Lecture 18 – Review, Conclusion

DSC 40A, Fall 2021 @ UC San Diego

Suraj Rampure, with help from **many others**

Announcements

- Homework 8 is due **tomorrow 12/3 at 11:59pm**.
- A recording of Discussion 8 (probability review) is posted on the course website and on Campuswire.
- Fill out CAPEs + the End-of-Quarter survey. If 90% of the class does both, everyone gets 0.5% extra credit added to their final course grade.
    - Deadline: Monday at 8am.
- The Final Exam is on **Wednesday 12/8 from 11:30AM–2:30PM**.
    - You'll take the exam remotely by downloading a PDF from Gradescope and submitting your answers as a PDF by the deadline.
    - Open internet, but no Googling for the answers, and **no collaboration**.
    - More details to come this weekend.

Final preparation → cumulative!!!

- Review the solutions to previous homeworks and groupworks.
  - All except Homework 8 are up.
- Identify which concepts are still iffy. Re-watch lecture, post on Campuswire, come to office hours.
  - **We have many office hours between now and the exam.**
- Look at the past exams at `https://dsc40a.com/resources`.
  - Watch the probability review discussion.
- Study in groups.
- Make a “cheat sheet”.

Agenda

- High-level summary of the course.

  → Midterm review session has more!

- Review problems.

- Conclusion.

What was this course about?

Part 1: Supervised learning (Lectures 1-10)

The "learning from data" recipe to make predictions:

1. Choose a **prediction rule**. We've seen a few:
   - Constant: $H(x) = h$.
   - Simple linear: $H(x) = w_0 + w_1 x$.
   - Multiple linear: $H(x) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)}$.

   linear algebra
   feature eng.
2. Choose a **loss function**.
   - Absolute loss: $L(h, y) = |y - h|$.
   - Squared loss: $L(h, y) = (y - h)^2$.
   - 0-1 loss, UCSD loss, etc.
3. Minimize **empirical risk** to find optimal parameters.
   - Algebraic arguments.
   - Calculus (including vector calculus).
   - Gradient descent.

   average of loss f'n over data set

constant pred, abs loss:

$$R_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |y_i - h|$$

simple linear, sq loss:

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

multiple linear, sq loss:

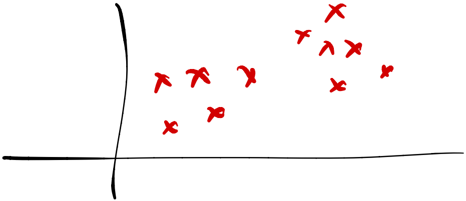
$$R_{sq}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

obs vector

design matrix

Part 1: Unsupervised learning (Lectures 10-11)

- When learning how to fit prediction rules in Lectures 1-10, we were performing **supervised machine learning**.
- In Lectures 10 and 11, we discussed ***k*-Means Clustering**, an **unsupervised machine learning** method.
  - Supervised learning: there is a "right answer" that we are trying to predict.
  - Unsupervised learning: there is no right answer, instead we're trying to find patterns in the structure of the data.



inertia

centroids

⋮

Part 2: Probability fundamentals (Lectures 11-12)

- If all outcomes in the **sample space** S are equally likely, then $P(A) = \frac{|A|}{|S|}$. $= \frac{\text{\# outcomes in } A}{\text{\# outcomes total}}$
- $\bar{A}$ is the **complement** of event A. $P(\bar{A}) = 1 - P(A)$.
- Two events A, B are **mutually exclusive** if they share no outcomes, i.e. they don't overlap. In this case, the probability that A happens or B happens is $P(A \cup B) = P(A) + P(B)$.
- More generally, for any two events, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. principle of inclusion/exclusion
- The probability that events A and B both happen is $P(A \cap B) = P(A)P(B|A)$.
  - $P(B|A)$ is the probability that B happens given that you know A happened.
  - Through re-arranging, we see that $P(B|A) = \frac{P(A \cap B)}{P(A)}$.

Part 2: Combinatorics (Lectures 13-14)

- A **sequence** is obtained by selecting k elements from a group of n possible elements with replacement, such that order matters.
  - Number of sequences: n^k.

- A **permutation** is obtained by selecting k elements from a group of n possible elements without replacement, such that order matters.
  - Number of permutations: $P(n,k) = \frac{n!}{(n-k)!}$.

8 people
3 person comm.

$\underset{\text{pres.}}{\underline{8}}$ $\underset{\text{vp}}{\underline{7}}$ $\underset{\text{secretary}}{\underline{6}}$

- A **combination** is obtained by selecting k elements from a group of n possible elements without replacement, such that order does not matter.
  - Number of combinations: $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.

