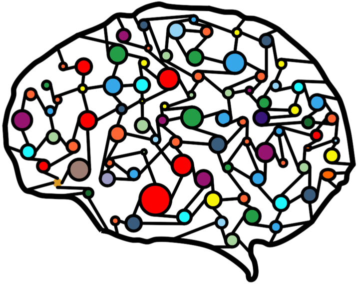
Lecture 23 – Naive Bayes



DSC 40A, Fall 2022 @ UC San Diego

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

Agenda

- Classification.
- Classification and conditional independence.
- Naive Bayes.

Recap: Bayes' theorem, independence, and conditional independence

- Bayes' theorem: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$.
- A and B are **independent** if $P(A \cap B) = P(A) \cdot P(B)$.
- A and B are **conditionally independent** given C if $P((A \cap B)|C) = P(A|C) \cdot P(B|C)$.
  - In general, there is no relationship between independence and conditional independence.

Example: avocados

You have a green-black avocado, and want to know if it is ripe.

color	ripeness
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	unripe
purple-black	ripe
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	ripe
green-black	unripe
purple-black	ripe

Question: Based on this data, would you predict that your avocado is ripe or unripe?

Example: avocados

You have a green-black avocado, and want to know if it is ripe. Based on this data, would you predict that your avocado is ripe or unripe?

color	**ripeness**
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	unripe
purple-black	ripe
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	ripe
green-black	unripe
purple-black	ripe

Strategy: Calculate two probabilities:

$$P(\text{ripe}|\text{green-black})$$

$$P(\text{unripe}|\text{green-black})$$

Then, predict the class with a **larger** probability.

Estimating probabilities

- We would like to determine $P(\text{ripe}|\text{green-black})$ and $P(\text{unripe}|\text{green-black})$ for all avocados in the universe.
- All we have is a single dataset, which is a **sample** of all avocados in the universe.
- We can estimate these probabilities by using sample proportions.

$$P(\text{ripe}|\text{green-black}) \approx \frac{\text{\# ripe green-black avocados in sample}}{\text{\# green-black avocados in sample}}$$

- Per the **law of large numbers** in DSC 10, larger samples lead to more reliable estimates of population parameters.

Example: avocados

You have a green-black avocado, and want to know if it is ripe. Based on this data, would you predict that your avocado is ripe or unripe?

color	**ripeness**
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	unripe
purple-black	ripe
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	ripe
green-black	unripe
purple-black	ripe

$P(\text{ripe}|\text{green-black}) = ?$

$P(\text{unripe}|\text{green-black}) = ?$

Example: avocados

By definition:

$$P(\text{ripe}|\text{green-black}) = \frac{P(\text{ripe,green-black})}{P(\text{green-black})}$$

There are 3 out of 11 rows that are (green-black, ripe). Thus,

$$P(\text{ripe,green-black}) = 3/11.$$

On the another, by the Law of total probability, we have:

$$P(\text{green-black}) = P(\text{green-black, ripe}) + P(\text{green-black, unripe}).$$

We count that there are 2 out of 11 rows that are (green-black, unripe). Thus,

$$P(\text{unripe,green-black}) = 2/11.$$

Example: avocados

Therefore:

$$P(\text{green-black}) = 5/11.$$

We have:

$$P(\text{ripe}|\text{green-black}) = \frac{3}{5} = 60\%$$

and:

$$P(\text{unripe}|\text{green-black}) = \frac{2}{5} = 40\%.$$

Bayes' theorem for classification

- Suppose that A is the event that an avocado has certain features, and B is the event that an avocado belongs to a certain class. Then, by Bayes' theorem:

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

- More generally:

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

- What's the point?
  - Usually, it's not possible to estimate $P(\text{class}|\text{features})$ directly from the data we have.
  - Instead, we have to estimate $P(\text{class})$, $P(\text{features}|\text{class})$, and $P(\text{features})$ separately.

