Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 8

Slides adapted from Chenhao Tan, Jordan Boyd-Graber, Chris Ketelsen

Logistics

- HW1 deadline is today
- HW2 will be available on Github today

Learning objectives

- Introduce logistic regression
- Introduce Naïve Bayes
- (Bonus) Understand generative models vs. discriminative models

Outline

Probabilistic classification

Logistic regression

Naïve Bayes

Outline

Probabilistic classification

Logistic regression

Naïve Bayes

Recap

Perceptron

- Learn weights $\boldsymbol{w}$ and b via the perceptron algorithm
- Predict $\hat{y}$ via $\hat{y} = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b)$
- $\hat{y} = \{-1, +1\}$

Recap

Do we want a prediction

- “human” or
- “It’s human, but human for 70%, and zombie for 30%”

Outline

Probabilistic classification

Logistic regression

Naïve Bayes

What are we talking about?

- Probabilistic classification: $P(Y|X)$
- Classification uses: ad placement, spam detection
- Building block of other machine learning methods

Logistic Regression: Definition

- Weight vector β_i
- Feature x_i
- "Bias" β_0

$$P(Y = 0|X) = \frac{1}{1 + \exp\left[\beta_0 + \sum_i \beta_i x_i\right]} \tag{1}$$

$$P(Y = 1|X) = 1 - P(Y = 0|X) \tag{2}$$

Logistic Regression: Definition

- Weight β_i
- Feature x_i
- For shorthand, we'll say that

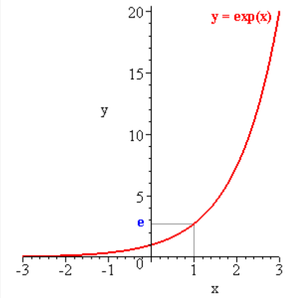
$$P(Y = 1|X) = \sigma(\beta_0 + \sum_i \beta_i x_i) \tag{3}$$

$$P(Y = 0|X) = 1 - \sigma(\beta_0 + \sum_i \beta_i x_i) \tag{4}$$

- Where $\sigma(z) = \frac{1}{1+exp[-z]}$

What's this "exp" doing?

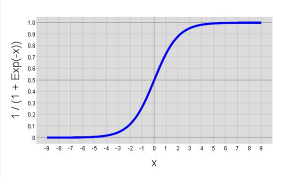
Exponential function



- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - It's the function whose derivative is itself

What's this "exp" doing?

Logistic function



- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - It's the function whose derivative is itself
- The "logistic" function is $\sigma(z) = \frac{1}{1+e^{-z}}$
- Always between 0 and 1.
  - Allows us to model probabilities
- it's "smooth"

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- What does $Y = 1$ mean?

Example 1: Empty Document?

$X = \{\}$

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- $Y = 1$: spam

Example 1: Empty Document?

$X = \{\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1]} =$
- $P(Y=1) = \frac{\exp[0.1]}{1+\exp[0.1]} =$

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- $Y = 1$: spam

Example 1: Empty Document?

$X = \{\}$

- $P(Y = 0) = \frac{1}{1+\exp[0.1]} = 0.48$
- $P(Y = 1) = \frac{\exp[0.1]}{1+\exp[0.1]} = 0.52$
- Bias β_0 encodes the prior probability of a class

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- $Y = 1$: spam

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- $Y = 1$: spam

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0+3.0]} =$
- $P(Y=1) = \frac{\exp[0.1-1.0+3.0]}{1+\exp[0.1-1.0+3.0]} =$
- Include bias, and sum the other weights

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- $Y = 1$: spam

Example 2

$X = \{\text{Mother}, \text{Nigeria}\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0+3.0]} = 0.11$
- $P(Y=1) = \frac{\exp[0.1-1.0+3.0]}{1+\exp[0.1-1.0+3.0]} = 0.89$
- Include bias, and sum the other weights

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
“viagra”	β_1	2.0
“mother”	β_2	-1.0
“work”	β_3	-0.5
“nigeria”	β_4	3.0

- $Y = 1$: spam

Example 3

$X = \{$Mother, Work, Viagra, Mother$\}$

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
“viagra”	β_1	2.0
“mother”	β_2	-1.0
“work”	β_3	-0.5
“nigeria”	β_4	3.0

- $Y = 1$: spam

Example 3

$X = \{$Mother, Work, Viagra, Mother$\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0-0.5+2.0-1.0]} =$
- $P(Y=1) = \frac{\exp[0.1-1.0-0.5+2.0-1.0]}{1+\exp[0.1-1.0-0.5+2.0-1.0]} =$
- Multiply feature presence by weight

Logistic Regression Example: Spam Classification

feature	coefficient	weight
bias	β_0	0.1
"viagra"	β_1	2.0
"mother"	β_2	-1.0
"work"	β_3	-0.5
"nigeria"	β_4	3.0

- $Y = 1$: spam

Example 3

$X = \{$Mother, Work, Viagra, Mother$\}$

- $P(Y=0) = \frac{1}{1+\exp[0.1-1.0-0.5+2.0-1.0]} = 0.60$
- $P(Y=1) = \frac{\exp[0.1-1.0-0.5+2.0-1.0]}{1+\exp[0.1-1.0-0.5+2.0-1.0]} = 0.40$
- Multiply feature presence by weight

How is Logistic Regression Used?

