Lecture 15 – Foundations of Probability

DSC 40A, Winter 2024

Announcements

- Midterm 1 grade released
  - Mean 23.82/40, Standard Deviation:8.6
  - Spring 2023 midterm: Mean 23.48/40, Standard Deviation: 8.3
  - Rubrics will be published today
- HW4 due this upcoming Friday, please start early.
  - HW4 Q6 is now a separate programming assignment on Gradescope
  - When submitting your program, please make sure its name is **”calculator.py”**
- HW5 published today and due next Wednesday.

Agenda

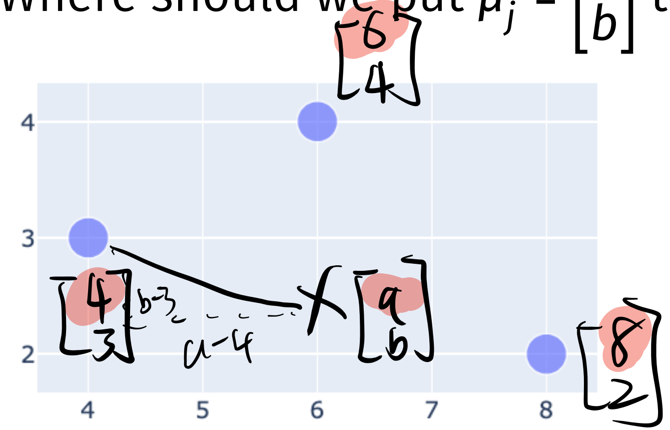
- Finish Clustering
- Probability: context and overview.
- Complement, addition, and multiplication rules.
- Conditional probability.

Why does k-Means work? (Step 3)

$C(\mu_j)$ = total squared distance of each data point $\vec{x}_i$ in group j to centroid μ_j

Suppose group j contains the points (4, 3), (6, 4), and (8, 2).

Where should we put $\mu_j = \begin{bmatrix} a \\ b \end{bmatrix}$ to minimize $C(\mu_j)$?



$$C(a,b) = (a-4)^2 + (b-3)^2 + (a-6)^2 + (b-4)^2 + (a-8)^2 + (b-2)^2$$

$$\text{distance} = \sqrt{(a-4)^2 + (b-3)^2}$$

$$\text{distance}^2 = (a-4)^2 + (b-3)^2$$

Why does k-Means work? (Step 3)

$$C(a,b) = (a-4)^2 + (b-3)^2$$

$$+ (a-6)^2 + (b-4)^2$$

$$+ (a-8^2) + (b-2)^2$$

$$\frac{\partial C}{\partial a} = 2(a-4) + 2\cdot(a-6) + 2(a-8) = 0$$

$$a-4 + a-6 + a-8 = 0$$

$$3a = 4+6+8$$

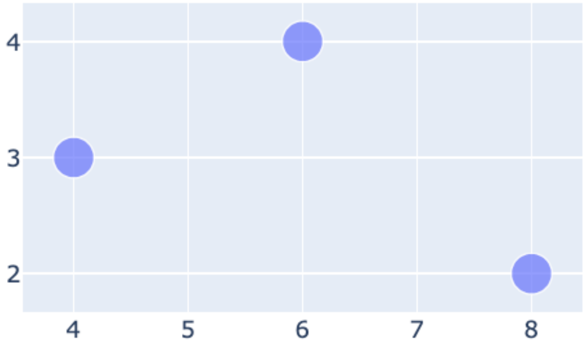
$$a = \frac{4+6+8}{3} = \text{mean}_x$$

Why does k-Means work? (Step 3)

$$C(\mu_j) = \text{total squared distance of each data point } \vec{x}_i \text{ in group } j \text{ to centroid } \mu_j$$

Suppose group j contains the points $(4, 3)$, $(6, 4)$, and $(8, 2)$.

Where should we put $\mu_j = \begin{bmatrix} a \\ b \end{bmatrix}$ to minimize $C(\mu_j)$?



Cost and empirical risk

- On the previous slide, we saw a function of the form

$$\begin{aligned} C(\mu_j) = C(a, b) &= (4 - a)^2 + (3 - b)^2 \\ &+ (6 - a)^2 + (4 - b)^2 \\ &+ (8 - a)^2 + (2 - b)^2 \end{aligned}$$

- $C(a, b)$ can be thought of as the sum of two separate functions, $f(a)$ and $g(b)$.
    - $f(a) = (4 - a)^2 + (6 - a)^2 + (8 - a)^2$ computes the total squared distance of each x_1 coordinate to a.
    - From earlier in the course, we know that $a^* = \frac{4+6+8}{3} = 6$ minimizes $f(a)$.

