A be a random variable representing the number on the first throw, and B be the number on the second throw. Let S be the sum of both throws. Compute $P(A = 2 \mid S = 2)$. 0

4. Under the same setup as the previous problem (the nonfair one), compute $P(A = 1 \mid S = 3)$. $0.02/0.04 = 0.5$

Possible modification: biases, detailed numbers…

Candidate 3

5. Suppose that the random variable X has the following distribution:

$$P(X = 0) = 0.3, P(X = 1) = 0.2, P(X = 2) = 0.5$$

Compute Var(X).

Possible modification: distribution, compute something else (E[X^3], E[X^2+X], …)

Computer Science

CSC380: Principles of Data Science

Wrap-up 2

Kyoungseok Jang

Announcement

- No office hours this Thursday and Friday (May 4^{th} and 5^{th})

- ML problem: Problem 11~18, maybe 19.
  - If we have 18 problems: 4 of them will have 15 points each, and the rest of them will have 10 points each.
  - If we have 19 problems: 2 of them will have 15 points each, and the rest of them will have 10 points each.
  - In any case,100 pts for the ML part.

Problem 11

- Cross validation (maybe change n or k to numbers)

11. You have a data set with n items and you want to evaluate neural network's performance. For each of the following methods, how many neural networks do you need to train, and how many training data points will each neural network be trained on?

(Except for the last retraining step)

1. Split data into 70% training and 30% test.
2. K fold cross validation.
3. Leave-one-out.

Answer:

1. once, 0.7n
2. K, (K-1)/K * n
3. n, n-1

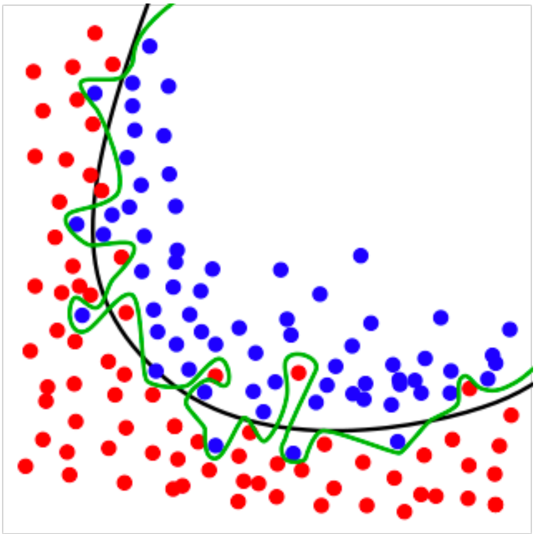
Challenges in ML

Why not learn a very complex function that can have 0 train set error and be done with it?

<u>**Extreme example:**</u> Let's memorize the data. To predict an unseen data, just guess a random label.

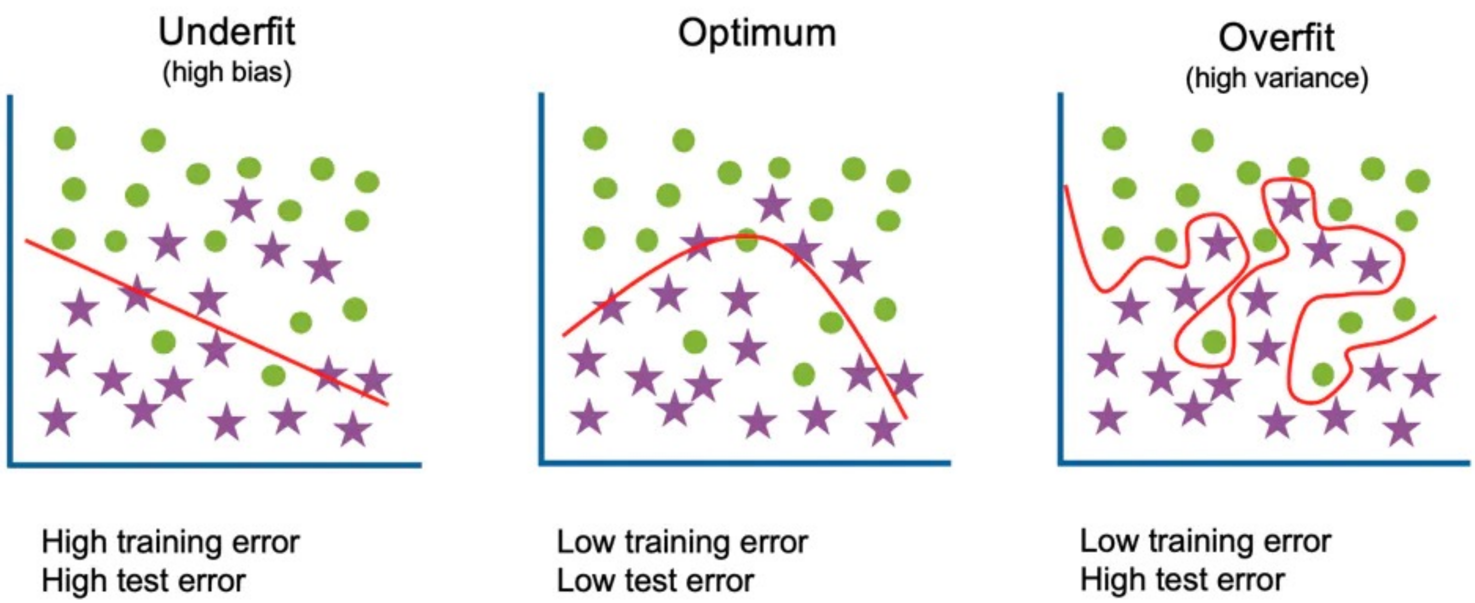
Doesn't generalize to unseen data – called *overfitting* the training data.

<u>**Solution:**</u> Fit the train set but don't "over-do" it. This is called **regularization**.



green: almost memorization
black: true decision boundary

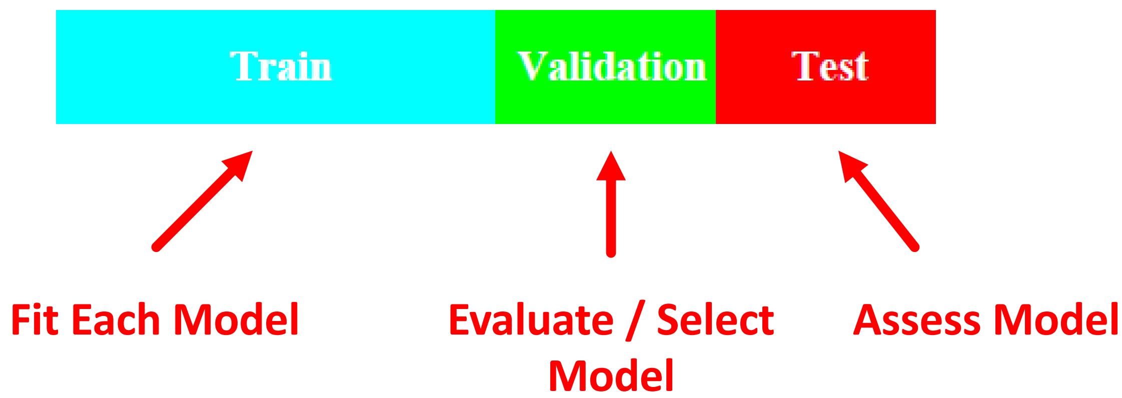
Overfitting vs Underfitting



Source: ibm.com

Model Selection / Assessment

Partition your data into Train-Validation-Test sets



- Ideally, Test set is kept in a "vault" and **only peek at it once model is selected**
- Small dataset: 50% Training, 25% Validation, 25% Test (very loose rule set by statisticians)
- For large data (say a few thousands), 80-10-10 is usually fine.

