CS 112: Computer System Modeling Fundamentals

Prof. Jenn Wortman Vaughan

May 12, 2011

Lecture 13

Reminders & Announcements

- Homework 4 has been posted on the website
- Midterms will be returned in section tomorrow

Today

- A “Naive Bayes” classifier for spam filtering
  (or, everything you need to know for homework 4)

Hypothesis Testing

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i\ P(D \mid H_i)$$

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i\ P(H_i \mid D) = \text{argmax}_i\ P(D \mid H_i)\ P(H_i)$$

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\arg\max_{\theta} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n;\, \theta)$$

- If X$_1$, …, X$_n$ are independent observations, then the ML estimate is

$$\arg\max_{\theta} \prod_{i=1}^{n} P(X_i = x_i; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log\big(P(X_i = x_i;\, \theta)\big)$$

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\arg\max_{\theta} f_{X_1,\ldots,X_n}(x_1,x_2,\ldots,x_n;\ \theta)$$

- If X$_1$, …, X$_n$ are independent observations, then the ML estimate is

$$\arg\max_{\theta} \prod_{i=1}^{n} f_{X_i}(x_i;\ \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log\left(f_{X_i}(x_i;\ \theta)\right)$$

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes $X_1, \ldots, X_n$ of n independent tosses of the coin

What is the maximum likelihood estimate of p?

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes X_1, …, X_n of n independent tosses of the coin

What is the maximum likelihood estimate of p?

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes $X_1, \ldots, X_n$ of n independent tosses of the coin

What is the maximum likelihood estimate of p?

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\#\,\text{times we observed heads}}{n}$$

Parameter Estimation

Suppose we would like to estimate the unknown parameters p_1, …, p_k of a multinomial (e.g., rolls of a die) based on n independent observations

What is the maximum likelihood estimate of each p_j?

Parameter Estimation

Suppose we would like to estimate the unknown parameters $p_1, \ldots, p_k$ of a multinomial (e.g., rolls of a die) based on n independent observations

What is the maximum likelihood estimate of each p_j?

$$\frac{\#\,\text{times we observed outcome } j}{n}$$

Classifying Spam

- Suppose that we would like to classify a new email message as either spam or not spam

Classifying Spam

- Suppose that we would like to classify a new email message as either spam or not spam

- We can represent the email message as a vector of features, e.g., presence or absence of the word "cash", or presence or absence of the recipient's name

Classifying Spam

- Suppose that we would like to classify a new email message as either spam or not spam

- We can represent the email message as a vector of features, e.g., presence or absence of the word “cash”, or presence or absence of the recipient’s name

- We can use previously labeled emails (also represented as feature vectors) to build a probabilistic model

Classifying Spam

- Suppose that we would like to classify a new email message as either spam or not spam

- We can represent the email message as a vector of features, e.g., presence or absence of the word "cash", or presence or absence of the recipient's name

- We can use previously labeled emails (also represented as feature vectors) to build a probabilistic model

- Using this model, we can calculate a MAP (or ML, if we want) hypothesis to classify the new email

What Do We Know?

- We have two hypotheses, H_1 (spam) and H_0 (not spam)

What Do We Know?

- We have two hypotheses, H_1 (spam) and H_0 (not spam)
- We have *d* pieces of data (features) about the new email,

$$F_1 = f_1, F_2 = f_2, \ldots, F_d = f_d$$

What Do We Know?

- We have two hypotheses, H_1 (spam) and H_0 (not spam)
- We have *d* pieces of data (features) about the new email,

$$F_1 = f_1, F_2 = f_2, \ldots, F_d = f_d$$

- The MAP hypothesis is the one that maximizes

$$\mathrm{P}(F_1 = f_1, \ldots, F_d = f_d \mid \mathrm{H}_i)\, \mathrm{P}(\mathrm{H}_i)$$

What Do We Know?

- We have two hypotheses, H_1 (spam) and H_0 (not spam)

- We have d pieces of data (features) about the new email,

$$F_1 = f_1, F_2 = f_2, \ldots, F_d = f_d$$

- The MAP hypothesis is the one that maximizes

$$P(F_1 = f_1, \ldots, F_d = f_d \mid H_i)\, P(H_i)$$

Our goal: Use the labeled emails to estimate this value for each hypothesis H_i so that we can find the MAP hypothesis

Step 1: Estimate the Prior

How can we estimate $P(H_i)$ from data?

Step 1: Estimate the Prior

How can we estimate $P(H_i)$ from data?

- For each previously labeled email *k*, let

$$X_k = \begin{cases} 1, \text{ if email is spam} \\ 0, \text{ otherwise} \end{cases}$$

Step 1: Estimate the Prior

How can we estimate P(H_i) from data?

- For each previously labeled email *k*, let

$$X_k = \begin{cases} 1, \text{ if email is spam} \\ 0, \text{ otherwise} \end{cases}$$

- If our emails are i.i.d., these are Bernoulli random variables with unknown parameter P(H_i) – can estimate this unknown parameter using maximum likelihood

Step 1: Estimate the Prior

How can we estimate P(H_i) from data?

- For each previously labeled email *k*, let

$$X_k = \begin{cases} 1, \text{ if email is spam} \\ 0, \text{ otherwise} \end{cases}$$

- If our emails are i.i.d., these are Bernoulli random variables with unknown parameter P(H_i) – can estimate this unknown parameter using maximum likelihood

$$P(H_i) = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Step 2: Make Some Assumptions

How can we estimate $\mathrm{P}(F_1 = f_1, \ldots, F_d = f_d \mid \mathrm{H}_i)$ from data?