Practical considerations

Initialization

- Depending on our initial centroids, k-Means may "converge" to a clustering that doesn't actually have the lowest possible inertia.
  - In other words, like gradient descent, k-Means can get caught in a **local minimum**.

- Some solutions:
  - Run k-Means several times, each with different randomly chosen initial centroids. Keep track of the inertia of the final result in each attempt. Choose the attempt with the lowest inertia.

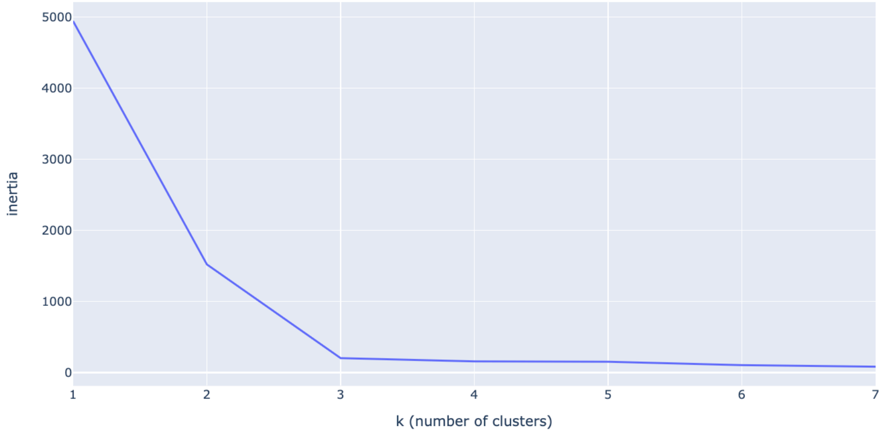
  - **k-Means++**: choose one initial centroid at random, and place other centroids far from all other centroids.

Choosing *k*

- Note that as *k* increases, inertia decreases.
    - Intuitively, as we add more centroids, the distance between each point and its closest centroid will drop.
- But the goal of clustering is to put data points into groups, and having a large number of groups may not be meaningful.
- This suggests a tradeoff between *k* and inertia.

