Lecture 19

Review, Final Thoughts

DSC 40A, Spring 2024

Announcements

- Homework 8 is due **tonight** (no slip days). Solutions will be released at midnight.
- The Final Exam is on **Saturday from 8-11AM**.
    - You will be assigned a seat, either in Center Hall 212 or 214.
    - 180 minutes, on paper, no calculators or electronics, but **you are allowed to bring two double-sided index cards (4 inches by 6 inches) of notes that you write by hand.**
- There is one more review session **tonight from 5-7PM in Center Hall 216**, on gradient descent and probability.
- **Tomorrow, from 4-9PM**, we have a study session in HDSI 123.
- If at least 90% of the class fills out both the End-of-Quarter Survey and SETs by 8AM on Saturday, then the entire class will have 2% of extra credit added to their overall grade. **As of this morning, we're only at 6~~2~~%.** 68%
- See more review videos on Ed.

Agenda

- High-level overview of the course.
- Old exam problems.
- Final thoughts.

What was this course about?

"Finding the best way to make predictions, using data."

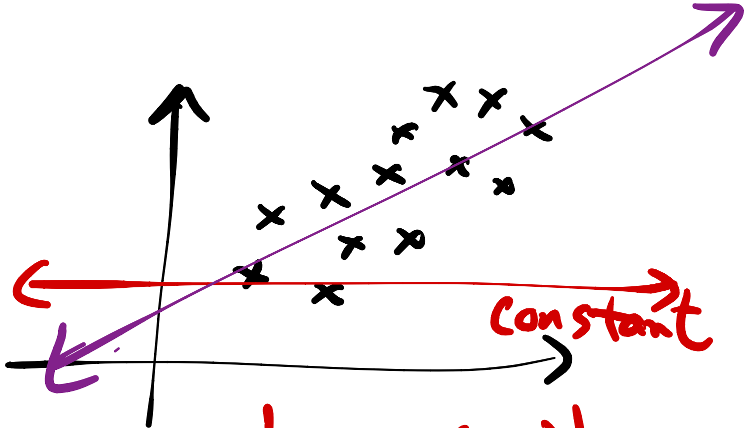
Part 1: Empirical risk minimization (Lectures 1-11)

1. Choose a model.

constant model : $H(x)=h$

simple linear regression : $H(x) = w_0 + w_1 x$

intercept (w_0), slope (w_1)



2. Choose a loss function.

squared loss : $(y_i - H(x_i))^2$ (actual − predicted)^2

absolute loss : $|y_i - H(x_i)|$

zero-one loss

relative squared loss

sinugod loss

infinity loss

⋮

3. Minimize average loss to find optimal model parameters.

"empirical risk"

$$R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h)^2 \xrightarrow{\text{calculus}} h^* = \text{Mean}(y_1, y_2, \dots, y_n)$$

"mean squared error"

$$R_{sq}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (w_0 + w_1 x_i)\right)^2 \xrightarrow{\text{calculus}} w_0^*, w_1^*$$

Why did we need linear algebra?

multiple linear regression: $H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \dots + w_d x^{(d)}$

$= \vec{w} \cdot \text{Aug}(\vec{x})$

To find $w_0^*, w_1^*, w_2^*, \dots, w_d^*$:

minimize $\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(w_0 + w_1 x_i^{(1)} + w_2 x_i^{(2)} + \dots + w_d x_i^{(d)} \right) \right)^2$

That looks ugly... linear algebra can help!!!

$$X = \begin{bmatrix} — \text{Aug}(\vec{x}_1) — \\ — \text{Aug}(\vec{x}_2) — \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & & x_2^{(d)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & & x_n^{(d)} \end{bmatrix}_{n \times (d+1)}$$

$$\vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\boxed{\text{minimize } \|\underbrace{\vec{y} - X\vec{w}}_{\vec{e}}\|^2}$$

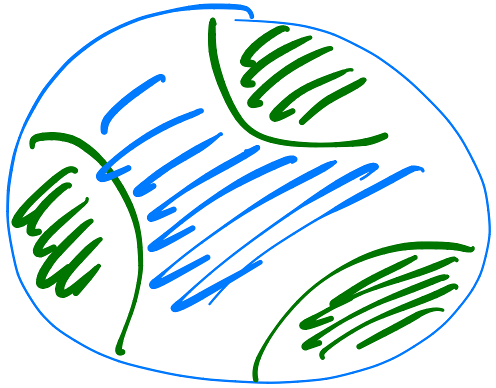
$\leftarrow$ normal equations!

Why gradient descent?