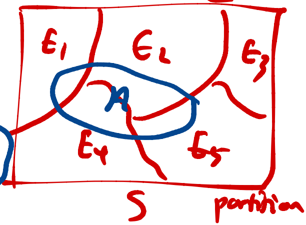
$\binom{8}{3}$

Part 2: The law of total probability and Bayes' theorem (Lecture 14)

- 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

$$P(A) = P(E_1) \cdot P(A|E_1) + P(E_2) \cdot P(A|E_2) + \ldots + P(E_k) \cdot P(A|E_k)$$

$$= \sum_{i=1}^{k} P(E_i) \cdot P(A|E_i)$$



- **Bayes' theorem** states that

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

- We often re-write the denominator $P(A)$ in Bayes' theorem using the law of total probability.

Part 2: Independence and conditional independence (Lecture 15)

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

$P(A \cap B) = P(A)P(B|A)$
$= P(B)P(A|B)$

- Two events A and B are **conditionally independent** if they are independent given knowledge of a third event, C.
  - Condition: $P((A \cap B)|C) = P(A|C) \cdot P(B|C)$.

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

- See pinned post on Campuswire for clarification.

Part 2: Naive Bayes (Lecture 16-17)

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

supervised ML technique!

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

- It uses Bayes' theorem:

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

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

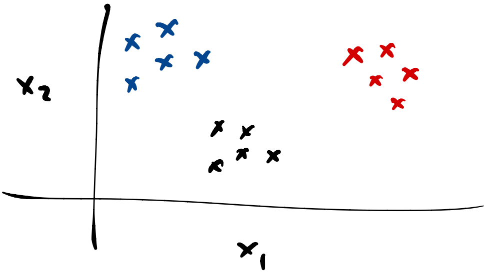
$$P(\text{features}|\text{class}) = P(\text{feature}_1|\text{class}) \cdot P(\text{feature}_2|\text{class}) \cdot \ldots$$

Review problems

Example: Clustering and combinatorics

- Suppose we have a dataset of 15 points, each with two features (x_1, x_2). In the dataset, there exist 3 "natural" clusters, each of which contain 5 data points.

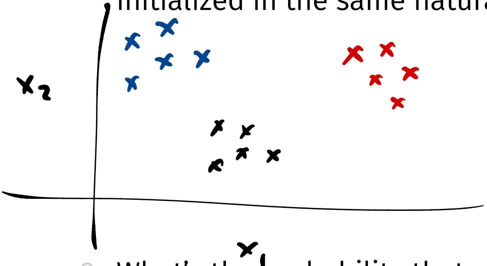
- Recall that in the k-Means Clustering algorithm, we initialize k centroids by choosing k points at random from our dataset. Suppose $k = 3$.



need to pick 3 points from 15 to initialize centroids

$\Rightarrow \binom{15}{3}$ ways to initialize

1. What's the probability that all three initial centroids are initialized in the same natural cluster?



$$\frac{3\binom{5}{3}}{\binom{15}{3}} = \frac{4}{14}\cdot\frac{3}{13}$$

2. What's the probability that all three initial centroids are initialized in different natural clusters?

blue cluster

red cluster

black cluster

$$\frac{5\cdot 5\cdot 5}{\binom{15}{3}}$$

equiv. sol:

$$\frac{\frac{15\cdot 10\cdot 5}{3!}}{\binom{15}{3}}$$

Proof that $\dfrac{3\binom{5}{3}}{\binom{15}{3}} = \dfrac{4}{14}\cdot\dfrac{3}{13}$:

$$\frac{3\binom{5}{3}}{\binom{15}{3}} = \frac{3\cdot\dfrac{5!}{2!\,\cancel{3!}}}{\dfrac{15!}{12!\,\cancel{3!}}} = \frac{\cancel{3}\cdot\cancel{5}\cdot 4\cdot 3}{\cancel{15}\cdot 14\cdot 13}$$

$$= \frac{4\cdot 3}{14\cdot 13}$$

as required.

Example: basketball

Suppose we have 6 basketball players who want to organize themselves into 3 basketball teams of 2 players each. Suppose we have three teams, “Team USA”, “Team China”, and “Team Lithuania”. How many ways can these teams be formed?

$$\binom{6}{2}\binom{4}{2}\binom{2}{2}$$

A ~~B~~ ~~C~~ D ~~E~~ ~~F~~

$$\underbrace{\binom{6}{2}}_{\text{USA}} \quad \underbrace{\binom{4}{2}}_{\text{China}} \quad \underbrace{\binom{2}{2}}_{\text{Lithuania}}$$

USA	China	Lithuania
BE	CF	AD
CF	AD	BE

Example: basketball, again

Suppose we have 6 basketball players who want to organize themselves into 3 basketball teams of 2 players each. Now,

suppose the teams are irrelevant, and all we care about is the unique pairings themselves. How many ways can these 6 players be split into 3 teams?