Tuning hyperparameters (e.g., k in k-NN)

Validation set method:

- For each hyperparamter $h \in H$
  - Train $\hat{f}$ on train set with h
  - Compute the error rate of $\hat{f}$ on validation set
- Choose the best performing hyperparameter h^*
- Use h^* to retrain the final model $\hat{f}^*$ with both train and validation set.
- Finally, evaluate $\hat{f}^*$ on test set to estimate its future performance.

hyperparameter: parameters of the model that are not trained automatically by ML algorithms.

parameters: those that are trained automatically (e.g., tree structures in decision tree)

Pro tip

- Do not use arithmetic grids; use geometric grids.

Don't	k = 1, 3, 5, 7, 9, …
Do	k = 1, 2, 4, 8, 16, …

Downside: How much do we trust the validation set?

Tuning hyperparameters

K-fold cross validation

- Randomly partition train set S into K disjoint sets; call them $\mathrm{fold}_1, \ldots, \mathrm{fold}_K$
- For each hyperparameter $h \in \{1, \ldots, H\}$
  - For each $k \in \{1, \ldots, K\}$
    - train $\hat{f}_k^h$ with $S \setminus \mathrm{fold}_\mathrm{k}$
    - measure error rate $e_{h,k}$ of $\hat{f}_k^h$ on fold_k
  - Compute the average error of the above: $\widehat{err}^h = \frac{1}{K} \sum_{k=1}^K e_{h,k}$
- Choose $\hat{h} = \arg\min_h \widehat{err}^h$
- Train $\hat{f}^*$ using S (all the training points) with hyperparameter $\hat{h}$
- Finally, evaluate $\hat{f}^*$ on test set to estimate its future performance.

K=10 is standard, but K=5 is okay, too

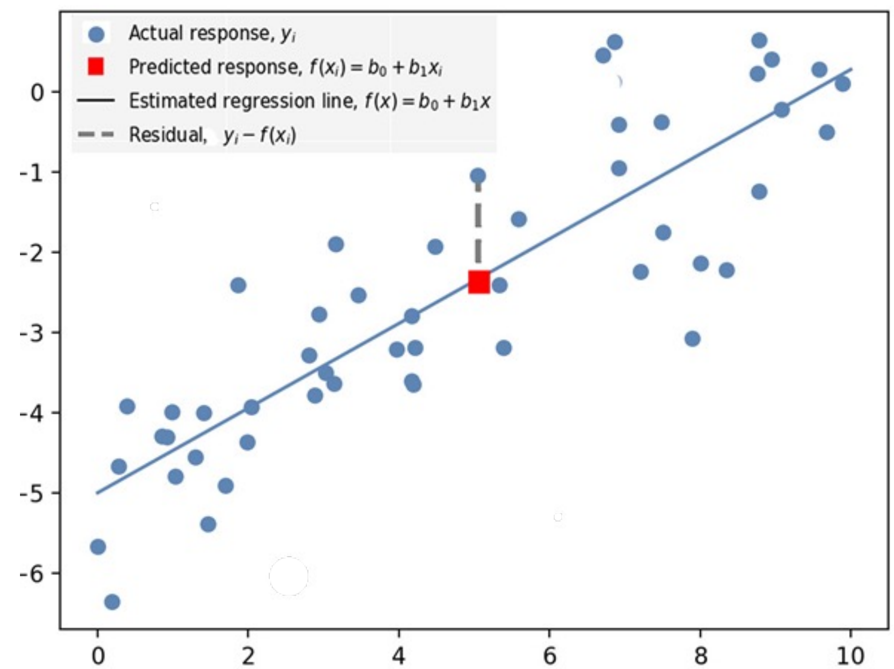
Leave one-out = m-fold cross validation (m: train set size)

$\Longrightarrow$ When (1) the dataset is small (2) ML algorithm's retraining time complexity is low (e.g., kNN)

Problem 13

- Linear Regression (5 points each)
- Suppose that we have the following 4 data points
  - $(x_1, y_1) = (0,1), (x_2, y_2) = (1,0), (x_3, y_3) = (2,1), (x_4, y_4) = (1,2)$
- And we have two linear regression models
  - Model 1: $y = 0 \cdot x + 1$
  - Model 2: $y = 1 \cdot x + 0$
- A) Calculate the residual of (x_1, y_1) with respect to Model 1
- B) Calculate the residual of (x_2, y_2) with respect to Model 2
- C) State which model is a better linear regression when the loss function is the sum-of-square residuals.
- C') Suppose that it is known that one of the two models is the ordinary least square solution. State which one is the OLS solution. (note: you don't need to know the OLS formula for this problem!)

Fitting Linear Regression



Intuition Find a line that is as *close as possible* to every training data point

The distance from each point to the line is the **residual**

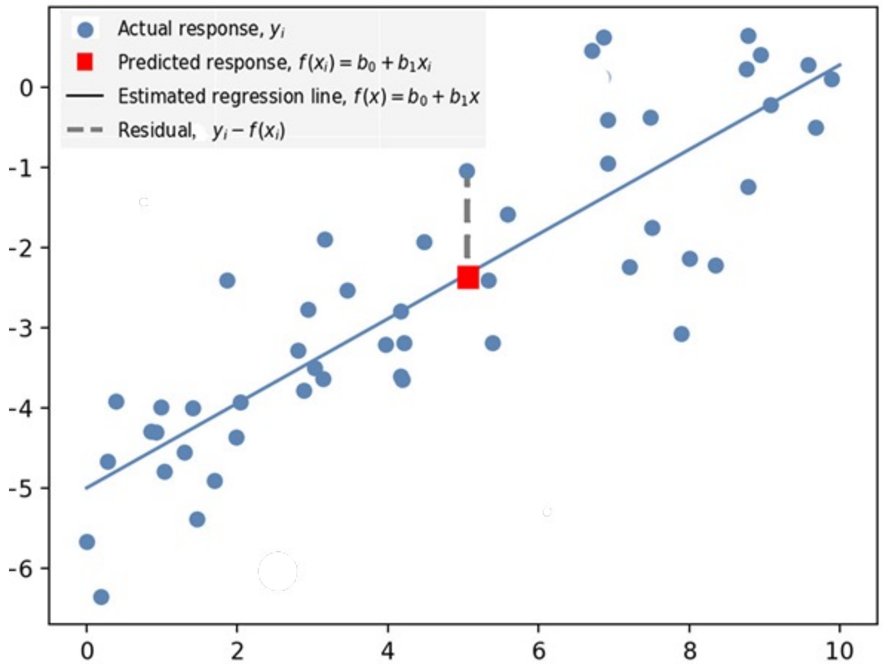
$$y - w^T x$$

Training Output **Prediction**

Let's find w that will minimize the residual!

https://www.activestate.com/resources/quick-reads/how-to-run-linear-regressions-in-python-scikit-learn/

Least Squares Solution



Functional Find a line that minimizes the sum of squared residuals!