Example: avocados

You have a green-black avocado, and want to know if it is ripe. Based on this data, would you predict that your avocado is ripe or unripe?

color	ripeness
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	unripe
purple-black	ripe
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	ripe
green-black	unripe
purple-black	ripe

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

Example: avocados

You have a green-black avocado, and want to know if it is ripe. Based on this data, would you predict that your avocado is ripe or unripe?

color	**ripeness**
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	unripe
purple-black	ripe
bright green	unripe
green-black	ripe
purple-black	ripe
green-black	ripe
green-black	unripe
purple-black	ripe

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

Shortcut: Both probabilities have the same denominator. The larger one is the one with the larger numerator.

$$P(\text{ripe}|\text{green-black})$$

$$P(\text{unripe}|\text{green-black})$$

Example: avocados

We can ignore the denominator $P(\text{features})$:

$$P(\text{class}|\text{features}) \propto P(\text{class}) \cdot P(\text{features}|\text{class}),$$

where $\propto$ means "proportional to". We have:

$$P(\text{ripe}|\text{green-black}) \propto P(\text{ripe}) \cdot P(\text{green-black}|\text{ripe})$$

$$P(\text{unripe}|\text{green-black}) \propto P(\text{unripe}) \cdot P(\text{green-black}|\text{unripe})$$

Example: avocados

First, we can do the approximation for the **priors**:

- There are 7 out of 11 rows having “ripe” labels:
  $P(\text{ripe}) = 7/11$.

- There are 4 out of 11 rows having “unripe” labels:
  $P(\text{unripe}) = 4/11$.

Example: avocados

First, we can do the approximation for the **priors**:

- There are 7 out of 11 rows having “ripe” labels: $P(\text{ripe}) = 7/11$.

- There are 4 out of 11 rows having “unripe” labels: $P(\text{unripe}) = 4/11$.

Second, we can do the approximation for the **posteriors**:

- Out of 7 rows with “ripe” labels, only 3 rows have “green-black”: $P(\text{green-black}|\text{ripe}) = 3/7$.

- Out of 4 rows with “unripe“ labels, only 2 rows have “green-black”: $P(\text{green-black}|\text{ripe}) = 2/4 = 1/2$.

Example: avocados

We have:

$$P(\text{ripe}|\text{green-black}) \propto P(\text{ripe}) \cdot P(\text{green-black}|\text{ripe}) = \frac{7}{11} \cdot \frac{3}{7} = \frac{3}{11}$$

$$P(\text{unripe}|\text{green-black}) \propto P(\text{unripe}) \cdot P(\text{green-black}|\text{unripe}) = \frac{4}{11} \cdot \frac{1}{2} = \frac{2}{11}$$

Example: avocados

We have:

$$P(\text{ripe}|\text{green-black}) \propto P(\text{ripe}) \cdot P(\text{green-black}|\text{ripe}) = \frac{7}{11} \cdot \frac{3}{7} = \frac{3}{11}$$

$$P(\text{unripe}|\text{green-black}) \propto P(\text{unripe}) \cdot P(\text{green-black}|\text{unripe}) = \frac{4}{11} \cdot \frac{1}{2} = \frac{2}{11}$$

We got a vector $(3/11, 2/11)$, that does **not** form a probability distribution yet. In general, given a non-negative vector $\vec{x} = (x_1, .., x_n) \in \mathbb{R}^n$, we can create a probability distribution $\vec{p} = (p_1, .., p_n)$ by dividing $\vec{x}$ by its sum of all elements:

$$\vec{p} = \frac{\vec{x}}{\|\vec{x}\|_1},$$

where $p_i = x_i / \sum_{j=1}^{n} x_j$.

Example: avocados

After normalization, we get (60%, 40%). Actually, for prediction, we do **not** even need to calculate the probability exactly. In this case, 3/11 > 2/11, thus:

$$P(\text{ripe}|\text{green-black}) > P(\text{unripe}|\text{green-black})$$

and we can conclude that given green-black color, it is likely that the avocado is ripe (i.e. the prediction is ripe).

Classification and conditional independence

Example: avocados, but with more features

color	softness	variety	ripeness
bright green	firm	Zutano	unripe
green-black	medium	Hass	ripe
purple-black	firm	Hass	ripe
green-black	medium	Hass	unripe
purple-black	soft	Hass	ripe
bright green	firm	Zutano	unripe
green-black	soft	Zutano	ripe
purple-black	soft	Hass	ripe
green-black	soft	Zutano	ripe
green-black	firm	Hass	unripe
purple-black	medium	Hass	ripe

You have a firm green-black Zutano avocado. Based on this data, would you predict that your avocado is ripe or unripe?

