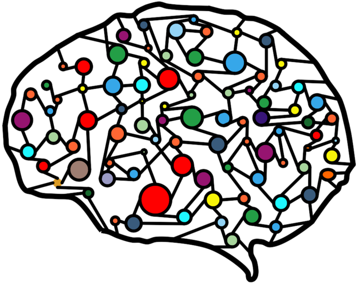
Lecture 27 – Course summary



DSC 40A, Fall 2022 @ UC San Diego

Dr. Truong Son Hy, with help from **many others**

Announcements

- Final exam is coming soon!
- Review the solutions to previous homeworks and groupworks.
- Identify which concepts are still iffy. Re-watch lecture and ask questions (now!).
- Look at the past exams at `https://dsc40a.com/resources`.
- Study in groups.
- Make a “cheat sheet”.
- Bring a calculator.
- Remember to submit The Course and Professor Evaluations (CAPE) – deadline December 2. If everyone submits CAPE, everyone will get a bonus percentage!

Final schedule

O	40A	Theor Fndtns of Data Sci I (4 Units)							Prerequisites | Resources | Evaluations		
			LE	A00	MWF	3:00p-3:50p	PCYNH	122	Hy, Truong Son		
		88107	DI	A01	M	5:00p-5:50p	PCYNH	122	Hy, Truong Son	9	115
			FI	12/03/2022	S	7:00p-9:59p	CSB	001			
O	40A	Theor Fndtns of Data Sci I (4 Units)							Prerequisites | Resources | Evaluations		
			LE	B00	MWF	4:00p-4:50p	PCYNH	122	Soleymani, Mahdi		
		88109	DI	B01	M	6:00p-6:50p	PCYNH	122	Soleymani, Mahdi	19	115
			FI	12/03/2022	S	7:00p-9:59p	CSB	002			

Time: December 3rd, 2022 – 7:00pm to 10:00pm (3 hours)
Location: CSB building – room 001 (for my section)

`https://act.ucsd.edu/scheduleOfClasses/scheduleOfClassesStudentResult.htm`

Please double-check!

Agenda

- Acknowledgements
- High-level summary of the course.
- Review problems.

Acknowledgements

Acknowledgements

Special thanks to:

- Mahdi Soleymani, the other instructor of the course.
- Pushkar Bhuse, the teaching assistant, and Weiyue (Larry) Li, Karthikeya Manchala, Aryaman Sinha, Yujia (Joy) Wang, Yuxin Guo, Vivian Lin, Shiv Sakthivel, Jessica Song, the tutors of the course.
- Justin Eldridge, Janine Tiefenbruck and Suraj Rampure, the instructors of the past courses for their helps.
- And to all of you, the students who attended, worked hard and gave us feedback to improve the course further.

What was this course about?

Part 1: Supervised learning

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: $H(x) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + ... + w_d x^{(d)}$.
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.

Part 1: Unsupervised learning

- When learning how to fit prediction rules, we were performing **supervised machine learning**.
- Then, 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.

Part 2: Probability fundamentals

- If all outcomes in the **sample space** S are equally likely, then $P(A) = \frac{|A|}{|S|}$.
- $\bar{A}$ 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, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- 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

- 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: $P(n,k) = \frac{n!}{(n-k)!}$.

- 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: $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.

Part 2: The law of total probability and Bayes' theorem

- A set of events $E_1, E_2, ..., E_k$ 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 $E_1, E_2, ..., E_k$ is a partition of S, then

$$\begin{aligned} P(A) &= P(E_1) \cdot P(A|E_1) + P(E_2) \cdot P(A|E_2) + \dots + P(E_k) \cdot P(A|E_k) \\ &= \sum_{i=1}^{k} P(E_i) \cdot P(A|E_i) \end{aligned}$$

- **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

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

- 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

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

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

Summary

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.

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 $k = 3$.

Questions:

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

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

Let S denote the whole 2-dimensional space (i.e. sample space). Let E_1, E_2, and E_3 denote the clusters (i.e. partitions):

$$E_1 \cap E_2 = \emptyset, \quad E_2 \cap E_3 = \emptyset, \quad E_3 \cap E_1 = \emptyset$$

$$E_1 \cup E_2 \cup E_3 = S.$$

Let $\mu_1, \mu_2, \mu_3 \in S$ denote the initial centroids.

For simplicity, we assume S is bounded (i.e. S is the convex hull of all $\{(x_i, y_i)\}_{i=1}^{15}$). We have:

$$P(\mu_i \in E_i) = \frac{|E_i|}{|S|}.$$

1. What's the probability that all three initial centroids are initialized in the same natural cluster?
   We further assume **uniform** distribution: $P(\mu_i \in E_i) = 1/3$.
   Because of **independence** in sampling, the final result is:

$$P(\mu_1 \in E_1, \mu_2 \in E_2, \mu_3 \in E_3) = P(\mu_1 \in E_1)P(\mu_2 \in E_2)P(\mu_3 \in E_3) = \frac{1}{27}$$

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

$$P(\mu_1 \notin E_1, \mu_2 \notin E_2, \mu_3 \notin E_3) = P(\mu_1 \notin E_1)P(\mu_2 \notin E_2)P(\mu_3 \notin E_3) = \frac{8}{27}$$

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?

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?