Given: $\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$

Compute:

$$w^* = \arg\min_{w} \sum_{i=1}^{m} \left(y^{(i)} - w^T x^{(i)}\right)^2$$

Least squares regression

https://www.activestate.com/resources/quick-reads/how-to-run-linear-regressions-in-python-scikit-learn/

Problem 14

- K-means clustering (15 points)

- Suppose that we have the following 4 data points
  - $x_1 = (5,4), x_2 = (0,1), x_3 = (1,0), x_4 = (4,5)$
- Starting from the initial centroids $y_1 = (0,0)$ and $y_2 = (5,5)$, run the k-means clustering algorithm until the centroids don't move anymore. State your final clustering result (which means, state which points are in the same cluster)

k-means clustering

Input: k: num. of clusters, $S = \{x_1, \ldots, x_n\}$

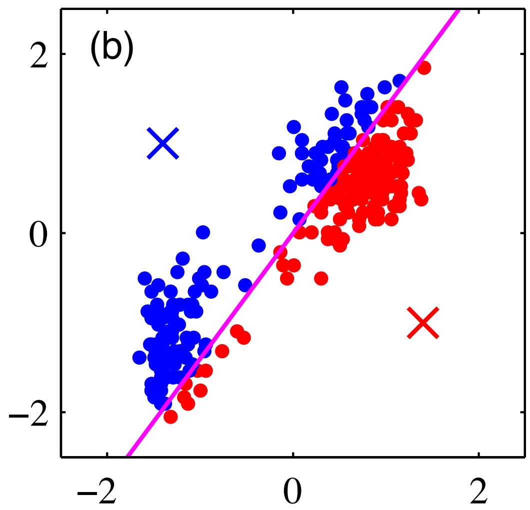
[Initialize] Pick $c_1, \ldots, c_k$ as randomly selected points from S (see next slides for alternatives)

For t=1,2,…,max_iter

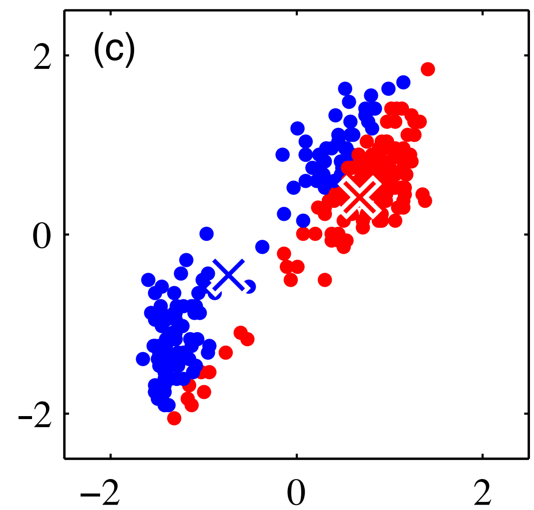
- ***[Assignments]*** $\forall x \in S, \quad a_t(x) = \arg\min_{j \in [k]} \left\| x - c_j \right\|_2^2$
- If $\mathrm{t} \neq 1 \quad \text{AND} \quad a_t(x) = a_{t-1}(x), \forall x \in S$
  - break
- ***[Centroids]*** $\forall j \in [k], \quad c_j \leftarrow \text{average}(\ \{x \in S: a_t(x)\ = j\}\ )$

Output: $c_1, \ldots, c_k$ and $\{a_t(x_i)\}_{i \in [n]}$

Iteration 1

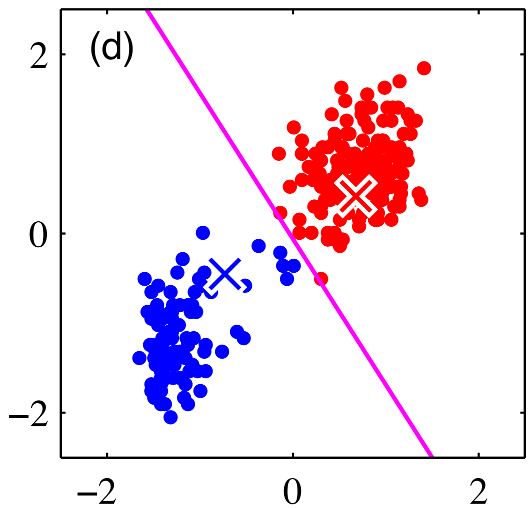


(A) update the cluster assignments.

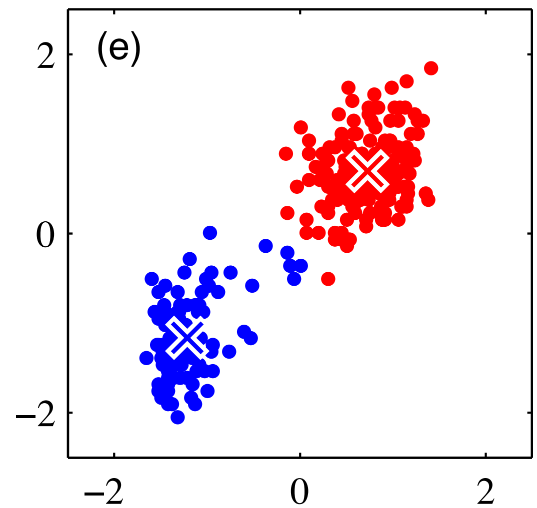


(B) Update the centroids $\{c_j\}$

Iteration 2

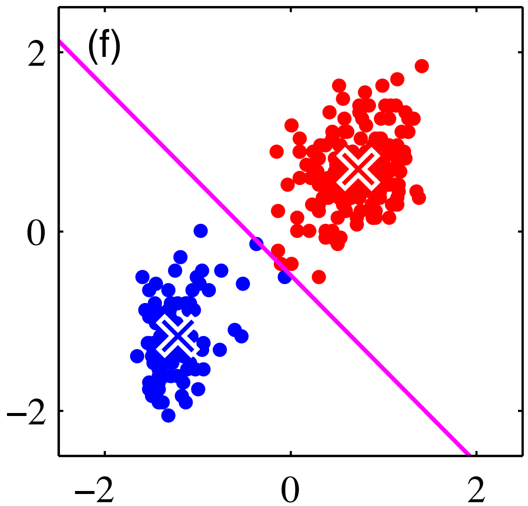


(A) update the cluster assignments.

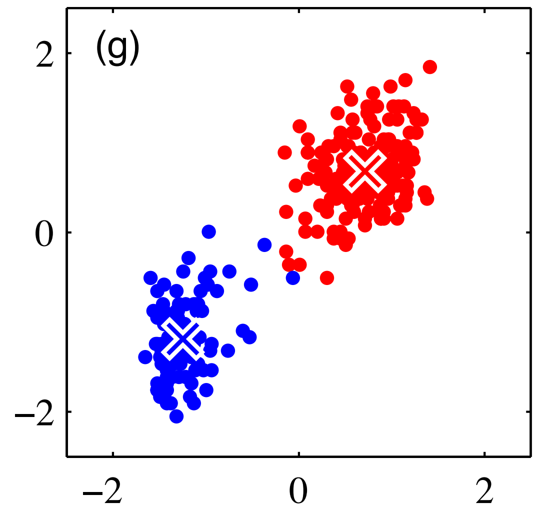


(B) Update the centroids $\{c_j\}$

Iteration 3

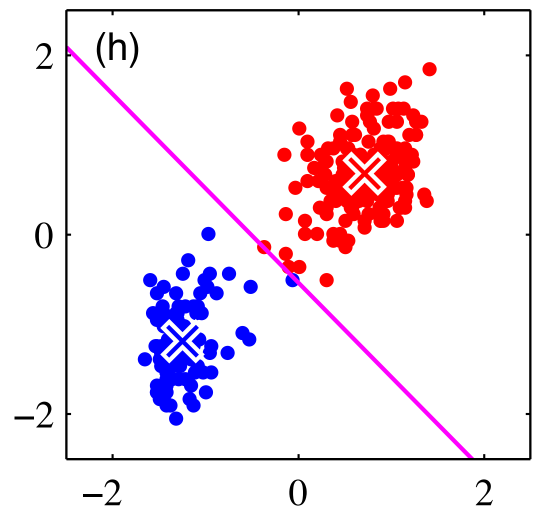


(A) update the cluster assignments.