Example: avocados, but with more features

color	softness	variety	ripeness
bright green	firm	Zutano	unripe
green-black	medium	Hass	ripe
purple-black	firm	Hass	ripe
green-black	medium	Hass	unripe
purple-black	soft	Hass	ripe
bright green	firm	Zutano	unripe
green-black	soft	Zutano	ripe
purple-black	soft	Hass	ripe
green-black	soft	Zutano	ripe
green-black	firm	Hass	unripe
purple-black	medium	Hass	ripe

You have a firm green-black Zutano avocado. Based on this data, would you predict that your avocado is ripe or unripe?

Strategy: Calculate $P(\text{ripe}|\text{features})$ and $P(\text{unripe}|\text{features})$ and choose the class with the **larger** probability.

$$P(\text{ripe}|\text{firm, green-black, Zutano})$$

$$P(\text{unripe}|\text{firm, green-black, Zutano})$$

Example: avocados, but with more features

color	softness	variety	ripeness
bright green	firm	Zutano	unripe
green-black	medium	Hass	ripe
purple-black	firm	Hass	ripe
green-black	medium	Hass	unripe
purple-black	soft	Hass	ripe
bright green	firm	Zutano	unripe
green-black	soft	Zutano	ripe
purple-black	soft	Hass	ripe
green-black	soft	Zutano	ripe
green-black	firm	Hass	unripe
purple-black	medium	Hass	ripe

You have a firm green-black Zutano avocado. Based on this data, would you predict that your avocado is ripe or unripe?

Issue: We have not seen a firm green-black Zutano avocado before.

This means that $P(\text{ripe}|\text{firm, green-black, Zutano})$ and $P(\text{unripe}|\text{firm, green-black, Zutano})$ are undefined.

A simplifying assumption

- We want to find $P(\text{ripe}|\text{firm, green-black, Zutano})$, but there are no firm green-black Zutano avocados in our dataset.

- Bayes' theorem tells us this probability is equal to

$$P(\text{ripe}|\text{firm, green-black, Zutano}) = \frac{P(\text{ripe}) \cdot P(\text{firm, green-black, Zutano}|\text{ripe})}{P(\text{firm, green-black, Zutano})}$$

- **Key idea:** **Assume** that features are **conditionally independent** given a class (e.g. ripe).

$$P(\text{firm, green-black, Zutano}|\text{ripe}) = P(\text{firm}|\text{ripe}) \cdot P(\text{green-black}|\text{ripe}) \cdot P(\text{Zutano}|\text{ripe})$$

A simplifying assumption

$P(\text{ripe}|\text{firm, green-black, Zutano}) \propto P(\text{ripe}) \cdot P(\text{firm, green-black, Zutano}|\text{ripe})$

$P(\text{firm, green-black, Zutano}|\text{ripe}) = P(\text{firm}|\text{ripe}) \cdot P(\text{green-black}|\text{ripe}) \cdot P(\text{Zutano}|\text{ripe})$

A simplifying assumption

$P(\text{ripe}|\text{firm, green-black, Zutano}) \propto P(\text{ripe}) \cdot P(\text{firm, green-black, Zutano}|\text{ripe})$

$P(\text{firm, green-black, Zutano}|\text{ripe}) = P(\text{firm}|\text{ripe}) \cdot P(\text{green-black}|\text{ripe}) \cdot P(\text{Zutano}|\text{ripe})$

Among 7 rows with label "ripe":

- Only 1 row with "firm": $P(\text{firm}|\text{ripe}) = 1/7$.
- 3 rows with "green-black": $P(\text{green-black}|\text{ripe}) = 3/7$.
- 2 rows with "Zutano": $P(\text{Zutano}|\text{ripe}) = 2/7$.

