University of Colorado **Boulder**

Department of Computer Science

CSCI 5622: Machine Learning

Chenhao Tan

Lecture 7: Logistic regression

Slides adapted from Chris Ketelsen, Jordan Boyd-Graber, and Noah Smith

Administrivia

- HW2 is out
- Grading issues for HW1

Learning Objectives

- Understand probabilistic classification
- Understand logistic regression
- Understand generative models vs. discriminative models

Naive Bayes

Outline

- Probabilistic classification
- Logistic regression
- Generative vs. Discriminative models

Outline

- Probabilistic classification
- Logistic regression
- Generative vs. Discriminative models

Recap

K-nearest neighbor

- Find $\mathcal{N}_K(\boldsymbol{x})$: the set of K training examples nearest to $\boldsymbol{x}$
- Predict $\hat{y}$ to be majority label in $\mathcal{N}_K(\boldsymbol{x})$
- Admits a probabilistic interpretation of class given data: $\underline{p(y = c \mid \boldsymbol{x})}$

Recap

K-nearest neighbor

- Find $\mathcal{N}_K(\boldsymbol{x})$: the set of K training examples nearest to $\boldsymbol{x}$
- Predict $\hat{y}$ to be majority label in $\mathcal{N}_K(\boldsymbol{x})$
- Admits a probabilistic interpretation of class given data: $p(y = c \mid \boldsymbol{x})$

Perceptron

- Learn weights $\boldsymbol{w}$ and b via the perceptron algorithm
- Predict $\hat{y}$ via $\hat{y} = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x} + b)$
- Has no probabilistic interpretation

Probabilistic Models

- hypothesis function $h : X \to Y$.

Probabilistic Models

- hypothesis function $h : X \to Y$.
  In this special case, we define h based on estimating a probabilistic model $\underline{P(X, Y)}$. $P(Y|X)$

Probabilistic Classification

Input: $S_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ training examples

$$y_i \in \{c_1, c_2, \ldots, c_J\}$$

Goal: $h : X \to Y$

P(y|x)

For each class c_j, estimate

$$P(y = c_j \mid \boldsymbol{x}, S_{\text{train}})$$

Assign to $\boldsymbol{x}$ the class with the highest probability

$$\hat{y} = h(\boldsymbol{x}) = \arg\max_{c} P(y = c \mid \boldsymbol{x}, S_{\text{train}})$$

Perceptron

$\hat{y} = \text{sign}(w \cdot x + b)$

Outline

- Probabilistic classification
- Logistic regression
- Generative vs. Discriminative models

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

$P(y|x)$

- Weight vector β_i
- Observations X_i
- "Bias" β_0 (like intercept in linear regression)

$P(y=0|x) + P(y=1|x) = 1$

$\sum_{j=1}^{c} P(y=c|x) = 1$

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

$$P(Y=1|X) = \frac{\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]} \quad (2)$$

Logistic Regression: Definition

- Weight vector β_i
- Observations X_i
- "Bias" β_0 (like intercept in linear regression)

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

$$P(Y=1|X) = \frac{\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]} \tag{2}$$

$$\beta_0 + \sum_i \beta_i X_i = \log \frac{P(Y=1|X)}{P(Y=0|X)}$$

Logistic Regression: Definition

- Weight vector β_i
- Observations X_i
- "Bias" β_0 (like intercept in linear regression)

$$\hat{y} = \arg\max_{c} P(Y=c|x)$$

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

$$P(Y=1|X) = \frac{\exp\left[\beta_0 + \sum_i \beta_i X_i\right]}{1+\exp\left[\beta_0 + \sum_i \beta_i X_i\right]} = \frac{1}{1+\exp(-\beta_0-\sum_i \beta_i X_i)} \tag{2}$$

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

$$\boxed{\beta_0 + \sum_i \beta_i X_i} = \log \frac{P(Y=1|X)}{P(Y=0|X)} \quad 0$$

What is the decision boundary?

$$\beta_0 + \sum_i \beta_i X_i = 0$$

$$y = \begin{cases} 1 & \beta_0 + \sum_i \beta_i x_i > 0 \\ 0 & \beta_0 + \sum_i \beta_i x_i < 0 \end{cases}$$