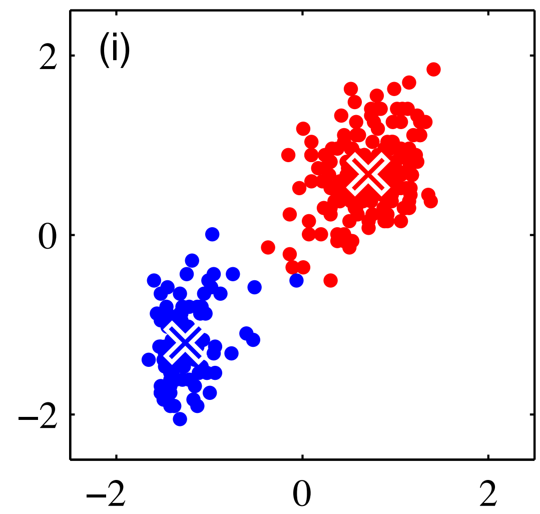


(B) Update the centroids $\{c_j\}$

Iteration 4



(A) update the cluster assignments.



(B) Update the centroids $\{c_j\}$

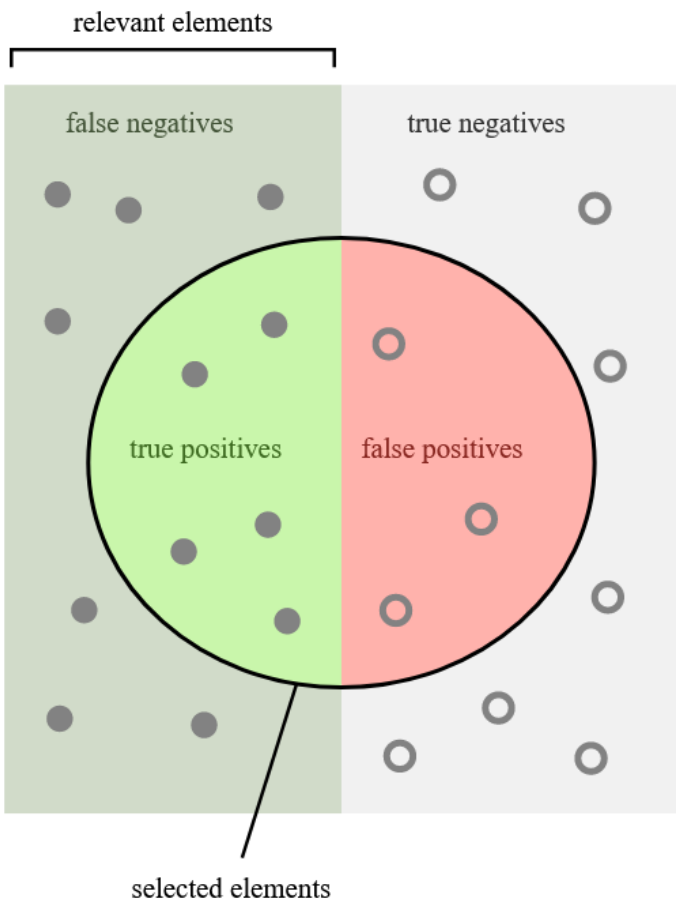
Problem 15

- Precision Recall F1 (maybe it will be good for you to remember the definitions)

14. There are 1000 photos in a data set, 100 of which are of fish. A fish detector applied on the data set claimed 120 photos as fish, of which 80 photos are truly fish. What is the precision and recall of the fish detector?

- Maybe change number
- Maybe I will ask you F1 score additionally

Evaluating Classifiers



For binary classifiers we evaluate a couple standard metrics,

How many selected items are relevant?

Precision = 

$$\frac{\text{True Positives}}{\text{True Positives} \quad + \text{False Positives}}$$

How many relevant items are selected?

Recall = 

$$\frac{\text{True Positives}}{\text{True Positives} \quad + \text{False Negatives}}$$

Evaluating Classifiers

Tuning with precision vs. recall can be tricky, so we use F1 score,

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

- This is the *harmonic mean* of precision and recall
  - min(x,y) <= harmonic_mean(x,y) <= geometric_mean(x,y) <= arithmetic_mean(x,y) <= max(x,y)

$$\frac{1}{\frac{1}{2}(\frac{1}{x} + \frac{1}{y})} \qquad \sqrt{xy} \qquad \frac{1}{2}(x+y)$$

- Can be very sensitive to *class imbalance* (num. positives vs negative)

- Gives equal importance to precision and recall – F1 may not be best when you care about one more than the other (e.g., in medical tests we care about recall)

Problem 16

- Suppose we trained a support vector machine classifier with a linearly separable dataset. Our model has the slope and intercept as $w = (1,1), b = 2$.
- Q1) What is the prediction of your classifier when your input is
  - $x^* = (0,0)$?
  - $x^* = (3, -6)$?
- Q2) Find out a support vector from these candidates (hint: $y^{(i)} = \pm 1$)

  (1) $x^{(1)} = (1,1)$ (2) $x^{(2)} = (2, -3)$ (3) $x^{(3)} = (4, -4)$

Problem 17

- Decision Tree (15 points)

19. Suppose we have the following 7 data points where the first column is the label:

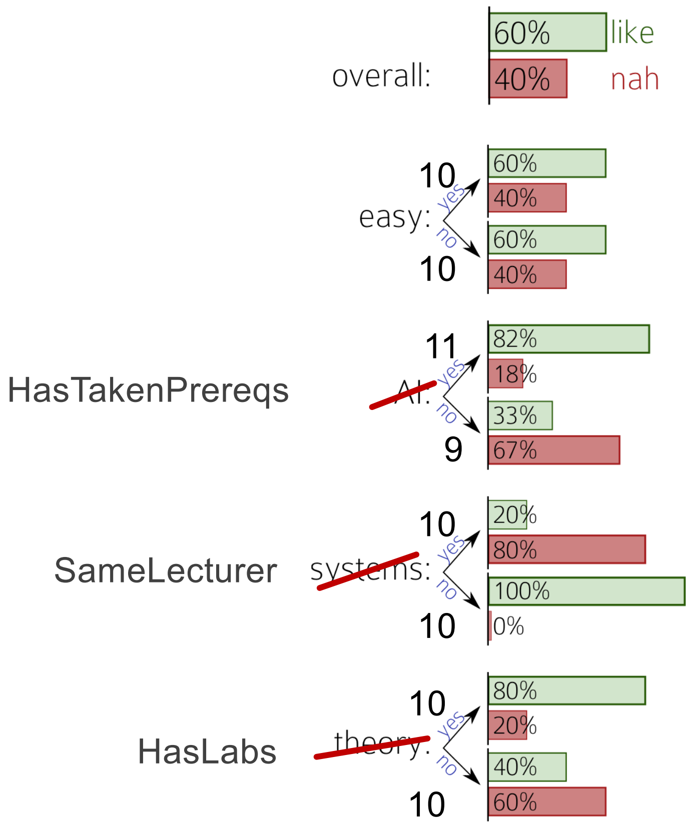
label	f_1	f_2	f_3
$+$	1	1	0
$+$	0	1	1
$+$	1	0	0
$+$	1	0	0
$-$	0	0	1
$-$	1	0	1
$-$	0	1	1

- We want to build a depth-2 decision tree.
- A) Report which feature should be chosen as the root node. (10 pts) (Last year final exam problem)
- B) Let's call the feature on the root node as f^*. Report the next node for $f^* = 1$ subset. (5 pts)

Algorithm 1 DecisionTreeTrain(*data*, *remaining features*)

