DSC 40A

Theoretical Foundations of Data Science I

Announcements

- Homework 7 was released and due 12/3 – no slip days.
- Gal's OH today at 4:15 instead of Wednesday.

Question

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!
If the direct link doesn't work, click the "Lecture Questions" link in the top right corner of dsc40a.com.

Last Week

- We defined Bayes' Theorem:

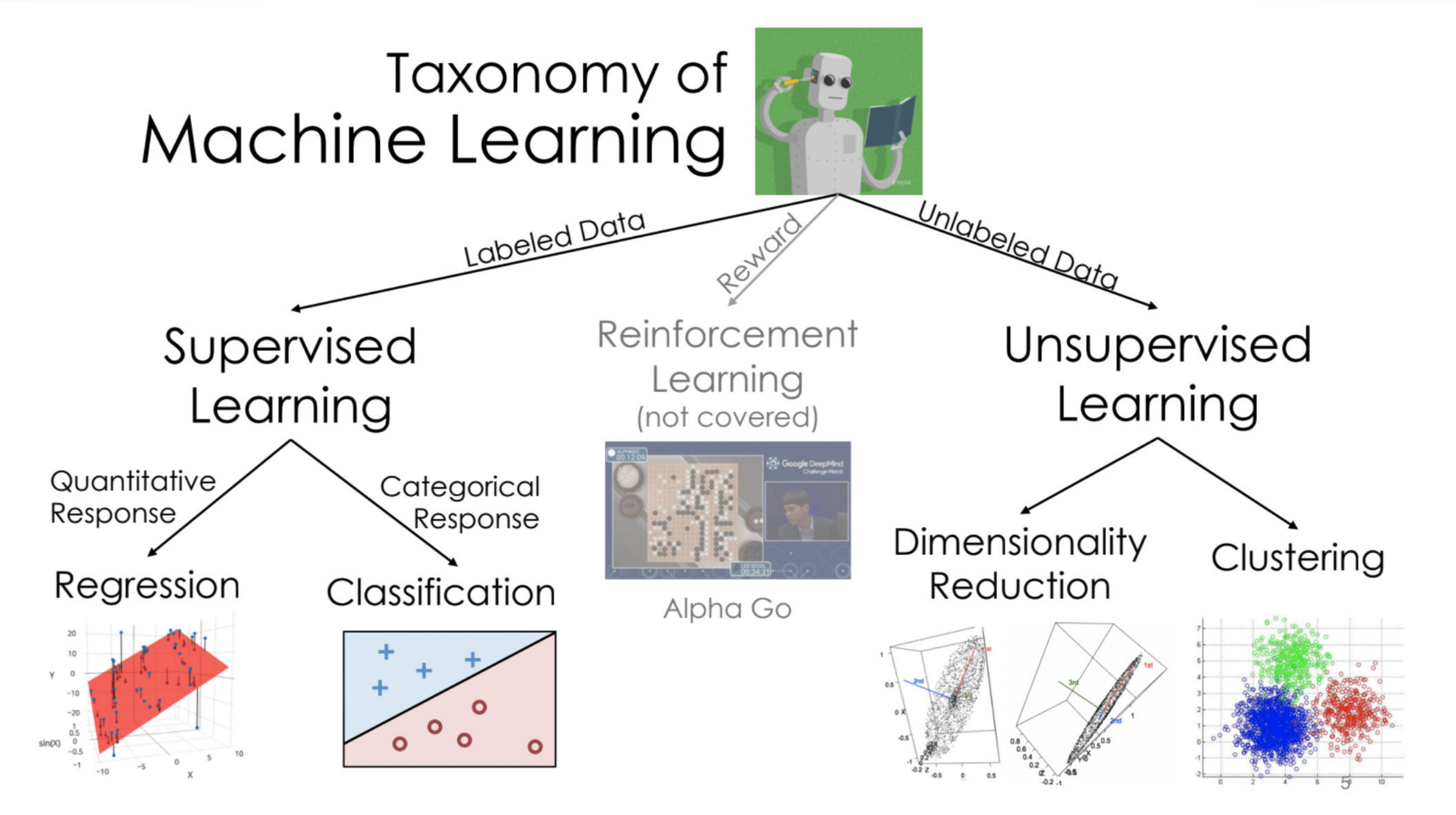
$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

- Bayes' Theorem describes how to update the probability of one event given that another has occurred.

