## 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 | You have a green-black avocado. Based on this data, would |           |                        |   |  |  |  |
|--------------|----------|---|-----------|------------------------|---|--|--|--|
| bright green | unripe   | you predict that your avocado is ripe or unripe?          |           |                        |   |  |  |  |
| green-black  | ripe     |   |           |                        |   |  |  |  |
| purple-black | ripe     |   |           |                        | 1 |  |  |  |
| green-black  | unripe   | ,   | Which cla | ass would you predict? |   |  |  |  |
| purple-black | ripe     |   | _         | ripe                   |   |  |  |  |
| bright green | unripe   |   | D. (      | unripe                 |   |  |  |  |
| green-black  | ripe     |   |           |                        |   |  |  |  |
| purple-black | ripe     |   |           |                        |   |  |  |  |
| green-black  | ripe     |   |           |                        |   |  |  |  |
| green-black  | unripe   |   |           |                        |   |  |  |  |
| purple-black | ripe     |   |           |                        |   |  |  |  |

## Example

| Color        | Ripeness | You have a green-black avocado. Based on this data, would  |
|--------------|----------|--|
| bright green | unripe   | you predict that your avocado is ripe or unripe?           |
| green-black  | ripe     |  |
| purple-black | ripe     | Strategy: Calculate two probabilities:                     |
| green-black  | unripe   |  |
| purple-black | ripe     | P(ripe   green-black)                                      |
| bright green | unripe   |  |
| green-black  | ripe     | P(unripe   green-black)                                    |
| purple-black | ripe     |  |
| green-black  | ripe     | Then choose the class according to the larger of these two |
| green-black  | unripe   | probabilities.   |
| purple-black | ripe     |  |

#### 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

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

#### 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

| Color        | Ripeness | You have a green-black avocado. Based on this data, would |
|--------------|----------|---|
| bright green | unripe   | you predict that your avocado is ripe or unripe?          |
| green-black  | ripe     | P(features class) * P(class)                              |
| purple-black | ripe     | $P(\text{class} \text{features}) = \frac{1}{2}$           |
| green-black  | unripe   | P(features)   |
| purple-black | ripe     |   |
| bright green | unripe   |   |
| green-black  | ripe     |   |
| purple-black | ripe     |   |
| green-black  | ripe     |   |
| green-black  | unripe   |   |
| purple-black | ripe     |   |

| Color        | Ripeness | You have a green-black avocado. Based on this data, would |
|--------------|----------|---|
| bright green | unripe   | you predict that your avocado is ripe or unripe?          |
| green-black  | ripe     | P(features class) * P(class)                              |
| purple-black | ripe     | $P(\text{class} \text{features}) = \frac{1}{2}$           |
| green-black  | unripe   | P(features)   |
| purple-black | ripe     |   |
| bright green | unripe   |   |
| green-black  | ripe     |   |
| purple-black | ripe     |   |
| green-black  | ripe     |   |
| green-black  | unripe   |   |
| purple-black | ripe     |   |

| Color        | Ripeness | You have a green-black avocado. Based on this data, would |  |  |  |
|--------------|----------|---|--|--|--|
| bright green | unripe   | you predict that your avocado is ripe or unripe?          |  |  |  |
| green-black  | ripe     | P(features class) * P(class)                              |  |  |  |
| purple-black | ripe     | P(c ass teat  res ) =                                     |  |  |  |
| green-black  | unripe   | P(features) = P(features)                                 |  |  |  |
| purple-black | ripe     |   |  |  |  |
| bright green | unripe   | Shortcut: Both probabilities have same denominator. To    |  |  |  |
| green-black  | ripe     | find larger one, choose one with larger numerator.        |  |  |  |
| purple-black | ripe     | P(ring Largen black)                                      |  |  |  |
| green-black  | ripe     | P(ripe   green-black)                                     |  |  |  |
| green-black  | unripe   |   |  |  |  |
| purple-black | ripe     | P(unripe   green-black)                                   |  |  |  |

## More Features

| Color        | Softness | Variety | Ripeness | You have a firm green-black Zutano     |
|--------------|----------|---------|----------|--|
| bright green | firm     | Zutano  | unripe   | avocado. Based on this data, would you |
| green-black  | medium   | Hass    | ripe     | predict that your avocado is ripe or   |
| purple-black | firm     | Hass    | ripe     | unripe?                                |
| 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     |  |

| Color        | Softness | Variety | Ripeness | You have a firm green-black Zutano     |
|--------------|----------|---------|----------|--|
| bright green | firm     | Zutano  | unripe   | avocado. Based on this data, would you |
| green-black  | medium   | Hass    | ripe     | predict that your avocado is ripe or   |
| purple-black | firm     | Hass    | ripe     | unripe?                                |
| green-black  | medium   | Hass    | unripe   | Strategy: Calculate two probabilities: |
| purple-black | soft     | Hass    | ripe     |  |
| bright green | firm     | Zutano  | unripe   | P(ripe   firm, green-black, Zutano)    |
| green-black  | soft     | Zutano  | ripe     |  |
| purple-black | soft     | Hass    | ripe     | P(unripe   firm, green-black, Zutano)  |
| green-black  | soft     | Zutano  | ripe     |  |
|              |          |         |          |  |

green-black

purple-black

firm

medium

Hass

Hass

unripe

ripe

Then choose the class according to the

larger of these two probabilities.