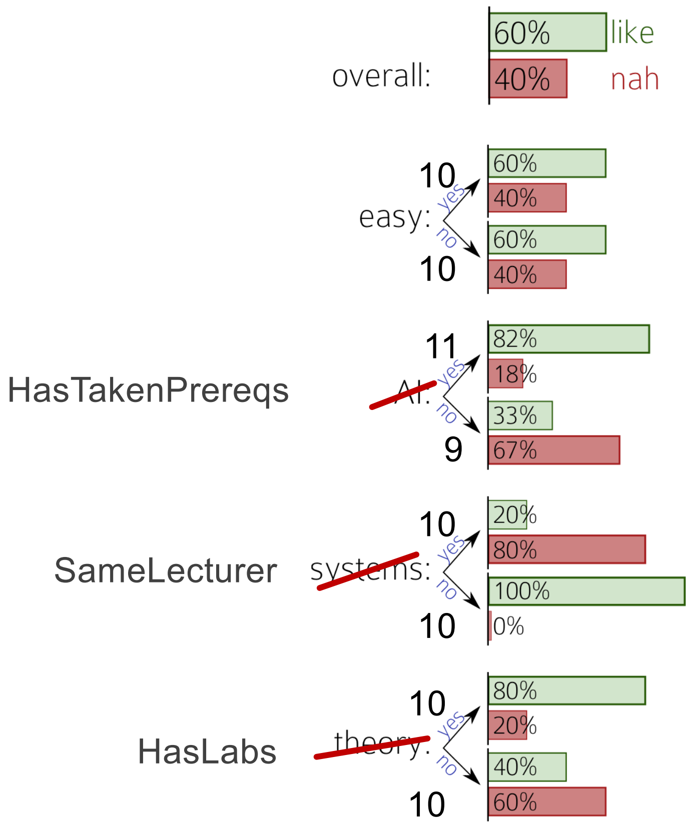
```
guess ← most frequent answer in data                     // default answer for this data
if the labels in data are unambiguous then               <= i.e., all data points have the same label
    return Leaf(guess)                                   // base case: no need to split further
else if remaining features is empty then
    return Leaf(guess)                                   // base case: cannot split further
else                                                     // we need to query more features
    for all f ∈ remaining features do                    <= there is no point in adding a feature
                                                            that appeared in its parent!
        NO ← the subset of data on which f=no
        YES ← the subset of data on which f=yes
        score[f] ← ( # of majority vote answers in NO    <= answer = label
                   + # of majority vote answers in YES ) / size(data)

    end for
    f ← the feature with maximal score(f)
    NO ← the subset of data on which f=no
    YES ← the subset of data on which f=yes
    left ← DecisionTreeTrain(NO, remaining features \ {f})
    right ← DecisionTreeTrain(YES, remaining features \ {f})
    return Node(f, left, right)
end if
```

How to construct a tree



		Prereqs	Lecturer	HasLabs	
Rating	**Easy?**	~~**AI?**~~	~~**Sys?**~~	~~**Thy?**~~	**Morning?**
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

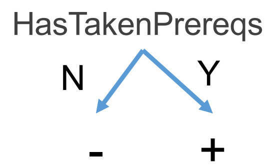
How to construct a tree



Baseline: majority vote classifer

Q: What is the train set accuracy? 0.60

Suppose we place the node HasTakenPrereqs at the root.
Set the prediction at each leaf node as the majority vote.

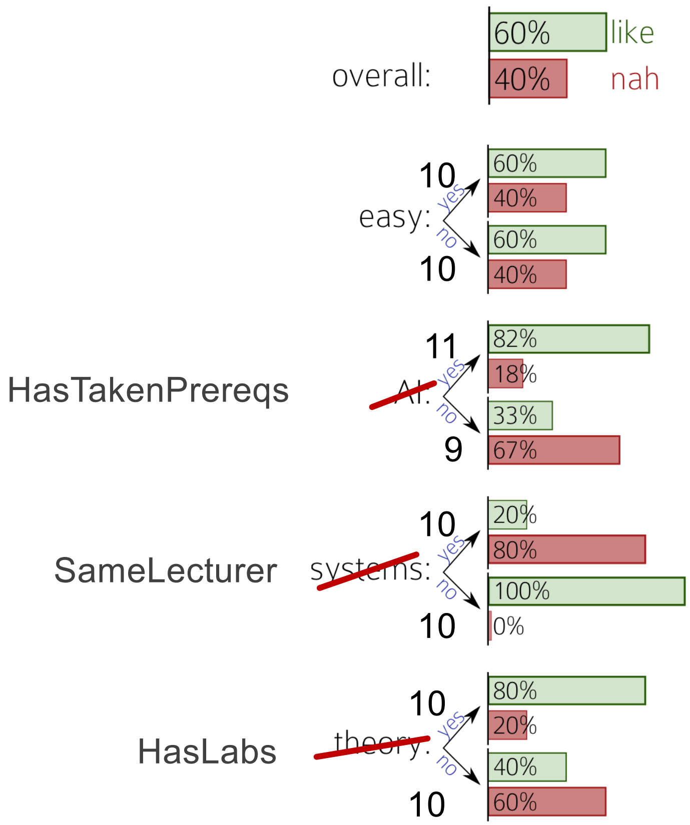


What is the train set accuracy now?

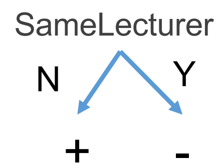
$$\frac{9}{20}\cdot\frac{6}{9}+\frac{11}{20}\cdot\frac{9}{11}=\frac{15}{20}=0.75$$ improved!

$$\frac{6}{20}+\frac{9}{20}=\frac{15}{20}=0.75$$ Alternative way to calculate

How to construct a tree



Suppose placing the node SameLecturer at the root.



What is the train set accuracy now?

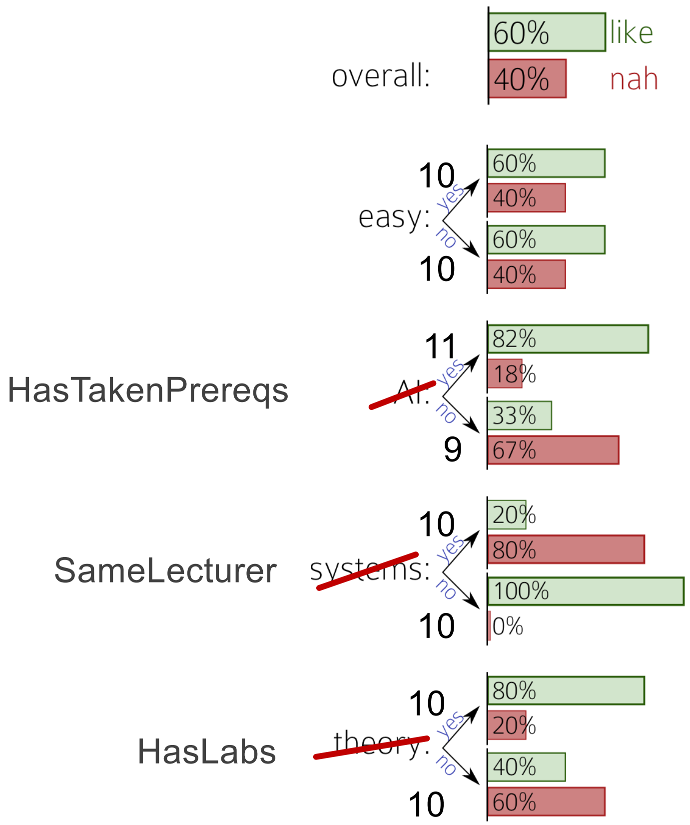
$$\frac{10}{20} \cdot \frac{10}{10} + \frac{10}{20} \cdot \frac{8}{10} = \frac{18}{20} = 0.9$$

even better!

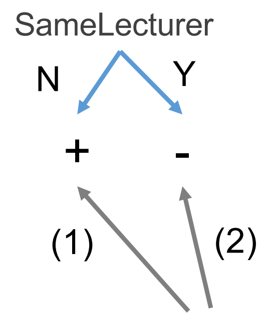
What would you do to build a depth-1 tree?

try out each feature and choose the one that leads to the largest accuracy!