We will run into functions that can't be minimized using calculus or linear algebra

⇒ but! we do know their derivative

⇒ convexity: convex functions only have one minimum, which is

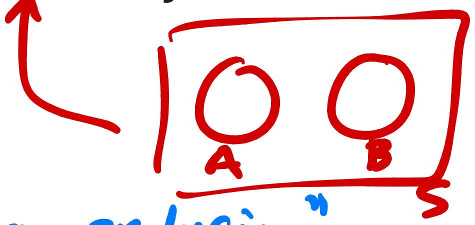


GLOBAL!

$= \frac{\#\text{ outcomes in } A}{\#\text{ total outcomes}}$

Part 2: Probability fundamentals (Lecture 12)

- If all outcomes in the **sample space** S are equally likely, then $\mathbb{P}(A) = \frac{|A|}{|S|}$.
- $\bar{A}$ is the **complement** of event A. $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$.
- Two events A, B are **mutually exclusive** if they share no outcomes, i.e. they don't overlap: $\mathbb{P}(A \cap B) = 0$.



- For any two events, the probability that A happens or B happens is $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ "principle of inclusion-exclusion"
- The probability that events A and B both happen is $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$.
  - $\mathbb{P}(B|A)$ is the probability that B happens, given that you know A happened.
  - Through re-arranging, we see that $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$, which is the definition of conditional probability.

multiplication rule

Part 2: Combinatorics (Lectures 13-14)

Suppose we want to select k elements from a group of n possible elements. The following table summarizes whether the problem involves sequences, permutations, or combinations, along with the number of relevant orderings.

	Yes, order matters	No, order doesn't matter
With replacement Repetition allowed	$\boxed{n^k}$ possible **sequences**	more complicated: watch **this video**
Without replacement Repetition not allowed	$\boxed{\dfrac{n!}{(n-k)!}}$ possible **permutations**	$\boxed{\dbinom{n}{k}}$ **combinations**

Part 2: The law of total probability and Bayes' Theorem (Lectures 15 and 16)

- A set of events $E_1, E_2, \ldots, E_k$ is a **partition** of S if each outcome in S is in exactly one E_i.
- The **law of total probability** states that if A is an event and $E_1, E_2, \ldots, E_k$ is a partition of S, then: $\mathbb{P}(A) = \mathbb{P}(A \cap E_2) + \mathbb{P}(A \cap E_3) + \mathbb{P}(A \cap E_4) + \mathbb{P}(A \cap E_5) + \mathbb{P}(A \cap E_1)$

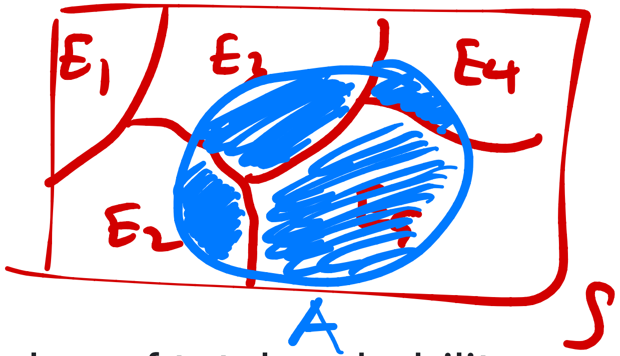
multiplication rule

$$\mathbb{P}(A) = \mathbb{P}(E_1)\mathbb{P}(A|E_1) + \mathbb{P}(E_2)\mathbb{P}(A|E_2) + \ldots + \mathbb{P}(E_k)\mathbb{P}(A|E_k) = \sum_{i=1}^{k} \mathbb{P}(E_i)\mathbb{P}(A|E_i)$$

- **Bayes' Theorem** states that:

example: shake shack, in-n-out, 5 guys

$$\underbrace{\mathbb{P}(B|A)}_{\text{new}} = \frac{\overbrace{\mathbb{P}(B)}^{\text{old}}\mathbb{P}(A|B)}{\mathbb{P}(A)}$$



- We often re-write the denominator $\mathbb{P}(A)$ in Bayes Theorem' using the law of total probability.

Part 2: Independence and conditional independence (Lectures 15-16)

- Two events A and B are **independent** when knowledge of one event does not change the probability of the other event.
  - Equivalent conditions: $\mathbb{P}(B|A) = \mathbb{P}(B)$, $\mathbb{P}(A|B) = \mathbb{P}(A)$, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

equivalent

- Two events A and B are **conditionally independent** given event C if they are independent given the knowledge that event C happened.
  - Condition:

$$\mathbb{P}((A \cap B)|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$$

- In general, there is no relationship between independence and conditional independence.

lecture note

- Make sure you've read **this**!