Naïve Bayes Classifier

Today

- Using Bayes' Theorem to solve the classification problem



Preview: Bayes' Theorem for Classification

Bayes’ Theorem is very useful for classification problems, where we want to predict a class based on some features.

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

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

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

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

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)

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.

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

ain class

ures

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

Bayes' Theorem for Classification

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

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

ain class

ures

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

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?

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

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?

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

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

P(ripe | green-black)

P(unripe | green-black)

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?

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

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

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

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

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?

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

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?

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

Naïve 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 naïve 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 Naïve 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

DSC 40A

Theoretical Foundations of Data Science I

Question

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!
If the direct link doesn't work, click the "Lecture Questions" link in the top right corner of dsc40a.com.

Agenda

- How can we use naïve Bayes to classify text?

Last Time

- Classification is the problem of predicting a categorical variable.
- The Naïve Bayes algorithm gives a strategy for classifying data according to its features.
- It is naïve because it relies on an assumption of conditional independence of the features, for a given class.

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 naïve 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.

Naïve Bayes Classifier for Text Classification

Bayes' Theorem for Text Classification

Text classification problems include:

- sentiment analysis
  - positive and negative customer reviews
- determining genre
  - news articles, blog posts, etc.
- email foldering
  - promotions tab in Gmail
- **spam filtering**
  - **separating spam from ham (good, non-spam email)**

Features

Represent an email as a vector or array of features

$$(x_1, x_2, x_3, \ldots , x_n)$$

where *i* is an index into a dictionary of *n* possible words, and

$x_i = 1$ if word *i* is present in the email
$x_i = 0$ otherwise

Features

Called the "bag of words" model:

Ignores location of words within the email, and the frequency of words

Example:

usually n = 10,000 to 50,000 words in practice

Naïve Bayes Spam Classifier

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

To classify an email, we will use Bayes' Theorem to calculate the probability of it belonging to each class:

$P(\text{spam} \mid \text{features})$ and $P(\text{ham} \mid \text{features})$

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

Naïve Bayes Spam Classifier

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

Observe: the formulas for P(spam | features) and P(ham | features) have the same denominator, P(features).

We can find the larger of the two probabilities by just comparing numerators.

P(features | spam)**P*(spam) vs. *P*(features | ham)**P*(ham)

Naive Bayes Spam Classifier

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

To use Bayes' Theorem, need to determine four quantities:

i. P(features | spam)
ii. P(features | ham)
iii. P(spam)
iv. P(ham)

Which of these probabilities should add to 1?

A. i, ii
B. iii, iv
C. both A and B
D. neither A nor B

Estimating Parameters with Training Data

parameter: $P(\text{spam})$ **estimate:** $\dfrac{\text{spam emails in training}}{\text{size of training set}}$

parameter: $P(\text{ham})$ **estimate:** $\dfrac{\text{\# ham emails in training set}}{\text{size of training set}}$

$P(\text{features} \mid \text{spam})$
$P(\text{features} \mid \text{ham})$

← harder to estimate

Assumption of Conditional Independence

To estimate P(features | spam) and P(features | ham), we assume that the probability of a word appearing in an email of a given class is not affected by other words in the email.

$P(x_1=0$ and $x_2=1$ and $x_3=1$ and ... | spam) **=** ← **assumed equal**

$P(x_1=0$ | spam)*$P(x_2=1$ | spam)*$P(x_3=1$ | spam)*...

Is this a reasonable assumption?