Logistic Regression: Definition

- Weight vector β_i
- Observations X_i
- For shorthand, we'll say that

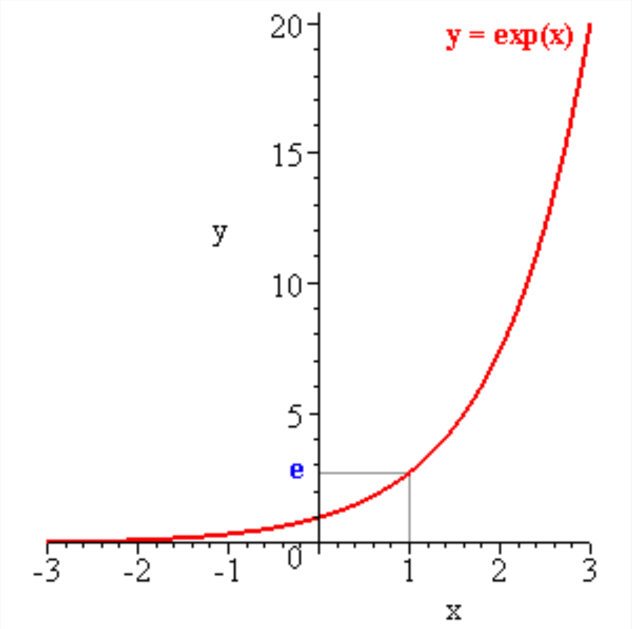
$$P(Y=1|X) = \sigma((\beta_0 + \sum_i \beta_i X_i)) = \frac{1}{1+exp(-z)} \quad (3)$$

$$P(Y=0|X) = 1 - \sigma((\beta_0 + \sum_i \beta_i X_i)) = \sigma(-\beta_0 - \sum_i \beta_i X_i) \quad (4)$$

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

What's this "exp" doing?

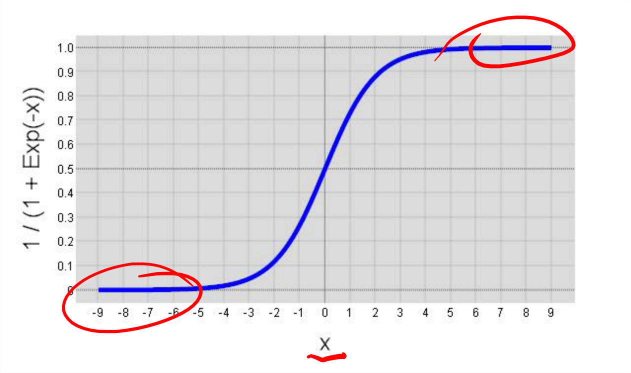
Exponential function



- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - e^x is the limit of compound interest formula as compounds become infinitely small
  - It's the function whose derivative is itself
- The "logistic" function is $\sigma(z) = \frac{1}{1+e^{-z}}$
- Looks like an "S"
- Always between 0 and 1.

What's this "exp" doing?

Logistic function



- $\exp[x]$ is shorthand for e^x
- e is a special number, about 2.71828
  - e^x is the limit of compound interest formula as compounds become infinitely small
  - It's the function whose derivative is itself
- The "logistic" function is $\sigma(z) = \frac{1}{1+e^{-z}}$
- Looks like an "S"
- Always between 0 and 1.
  - Allows us to model probabilities
  - Different from **linear** regression

Logistic Regression Example

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?

Spam

Example 1: Empty Document?

$X = \{\}$

$$P(Y=1 \mid \{\}) = \frac{1}{1+\exp(-0.1)} > 0.5$$

Logistic Regression Example

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?

$X = \{\}$

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

Logistic Regression Example

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

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

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

$\underline{-1} + \underline{3} + \underline{0.1} = \underline{2.1}$

$P(Y=1|X) = \frac{1}{1+\exp(-2.1)} > 0.5$

Logistic Regression Example

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

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

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

Logistic Regression Example

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

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$\}$

$-1 + -0.5 + 2 - 1 \quad -1 + 0.1 = -0.4$

Logistic Regression Example

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$\}$

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

Logistic Regression Example

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

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

- $P(Y=0) = \frac{1}{1+\exp\left[0.1-1.0-0.5+2.0-1.0\right]} = 0.60$
- $P(Y=1) = \frac{\exp\left[0.1-1.0-0.5+2.0-1.0\right]}{1+\exp\left[0.1-1.0-0.5+2.0-1.0\right]} = 0.40$
- 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 $P(y|\beta, x)$
- Find the set of weights $\vec{\beta}$ that maximize the conditional likelihood on training data (next week)
- **Intuition**: higher weights mean that this feature implies that this feature is a good feature for the positive class

Outline

- Probabilistic classification
- Logistic regression
- Generative vs. Discriminative models

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
- General cookbook for building conditional probability distributions

Generative

Model joint probability $p(x, y)$ including the data x.