How to construct a tree



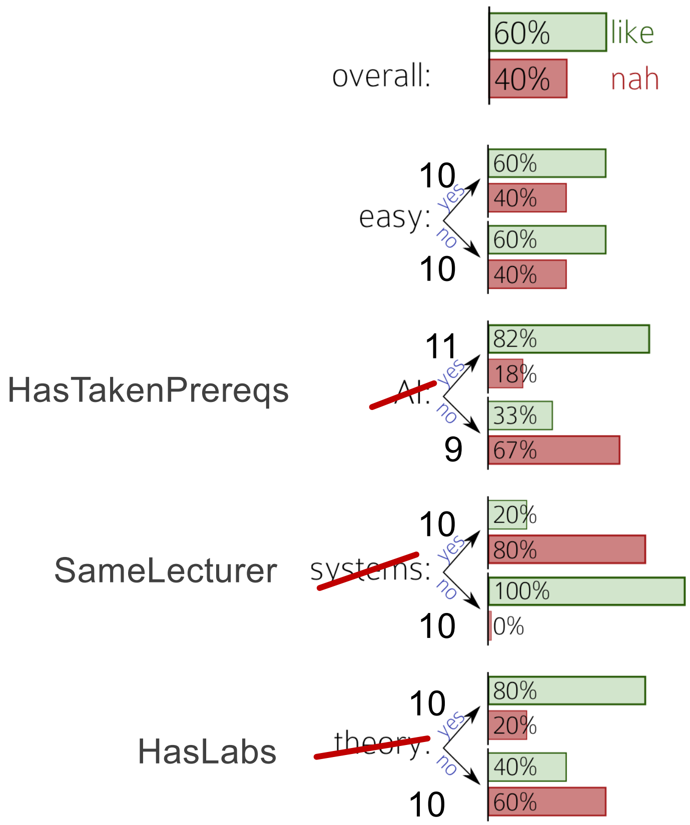
What about depth 2?

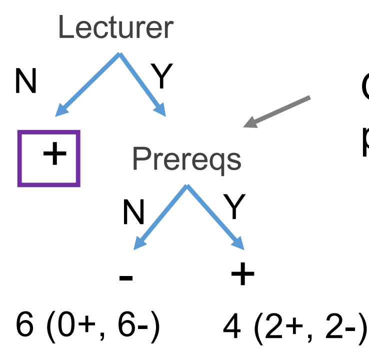


Which nodes to put at each leaf node?

Focus on (2). Try placing HasTakenPrereqs

How to construct a tree





Q: How many training data points fall here? 10

Q: How many training data points arrive at these two leaves? How many for each label?

Q: what prediction should we use for each leaf?

Q: What is the train set accuracy, conditioning on SameLecturer=Y?

'local' train set accuracy $\frac{6}{10} \cdot \frac{6}{6} + \frac{4}{10} \cdot \frac{2}{4} = \frac{8}{10}$

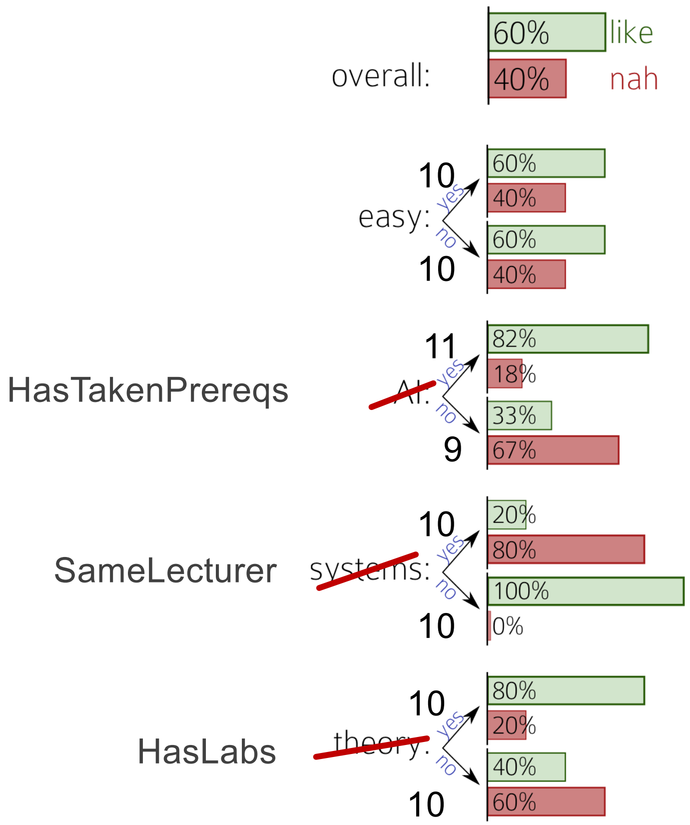
Try all the other nodes and pick the one with the largest acc.!

Then, repeat the same for SameLecturer=N branch!

=> but this has 1 local train set acc. So leave it be!

Move onto expanding nodes at depth 2!

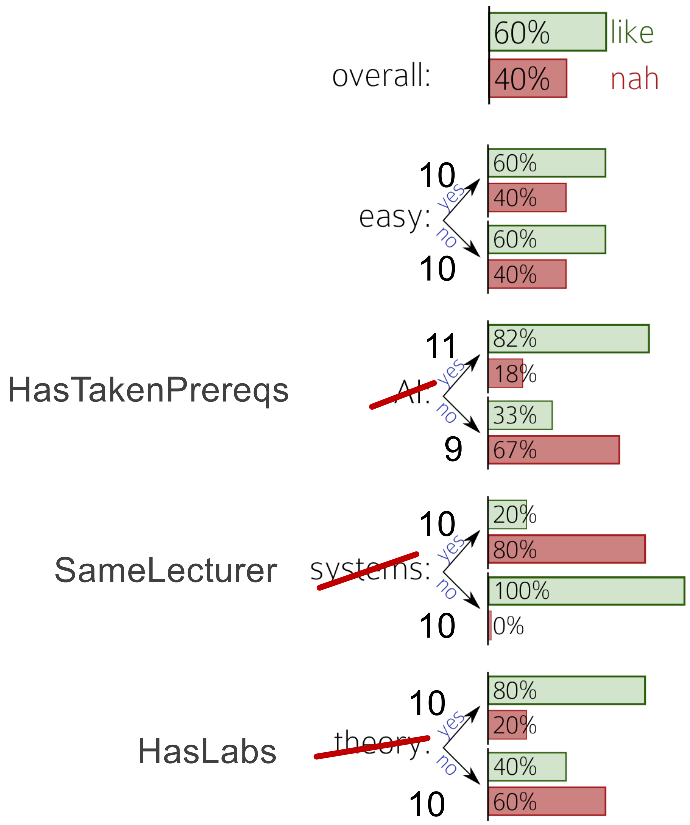
How to construct a tree



Prereqs Lecturer HasLabs

Rating	Easy?	~~AI?~~	~~Sys?~~	~~Thy?~~	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

How to construct a tree



Overall idea:
1. Set the root node as a leaf node.
2. Grab a leaf node for which its 'local' train accuracy is not 1.
3. Find a feature that maximizes the 'local' train accuracy and replace the leaf node with a node with that feature; add leaf nodes and set their predictions by majority vote.
4. Repeat 2-3.