Estimating Parameters with Training Data

$P(x_1=0 \text{ and } x_2=1 \text{ and } x_3=1 \text{ and } \ldots \mid \text{spam})$ **=** ⟵ **assumed equal**

$$P(x_1=0 \mid \text{spam}) * P(x_2=1 \mid \text{spam}) * P(x_3=1 \mid \text{spam}) * \ldots$$

parameter: $P(x_1=0 \mid \text{spam})$

estimate:

$$\frac{\text{\# spam emails in training set not containing the first word in the dictionary}}{\text{\# spam emails in training set}}$$

Estimating Parameters with Training Data

$P(x_1=0 \text{ and } x_2=1 \text{ and } x_3=1 \text{ and } \ldots \mid \text{spam})$ **=** ← **assumed equal**

$P(x_1=0 \mid \text{spam}) * P(x_2=1 \mid \text{spam}) * P(x_3=1 \mid \text{spam}) * \ldots$

parameter: $P(x_1=0 \mid \text{spam})$

estimate: $\dfrac{\text{\# spam emails in training set not containing the first word in the dictionary}}{\text{\# spam emails in training set}}$

Estimating Parameters with Training Data

$P(x_1=0$ and $x_2=1$ and $x_3=1$ and ... | spam) **=** ← **assumed equal**

$P(x_1=0 \mid \text{spam}) * P(x_2=1 \mid \text{spam}) * P(x_3=1 \mid \text{spam}) * \ldots$

parameter: $P(x_1=0 \mid \text{spam})$

estimate:

$$\frac{\text{\# spam emails in training set not containing the first word in the dictionary}}{\text{\# spam emails in training set}}$$

Can term-by-term estimate P(features|class)

Naïve Bayes Spam Classifier: Recap

Bayes' Theorem shows how to calculate P(spam | features) and P(ham | 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 words given the class.

Estimate each term in the numerator based on the training data.

Select class based on whichever has the larger numerator.

Modifications and Extensions

- features are pairs (or longer sequences) of words rather than individual words
  - better captures dependencies between words
  - less naïve
  - much bigger feature space
    - n words → n^2 pairs of words

- features are the number of occurrences of each word
  - captures low-frequency vs. high-frequency words

- smoothing
  - better handling of previously unseen words

Smoothing

$P(x_1=0$ and $x_2=1$ and $x_3=1$ and ... | spam) **=**
$P(x_1=0$ | spam)*$P(x_2=1$ | spam)*$P(x_3=1$ | spam)*...

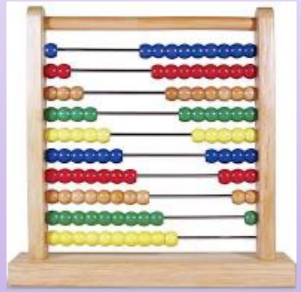
Dictionary

1. a
2. aardvark
3. abacus
4. abandon
5. abate

...

Suppose you are classifying an email containing the word "abacus," which does not appear in any emails in your training data. For this new email's features, what is ***P*(features | spam)** according to a naive Bayes classifier?

A. P(features | spam) = undefined
B. P(features | spam) = 0
C. P(features | spam) = 1/n
D. P(features | spam) = 1



Smoothing

$P(x_1\!=\!0 \text{ and } x_2\!=\!1 \text{ and } x_3\!=\!1 \text{ and ... } |\text{ ham}) =$

$P(x_1\!=\!0 \mid \text{ham})*P(x_2\!=\!1 \mid \text{ham})*P(x_3\!=\!1 \mid \text{ham})*\ldots$

Dictionary

1. a
2. aardvark
3. abacus
4. abandon
5. abate

...

Suppose you are classifying an email containing the word "abacus," which does not appear in any emails in your training data. For this new email's features, what is ***P*(features | ham)** according to a naïve Bayes classifier?