Thus:

$$P(\text{firm, green-black, Zutano}|\text{ripe}) = \frac{6}{7^3}.$$

Therefore:

$$P(\text{ripe}|\text{firm, green-black, Zutano}) \propto \frac{7}{11} \cdot \frac{6}{7^3} = \frac{6}{539}.$$

Example: avocados, but with more features

color	softness	variety	ripeness
bright green	firm	Zutano	unripe
green-black	medium	Hass	ripe
purple-black	firm	Hass	ripe
green-black	medium	Hass	unripe
purple-black	soft	Hass	ripe
bright green	firm	Zutano	unripe
green-black	soft	Zutano	ripe
purple-black	soft	Hass	ripe
green-black	soft	Zutano	ripe
green-black	firm	Hass	unripe
purple-black	medium	Hass	ripe

You have a firm green-black Zutano avocado. Based on this data, would you predict that your avocado is ripe or unripe?

$$P(\text{unripe}|\text{firm, green-black, Zutano}) = \frac{P(\text{unripe}) \cdot P(\text{firm, green-black, Zutano}|\text{unripe})}{P(\text{firm, green-black, Zutano})}$$

Let's calculate for "unripe" label!

More calculation

Among 4 rows with label "unripe":

- 3 rows with "firm": $P(\text{firm}|\text{unripe}) = 3/4$.
- 2 rows with "green-black": $P(\text{green-black}|\text{unripe}) = 2/4 = 1/2$.
- 2 rows with "Zutano": $P(\text{Zutano}|\text{unripe}) = 2/4 = 1/2$.

Thus:

$$P(\text{firm, green-black, Zutano}|\text{unripe}) = \frac{12}{4^3}.$$

Therefore:

$$P(\text{unripe}|\text{firm, green-black, Zutano}) \propto \frac{4}{11} \cdot \frac{12}{4^3} = \frac{3}{44}.$$

Conclusion

- The numerator of $P(\text{ripe}|\text{firm, green-black, Zutano})$ is $\frac{6}{539}$.
- The numerator of $P(\text{unripe}|\text{firm, green-black, Zutano})$ is $\frac{3}{44}$.
  - Both probabilities have the same denominator, $P(\text{firm, green-black, Zutano})$.
  - Since we're just interested in seeing which one is larger, we can ignore the denominator and compare numerators.
- Since the numerator for unripe is **larger** than the numerator for ripe, we **predict that our avocado is unripe**.

Naive Bayes

Naive Bayes classifier

- We want to predict a class, given certain features.
- Using Bayes' theorem, we write

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

- For each class, we compute the numerator using the **naive assumption of conditional independence of features given the class**.
- We estimate each term in the numerator based on the training data.
- We predict the class with the largest numerator.
  - Works if we have multiple classes, too!

na·ive

/nīˈēv/

adjective

(of a person or action) showing a lack of experience, wisdom, or judgment.
"the rather naive young man had been totally misled"

- (of a person) natural and unaffected; innocent.
  "Andy had a sweet, naive look when he smiled"
  Similar: innocent unsophisticated artless ingenuous inexperienced
- of or denoting art produced in a straightforward style that deliberately rejects sophisticated artistic techniques and has a bold directness resembling a child's work, typically in bright colors with little or no perspective.

Example: comic characters

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

My favorite character is a male Marvel character. Using Naive Bayes, would we predict that my favorite character is bad, good, or neutral?

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

$$P(\text{bad}|\text{male, Marvel}) \propto P(\text{bad}) \cdot P(\text{male, Marvel}|\text{bad})$$

$$P(\text{male, Marvel}|\text{bad}) = P(\text{male}|\text{bad}) \cdot P(\text{Marvel}|\text{bad})$$

$$P(\text{bad}) = \frac{5}{10}$$

$$P(\text{male}|\text{bad}) = \frac{3}{5}$$

$$P(\text{Marvel}|\text{bad}) = \frac{2}{5}$$

$$P(\text{bad}|\text{male, Marvel}) \propto \frac{5 \cdot 3 \cdot 2}{10 \cdot 5 \cdot 5} = \frac{3}{25}$$

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

$$P(\text{good}|\text{male, Marvel}) \propto P(\text{good}) \cdot P(\text{male, Marvel}|\text{good})$$

$$P(\text{male, Marvel}|\text{good}) = P(\text{male}|\text{good}) \cdot P(\text{Marvel}|\text{good})$$

$$P(\text{good}) = \frac{4}{10}$$

$$P(\text{male}|\text{good}) = \frac{2}{4}$$

$$P(\text{Marvel}|\text{good}) = \frac{3}{4}$$

$$P(\text{good}|\text{male, Marvel}) \propto \frac{4 \cdot 2 \cdot 3}{10 \cdot 4 \cdot 4} = \frac{3}{20}$$

