DSC 40A

Theoretical Foundations of Data Science I

Last Time

- Recall that A and B are independent if

$$P(A \text{ and } B) = P(A) * P(B)$$

- A and B are conditionally independent given C if

$$P((A \text{ and } B)|C) = P(A|C) * P(B|C)$$

- Given that C occurs, this says that A and B are independent of one another.

In This Video

- Using Bayes' Theorem to solve the classification problem

Classification

- Making predictions based on examples (training data)
- Response variable is categorical
- Categories are called *classes*
- Examples:
  - decide whether patient has kidney disease
  - identify handwritten digits
  - determine whether an avocado is ripe
  - predict whether credit card activity is fraudulent

Example

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

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

Which class would you predict?

A. ripe

B. unripe

Example

sample

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

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

Strategy: Calculate two probabilities:

P(ripe | green-black) $\approx$ # ripe green-black in sample / # green-black in sample

population parameter

sample statistic

P(unripe | green-black)

Then choose the class according to the **larger** of these two probabilities.

Bayes' Theorem for Classification

Bayes' Theorem gives another strategy for predicting the class given features.

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

B = belonging to a certain class
A = having certain features

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

directly estimate

Bayes' Theorem for Classification

Bayes' Theorem gives another strategy for predicting the class given features.

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

B = belonging to a certain class
A = having certain features

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

Can all be estimated from the training data

Avocado Ripeness

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

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

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

ripe green-black

$P(\text{unripe}/\text{green-black}) = \frac{2}{5}$

$P(\text{green-black}/\text{ripe}) \quad \frac{3}{7}$

$P(\text{ripe}) \quad \frac{7}{11}$

$P(\text{green-black}) \quad \frac{5}{11}$

$= \frac{3/7 * 7/11}{5/11} = \frac{3}{5}$

Avocado Ripeness

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

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

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

Avocado Ripeness

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

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

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

ripe, unripe

$P(\text{features}) = \frac{5}{11}$

Shortcut: Both probabilities have same denominator. To find larger one, choose one with larger numerator.

P(ripe | green-black) prop. to $\frac{3}{7} * \frac{7}{11} = \frac{3}{11}$

P(unripe | green-black) prop. to $\frac{2}{4} * \frac{4}{11} = \frac{2}{11}$

3 features

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?

Avocado Ripeness

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 two probabilities:

P(ripe | firm, green-black, Zutano)

P(unripe | firm, green-black, Zutano)

Then choose the class according to the **larger** of these two probabilities.

Avocado Ripeness

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?

Problem: We have not seen an avocado with all these features. Both probabilities will be undefined.

P(ripe | firm, green-black, Zutano)

P(unripe | firm, green-black, Zutano)

Avocado Ripeness

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?

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

Solution: Use Bayes' Theorem, plus a simplifying assumption, to calculate the two numerators.

Avocado Ripeness

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?

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

Simplifying assumption: Within a given class, the features are independent.

P(firm, green-black, Zutano | ripe) =
P(firm | ripe)*P(green-black | ripe)*P(Zutano | ripe)

Conditional Independence

- Recall that A and B are independent if

$$P(A \text{ and } B) = P(A) * P(B)$$

- A and B are conditionally independent given C if

$$P((A \text{ and } B)|C) = P(A|C) * P(B|C) \ldots$$

features | class

- Given that C occurs, this says that A and B are independent of one another.

Avocado Ripeness

$\frac{1}{7} \cdot \frac{3}{7} \cdot \frac{2}{7} \cdot \frac{7}{11} = \frac{6}{539}$

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?

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

ripe

firm green-black Zutano

assumption says

$\frac{7}{11}$

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

$= \frac{1}{7} * \frac{3}{7} * \frac{2}{7}$

Avocado Ripeness

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?

Assuming conditional independence of features given the class, calculate P(firm, green-black, Zutano | unripe).

A. 0
B. 1/4
C. 3/16
D. 1 - (1/7*3/7*2/7)

Avocado Ripeness

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?

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

assumption

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

$\frac{3}{4} \times \frac{2}{4} * \frac{2}{4} = \frac{3}{16}$

Naive Bayes Algorithm

- Bayes' Theorem shows how to calculate P(class | features).

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

- Rewrite the numerator, using the naive assumption of conditional independence of features given the class.
- Estimate each term in the numerator based on the training data.
- Select class based on whichever has the larger numerator.

Summary

- The Naive Bayes algorithm gives a strategy for classifying data according to its features.
- It relies on an assumption of conditional independence of the features.
- **Next time:** application to text classification