| green-black  | medium | Hass   | ripe   | predict that your avocado is ripe or        |
|--------------|--------|--------|--------|---|
| purple-black | firm   | Hass   | ripe   | unripe?                                     |
| green-black  | medium | Hass   | unripe | Problem: We have not seen an avocado        |
| purple-black | soft   | Hass   | ripe   | with all these features. Both probabilities |
| bright green | firm   | Zutano | unripe | will be undefined.                          |

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

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

You have a firm green-black Zutano avocado. Based on this data, would you Zutano unripe

**Ripeness** 

ripe

ripe

ripe

ripe

unripe

Color

bright green

green-black

purple-black

green-black

green-black

purple-black

Softness

firm

soft

soft

soft

firm

medium

**Variety** 

Zutano

Hass

Zutano

Hass

Hass

**Ripeness** 

Color

**Softness** 

**Variety** 

You have a firm green-black Zutano

|              |        | •      | •      | Tod Have a mini groom black Zatano               |
|--------------|--------|--------|--------|--|
| bright green | firm   | Zutano | unripe | avocado. Based on this data, would you           |
| green-black  | medium | Hass   | ripe   | predict that your avocado is ripe or             |
| purple-black | firm   | Hass   | ripe   | unripe?  |
| green-black  | medium | Hass   | unripe | P(class features) = P(features class) * P(class) |
| purple-black | soft   | Hass   | ripe   | P(features)                                      |
| bright green | firm   | Zutano | unripe |  |
| green-black  | soft   | Zutano | ripe   | Solution: Use Bayes' Theorem, plus a             |
| purple-black | soft   | Hass   | ripe   | simplifying assumption, to calculate the         |
| green-black  | soft   | Zutano | ripe   | two numerators.                                  |
| green-black  | firm   | Hass   | unripe |  |
| purple-black | medium | Hass   | ripe   |  |

You have a firm green-black Zutano

class, the features are independent.

P(firm | ripe)\*P(green-black | ripe)\*P(Zutano | ripe)

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

avocado Rasad on this data would you

P(features)

| bright green | TIrm   | Zutano | unripe | avocado. Dased on triis data, would you          |
|--------------|--------|--------|--------|--|
| green-black  | medium | Hass   | ripe   | predict that your avocado is ripe or             |
| purple-black | firm   | Hass   | ripe   | unripe?  |
|              |        |        |        |  |
| green-black  | medium | Hass   | unripe | P(class footunes) = P(features class) * P(class) |

P(class|leatures) purple-black soft Hass ripe bright green firm Zutano unripe Simplifying assumption: Within a given

ripe

ripe

ripe

ripe

unripe

**Ripeness** 

Color

green-black

purple-black

green-black

green-black

purple-black

Softness

soft

soft

soft

firm

medium

**Variety** 

Zutano

Hass

Zutano

Hass

Hass

#### 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

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

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

| Color        | Softness | Variety | Ripeness | You have a firm green-black Zutano               |
|--------------|----------|---------|----------|--|
| bright green | firm     | Zutano  | unripe   | avocado. Based on this data, would you           |
| green-black  | medium   | Hass    | ripe     | predict that your avocado is ripe or             |
| purple-black | firm     | Hass    | ripe     | unripe?  |
| green-black  | medium   | Hass    | unripe   | P(class features) = P(features class) * P(class) |
| purple-black | soft     | Hass    | ripe     | P(features)                                      |
| 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     |  |

**Ripeness** 

Color

**Softness** 

**Variety** 

You have a firm green-black Zutano

| bright green | firm   | Zutano | unripe | avocado. Based on this data, would you       |
|--------------|--------|--------|--------|--|
| green-black  | medium | Hass   | ripe   | predict that your avocado is ripe or         |
| purple-black | firm   | Hass   | ripe   | unripe?                                      |
| green-black  | medium | Hass   | unripe |  |
| purple-black | soft   | Hass   | ripe   | Assuming conditional independence of         |
| bright green | firm   | Zutano | unripe | features given the class, calculate          |
| green-black  | soft   | Zutano | ripe   | P(firm, green-black, Zutano   unripe).  A. 0 |
| purple-black | soft   | Hass   | ripe   | B. 1/4                                       |
| green-black  | soft   | Zutano | ripe   | C. 3/16                                      |
| green-black  | firm   | Hass   | unripe | D. 1 - (1/7*3/7*2/7)                         |
| purple-black | medium | Hass   | ripe   |  |

| Color        | Softness | Variety | Ripeness | You have a firm green-black Zutano avocado. Based on this data, would you predict that your avocado is ripe or unripe? $P({\rm class} {\rm features}) = \frac{P({\rm features} {\rm class}) * P({\rm class})}{P({\rm features})}$ |
|--------------|----------|---------|----------|---|
| 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     |   |

### 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