- Given a set of weights $\vec{\beta}$, we know how to compute the conditional likelihood $P(y|\vec{\beta}, x)$
- Find the set of weights $\vec{\beta}$ that maximize the conditional likelihood on training data (next lecture)
- **Intuition**: higher weights mean that this feature implies that this feature is a good feature for the positive class

Outline

Probabilistic classification

Logistic regression

Naïve Bayes

Spam Classification

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

Goal: Estimate $P(Y|X)$

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(Y)$: Prior, the probability that any email belongs to Y

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(Y)$: Prior, the probability that any email belongs to Y

HAM	SPAM	SPAM	SPAM	HAM
work buy money	nigeria opportunity viagra	fly buy nigeria	money buy fly	fly home nigeria

- $P(Y = SPAM) = \frac{3}{5}$
- $P(Y = HAM) = \frac{2}{5}$
- The fraction of spams in the training data

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X)$: Evidence, the probability that we encounter X independent of Y

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X)$: Evidence, the probability that we encounter X independent of Y

When classifying SPAM vs. HAM, then

$$\frac{P(X|Y = \text{SPAM})P(Y = \text{SPAM})}{P(X)} \text{ vs } \frac{P(X|Y = \text{HAM})P(Y = \text{HAM})}{P(X)}$$

The denominator does not affect the decision of the estimated class

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X|Y)$: Class likelihood, given class Y, the probability that X is observed
- Given assumptions about the nature of SPAM/HAM emails, the probability that we observe this particular email

Bayesian Classifiers

What is different from logistic regression?
We model $P(Y|X)$ using **Bayes Rule** i.e.,

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(X|Y)$: Class likelihood, given class Y, the probability that X is observed
- Given assumptions about the nature of SPAM/HAM emails, the probability that we observe this particular email

How do we estimate $P(X|Y)$?

The Naïve Bayes Assumption

We make the following assumption on $P(X|Y)$:

$$P(X|Y) = \prod_{j=1}^{N} P(x_j|Y)$$

i.e., features X are **independent** given class Y

- x_j: each word in a document
- N: Number of words in a document

The Naïve Bayes Assumption

We make the following assumption on $P(X|Y)$:

$$P(X|Y) = \prod_{j=1}^{N} P(x_j|Y)$$

i.e., features X are **independent** given class Y

- x_j: each word in a document
- N: Number of words in a document

In reality, this is not true e.g.,
$P(X = \{\text{peanut, butter}\}|SPAM)$

Do the words “peanut” and “butter” occur independent to each other? i.e.,
$P(X = \{\text{peanut, butter}\}|SPAM) = P(X = \{\text{peanut}\}|SPAM)P(X = \{\text{butter}\}|SPAM)$

Classifying Unseen Examples using Naïve Bayes Classifier

Training Data:

HAM	SPAM	SPAM	SPAM	HAM
work	nigeria	fly	money	fly
buy	opportunity	buy	buy	home
money	viagra	nigeria	fly	nigeria

Unseen Example: $X =$ work, nigeria

$$
\begin{aligned}
P(Y = HAM|X) &\propto P(X|Y = HAM)P(Y = HAM) \\
&= P(\text{work}|Y = HAM)P(\text{nigeria}|Y = HAM) \cdot \frac{2}{5} \\
&= \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{2}{5}
\end{aligned}
$$

Naive Bayes Classifier: More examples

What about this case:

- want to identify the type of fruit given a set of features: color, shape and size
- color: red, green, yellow or orange (discrete)
- shape: round, oval or long+skinny (discrete)
- size: diameter in inches (continuous)

Naive Bayes Classifier: More examples

Conditioned on type of fruit, these features are not necessarily independent:

Given category “apple,” the color “green” has a higher probability given “size < 2”:

$$P(green \mid size < 2,\ apple) > P(green \mid apple)$$

Generative vs. Discriminative Models

Discriminative

Model only conditional probability $p(Y|X)$, excluding the data X.

Logistic regression

- Logistic: A special mathematical function it uses
- Regression: Combines a weight vector with observations to create an answer

Generative

Model joint probability $p(X, Y)$ including the data X.

Naïve Bayes

- Uses Bayes rule to reverse conditioning $p(X|Y) \to p(Y|X)$
- Naïve because it ignores joint probabilities within the data distribution

Contrasting Naïve Bayes and Logistic Regression

- Naïve Bayes is easier for learning
- Naïve Bayes works better on smaller datasets
- Logistic regression works better on medium-sized datasets
- On huge datasets, both algorithms perform about the same (data always win)
- The Naïve Bayes assumption

- Next Monday, we will cover it more in depth