Lecture 18

Review, Final Thoughts

DSC 40A, Summer 2024

Announcements

- Homework 8 is due **tonight**. Solutions will be released at midnight.
- The Final Exam is **tomorrow, September 6th from 11:30AM-2:30PM** in WLH 2113.
- 180 minutes, on paper, no calculators or electronics.
  - **You are allowed to bring two double-sided index cards (4 inches by 6 inches) of notes that you write by hand (no iPad).**
- Content: All lectures (including this week), homeworks (including HW 8), and groupworks.
- Prepare by practicing with old exam problems at practice.dsc40a.com.
- Office hours this afternoon - come through and study!
- 16.67% right now, far away!
  If 90% of the class fills out both the SETs and the Final Survey by **Friday 8AM**, everyone gets 2% extra credit.

16 students

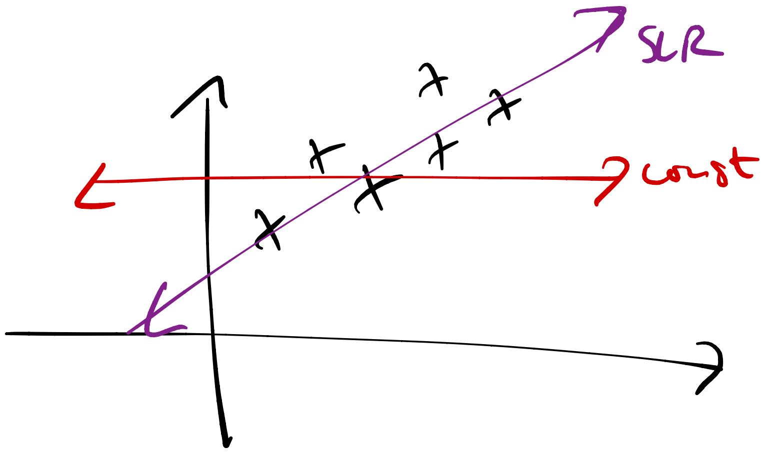
Agenda

- 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 : $H(x) = h$

intercept

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

slope

2. Choose a loss function.

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

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

0-1 loss

relative squared loss

Jack loss (midterm)

infinity loss

"empirical risk"

3. Minimize average loss to find optimal model parameters.

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

"mean squared error"

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

Why did we need Linear Algebra?

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

other features: $x^{(1)} x^{(2)^2}$

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

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

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

This looks tough. Linear algebra can help!

$$\vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}_{(d+1)\times 1} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$$

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

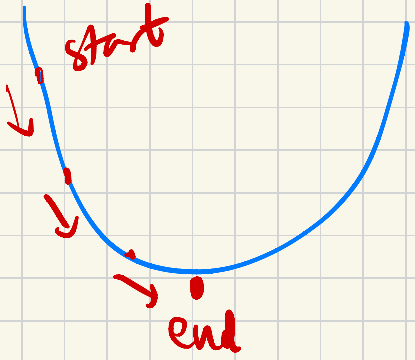
$$\boxed{\text{minimize } \underbrace{\|\vec{y} - X\vec{w}\|}_{\vec{e}}} \rightarrow \text{normal equations}$$

Why do gradient descent?

We might encounter functions that we can't minimize with calculus or linear algebra.

$\Rightarrow$ but! we do know their derivatives.

$\Rightarrow$ Convexity: convex functions have only one minimum, which is GLOBAL!



$\mathbb{P}$(exactly 2H) $\neq$ {0H, 1H, 2H, 3H} $\rightarrow$ {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} $= P = \frac{3}{8}$ ✓

~~ok~~ not equally likely

fair coin: equally likely

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. $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$. "at least 1" $\Rightarrow$ $1 - \mathbb{P}(0)$
- 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 $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. $- \mathbb{P}(A \cap B)$ $\Leftrightarrow$ "inclusion-exclusion" $\mathbb{P}(B)$ // if indp
- 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.

if indp $\rightarrow$ $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

multiplication rule

$\mathbb{P}($exactly 6 H in 10 flips for biased coin: $\frac{1}{3}$ H, $\frac{2}{3}$ T$)$

$= (\#$ ways of 6H in 10 flips$) \cdot \mathbb{P}($one outcome$)$

$$= \binom{10}{6} \qquad \left(\frac{1}{3}\right)^6 \cdot \left(\frac{2}{3}\right)^{10-6}$$

$\hookrightarrow$ $\mathbb{P}($exactly 2H in 3 flips of a fair coin$)$

$$= \binom{3}{2} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^1 = \binom{3}{2} \cdot \left(\frac{1}{2}\right)^3 = \frac{\binom{3}{2}}{8} = \frac{3}{8} \quad \checkmark$$

$\frac{\# \text{ favorable}}{\# \text{ possible}} = P(\text{fav})$

seq / comb

seq / comb

Part 2: Combinatorics (Lectures 13-14)

population

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	$\underbrace{n \; n \; n \; \cdots \; n}_{k \text{ times}}$ $\boxed{n^k}$ possible **sequences**	more complicated: watch **this video** domino example
Without replacement Repetition not allowed	$\boxed{\frac{n!}{(n-k)!}}$ possible **permutations**	$\boxed{\binom{n}{k}}$ **combinations**

$\underline{52} \; \underline{51} \; \underline{50} \; \underline{49} \rightarrow \frac{52!}{48!}$

! draw the boxes !

