# 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 avoca
bright green	unripe	you predict that your avocado
green-black	ripe	
purple-black	ripe	
green-black	unripe	Which class wou
purple-black	ripe	A. ripe
bright green	unripe	B. unripe
green-black	ripe	
purple-black	ripe	
green-black	ripe	
aroon blook	Luprino	

purple-black

ripe

do. Based on this data, would is ripe or unripe? uld you predict? green-black unripe



**Ripeness** 

unripe

ripe

Color

bright green

green-black

# Example

you predict that your avocado is ripe or unripe?

You have a green-black avocado. Based on this data, would

purple-black	ripe	Strategy: Calculate two probabilities:
green-black	unripe	14 rine green-black
purple-black	ripe	P(ripe   green-black) $\approx$ # ripe green-black  H green - black in
bright green	unripe	population parameter # green -black in cample
green-black	ripe	P(unripe   green-black)  H green -black in cample
purple-black	ripe	sample statistic
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

You have a green-black avocado. Based on this data, would

**Ripeness** 

Color

bright green	unripe	you predict that your avocado is ripe or unripe?
green-black	ripe	P(features class) * P(class)
purple-black	ripe	P(class teatures) =
green-black	unripe	P(features)
purple-black	ripe	ripe gren-black
bright green	unripe	P(gren-plack ripe) Price
green-black	ripe	Physical acoust 1 1
purple-black	ripe	P(unripe/green-black) 3
green-black	ripe	= 12 P(green-black) I
green-black	unripe	$5 - \frac{3}{4} \times \frac{3}{11}$
purple-black	ripe	5/11 [5]

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(class teatures) =						
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							

you predict that your avocado is ripe or unripe?

Color

bright green

**Ripeness** 

unripe

You have a green-black avocado. Based on this data, would

green-black	ripe	P(features class) * P(class)		
purple-black	ripe	P(class features) =		
green-black	unripe	P(features) = 1		
purple-black	ripe	Who will be		
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(ripe   green-black) Pop. to 3 + 4 (1)		
green-black 🕶	ripe	r(lipe   green-black)		
green-black	unripe	Diversing Lawrence blooks and to ZXX = 2/		
purple-black	ripe	P(unripe   green-black) $prop. to \frac{2}{4} + \frac{4}{11} = \frac{7}{11}$		

## eatures

3	Feat	wes	More	Fe
Color	Softness	Variety	Ripeness	Yo
bright green	firm	Zutano	unripe	ave
green-black	medium	Hass	ripe	pre
purple-black	firm	Hass	ripe	un
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	

ou have a firm green-black Zutano ocado. Based on this data, would you edict that your avocado is ripe or 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	firm	Hass	unripe	Then choose the class according to the
purple-black	medium	Hass	ripe	larger of these two probabilities.

**Ripeness** 

unripe

ripe

ripe

ripe

ripe

unripe

Color

bright green

green-black

purple-black

green-black

green-black

purple-black

Softness

medium

firm

soft

soft

firm

medium

**Variety** 

Zutano

Hass

Hass

Zutano

Hass

Hass

You have a firm green-black Zutano

predict that your avocado is ripe or

avocado. Based on this data, would you

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

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

purple-black	firm	Hass	ripe	unipe?
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.
green-black	soft	Zutano	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	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	

P(class|features) =

P(features|class) \* P(class)

P(features)

Simplifying assumption: Within a given

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

class, the features are independent.

P(firm, green-black, Zutano | 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

unripe? purple-black firm Hass ripe

unripe

unripe

ripe

ripe

ripe

ripe

ripe

unripe

Hass

Hass

Zutano

Zutano

Hass

Zutano

Hass

Hass

green-black

purple-black

bright green

green-black

purple-black

green-black

green-black

purple-black

medium

soft

firm

soft

soft

soft

firm

medium

### 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.

Δ	vocad	o Ripeness $\frac{1}{2}$
	Ripeness	You have a firm green-black Zutano
	unripe	avocado. Based on this data, would
	rine	predict that your avocado is ripe or

Color **Softness Variety** Zutano bright green firm Hass medium ripe unripe? Hass ripe firm <sub>4</sub> medium Hass unripe

Hass

Hass

green-black

purple-black

firm

medium

on this data, would you vocado is ripe or green-black? purple-black green-black P(class|features) =soft purple-black Hass ripe bright green firm Zutano unripe green-black soft Zutano • ripe purple-black soft Hass ripe Zutano\* green-black \* soft ripe

unripe

ripe

P(features|class) \* P(class)P(features)

**Ripeness** 

Color

**Softness** 

**Variety** 

You have a firm green-black Zutano

avocado. Based on this data, would vou

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	

You have a firm green-black Zutano **Variety Ripeness** avocado. Based on this data, would you unripe Zutano V predict that your avocado is ripe or ripe unripe? ripe P(features|class) \* P(class)unripe P(class|features) =P(features)ripe P(firm/unripe) x p(green unripe)

P(Tutano)

y p(Zutano) unripe ripe ripe

ripe unripe ripe

firm Hass medium Hass soft Hass firm -Zutano soft Zutano soft Hass Zutano soft Hass firm medium Hass

Hass

Softness

medium

firm

Color

bright green

green-black

purple-black

green-black 、

purple-black

bright green

green-black

purple-black

green-black

green-black\*

purple-black

## 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