

Lecture 23 – Classification and Conditional Independence



DSC 40A, Fall 2022 @ UC San Diego

Mahdi Soleymani, with help from [many others](#)

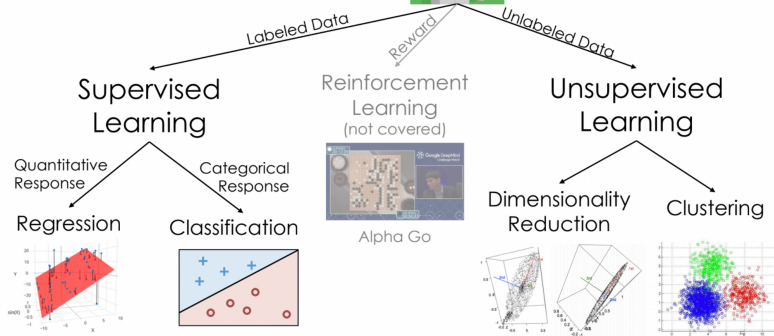
Agenda

- ▶ Classification.
- ▶ Classification and conditional independence.

Classification

Taxonomy of machine learning

Taxonomy of Machine Learning



Classification problems

- ▶ Like with regression, we're interested in making predictions based on data we've already collected (called **training data**).
- ▶ The difference is that the response variable is **categorical**.
- ▶ Categories are called **classes**.
- ▶ Example classification problems:
 - ▶ Deciding whether a patient has kidney disease.
 - ▶ Identifying handwritten digits.
 - ▶ Determining whether an avocado is ripe.
 - ▶ Predicting whether credit card activity is fraudulent.

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

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

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

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

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

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.

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

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

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).

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

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?

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

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?

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

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 $\frac{6}{88}$.
 - ▶ 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**.

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.