Naive Bayes Classification

Guido Sanguinetti

Informatics 2B— Learning and Data Lecture 6
10 February 2012

Back to the fish

x	$\hat{P}(x\|M)$	$\hat{P}(x\|F)$
4	0.00	0.02
...	...	...
18	0.01	0.00
19	0.01	0.00
20	0.00	0.00

1. What is the value of $P(M|X = 4)$?
2. What is the value of $P(F|X = 18)$?
3. You observe data point $x = 20$. To which class should it be assigned?

Overview

Today's lecture

- The curse of dimensionality
- Naive Bayes approximation
- Introduction to text classification

Recap: Bayes' Theorem and Pattern Recognition

- Let $C = c_1, \ldots, c_K$ denote the class and $X = \mathbf{x}$ denote the input feature vector
- Classify $\mathbf{x}$ as the class with the maximum posterior probability:

$$c^* = \arg\max_{c_k} P(c_k \mid \mathbf{x})$$

- Re-express this conditional probability using Bayes' theorem:

$$\overbrace{P(c_k \mid \mathbf{x})}^{\text{posterior}} = \frac{\overbrace{P(\mathbf{x} \mid c_k)}^{\text{likelihood}} \overbrace{P(c_k)}^{\text{prior}}}{P(\mathbf{x})}$$

The curse of dimensionality

- Fish example: we constructed a histogram of lengths (m bins)

The curse of dimensionality

- Fish example: we constructed a histogram of lengths (m bins)
- Imagine the input is (length, weight): we need a 2-d histogram ($m \times m$ bins)

The curse of dimensionality

- Fish example: we constructed a histogram of lengths (m bins)
- Imagine the input is (length, weight): we need a 2-d histogram ($m \times m$ bins)
- And if we have a third feature, such as circumference: m^3 bins

The curse of dimensionality

- Fish example: we constructed a histogram of lengths (m bins)
- Imagine the input is (length, weight): we need a 2-d histogram ($m \times m$ bins)
- And if we have a third feature, such as circumference: m^3 bins
- The space of inputs grows exponentially with the number of dimensions. Bellman termed this the *curse of dimensionality*

Weather Example

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	NO
sunny	hot	high	true	NO
overcast	hot	high	false	YES
rainy	mild	high	false	YES
rainy	cool	normal	false	YES
rainy	cool	normal	true	NO
overcast	cool	normal	true	YES
sunny	mild	high	false	NO
sunny	cool	normal	false	YES
rainy	mild	normal	false	YES
sunny	mild	normal	true	YES
overcast	mild	high	true	YES
overcast	hot	normal	false	YES
rainy	mild	high	true	NO

Weather data summary

Counts:

Outlook			Temperature			Humidity			Windy			Play	
	Y	N		Y	N		Y	N		Y	N	Y	N
sunny	2	3	hot	2	2	high	3	4	F	6	2	9	5
overc	4	0	mild	4	2	norm	6	1	T	3	3		
rainy	3	2	cool	3	1								

Weather data summary

Counts:

Outlook			Temperature			Humidity			Windy			Play	
	Y	N		Y	N		Y	N		Y	N	Y	N
sunny	2	3	hot	2	2	high	3	4	F	6	2	9	5
overc	4	0	mild	4	2	norm	6	1	T	3	3		
rainy	3	2	cool	3	1								

Relative frequencies:

Outlook			Temperature			Humidity			Windy			Play	
	Y	N		Y	N		Y	N		Y	N	Y	N
s	2/9	3/5	h	2/9	2/5	h	3/9	4/5	F	6/9	2/5	9/14	5/14
o	4/9	0/5	m	4/9	2/5	n	6/9	1/5	T	3/9	3/9		
r	3/9	2/5	cl	3/9	1/5								

Weather data summary

Counts:

Outlook	Y	N	Temperature	Y	N	Humidity	Y	N	Windy	Y	N	Play Y	Play N
sunny	2	3	hot	2	2	high	3	4	F	6	2	9	5
overc	4	0	mild	4	2	norm	6	1	T	3	3		
rainy	3	2	cool	3	1								

Relative frequencies:

Outlook	Y	N	Temperature	Y	N	Humidity	Y	N	Windy	Y	N	Play Y	Play N
s	2/9	3/5	h	2/9	2/5	h	3/9	4/5	F	6/9	2/5	9/14	5/14
o	4/9	0/5	m	4/9	2/5	n	6/9	1/5	T	3/9	3/9		
r	3/9	2/5	cl	3/9	1/5								

We are given the following test example:

	Outlook	Temp.	Humidity	Windy	Play
$\mathbf{x}^1$	sunny	cool	high	true	?