$\sum_y P(x,y) = P(x)$

Naïve Bayes

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

The Naïve Bayes classifier

- The Naïve Bayes classifier is a probabilistic classifier.
- We compute the probability of a document d being in a class c as follows:

$$P(c|d) \propto P(c, d) = P(c) \prod_{1 \leq i \leq n_d} P(w_i|c)$$

$$= \underbrace{P(c)\, P(d|c)}$$

$$\prod_{1 \leq i \leq n_d} P(w_i|c)$$

the class is happening

happening is the class

$$P(d|c) = P(w_1 \cdots, w_{n_d} | c)$$

$$= P(w_1|c) \cdots P(w_{n_d}|C)$$

$$= \prod_{1 \leq i \leq n_d} P(w_i|c)$$

The Naïve Bayes classifier

- The Naïve Bayes classifier is a probabilistic classifier.
- We compute the probability of a document d being in a class c as follows:

$$P(c|d) \propto P(c,d) = P(c) \prod_{1 \le i \le n_d} P(w_i|c)$$

The Naïve Bayes classifier

- The Naïve Bayes classifier is a probabilistic classifier.
- We compute the probability of a document d being in a class c as follows:

$$P(c|d) \propto P(c, d) = P(c) \prod_{1 \leq i \leq n_d} P(w_i|c)$$

- n_d is the length of the document. (number of tokens)
- $P(w_i|c)$ is the conditional probability of term w_i occurring in a document of class c
- $P(w_i|c)$ as a measure of how much evidence w_i contributes that c is the correct class.
- $P(c)$ is the prior probability of c.
- If a document's terms do not provide clear evidence for one class vs. another, we choose the c with higher $P(c)$.

Maximum a posteriori class

- Our goal is to find the “best” class.
- The best class in Naïve Bayes classification is the most likely or *maximum a posteriori (MAP) class* c_{MAP}:

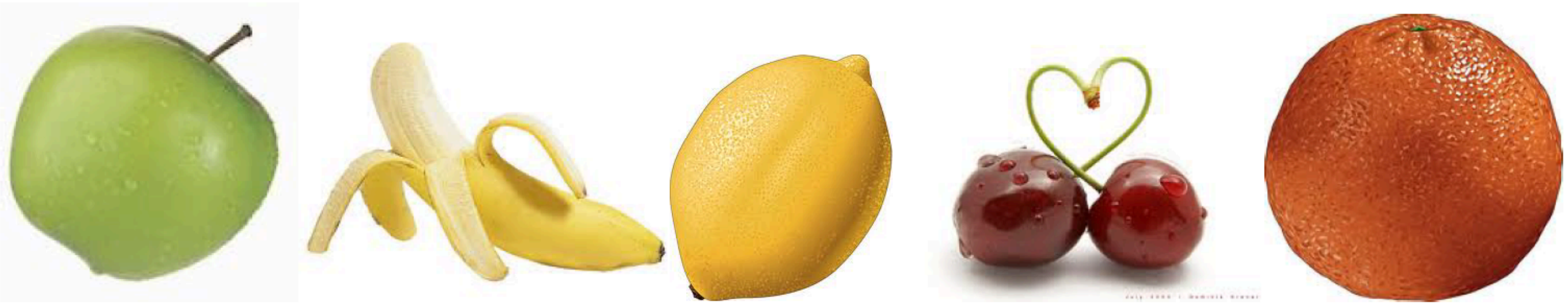
$$c_{\text{MAP}} = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j|d) = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j) \prod_{1 \le i \le n_d} \hat{P}(w_i|c_j)$$

- We write $\hat{P}$ for P since these values are *estimates* from the training set.

