Statistical Learning,
Naïve Bayes

CSCI 347, Data Mining

Statistical Learning

- Use all of the attributes to predict a class value
- Assumptions:
  - Attributes are equally important in predicting
  - Attributes are independent
- Independence assumption is never correct, but this scheme works well in practice

Basic Bayes's Rule

Probability of event *H* given evidence *E:*

$$Pr[H \mid E] = \frac{Pr[E \mid H]Pr[H]}{Pr[E]}$$

H is the hypothesis

E is the evidence

Pr[X] – probability that X occurs

PR[X|Y] – probability that X occurs given that Y occurred

Bayes's Rule

Typically, multiple attributes are used to predict a class value

For E = E1 & E2 & … & En

Probability of event *H* given evidence *E:*

$$Pr[H \mid E] = \frac{Pr[E1 \mid H] * Pr[E2 \mid H] * \ldots * Pr[En \mid H] * Pr[H]}{Pr[E]}$$

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/ 14	5/ 14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	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

Bayes' Rule

Probability of event H given evidence E:

$$Pr[H \mid E] = \frac{Pr[E \mid H]Pr[H]}{Pr[E]}$$

Evidence:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

$$Pr[H \mid E] = \frac{Pr[E_1 \mid H] * Pr[E_2 \mid H] * \cdots * Pr[E_n \mid H] * Pr[H]}{Pr[E]}$$

Weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← ***Evidence E***

$$Pr[yes \mid E] = Pr[Outlook = Sunny \mid yes]$$

$$\times Pr[Temperature = Cool \mid yes]$$

$$\times Pr[Humidity = High \mid yes]$$

$$\times Pr[Windy = True \mid yes]$$

$$\times \frac{Pr[yes]}{Pr[E]} \qquad = \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{Pr[E]}$$

Probability of class "yes"

Probabilities for weather data

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Likelihood of the two classes

For "yes" = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

For "no" = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Conversion into a probability by normalization:

P("yes") = 0.0053 / (0.0053 + 0.0206) = 0.205

P("no") = 0.0206 / (0.0053 + 0.0206) = 0.795

The "Zero-Frequency Problem"

If an attribute value doesn't occur with every class value the entire probability will be 0
(e.g. "Outlook = overcast" for class "no")

Laplace estimator: add 1 to the count for every attribute value-class combination

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	~~2~~ 3	~~3~~ 4	Hot	~~2~~ 3	~~2~~ 3	High	~~3~~ 4	4 5	False	~~6~~ 7	~~2~~ 3	~~9~~ 10	~~5~~ 6
Overcast	4 5	~~0~~ 1	Mild	4 5	~~2~~ 3	Normal	~~6~~ 7	~~1~~ 2	True	~~3~~ 4	~~3~~ 4		
Rainy	~~3~~ 4	~~2~~ 3	Cool	~~3~~ 4	~~1~~ 2								
Sunny	~~2/9~~ 3/12	~~3/5~~	Hot	~~2/9~~	~~2/5~~	High	~~3/9~~	~~4/5~~	False	~~6/9~~	~~2/5~~	~~9/14~~ 10/16	~~5/14~~ 6/16
Overcast	~~4/9~~ 5/12	~~0/5~~	Mild	~~4/9~~	~~2/5~~	Normal	~~6/9~~	~~1/5~~	True	~~3/9~~	~~3/5~~		
Rainy	~~3/9~~ 4/12	~~2/5~~	Cool	~~3/9~~	~~1/5~~								

Outlook	Temp	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

Bayes Theorem Example

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people of your age. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

Bayes Rule

$$Pr[H \mid E] = \frac{Pr[E \mid H]Pr[H]}{Pr[E]}$$

- We are given:

  P(test | disease) = 0.99
  P(~test | ~disease) = 0.99
  P(disease) = 0.0001

  Hypothesis (H) is disease
  Evidence (E) is test

$$P(\text{disease} \mid \text{test}) = \frac{P(\text{test} \mid \text{disease}) * P(\text{disease})}{P(\text{test} \mid \text{disease}) * P(\text{disease}) + P(\text{test} \mid \sim\text{disease}) * P(\sim\text{disease})}$$

$$= \frac{0.99*0.0001}{0.99*0.0001 + 0.01*0.9999}$$

$$= 0.009804$$

So, the chance that you actually have the disease, given that you tested positive is less than 1%