Naive Bayes

- Write the likelihood as a joint distribution of the d components of $\mathbf{x}$

$$P(\mathbf{x} \mid c_k) = P(x_1, x_2, \ldots, x_d \mid c_k)$$

Naive Bayes

- Write the likelihood as a joint distribution of the d components of $\mathbf{x}$

$$P(\mathbf{x} \mid c_k) = P(x_1, x_2, \ldots, x_d \mid c_k)$$

- **Naive Bayes**: Assume the components of the input feature vector are independent:

$$P(x_1, x_2, \ldots, x_d \mid c_k) \simeq P(x_1 \mid c_k) P(x_2 \mid c_k) \ldots P(x_d \mid c_k)$$
$$= \prod_{i=1}^{d} P(x_i \mid c_k)$$

Naive Bayes

- Write the likelihood as a joint distribution of the d components of $\mathbf{x}$

$$P(\mathbf{x} \mid c_k) = P(x_1, x_2, \ldots, x_d \mid c_k)$$

- **Naive Bayes**: Assume the components of the input feature vector are independent:

$$\begin{aligned} P(x_1, x_2, \ldots, x_d \mid c_k) &\simeq P(x_1 \mid c_k) P(x_2 \mid c_k) \ldots P(x_d \mid c_k) \\ &= \prod_{i=1}^{d} P(x_i \mid c_k) \end{aligned}$$

- Weather example:

$$\begin{aligned} P(O, T, H, W \mid \text{Play}) \simeq\ & P(O \mid \text{Play}) \cdot P(T \mid \text{Play}) \\ & \cdot P(H \mid \text{Play}) \cdot P(W \mid \text{Play}) \end{aligned}$$

Naive Bayes Approximation

- Take d 1-dimensional distributions rather than a single d-dimensional distribution

Naive Bayes Approximation

- Take d 1-dimensional distributions rather than a single d-dimensional distribution
- If each dimension can take m different values, this results in md relative frequencies rather than m^d

Naive Bayes Approximation

- Take d 1-dimensional distributions rather than a single d-dimensional distribution
- If each dimension can take m different values, this results in md relative frequencies rather than m^d
- Re-express Bayes' theorem:

$$P(c_k \mid \mathbf{x}) = \frac{P(\mathbf{x}|c_k)P(c_k}{P(\mathbf{x})}$$

$$= \frac{\prod_{i=1}^{d} P(x_i \mid c_k)P(c_k)}{\prod_{i=1}^{d} P(x_i)}$$

$$\propto P(c_k) \prod_{i=1}^{d} P(x_i \mid c_k)$$

$$c^* = \arg\max_c P(c \mid \mathbf{x})$$

Naive Bayes: Weather Example

- We need much more training data to estimate directly $P(O, T, H, W \mid \text{Play})$ using relative frequencies (since most combinations of the input variables are not observed)

Naive Bayes: Weather Example

- We need much more training data to estimate directly $P(O, T, H, W \mid \text{Play})$ using relative frequencies (since most combinations of the input variables are not observed)
- The training data does let us estimate $P(O \mid \text{Play})$ $P(T \mid \text{Play})$ $P(H \mid \text{Play})$ $P(W \mid \text{Play})$, using relative frequencies

Naive Bayes: Weather Example

- We need much more training data to estimate directly $P(O, T, H, W \mid \text{Play})$ using relative frequencies (since most combinations of the input variables are not observed)
- The training data does let us estimate $P(O \mid \text{Play})$ $P(T \mid \text{Play})$ $P(H \mid \text{Play})$ $P(W \mid \text{Play})$, using relative frequencies
- For test data

	Outlook	Temp.	Humidity	Windy	Play
$\mathbf{x}^1$	sunny	cool	high	true	?

$$P(O = s \mid \text{play} = Y) = 2/9 \qquad P(O = s \mid \text{play} = N) = 3/5$$
$$P(T = c \mid \text{play} = Y) = 3/9 \qquad P(T = c \mid \text{play} = N) = 1/5$$
$$P(H = h \mid \text{play} = Y) = 3/9 \qquad P(O = s \mid \text{play} = N) = 4/5$$
$$P(W = t \mid \text{play} = Y) = 3/9 \qquad P(W = t \mid \text{play} = N) = 3/5$$

Naive Bayes Classification: Weather Example

$$
\begin{aligned}
P(\text{play} = Y \mid \mathbf{x}) \propto P(\text{play} = Y) \cdot [&P(O = s \mid \text{play} = Y) \\
&\cdot P(T = c \mid \text{play} = Y) \cdot P(H = h \mid \text{play} = Y) \\
&\cdot P(W = t \mid \text{play} = Y)] \\
= \frac{9}{14} \cdot \left[\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}\right] &= 0.0053
\end{aligned}
$$