Step 2: Make Some Assumptions

How can we estimate P($F_1 = f_1, \ldots, F_d = f_d \mid \mathrm{H}_i$) from data?

- We could treat this as a multinomial and use maximum likelihood here too

Step 2: Make Some Assumptions

How can we estimate P($F_1 = f_1, \ldots, F_d = f_d \mid \mathrm{H}_i$) from data?

- We could treat this as a multinomial and use maximum likelihood here too… Why is this a bad idea?

Step 2: Make Some Assumptions

How can we estimate P($F_1 = f_1$, …, $F_d = f_d$ | H_i) from data?

- We could treat this as a multinomial and use maximum likelihood here too… Why is this a bad idea?
- Instead, we make the Naive Bayes assumption that all feature values are conditionally independent given H_i

Step 2: Make Some Assumptions

How can we estimate P($F_1 = f_1$, …, $F_d = f_d$| H$_i$) from data?

- We could treat this as a multinomial and use maximum likelihood here too… Why is this a bad idea?
- Instead, we make the Naive Bayes assumption that all feature values are conditionally independent given H$_i$

$$P(F_1 = f_1, \ldots, F_d = f_d \mid H_i) = \prod_{j=1}^{d} P(F_j = f_j \mid H_i)$$

Step 2: Make Some Assumptions

How can we estimate P($F_1 = f_1$, …, $F_d = f_d$ | H$_i$) from data?

- We could treat this as a multinomial and use maximum likelihood here too… Why is this a bad idea?

- Instead, we make the Naive Bayes assumption that all feature values are conditionally independent given H$_i$

$$P(F_1 = f_1, \ldots, F_d = f_d \mid H_i) = \prod_{j=1}^{d} P(F_j = f_j \mid H_i)$$

- For each i, have to estimate d parameters instead of 2^d-1

Step 3: Estimate the Feature Probabilities

How can we estimate $\mathrm{P}(F_j = f_j \mid \mathrm{H}_i)$ from data?

Step 3: Estimate the Feature Probabilities

How can we estimate P($F_j = f_j$ | H$_i$) from data?

- We can use maximum likelihood again

Step 3: Estimate the Feature Probabilities

How can we estimate P($F_j = f_j$ | H$_i$) from data?

- We can use maximum likelihood again

$$P(F_j = f_j) = \frac{\#\ \text{examples with feature } j = f_j}{\#\text{examples}}$$

Step 3: Estimate the Feature Probabilities

How can we estimate P($F_j = f_j$ | H$_i$) from data?

- We can use maximum likelihood again

$$P(F_j = f_j \mid H_i) = \frac{\#\text{examples w/ feature } j = f_j \text{ and label } = H_i}{\#\text{examples w/ label} = H_i}$$

Step 3: Estimate the Feature Probabilities

How can we estimate P($F_j = f_j$ | H$_i$) from data?

- We can use maximum likelihood again

$$P(F_j = f_j \mid H_i) = \frac{\#\text{examples w/ feature } j = f_j \text{ and label } = H_i}{\#\text{examples w/ label} = H_i}$$

Problem: What happens if we don’t observe an example with a particular feature value and label together??

Step 3: Estimate the Feature Probabilities

How can we estimate P($F_j = f_j$ | H$_i$) from data?

- We can use maximum likelihood with smoothing

$$P(F_j = f_j \mid H_i) = \frac{(\#\text{examples w/ feature } j = f_j \text{ and label } = H_i) + 1}{(\#\text{examples w/ label} = H_i) + 2}$$

The Naive Bayes Classifier

- For each *i*, calculate

$$P(F_1 = f_1,\ldots,F_d = f_d \mid H_i)P(H_i)$$

The Naive Bayes Classifier

- For each *i*, calculate

$$\boxed{P(F_1 = f_1, \ldots, F_d = f_d \mid H_i)}P(H_i)$$

Naive Bayes Independence Assumption

The Naive Bayes Classifier

- For each *i*, calculate

$$\prod_{j=1}^{d} P(F_j = f_j \mid H_i)\, P(H_i)$$

The Naive Bayes Classifier

- For each *i*, calculate

$$\prod_{j=1}^{d} P(F_j = f_j \mid H_i)\, \boxed{P(H_i)}$$

ML estimate

The Naive Bayes Classifier

- For each *i*, calculate

$$\prod_{j=1}^{d} \underbrace{\boxed{P(F_j = f_j \mid H_i)}}_{\text{ML estimate using smoothing}} \underbrace{\boxed{P(H_i)}}_{\text{ML estimate}}$$

The Naive Bayes Classifier

- For each *i*, calculate

$$\prod_{j=1}^{d} \underbrace{\boxed{P(F_j = f_j \mid H_i)}}_{\text{ML estimate using smoothing}} \; \underbrace{\boxed{P(H_i)}}_{\text{ML estimate}}$$

- The MAP hypothesis is the one that maximizes this

Example: Classifying Email

“Jenn”	“cash”	“viagra”	spam
1	0	0	0
1	1	0	0
0	0	0	0
0	0	1	1
0	1	0	1

1	0	0	???

Example: Classifying Email

- What if we wanted to estimate the probability that this email is spam?