The “elbow” method

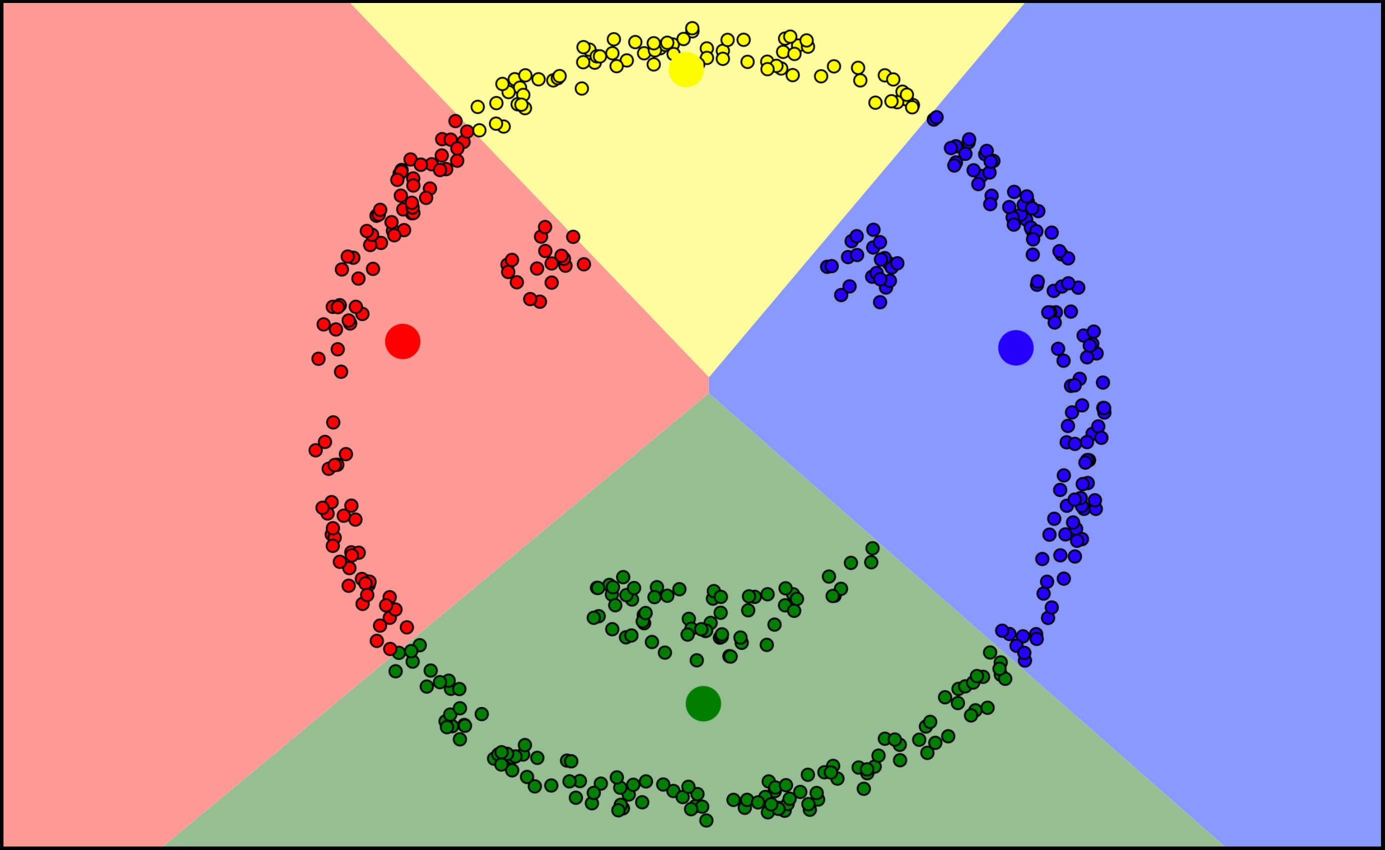
- Strategy: run k-Means Clustering for many choices of k (e.g. $k = 1, 2, 3, ..., 8$).
- Compute the value of inertia for each resulting set of centroids.
- Plot a graph of inertia vs k.
- Choose the value of k that appears at an “elbow”.



See the notebook for a demo.

Low inertia isn’t everything!

- Even if k-Means works as intended and finds the choice of centroids that minimize inertia, the resulting clustering may not look “right” to us humans.
  - Recall, inertia measures the total squared distance to centroids.
  - This metric doesn’t always match our intuition.
- Let’s look at some examples at `https://tinyurl.com/40akmeans`.
  - Go to “I’ll Choose” and “Smiley Face”. Good luck!



Other clustering techniques

- k-Means Clustering is just one way to cluster data.
- There are many others, each of which work differently and produce different kinds of results.
- Another common technique: **agglomerative clustering**.
  - High level: start out with each point being in its own cluster. Repeatedly combine clusters until only k are left.
- Check out this chart.

Probability: context and overview

From Lecture 1: course overview

Part 1: Learning from Data

- Summary statistics and loss functions; mean absolute error and mean squared error.
- Linear regression (incl. linear algebra).
- Clustering.

Part 2: Probability

- Probability fundamentals. Set theory and combinatorics.
- Conditional probability and independence.
- Naïve Bayes (uses concepts from both parts of the class).

Why do we need probability?

- So far in this class, we have made predictions based on a dataset.
- This dataset can be thought of as a **sample** of some population.
- For a prediction rule to be useful in the future, the sample that was used to create the prediction rule needs to look similar to samples that we'll see in the future.

Probability and statistics

Statistical inference

Given observed data, we want to know how it was generated or where it came from, for the purposes of

- predicting outcomes for other data generated from the same source.
- knowing how different our sample could have been.
- drawing conclusions about our entire population and not just our observed sample (i.e. generalizing).

Probability

Given a certain model for data generation, what kind of data do you expect the model to produce? How similar is it to the data you have?

- Probability is the tool to answer these questions.
- You need probability to do statistics, and vice versa.
- Example: Is my coin fair?

Terminology

- An **experiment** is some process whose outcome is random (e.g. flipping a coin, rolling a die).
- A **set** is an unordered collection of items. $|A|$ denotes the number of elements in set A.
- A **sample space**, S, is the set of all possible outcomes of an experiment.
  - Could be finite or infinite!
- An **event** is a subset of the sample space, or a set of outcomes.
  - Notation: $E \subseteq S$.