A. P(features | ham) = undefined
B. P(features | ham) = 0
C. P(features | ham) = 1/n
D. P(features | ham) = 1

Smoothing

P(features | spam) = 0 and P(features | ham) = 0

Tiebreaker: randomly select one of the classes?

Smoothing

P(features | spam) = 0 and P(features | ham) = 0

Tiebreaker: randomly select one of the classes?

Better solution: make sure probabilities can't be zero

Key idea: just because you've never seen something happen doesn't mean it's impossible

Estimating Parameters with Training Data

parameter:		estimate:	Without Smoothing	With Smoothing
parameter:	P(spam)	estimate:	$\frac{\#spam}{\#spam + \#ham}$	$\frac{\#spam + 1}{\#spam + 1 + \#ham + 1}$
parameter:	P(ham)	estimate:	$\frac{\#ham}{\#spam + \#ham}$	$\frac{\#ham + 1}{\#spam + 1 + \#ham + 1}$

Estimating Parameters with Training Data

parameter:

$P(x_i{=}1 \mid \text{spam})$

estimate:

$$\frac{\#\text{spam containing word } i}{\#\text{spam containing word } i + \#\text{spam not containing word } i}$$

Without Smoothing

$$\frac{(\#\text{spam containing word } i) + 1}{(\#\text{spam containing word } i) + 1 + (\#\text{spam not containing word } i) + 1}$$

With Smoothing

Similarly for other parameters $P(x_i{=}0 \mid \text{spam})$, $P(x_i{=}1 \mid \text{ham})$, $P(x_i{=}1 \mid \text{ham})$.

Smoothing

$P(x_1{=}0 \text{ and } x_2{=}1 \text{ and } x_3{=}1 \text{ and } \ldots \mid \text{ham}) =$
$P(x_1{=}0 \mid \text{ham}) * P(x_2{=}1 \mid \text{ham}) * P(x_3{=}1 \mid \text{ham}) * \ldots$

<u>Dictionary</u>

1. a
2. aardvark
3. abacus
4. abandon
5. abate

...

Suppose you are classifying an email containing the word "abacus," which does not appear in any emails in your training data. What is $\boldsymbol{P(x_3{=}\ 1 \mid \text{ham})}$ according to a naïve Bayes classifier ***with smoothing***?

A. $P(x_3{=}\ 1 \mid \text{ham}) = 0$
B. $P(x_3{=}\ 1 \mid \text{ham}) = 1/2$
C. $P(x_3{=}\ 1 \mid \text{ham}) = 1/(\text{total \#ham} + 1)$
D. $P(x_3{=}\ 1 \mid \text{ham}) = 1/(\text{total \#ham} + 2)$
E. $P(x_3{=}\ 1 \mid \text{ham}) = 1/(\text{total \#ham} + \text{total \#spam} + 2)$

Smoothing

Suppose your training set includes only six emails, all of which are ham. For a new email, how will we estimate *P*(ham)

	without smoothing?	**with smoothing?**
A.	P(ham) = 0	P(ham) = 1/8
B.	P(ham) = 0	P(ham) = 6/7
C.	P(ham) = 1	P(ham) = 1/8
D.	P(ham) = 1	P(ham) = 6/7
E.	P(ham) = 1	P(ham) = 7/8

Smoothing

Suppose your training set includes only six emails, all of which are ham. For a new email, how will we estimate *P*(ham)

	without smoothing?	**with smoothing?**
A.	P(ham) = 0	P(ham) = 1/8
B.	P(ham) = 0	P(ham) = 6/7
C.	P(ham) = 1	P(ham) = 1/8
D.	P(ham) = 1	P(ham) = 6/7
E.	P(ham) = 1	P(ham) = 7/8

Is it still true that *P*(ham) + *P*(spam) = 1 with smoothing?

Summary

- The Naive Bayes algorithm is useful for text classification.
- The bag of words model treats each word in a large dictionary as a feature.
- Smoothing is one modification that allows for better predictions when there are words that have never been seen before.