Naive Bayes Classifier: More examples

This works because the coin flips are independent given the coin parameter. 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

Using chain rule,

$$P(apple \mid green, round, size = 2)$$

Naive Bayes Classifier: More examples

Using chain rule,

$$
\begin{aligned}
&P(apple \mid green, round, size = 2) \\
&\quad = \frac{P(green, round, size = 2 \mid apple) P(apple)}{\sum_{fruits} P(green, round, size = 2 \mid fruit\, j) P(fruit\, j)} \\
&\quad \propto P(green \mid round, size = 2, apple) P(round \mid size = 2, apple) \\
&\qquad \times P(size = 2 \mid apple) P(apple)
\end{aligned}
$$

$$\frac{P(c,d) = P(c)P(d \mid c)}{\sum_{c \in fruits} P(c,d)}$$

But computing conditional probabilities is hard! There are many combinations of $(color, shape, size)$ for each fruit.

Naive Bayes Classifier: More examples

Idea: assume conditional independence for all features given class,

$$P(green \mid round, size = 2, apple) = P(green \mid apple)$$
$$P(round \mid green, size = 2, apple) = P(round \mid apple)$$
$$P(size = 2 \mid green, round, apple) = P(size = 2 \mid apple)$$

Naive Bayes Classifier: More examples

Idea: assume conditional independence for all features given class,

$$P(green \mid round, size = 2, apple) = P(green \mid apple)$$
$$P(round \mid green, size = 2, apple) = P(round \mid apple)$$
$$P(size = 2 \mid green, round, apple) = P(size = 2 \mid apple)$$

$$P(apple \mid green, round, size = 2) \propto P(apple)P(green \mid apple)P(round \mid apple)P(size = 2 \mid apple)$$

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":

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

Contrasting Naïve Bayes and Logistic Regression

$$c_{\text{MAP}} = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j|d) = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j) \prod_{1 \le i \le n_d} \hat{P}(w_i|c_j)$$

Logistic $C = \arg\max_{c} \hat{P}(c|d) = \arg\max_{c} c \cdot P(c=1|d) + (1-c) P(c=0|d)$

$= (\beta_0 + \sum_i \beta_i x_i) > 0$

$C_{map} = \arg\max_{c_j \in C} \log \hat{P}(c_j) + \sum_{i=1}^{d} \log \hat{P}(w_i|c_j)$

$\beta_0 = \log \hat{P}(c_j)$

$\max_{c_j} \log P(c_j)$

$\beta_0 = \log \hat{P}(1) - \log \hat{P}(0) = \log \frac{\hat{P}(1)}{\hat{P}(0)}$

Contrasting Naïve Bayes and Logistic Regression

$$c_{\mathrm{MAP}} = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j|d) = \arg\max_{c_j \in \mathbb{C}} \; \hat{P}(c_j) \prod_{1 \leq i \leq n_d} \hat{P}(w_i|c_j)$$

$$\arg\max_{c_j \in \mathbb{C}} \; [\ln \hat{P}(c_j) + \sum_{1 \leq i \leq n_d} \ln \hat{P}(w_i|c_j)]$$

Contrasting Naïve Bayes and Logistic Regression

$$c_{\text{MAP}} = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j|d) = \arg\max_{c_j \in \mathbb{C}} \; \hat{P}(c_j) \prod_{1 \le i \le n_d} \hat{P}(w_i|c_j)$$

$$\arg\max_{c_j \in \mathbb{C}} \, [\, \ln \hat{P}(c_j) + \sum_{1 \le i \le n_d} \ln \hat{P}(w_i|c_j)]$$

Naïve Bayes is a special case of logistic regression that uses Bayes rule and conditional probabilities to set these weights

Contrasting Naïve Bayes and Logistic Regression

$$c_{\text{MAP}} = \arg\max_{c_j \in \mathbb{C}} \hat{P}(c_j|d) = \arg\max_{c_j \in \mathbb{C}} \; \hat{P}(c_j) \prod_{1 \le i \le n_d} \hat{P}(w_i|c_j)$$

$$\arg\max_{c_j \in \mathbb{C}} \left[\ln \hat{P}(c_j) + \sum_{1 \le i \le n_d} \ln \hat{P}(w_i|c_j) \right]$$

Naïve Bayes is a special case of logistic regression that uses Bayes rule and conditional probabilities to set these weights

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)
- Logistic regression allows for arbitrary features (biggest difference!)

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)
- Logistic regression allows for arbitrary features (biggest difference!)
- Don’t need to memorize (or work through) previous slide—just understand that naïve Bayes is a special case of logistic regression