Part 2: Naïve Bayes (Lectures 17-18)

- In classification, our goal is to predict a discrete category, called a **class**, given some features.
- The **Naïve Bayes** classifier works by estimating the numerator of $\mathbb{P}(\text{class}|\text{features})$ for all possible classes.
- It uses Bayes' Theorem:

e.g. ripe, unripe

$$\mathbb{P}(\text{class}|\text{features}) = \frac{\mathbb{P}(\text{class}) \cdot \mathbb{P}(\text{features}|\text{class})}{\mathbb{P}(\text{features})}$$

Zutano, firm, green-black

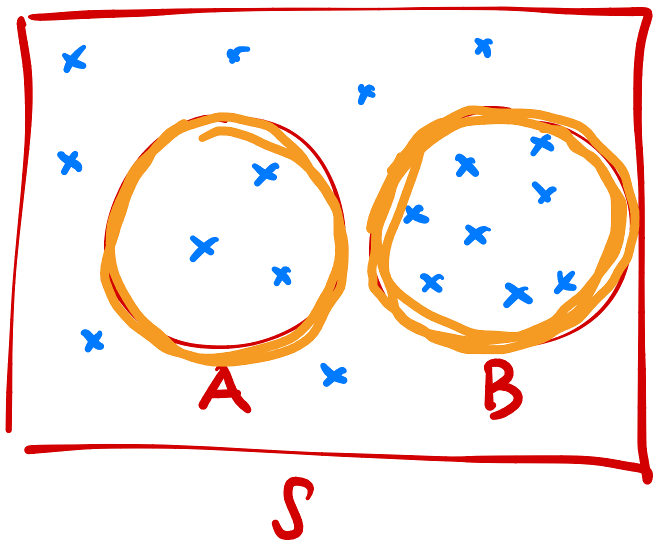
- It also uses a "naïve" simplifying assumption, that **features are conditionally independent given a class**:

$$\mathbb{P}(\text{features}|\text{class}) = \mathbb{P}(\text{feature}_1|\text{class}) \cdot \mathbb{P}(\text{feature}_2|\text{class}) \cdot \ldots$$

Practice problems

Spring 2023 Midterm Exam 2, Problem 6.2

The events A and B are mutually exclusive, or disjoint. More generally, for **any** two disjoint events A and B, show how to express $\mathbb{P}(\bar{A}|(A \cup B))$ in terms of $\mathbb{P}(A)$ and $\mathbb{P}(B)$ **only**.



$$\mathbb{P}(\bar{A}|(A \cup B)) = \frac{\mathbb{P}(B)}{\mathbb{P}(A \cup B)} = \boxed{\frac{\mathbb{P}(B)}{\mathbb{P}(A) + \mathbb{P}(B)}}$$

Fall 2021 Final Exam, Problem 8

Billy brings you back to Dirty Birds, the restaurant where he is a waiter. He tells you that Dirty Birds has 30 different flavors of chicken wings, 18 of which are 'wet' (e.g. honey garlic) and 12 of which are 'dry' (e.g. lemon pepper).

Each time you place an order at Dirty Birds, you get to pick 4 different flavors. The order in which you pick your flavors does not matter.

Part 1: How many ways can we select 4 flavors in total?

$$\binom{30}{4}$$

Part 2: How many ways can we select 4 flavors in total such that we select an equal number of wet and dry flavors? 2 wet AND 2 dry

$$\binom{18}{2} \times \binom{12}{2}$$

18 wet 12 dry

Part 3: Billy tells you he'll surprise you with 4 different flavors, randomly selected from the 30 flavors available. What's the probability that he brings you at least one wet flavor and and least one dry flavor?

$$\frac{\text{\# combinations w/ at least one wet and one dry flavor}}{\text{\# combinations of 4 flavors}}$$

$$\boxed{\frac{\binom{30}{4} - \binom{18}{0}\binom{12}{4} - \binom{18}{4}\binom{12}{0}}{\binom{30}{4}}}$$

① complement

total − # no wet − # no dry

$$\underbrace{\binom{30}{4} - \binom{18}{0}\binom{12}{4} - \binom{18}{4}\binom{12}{0}}$$

② directly

1w 3d

2w 2d

3w 1d

$$\underbrace{\binom{18}{1}\binom{12}{3} + \binom{18}{2}\binom{12}{2} + \binom{18}{3}\binom{12}{1}}$$

Part 4: Suppose you go to Dirty Birds once a day for 7 straight days. Each time you go there, Billy brings you 4 different flavors, randomly selected from the 30 flavors available. What's the probability that on at least one of the 7 days, he brings you all wet flavors or all dry flavors? (Note: All 4 flavors for a particular day must be different, but it is possible to get the same flavor on multiple days.)