- Using permutations:

$$\frac{6!}{2!2!2!} = 120$$

- Using combinations:

$$\binom{6}{2} \cdot \binom{4}{2} \cdot \binom{2}{2} = \frac{6!}{4!2!} \cdot \frac{4!}{2!2!} \cdot 1 = \frac{6!}{2!2!2!} = 120$$

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?

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?

$$\frac{120}{3!} = 20$$

Example: Lottery

When you buy a Powerball ticket, you select 5 different white numbers from among the numbers 1 through 59 (order of the selection does not matter), and one red number from among the numbers 1 through 35. How many different Powerball tickets can you buy?

Example: Lottery

When you buy a Powerball ticket, you select 5 different white numbers from among the numbers 1 through 59 (order of the selection does not matter), and one red number from among the numbers 1 through 35. How many different Powerball tickets can you buy?

If you check out the Powerball web site you will see that you need to select 5 distinct white numbers, so you can do this $\binom{59}{5} = 5,006,386$ ways. Then you can pick the red number $\binom{35}{1} = 35$ ways so the total number of tickets is:

$$\binom{59}{5} \cdot \binom{35}{1} = 5,006,386 \cdot 35 = 175,223,510.$$

Example: Mixed counting problems

An bag contains 15 marbles of which 10 are red and 5 are white. 4 marbles are selected from the bag.

(a) How many (different) samples (of size 4) are possible?

Example: Mixed counting problems

An bag contains 15 marbles of which 10 are red and 5 are white. 4 marbles are selected from the bag.

(a) How many (different) samples (of size 4) are possible?

$$\binom{15}{4} = 1,365$$

(b) How many samples (of size 4) consist entirely of red marbles?

Example: Mixed counting problems

An bag contains 15 marbles of which 10 are red and 5 are white. 4 marbles are selected from the bag.

(a) How many (different) samples (of size 4) are possible?

$$\binom{15}{4} = 1,365$$

(b) How many samples (of size 4) consist entirely of red marbles?

$$\binom{10}{4} = 210$$

(c) How many samples have 2 red and 2 white marbles?

Example: Mixed counting problems

An bag contains 15 marbles of which 10 are red and 5 are white. 4 marbles are selected from the bag.

(a) How many (different) samples (of size 4) are possible?

$$\binom{15}{4} = 1,365$$

(b) How many samples (of size 4) consist entirely of red marbles?

$$\binom{10}{4} = 210$$

(c) How many samples have 2 red and 2 white marbles?

$$\binom{10}{2} \cdot \binom{5}{2} = 45 \cdot 10 = 450$$

Example: Mixed counting problems

(d) How many samples (of size 4) have exactly 3 red marbles?

Example: Mixed counting problems

(d) How many samples (of size 4) have exactly 3 red marbles?

$$\binom{10}{3} \cdot \binom{5}{1} = 120 \cdot 5 = 600$$

(e) How many samples (of size 4) have at least 3 red?

Example: Mixed counting problems

(d) How many samples (of size 4) have exactly 3 red marbles?

$$\binom{10}{3} \cdot \binom{5}{1} = 120 \cdot 5 = 600$$

(e) How many samples (of size 4) have at least 3 red?

The answer is the number of samples with 3 red plus the number of samples with 4 red:

$$\binom{10}{3} \cdot \binom{5}{1} + \binom{10}{4} \cdot \binom{5}{0} = 120 \cdot 5 + 210 \cdot 1 = 810$$

Example: Mixed counting problems

(f) How many samples (of size 4) contain at least one red marble?

Example: Mixed counting problems

(f) How many samples (of size 4) contain at least one red marble?

One answer is "the number with exactly 1" + "the number with exactly 2" + "the number with exactly 3" + "the number with exactly 4":

$$\binom{10}{1} \cdot \binom{5}{3} + \binom{10}{2} \cdot \binom{5}{2} + \binom{10}{3} \cdot \binom{5}{1} + \binom{10}{4} \cdot \binom{5}{0}$$

$= 10 \cdot 10 + 45 \cdot 10 + 120 \cdot 5 + 210 \cdot 1 = 100 + 450 + 600 + 210 = 1,360$

But it is a slow computation!

Example: Mixed counting problems

(f) How many samples (of size 4) contain at least one red marble?

One answer is "the number with exactly 1" + "the number with exactly 2" + "the number with exactly 3" + "the number with exactly 4":

$$\binom{10}{1}\cdot\binom{5}{3}+\binom{10}{2}\cdot\binom{5}{2}+\binom{10}{3}\cdot\binom{5}{1}+\binom{10}{4}\cdot\binom{5}{0}$$

$$= 10\cdot 10+45\cdot 10+120\cdot 5+210\cdot 1 = 100+450+600+210 = 1,360$$

But it is a slow computation! The faster answer is the total number of samples minus the number of samples with no red marbles:

$$\binom{15}{4}-\binom{10}{0}\cdot\binom{5}{4} = 1,365-5 = 1,360.$$

Example: 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.

We can start the bitstring as 00 or 11. The rest of the string does not matter. Because, there are 4 different ways to start the bitstring: $\{00, 01, 10, 11\}$. The final result is:

$$\frac{1}{2}$$

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.

The first two bits must be two zeros, because there are exactly four 0s. There is also exactly one bit 1. We can only put bit 1 among the 3rd, 4th and 5th position. Thus, only 3 possibilities:

$$\{00100, 00010, 00001\}$$

The final result is:

$$\frac{3}{32}.$$