Problem 18

- (k-NN) Suppose that you are given the following 5 data points:

Label y	Feature x
+	(0,1)
+	(1,2)
+	(2,2)
-	(3,3)
-	(5,6)

- And you trained 1-nearest neighborhood classifier and 3-nearest neighborhood classifier with Euclidean distance as the distance measure (without standardization). Given the test point $x^* = (4,2)$, state the prediction result of
- A) 1- nearest neighborhood classifier
- B) 3- nearest neighborhood classifier

k-nearest neighbor: main concept

- Train set: $S = \{ (x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$

- **<u>Idea</u>**: given a new, unseen data point x, its label should resemble the labels of **nearby points**

- What function?
  - Input: $x \in \mathbb{R}^d$
  - From S, find the k nearest points to x from S; call it $N(x)$

    E.g., Euclidean distance
  - Output: the majority vote of $\{y_i : i \in N(x)\}$
    - For regression, take the average label.

Make sure features are scaled fairly

- Features having different scale can be problematic. (e.g., weights in lbs vs shoe size)
- [Definition] **Standardization**
  - For each feature f, compute $\mu_f = \frac{1}{m}\sum_{i=1}^{m} x_f^{(i)}$, $\sigma_f = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_f^{(i)} - \mu_f\right)^2}$
  - Then, transform the data by $\forall f \in \{1, \ldots, d\}, \forall i \in \{1, \ldots, m\},\ x_f^{(i)} \leftarrow \frac{x_f^{(i)} - \mu_f}{\sigma_f}$

after transformation, each feature has mean 0 and variance 1

- Be sure to keep the "standardize" function and apply it to the test points.
  - Save $\{(\mu_f, \sigma_f)\}_{f=1}^{d}$
  - For test point x^*, apply $x_f^* \leftarrow \frac{x_f^* - \mu_f}{\sigma_f}, \forall f$

Problem 12 (a)

- Naïve Bayes Classifier (15 point)
- A) (7 points) Count the number of parameters

12. In a Bernoulli Naive Bayes model with k classes and v binary features, how many parameters are needed? Clarification: each feature is modeled by separate Bernoulli models (i.e., not a shared model).

  - Possible variation: Bernoulli → Gaussian, maybe I could explicitly ask you how many class-prior parameters and likelihood parameters for this model.

Naïve Bayes Classifier: Class prior parameters

For the class prior distribution, take categorical distribution.

$$y \sim \text{Categorical}(\pi), \qquad \pi \in \mathbb{R}^C, \pi_c \geq 0, \sum_c \pi_c = 1$$

$\Rightarrow p(y = c) = \pi_c$

$\Rightarrow$ ***C-1*** *parameters for the 'class prior distribution'*

Naïve Bayes Classifier

For real-valued features we can use Normal distribution:

$$p(x \mid y = c) = \prod_{d=1}^{D} \mathcal{N}(x_d \mid \underbrace{\mu_{cd}, \sigma_{cd}^2}_{\text{Parameters of featured for class } c})$$

quiz candidate
Q: how many parameters?

For binary features $x_d \in \{0,1\}$ can use Bernoulli distributions:

$$p(x \mid y = c) = \prod_{d=1}^{D} \text{Bernoulli}(x_d \mid \theta_{cd})$$

quiz candidate
Q: how many parameters?

"Coin bias" for d^{th} feature and class c

- K-valued discrete features: use Categorical.
- Can mix-and-match, e.g. some discrete, some continuous features

$$p(x \mid y = c) = \prod_{d=1}^{D'} \text{Bernoulli}(x_d \mid \theta_{cd}) \prod_{d=D'+1}^{D} \mathcal{N}(x_d \mid \mu_{cd}, \sigma_{cd}^2)$$

Problem 12 (b)

- Suppose you trained Bernoulli Naïve Bayes Classifier (all features are binary) with the binary label and 2 binary features, x_1, x_2 and we got the parameters as follows:
  - Class-prior parameter: $\phi = p(y = 1) = 0.25$,
  - Likelihood parameters: $\theta = (\theta_{01}, \theta_{02}, \theta_{11}, \theta_{12})$
    - $\theta_{12} = p(x_2 = 1|y = 1) = 0.9$, $\theta_{11} = p(x_1 = 1|y = 1) = 0.8$
    - $\theta_{02} = p(x_2 = 1|y = 0) = 0.2, \theta_{01} = p(x_1 = 1|y = 0) = 0.3$
- What is the prediction result of the test point $x^* = (1,1)$?

Recall: Naïve Bayes Classifier

- Ex) Classifier that predicts cancer using two features
  - Feature $x^{(i)}$: (Smoke?, Drug?) (D=2)
  - Label $y^{(i)}$: (Cancer?)
- Suppose you trained Bernoulli Naïve Bayes Classifier (all features are binary) with the binary label (C=2).
- You will get 5 parameters
  - C-1 + CD= 5.
  - Class prior parameter: C-1 = 1
  - Likelihood parameter: $C * D = 4$

Recall: Naïve Bayes Classifier

- Prediction: When input $\mathrm{x} = (x_1, x_2) = (1,1)$, your prediction is
- $\hat{y} = \arg\max_{c \in \{0,1\}} p(y = c, x; \phi, \theta) = \max(p(y = 0, x; \phi, \theta), p(y = 1, x; \phi, \theta))$
- $p(y = 1, x; \phi, \theta) = p(y = 1; \phi)p(x_1 = 1, x_2 = 1|y = 1; \theta)$
  $= p(y = 1; \phi)p(x_1 = 1\ |y = 1; \theta)p(x_2 = 1\ |y = 1; \theta) = 0.25 * 0.9 * 0.8$
  $= 0.18$
- $p(y = 0, x; \phi, \theta) = (1 - 0.25) * 0.2 * 0.3 = 0.045$
- 0.18>0.045, so $\hat{y} = 1$

Additional candidates

- Some problems might be too lengthy and time-consuming.
- I might be able to remove some sub-problems and add one problem from the candidates on the next slides.

Additional Candidate 1

- Neural network – number of parameters

17. A fully-connected feedforward neural network has input $x \in \mathbb{R}^{10}$, a first hidden ReLU layer with 6 units, a second hidden ReLU layer with 5 units, and a single sigmoid output unit. How many parameters are there in the neural network? Don't forget the bias parameters. You need to answer the exact number, not equations.

Additional Candidate 2

- SVM and Kernel

15. We are given a train set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$. With the SVM dual formulation's solution $\alpha \in \mathbb{R}^m$, the learned function can be written as $f(x) = \sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)})^\top x$ for a test point x. Let s be the number of support vectors. Say we would like to train a nonlinear SVM (dual formulation) with a basis function $\phi(x) \in \mathbb{R}^\infty$ for which there exists a known kernel function $k(x, x')$.
(i) Write down the learned function $f(x)$ based on the kernel function.
(ii) To evaluate $f(x)$ for a test point x, how many evaluations of kernel function do we need to perform?