$\square$ at least once $\xrightarrow[\text{complement}]{}$ $\square$ never happens

$\square$ never happens : $\mathbb{P}$(never get all wet or all dry flavors)

$= \mathbb{P}\left(\text{don't get all wet/all dry day 1}\right) \times \mathbb{P}\left(\text{don't get all wet/all dry day 2}\right) \times \cdots$

$= \mathbb{P}\left(\text{don't get all wet/all dry day 1}\right)^7 = \underbrace{\mathbb{P}\left(\text{at least 1 wet, 1 dry day 1}\right)^7}_{\text{last slide}}$

$\square$ at least once $\xrightarrow[\text{complement}]{}$ $\square$ never happens

$\square$ never happens : $\mathbb{P}(\text{never get all wet or all dry flavors})$

$= \mathbb{P}\left(\begin{matrix}\text{don't get} \\ \text{all wet/all dry}\end{matrix} \text{ day } 1\right) \times \mathbb{P}\left(\begin{matrix}\text{don't get} \\ \text{all wet/all dry}\end{matrix} \text{ day } 2\right) \times \cdots$

$= \mathbb{P}\left(\begin{matrix}\text{don't get} \\ \text{all wet/all dry}\end{matrix} \text{ day } 1\right)^7 = \underbrace{\mathbb{P}\left(\begin{matrix}\text{at least 1 wet,} \\ \text{1 dry}\end{matrix} \text{ day } 1\right)^7}_{\text{last slide}}$

$\Rightarrow$ probability that at least once, we get all wet, all dry:

$$1 - \mathbb{P}\left(\begin{matrix}\text{at least 1 wet,} \\ \text{1 dry}\end{matrix} \text{ day } 1\right)^7$$

$$= 1 - \left[\frac{\binom{30}{4} - \binom{18}{0}\binom{12}{4} - \binom{18}{4}\binom{12}{0}}{\binom{30}{4}}\right]^7$$

Fall 2021 Final Exam, Problem 9

In this question, we'll consider the phone number $\underbrace{6789998212}_{\text{10 digits}}$ (mentioned in Soulja Boy's 2008 classic, "Kiss Me thru the Phone").

Part 1: How many permutations of 6789998212 are there?

$$\frac{10!}{3!\,2!\,2!} = \binom{10}{3}\binom{7}{1}\binom{6}{2}\binom{4}{1}\binom{3}{2}\binom{1}{1}$$

Part 2: How many permutations of 6789998212 have all three 9s next to each other?

$$\frac{8!}{2!\,2!}$$

6
7
88
X
22
1

Part 3: How many permutations of 6789998212 end with a 1 and start with a 6?

$$\frac{8!}{2!\,3!\,2!}$$

$\boxed{6}$ $\underbrace{\qquad\qquad\qquad}_{8 \text{ positions}}$ $\boxed{1}$

7
88
999
22

Part 4: How many different 3 digit numbers with unique digits can we create by selecting digits from 6789998212?

$$\underline{6}\ \underline{5}\ \underline{4} = \boxed{120}$$

$$\binom{6}{3} 3!$$

6
7
8
9
1
2

"stars and bars"

Example: Candy

I have 9 identical pieces of candy. How many ways can I distribute the 9 pieces of candy to 4 of my friends?

A	B	C	D
9	0	0	0
5	4	0	0
1	3	1	4
⋮			

x : candy

| : dividers

xx | xxx | x | xxx

A: 2 B: 3 C: 1 D: 3

How many permutations?

$$\binom{9+3}{3} = \binom{9+3}{9}$$

Final thoughts

Learning objectives

On the first day of the quarter, we told you that after taking DSC 40A, you would:

- understand the basic principles underlying almost every machine learning and data science method. true!
- be better prepared for the math in upper division: calculus, linear algebra, and probability. definitely true!

What's next?

In DSC 40A, we just scratched the surface of the theory behind data science. In future courses, you'll build upon your knowledge from DSC 40A, and will learn:

DSC 140A

- More **supervised learning**, e.g. logistic regression, decision trees, neural networks.
- **Unsupervised learning**, e.g. clustering, PCA.
- More **probability**, e.g. random variables, distributions, stochastic processes.
- More **connections** between all of these areas, e.g. the relationship between probability and linear regression.
- More practical tools.

DSC 80

Thank you!

This course would not have been possible without our 12 tutors.

Jack Determan	Javier Ponce
Yosen Lin	Harshita Saha
Utkarsh Lohia	Candus Shi
Zoe Ludena	Charlie Sun
Mert Ozer	Nicholas Swetlin
Varun Pabreja	Benjamin Xue

You can contact them with questions at dsc40a.com/staff.

Congrats on (almost) finishing DSC 40A!

Good luck on the final, and **please keep in touch!**

rampure@umich.edu