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

$$\frac{\binom{6}{2}\binom{4}{2}\binom{2}{2}}{3!} = \frac{\frac{6 \cdot 5}{2} \cdot \cancel{6}}{\cancel{6}} = \boxed{15}$$

Example: high school

A certain high school has 80 students: 20 freshmen, 20 sophomores, 20 juniors, and 20 seniors. If a random sample of 20 students is drawn without replacement, what is the probability that the sample contains 5 students in each grade level?

\# possible samples : $\binom{80}{20}$

juniors

freshmen

$$\frac{\binom{20}{5}\binom{20}{5}\binom{20}{5}\binom{20}{5}}{\binom{80}{20}}$$

Example: high school, again

A certain high school has 80 students: 20 freshmen, 20 sophomores, 20 juniors, and 20 seniors. If a random sample of 20 students is drawn with replacement, what is the probability that all students in the sample are from the same grade level?

① $p(\text{all freshmen}) = \left(\frac{1}{4}\right) \cdot \left(\frac{1}{4}\right) \cdots = \left(\frac{1}{4}\right)^{20}$

$p(\text{all same}) = 4 \cdot \left(\frac{1}{4}\right)^{20} = \boxed{\left(\frac{1}{4}\right)^{19}}$ final ans.

$\square$ 1/4 1/4 1/4 1/4 _ - - - -

all 19 people need to match first

② $p(\text{all freshmen}) = \dfrac{20^{20}}{80^{20}} = \left(\frac{1}{4}\right)^{20}$

samples w/ only freshmen

total samples

Example: bitstrings

e.g. 01101 $\quad 2^5$ bitstrings

What is the probability of a randomly generated bitstring of length 5 having the same first two bits? Assume that each bit is equally likely to be a 0 or a 1.

easy way

$\_\_ \; \frac{1}{2} \; \cancel{\_\_ \; \_\_ \; \_\_} = \boxed{\frac{1}{2}}$

just need second bit to match first

hard way

2 poss:

$\boxed{\underline{0}\;\underline{0}}\; \underline{2}\cdot\underline{2}\cdot\underline{2} = 8$ poss.

$\boxed{\underline{1}\;\underline{1}}\; \underline{2}\cdot\underline{2}\cdot\underline{2} = 8$ poss

$\frac{8+8}{32} = \boxed{\frac{1}{2}}$

Example: bitstrings, again

What is the probability of a randomly generated bitstring of length 5 having the same first two bits, if we know that the bitstring has exactly four 0s? Assume that each bit is equally likely to be a 0 or a 1.

00001
00010
00100
01000
10000

given one of these 5

3 of these 5 have the same first 2

$\Rightarrow \boxed{\frac{3}{5}}$

Conclusion

Learning objectives

At the start of the quarter, we told you that by the end of DSC 40A, you'll...

- understand the basic principles underlying almost every machine learning and data science method.
- be better prepared for the math in upper division: vector calculus, linear algebra, and probability.
- be able to tackle problems such as:
  - How do we know if an avocado is going to be ripe before we eat it?
  - How do we teach a computer to read handwritten text?
  - How do we predict a future data scientist's salary?

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:

- More supervised learning.
  - Logistic regression, decision trees, neural networks, etc.
- More unsupervised learning.
  - Other clustering techniques, PCA, etc.
- More probability.
  - Random variables, distributions, etc.
- More connections between all of these areas.
  - For instance, you'll learn how probability is related to linear regression.
- More practical tools.

Note on grades

Fall 2016

Class	Title	Un.	Gr.
CHEM 1A	General Chemistry	3	B-
CHEM 1AL	General Chemistry Laboratory	1	C+
COMPSCI 61A	The Structure and Interpretation of Computer Programs	4	B+
COMPSCI 70	Discrete Mathematics and Probability Theory	4	A
COMPSCI 195	Social Implications of Computer Technology	1	P
MATH 1A	Calculus	4	A+

Spring 2017

Class	Title	Un.	Gr.
COMPSCI 61B	Data Structures	4	B+
COMPSCI 97	Field Study	1	P
COMPSCI 197	Field Study	1	P
ELENG 16A	Designing Information Devices and Systems I	4	B-
MATH 110	Linear Algebra	4	C
MATH 128A	Numerical Analysis	4	B+

Moral of the story: good grades aren’t everything.

Thank you!

- This course would not have been possible without our TA: Harpreet Singh.
- It also would not have been possible without our 6 tutors: Jianming Geng, Yujian (Ken) He, Shiv Sakthivel, Aryaman Sinha, Luning Yang, and Sheng Yang.
- You can contact them with any questions at **dsc40a.com/staff**.

Theoretical Foundations of Data Science (Part 1)