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

" covering S fully w/o overlap"

- 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: $\mathbb{P}(A) = \mathbb{P}(A \cap E_1) + \mathbb{P}(A \cap E_2) \cdots + \mathbb{P}(A \cap E_k)$

multiplication rule

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

→ rearranging multiplication rule

- **Bayes' Theorem** states that:

ex: burgers: Shake shack, in n out, 5 guys → $\mathbb{P}$(shakeshack | correctly)

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

use a tree to visualize the process

- 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.

tack on "$|C$"

  - 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**!

→ 5 minutes

Part 2: Naïve Bayes (Lecture 17, 18-ish)

- In classification, our goal is to predict a discrete category, called a **class**, given some features.

  └ with probability

- The **Naïve Bayes** classifier works by estimating the numerator of $\mathbb{P}(\text{class}|\text{features})$ for all possible classes.

- It uses Bayes' Theorem:

ripe | unripe

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

Zutano, firm, green-black

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

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

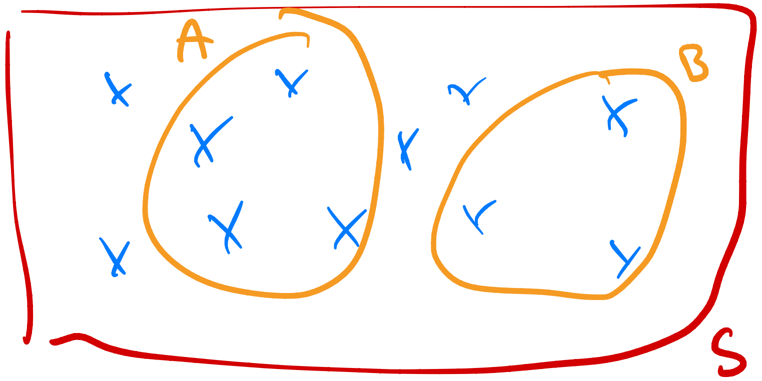
Smooth to avoid multiplying by 0.

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**.

$$\mathbb{P}(\bar{A} \mid A \cup B) = \frac{\mathbb{P}(\bar{A} \cap (A \cup B))}{\mathbb{P}(A \cup B)}$$



Bad $\underbrace{\mathbb{P}(B|\bar{A})}_{?}\,\mathbb{P}(\bar{A})$

$$= \frac{\mathbb{P}(\overset{0}{\cancel{(\bar{A} \cap A)}}) + \mathbb{P}((\bar{A} \cap B))}{\mathbb{P}(A) + \mathbb{P}(B) - \overset{0}{\cancel{\mathbb{P}(A \cap B)}}}$$

disjoint

mult rule

$\overset{1}{\cancel{\mathbb{P}(\bar{A}|B)}}\,\mathbb{P}(B) = \mathbb{P}(B)$

$$= \frac{\mathbb{P}(\bar{A} \cap B)}{\mathbb{P}(A) + \mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(A) + \mathbb{P}(B)}$$

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. Combinations $\binom{n}{k}$

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

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

\# wet flavors \# dry flavors

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

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

2 wet AND 2 dry

18 W 12 D

? ~~$18 \cdot 12 \cdot \binom{28}{2}$~~ 81648 doesn't match

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?

$$\frac{\text{\# comb w/ at least 1 wet and at least 1 dry flavor}}{\text{\# comb of 4 flavors: } \binom{30}{4}}$$

$\cancel{\binom{18}{3}}$? $\binom{12}{3} \cdot 4$

poker example

① complement

\# total − \# no wet (= all dry) − \# no dry (= all wet)

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

23 850

② directly

1 W 3D

2 W 2 D

3 W 1 D

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

23 850

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)

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

$= \mathbb{P}$(don't get all wet/dry day 1$)^7$

$= \mathbb{P}$(at least 1 wet / 1 dry, day 1$)^7$

last slide

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

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

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

$= \mathbb{P}$(don't get all wet/dry day 1$)^7$

$= \underline{\mathbb{P}(\text{at least 1 wet / 1 dry, day 1})^7}$

last slide

$\Rightarrow$ prob at least once we get all wet or all dry =

$1 - \mathbb{P}(\text{at least 1 wet 1 dry})^7$

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

Fall 2021 Final Exam, Problem 9

In this question, we'll consider the phone number 6789998212 (mentioned in Soulja Boy's 2008 classic, "Kiss Me thru the Phone").

Part 1: How many permutations of 6789998212 are there?

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

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?

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

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

7
88
999
22

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

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

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

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 $\rightarrow$ how many permutations?

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

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

Final thoughts

Learning objectives

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

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

What's next?

frameworks < ERM / gradient descent / probability

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

loss functions → $L(y_i, H(x_i))$

- More **supervised learning**, e.g. logistic regression, decision trees, neural networks.
- **Unsupervised learning**, e.g. clustering, PCA. linear algebra

  unlabeled
- 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 tutors.

Jack Determan – DSC minor

Owen Miller

Zoe Ludena

Suraj

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!**

nskh@umich.edu