Data Mining with Weka

Using probabilities

Ian H. Witten

Using probabilities

(OneR: One attribute does all the work)

Opposite strategy: use *all* the attributes

"Naïve Bayes" method

- Two assumptions: Attributes are
  - *equally important a priori*
  - *statistically independent (given the class value)*

    i.e., knowing the value of one attribute says nothing about the value of another *(if the class is known)*
- Independence assumption is never correct!
- But … often works well in practice

Using probabilities

Probability of event *H* given evidence *E*

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

class instance

- Pr[*H*] is *a priori* probability of *H*
  - *Probability of event before evidence is seen*
- Pr[*H* | *E*] is *a posteriori* probability of *H*
  - *Probability of event after evidence is seen*
- "Naïve" assumption:
  - *Evidence splits into parts that are independent*

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

Thomas Bayes, British mathematician, 1702 –1761

Using probabilities

Outlook			Temperature			Humidity			Wind			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	Wind	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

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

Using probabilities

Outlook			Temperature			Humidity			Wind			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								

A new day:

Outlook	**Temp.**	**Humidity**	**Wind**	**Play**
Sunny	Cool	High	True	?

Likelihood of the two classes

For "yes" = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For "no" = 3/5 × 1/ × 4/5 × 3/5 × 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

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

Using probabilities

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	True	?

Evidence E

$$\Pr[yes \mid E] = \Pr[Outlook = Sunny \mid yes]$$
$$\times \Pr[Temperature = Cool \mid yes]$$
$$\times \Pr[Humidity = High \mid yes]$$
$$\times \Pr[Windy = True \mid yes]$$
$$\times \frac{\Pr[yes]}{\Pr[E]}$$

Probability of class "yes"

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

Using probabilities

Use Naïve Bayes

- Open file weather.nominal.arff
- Choose Naïve Bayes method (bayes>NaiveBayes)
- Look at the classifier
- Avoid zero frequencies: start all counts at 1

Using probabilities

- "Naïve Bayes": all attributes contribute equally and independently
- Works surprisingly well
  - even if independence assumption is clearly violated
- Why?
  - classification doesn't need accurate probability estimates

    so long as the greatest probability is assigned to the correct class
- Adding redundant attributes causes problems

  (e.g. identical attributes) → *attribute selection*