Naive Bayes Classification: Weather Example

$$
\begin{aligned}
P(\text{play} = Y \mid \mathbf{x}) \propto\ & P(\text{play} = Y) \cdot [P(O = s \mid \text{play} = Y) \\
& \cdot P(T = c \mid \text{play} = Y) \cdot P(H = h \mid \text{play} = Y) \\
& \cdot P(W = t \mid \text{play} = Y)] \\
=\ & \frac{9}{14} \cdot \left[\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}\right] = 0.0053 \\
P(\text{play} = N \mid \mathbf{x}) \propto\ & P(\text{play} = N) \cdot [P(O = s \mid \text{play} = N) \\
& \cdot P(T = c \mid \text{play} = N) \cdot P(H = h \mid \text{play} = N) \\
& \cdot P(W = t \mid \text{play} = N)] \\
=\ & \frac{5}{14} \cdot \left[\frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}\right] = 0.0206
\end{aligned}
$$

Naive Bayes Classification: Weather Example

$$\begin{aligned}
P(\text{play} = Y \mid \mathbf{x}) \propto\ & P(\text{play} = Y) \cdot [P(O = s \mid \text{play} = Y) \\
& \cdot P(T = c \mid \text{play} = Y) \cdot P(H = h \mid \text{play} = Y) \\
& \cdot P(W = t \mid \text{play} = Y)] \\
=\ & \frac{9}{14} \cdot \left[\frac{2}{9} \cdot \frac{3}{9} \cdot \frac{3}{9} \cdot \frac{3}{9}\right] = 0.0053
\end{aligned}$$

$$\begin{aligned}
P(\text{play} = N \mid \mathbf{x}) \propto\ & P(\text{play} = N) \cdot [P(O = s \mid \text{play} = N) \\
& \cdot P(T = c \mid \text{play} = N) \cdot P(H = h \mid \text{play} = N) \\
& \cdot P(W = t \mid \text{play} = N)] \\
=\ & \frac{5}{14} \cdot \left[\frac{3}{5} \cdot \frac{1}{5} \cdot \frac{4}{5} \cdot \frac{3}{5}\right] = 0.0206
\end{aligned}$$

$P(\text{play} = Y \mid \mathbf{x}) < P(\text{play} = N \mid \mathbf{x})$, so classify $\mathbf{x}$ as play $= N$

Text Classification

Identifying Spam

Spam?

I got your contact information from your country's information directory during my desperate search for someone who can assist me secretly and confidentially in relocating and managing some family fortunes.

Identifying Spam

Spam?

Dear Dr. Steve Renals,
The proof for your article, Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition, is ready for your review. Please access your proof via the user ID and password provided below. Kindly log in to the website within 48 HOURS of receiving this message so that we may expedite the publication process.

Identifying Spam

Spam?

Congratulations to you as we bring to your notice, the results of the First Category draws of THE HOLLAND CASINO LOTTO PROMO INT. We are happy to inform you that you have emerged a winner under the First Category, which is part of our promotional draws.

Identifying Spam

Question

How can we identify an email as spam automatically?

Identifying Spam

Question

How can we identify an email as spam automatically?

Text classification: classify email messages as spam or non-spam (ham), based on the words they contain

Text Classification using Bayes Theorem

- Document D, with class c_k
- Classify D as the class with the highest posterior probability:

$$P(c_k \mid D) = \frac{P(D \mid c_k)P(c_k)}{P(D)} \propto P(D \mid c_k)P(c_k)$$

Text Classification using Bayes Theorem

- Document D, with class c_k
- Classify D as the class with the highest posterior probability:

$$P(c_k \mid D) = \frac{P(D \mid c_k)P(c_k)}{P(D)} \propto P(D \mid c_k)P(c_k)$$

- How do we represent D? How do we estimate $P(D \mid c_k)$?

Text Classification using Bayes Theorem

- Document D, with class c_k
- Classify D as the class with the highest posterior probability:

$$P(c_k \mid D) = \frac{P(D \mid c_k)P(c_k)}{P(D)} \propto P(D \mid c_k)P(c_k)$$

- How do we represent D? How do we estimate $P(D \mid c_k)$?
- **Bernoulli document model:** a document is represented by a binary feature vector, whose elements indicate absence or presence of corresponding word in the document

Text Classification using Bayes Theorem

- Document D, with class c_k
- Classify D as the class with the highest posterior probability:

$$P(c_k \mid D) = \frac{P(D \mid c_k)P(c_k)}{P(D)} \propto P(D \mid c_k)P(c_k)$$

- How do we represent D? How do we estimate $P(D \mid c_k)$?
- **Bernoulli document model:** a document is represented by a binary feature vector, whose elements indicate absence or presence of corresponding word in the document
- **Multinomial document model:** a document is represented by an integer feature vector, whose elements indicate frequency of corresponding word in the document

Summary

- Naive Bayes approximation
- Example: classifiying multidimensional data using Naive Bayes
- Next lecture: Text classification using Naive Bayes