Probability distributions

- A probability distribution, p, describes the **probability** of each outcome s in a sample space S.
  - The probability of each outcome must be between 0 and 1: $0 \le p(s) \le 1$.
  - The sum of the probabilities of each outcome must be exactly 1: $\sum_{s \in S} p(s) = 1$.
- The probability of an **event** is the sum of the probabilities of the outcomes in the event.
  - $P(E) = \sum_{s \in E} p(s)$.

Example: probability of rolling an even number on a 6-sided die

Equally-likely outcomes

- If S is a sample space with n possible outcomes, and all outcomes are equally-likely, then the probability of any one outcome occurring is $\frac{1}{n}$.

- The probability of an event E, then, is

$$P(E) = \frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n} = \frac{\text{\# of outcomes in E}}{\text{\# of outcomes in S}} = \frac{|E|}{|S|}$$

- **Example:** Flipping a coin three times.

Complement, addition, and multiplication rules

Complement rule

- Let A be an event with probability $P(A)$.

- Then, the event $\bar{A}$ is the **complement** of the event A. It contains the set of all outcomes in the sample space that are not in A.

- $P(\bar{A})$ is given by

$$P(\bar{A}) = 1 - P(A)$$

Addition rule

- We say two events are **mutually exclusive** if they have no overlap (i.e. they can't both happen at the same time).

- If A and B are mutually exclusive, then the probability that A or B happens is

$$P(A \cup B) = P(A) + P(B)$$

Principle of inclusion-exclusion

- If events A and B are not mutually exclusive, then the addition rule becomes more complicated.

- In general, if A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Discussion Question

Each day when you get home from school, there is a

- 0.3 chance your mom is at home.
- 0.4 chance your brother is at home.
- 0.25 chance that both your mom and brother are at home.

When you get home from school today, what is the chance that **neither** your mom nor your brother are at home?

a) 0.3
b) 0.45
c) 0.55
d) 0.7
e) 0.75

Multiplication rule and independence

- The probability that events A and B both happen is

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

- $P(B|A)$ means "the probability that B happens, given that A happened." It is a **conditional probability**.

- If $P(B|A) = P(B)$, we say A and B are **independent**.
  - Intuitively, A and B are independent if knowing that A happened gives you no additional information about event B, and vice versa.

  - For two independent events,

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

Example: rolling a die

Let’s consider rolling a fair 6-sided die. The results of each die roll are independent from one another.

- Suppose we roll the die once. What is the probability that the face is 1 and 2?
- Suppose we roll the die once. What is the probability that the face is 1 or 2?

Example: rolling a die

- Suppose we roll the die 3 times. What is the probability that the face 1 never appears in any of the rolls?

- Suppose we roll the die 3 times. What is the probability that the face 1 appears at least once?

Example: rolling a die

- Suppose we roll the die n times. What is the probability that only the faces 2, 4, and 5 appear?

- Suppose we roll the die twice. What is the probability that the two rolls have different faces?

Conditional probability

Conditional probability

- The probability of an event may **change** if we have additional information about outcomes.

- Starting with the multiplication rule, $P(A \cap B) = P(A)P(B|A)$, we have that

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

  assuming that $P(A) > 0$.

Example: pets

Suppose a family has two pets. Assume that it is equally likely that each pet is a dog or a cat. Consider the following two probabilities:

1. The probability that both pets are dogs given that **the oldest is a dog**.

2. The probability that both pets are dogs given that **at least one of them is a dog**.

Discussion Question

Are these two probabilities equal?

a) Yes, they're equal

b) No, they're not equal

Example: pets

Let’s compute the probability that both pets are dogs given that **the oldest is a dog**.

Example: pets

Let’s now compute the probability that both pets are dogs given that **at least one of them is a dog**.

Summary, next time

Summary

- Two events A and B are mutually exclusive if they share no outcomes (no overlap). In this case,

$$P(A \cup B) = P(A) + P(B).$$

- 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 conditional probability of B occurring, given that A occurs. If $P(B|A) = P(B)$, then events A and B are independent.

Next time

- More probability and introduction to combinatorics, the study of counting.
- **Important:** We've posted **many** probability resources on the resources tab of the course website. These will no doubt come in handy.
  - No more DSC 40A-specific readings, though the Probability Roadmap was written specifically for students of this course.