Outline

- Data Science Ethics and Fairness
- Course Recap
- **Additional Resources**
- Final Exam Overview

Data Science Competitions

Competitions can be a great way to hone your skills…

CodaLab

MACHINE HACK

topcoder™

Analytics Vidhya

kaggle

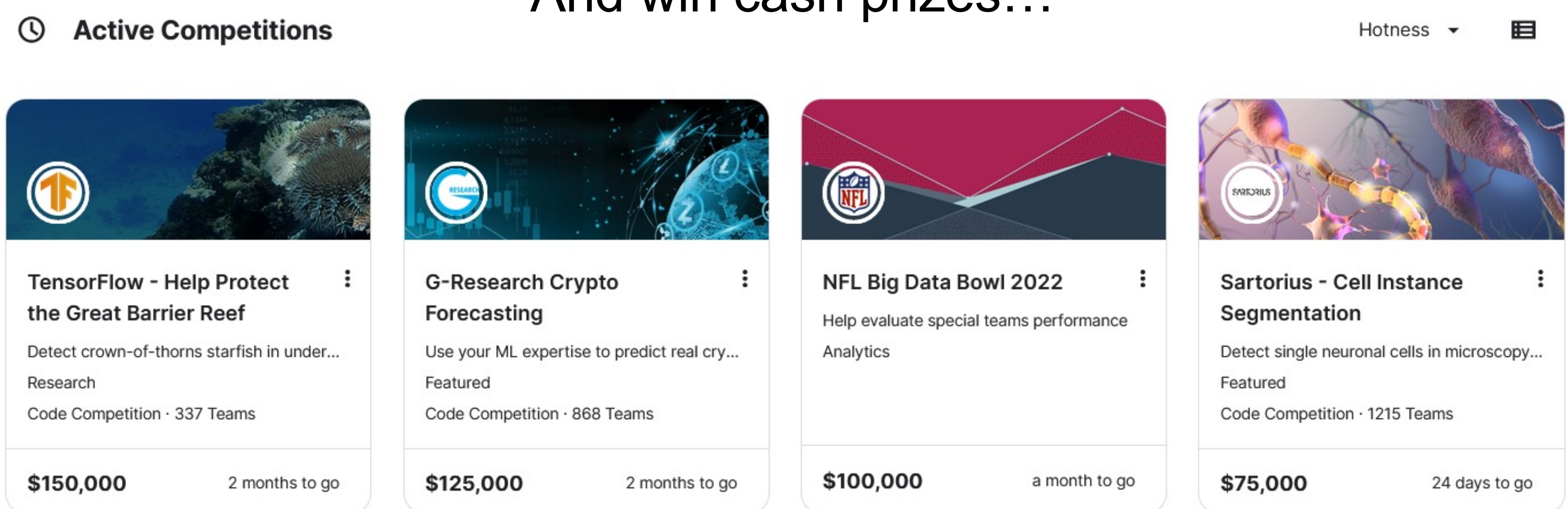
DRIVENDATA

IDA

INTERNATIONAL
DATA ANALYSIS OLYMPIAD

Data Science Competitions

And win cash prizes…



Can also be a great source for datasets to practice

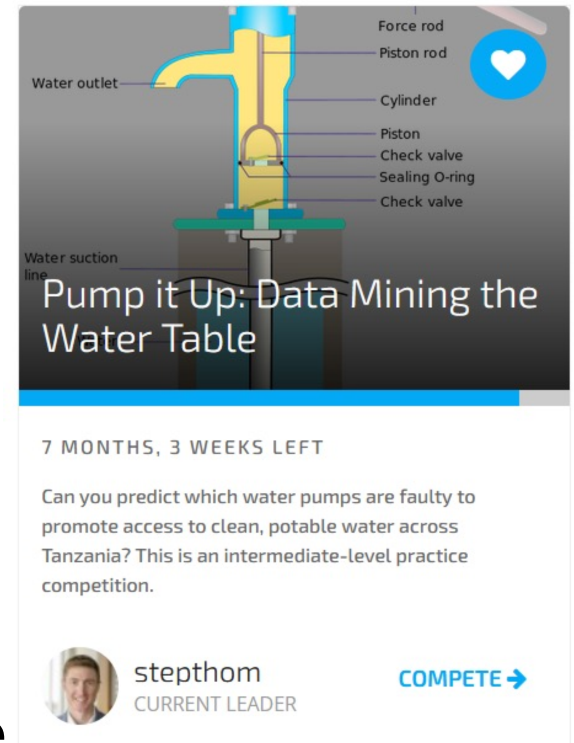
www.Kaggle.com

Data Science Competitions

Cash prizes aren’t the only goal…

DRIVENDATA

- Focuses on social impact
- Challenges last 2-3 months
- Real-world predictive problems
  - Detecting hateful content online
  - Predicting disease spread
  - Predicting damage from earthquakes
  - …
- Submissions are released as open source



Additional Relevant Courses

- CSC 480 : Principles of Machine Learning
- CSC 444 : Introduction to Data Visualization
- ISTA 457 : Neural Networks
- ESOC 330 : Digital Dilemmas : Privacy, Property, and Access
- MATH 574M : Statistical Machine Learning

Videos

3Blue1Brown

- Accessible videos on a variety of math topics
- Nicely produced, engaging graphics
- A number of ML / Data Science / Statistics topics covered

Steve Brunton – YouTube Channel

- More detailed videos on math / engineering topics
- Good linear algebra and machine learning videos
- Associated book,

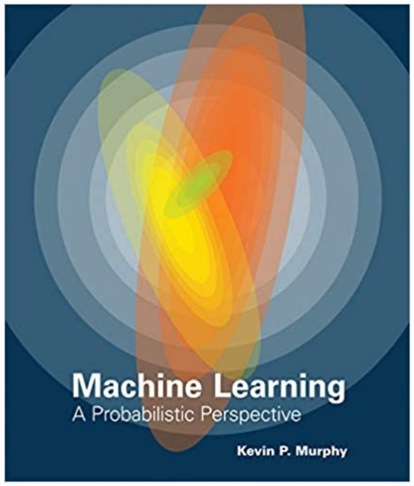
Data-Driven Science and Engineering : ML, Dynamical Systems, and control

Videos

MIT Open Courseware

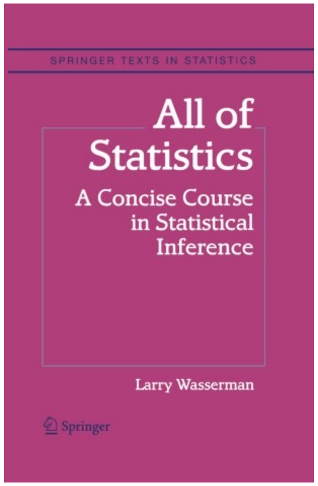
- Lots of topics freely available
- Excellent Linear Algebra course by Prof. Gilbert Strang (YouTube lectures)
- All assignments and exams available online

Textbooks



Murphy, K. "Machine Learning: A Probabilistic Perspective." MIT press, 2012

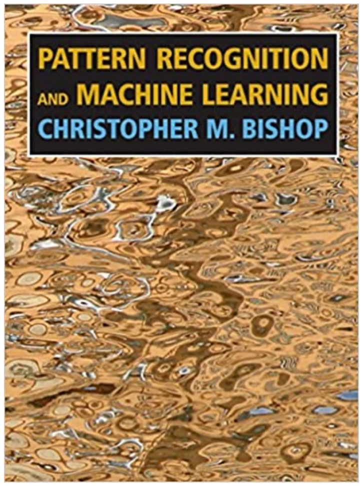
(UA Library)



Wasserman, L. “All of Statistics.” Springer, 2004

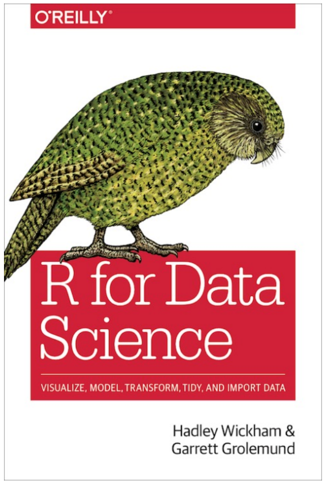
(Springer)

Textbooks



Bishop, C. "Pattern Recognition and Machine Learning." Springer, 2006

[(Microsoft)]



Wickham and Grolemund. “R for Data Science.” O’Reilly, 2016

(O’Reilly)

Outline

- Data Science Ethics and Fairness
- Course Recap
- Additional Resources
- Final Exam Overview

Final exam

- May 8th, 3:30 pm to 5:30 pm
- Cheat sheet: 2 pages, you can use both front and back (total 4 sides)
- Please tell me as soon as possible if you need DRC accommodations.

Thank you!

Good luck with your exam!