CS 383: Machine Learning

Prof Adam Poliak

Fall 2024

10/01/2024

Lecture 11

Announcements

HW03 is due Tuesday night

- **Reading quiz: Thursday**
  - Duame 9.3 (2 pages)

Updated schedule

No lecture on Thursday

No lecture Wednesday 10/09

Lecture tomorrow Wed 10/02

HW03 polynomial regression due tonight

HW04 naive Bayes due Tuesday 10/08 (it'll be a shorter assignment)

Midterm 1 on Thursday after fall break

Reading Quiz #4

1. How would you say $P(A, B)$ in words?

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) =$$

3. What is the Naive Bayes assumption? (circle one)
   (a) The label and the features are independent given the model
   (b) Features are independent given the label
   (c) Training examples are independent given their labels

4. If I want to predict the label (y) of an example based on its features ($\vec{x}$), which of the following expressions would I want to compute? (circle the best one)
   (a) $p(\vec{x}, y)$
   (b) $p(\vec{x} \mid y)$
   (c) $p(y \mid \vec{x})$

Reading Quiz #4

1. How would you say $P(A, B)$ in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) =$$

3. What is the Naive Bayes assumption? (circle one)
   (a) The label and the features are independent given the model
   (b) Features are independent given the label
   (c) Training examples are independent given their labels

4. If I want to predict the label (y) of an example based on its features ($\vec{x}$), which of the following expressions would I want to compute? (circle the best one)
   (a) $p(\vec{x}, y)$
   (b) $p(\vec{x} \mid y)$
   (c) $p(y \mid \vec{x})$

Reading Quiz #4

1. How would you say $P(A, B)$ in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$P(A, B) =$ **P(A) P(B|A)** **or** **P(B) P(A|B)**

3. What is the Naive Bayes assumption? (circle one)
   (a) The label and the features are independent given the model
   (b) Features are independent given the label
   (c) Training examples are independent given their labels

4. If I want to predict the label (y) of an example based on its features $(\vec{x})$, which of the following expressions would I want to compute? (circle the best one)
   (a) $p(\vec{x}, y)$
   (b) $p(\vec{x} \mid y)$
   (c) $p(y \mid \vec{x})$

Reading Quiz #4

1. How would you say $P(A, B)$ in words?

   Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

   $P(A, B) =$ **P(A) P(B|A)** **or** **P(B) P(A|B)**

3. What is the Naive Bayes assumption? (circle one)
   (a) The label and the features are independent given the model
   (b) $\boxed{\text{Features are independent given the label}}$
   (c) Training examples are independent given their labels

4. If I want to predict the label (y) of an example based on its features ($\vec{x}$), which of the following expressions would I want to compute? (circle the best one)
   (a) $p(\vec{x}, y)$
   (b) $p(\vec{x} \mid y)$
   (c) $p(y \mid \vec{x})$

Reading Quiz #4

1. How would you say $P(A, B)$ in words?

Probability of A and B

2. Based on class on Tuesday, what is Bayes rule?

$$P(A, B) = \quad \mathbf{P(A)\,P(B|A)} \quad \text{or} \quad \mathbf{P(B)\,P(A|B)}$$

3. What is the Naive Bayes assumption? (circle one)

   (a) The label and the features are independent given the model
   (b) Features are independent given the label
   (c) Training examples are independent given their labels

4. If I want to predict the label (y) of an example based on its features ($\vec{x}$), which of the following expressions would I want to compute? (circle the best one)

   (a) $p(\vec{x}, y)$
   (b) $p(\vec{x} \mid y)$
   (c) $p(y \mid \vec{x})$

Outline

Naive Bayes

Confusion Matrix

Components of a Bayes Model

Identify the evidence, prior, posterior, and likelihood in the equation below

$$\boxed{p(y=k|\boldsymbol{x}) = \frac{p(y=k)p(\boldsymbol{x}|y=k)}{p(\boldsymbol{x})}}$$

Components of a Bayes Model

Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Evidence: this is the data (features) we observe, which we think will help us predict the outcome we're interested in

Components of a Bayes Model

Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Prior: without seeing any evidence (data), what is our prior believe about each outcome (intuition: what is the outcome in the population as a whole?)

Components of a Bayes Model

Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Posterior: this is the quantity we are actually interested in. **Given* the evidence, what is the probability of the outcome?

Components of a Bayes Model

Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Likelihood: given an outcome, what is the probability of observing this set of features?

Naive Bayes Example

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

Test: powerful very fun

$p(+) = ?$

$p(-) = ?$

Naive Bayes Example

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

Test: powerful very fun

$$p(+) = \frac{p(pwerful\ |+)p(very|+)p(fun|+)p(+)}{p(powerful)p(very)p(fun)}$$

Naive Bayes Example

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

Test: predictable with no fun

$$p(+) = ? \qquad\qquad p(-) = ?$$

Naive Bayes Example

Test:
predictable with no fun

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

$$p(+) = \frac{p(predictable\,|+)p(with|+)p(no|+)p(fun|+)p(+)}{p(predictable)p(with)p(no)p(fun)}$$

$$p(-) = \frac{p(predictable\,|-)p(with|-)p(no|-)p(fun|-)p(-)}{p(predictable)p(with)p(no)p(fun)}$$

Laplacian Smoothing

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w\in V}(count(w,c)+1)} = \frac{count(w_i,c)+1}{\left(\sum_{w\in V} count(w,c)\right)+|V|}$$

Naive Bayes Example

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer

Test: predictable with no fun

$p(+) = ?$ $\qquad$ $p(-) = ?$

Outline

Naive Bayes

Confusion Matrix

Logistic Regression

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

2) What issues arise with making y real-valued?

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:

- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

Assumes some *ordering* of the outcomes that is probably not there!

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

Linear regression for classification

Case Study: you need to identify the medical condition of a patient in the emergency room on the basis of their symptoms.

Possible conditions (*y*) are:
- Stroke
- Drug overdose
- Epileptic seizure

1) If you were forced to use linear regression for this problem, how could you encode *y* to make it real-valued?

You could choose stroke=0, drug overdose=1, epileptic seizure=2 (or some permutation)

2) What issues arise with making y real-valued?

Assumes some *ordering* of the outcomes that is probably not there!

3) What if you just had two outcomes (i.e. stroke and drug overdose) -- why is linear regression still not a good choice?

The range of a linear function (i.e. *y* values) is $[-\infty, \infty]$, but we want $[0, 1]$