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

$$P(\text{neutral}|\text{male, Marvel}) \propto P(\text{neutral}) \cdot P(\text{male, Marvel}|\text{neutral})$$

$$P(\text{male, Marvel}|\text{neutral}) = P(\text{male}|\text{neutral}) \cdot P(\text{Marvel}|\text{neutral})$$

$$P(\text{neutral}) = \frac{1}{10}$$

$$P(\text{male}|\text{neutral}) = \frac{1}{1} = 1$$

$$P(\text{Marvel}|\text{neutral}) = \frac{1}{1} = 1$$

$$P(\text{neutral}|\text{male, Marvel}) \propto \frac{1}{10}$$

Example: comic characters

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

My other favorite character is a **male** Marvel character. Using Naive Bayes, would we predict that my favorite character is bad, good, or neutral? Good!

Example: comic characters

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

My other favorite character is a **female** Marvel character. What is the probability that this character is neutral?

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

$$P(\text{neutral}|\text{female, Marvel}) \propto P(\text{neutral}) \cdot P(\text{female, Marvel}|\text{neutral})$$

$$P(\text{female, Marvel}|\text{neutral}) = P(\text{female}|\text{neutral}) \cdot P(\text{Marvel}|\text{neutral})$$

$$P(\text{neutral}) = \frac{1}{10}$$

$$P(\text{female}|\text{neutral}) = \frac{0}{1} = 0$$

$$P(\text{Marvel}|\text{neutral}) = \frac{1}{1} = 1$$

$$P(\text{neutral}|\text{female, Marvel}) \propto 0$$

Uh oh...

- There are no neutral female characters in the data set.
- The estimate $P(\text{female}|\text{neutral}) \approx \frac{\text{\# female neutral characters}}{\text{\# neutral characters}}$ is 0.
- The estimated numerator, $P(\text{neutral}) \cdot P(\text{female, Marvel}|\text{neutral}) = P(\text{neutral}) \cdot P(\text{female}|\text{neutral}) \cdot P(\text{Marvel}|\text{neutral})$, is also 0.
- But just because there isn't a neutral female character in the data set, doesn't mean they don't exist!
- **Idea:** Adjust the numerators and denominators of our estimate so that they're never 0.

Smoothing

- **Without** smoothing:

$$P(\text{female}|\text{neutral}) \approx \frac{\text{\# female neutral}}{\text{\# female neutral} + \text{\# male neutral}}$$

$$P(\text{male}|\text{neutral}) \approx \frac{\text{\# male neutral}}{\text{\# female neutral} + \text{\# male neutral}}$$

- **With** smoothing:

$$P(\text{female}|\text{neutral}) \approx \frac{\text{\# female neutral} + 1}{\text{\# female neutral} + 1 + \text{\# male neutral} + 1}$$

$$P(\text{male}|\text{neutral}) \approx \frac{\text{\# male neutral} + 1}{\text{\# female neutral} + 1 + \text{\# male neutral} + 1}$$

- When smoothing, we add 1 to the count of every group whenever we're estimating a probability.

ALIGN	SEX	COMPANY
Bad	Male	Marvel
Neutral	Male	Marvel
Good	Male	Marvel
Bad	Male	DC
Good	Female	Marvel
Bad	Male	DC
Good	Male	DC
Bad	Male	Marvel
Good	Female	Marvel
Bad	Female	Marvel

$$P(\text{neutral}|\text{female, Marvel}) \propto P(\text{neutral}) \cdot P(\text{female, Marvel}|\text{neutral})$$

$$P(\text{female, Marvel}|\text{neutral}) = P(\text{female}|\text{neutral}) \cdot P(\text{Marvel}|\text{neutral})$$

$$P(\text{neutral}) = \frac{1}{10}$$

$$P(\text{female}|\text{neutral}) = \frac{1}{2}$$

$$P(\text{Marvel}|\text{neutral}) = \frac{2}{3}$$

$$P(\text{neutral}|\text{female, Marvel}) \propto \frac{1}{30}$$

Summary

Summary

- 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 simplifying assumption, that features are conditionally independent given a class:

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