Statistical Modeling Approach

Two assumptions

Attributes are

- equally important
- statistically independent

Statistical Modeling Approach

Knowledge about the value of a particular attribute doesn’t tell us anything about the value of another attribute (if the class is known)

Statistical Modeling Approach

Assumptions that are almost never correct

Scheme works well in practice!

Weather Data Set

Attributes:

Class Attribute

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Weather Data Counts

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								

Weather Data Set

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Weather Data Counts

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								

Weather Data Counts

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	Cool	High	True	?

We can use the Table as a Model

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	Cool	High	True	?

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	Cool	High	True	?

For Class=Yes

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

Likelihood for the class play tennis equals to Yes

Yes = 2/9 x 3/9 x 3/9 x 3/9 x 9/14 = 0.0053

Likelihood of the New Day Outcome

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

Likelihood of the two classes attribute Play can take

For each Class value (Yes and No)

Which class value is more likely?

Yes = 2/9 x 3/9 x 3/9 x3/9 x 9/14 = 0.0053

No = 3/5 x 1/5 x 4/5 x 3/5 x5/14 = 0.0206

Likelihood of the New Day Outcome

Convert into probabilities by normalization:

Prob (Class = Yes) = 0.0053/(0.0053 + 0.0206)= 0.205
Prob (Class = No) = 0.0206 / (0.0053 + 0.0206) = 0.795

Probability for NOT Play tennis is ~80%

Naïve Bayes's Rule

$$\Pr[H|E] = \frac{Pr[E|H]\,Pr[H]}{Pr[E]}$$

Naïve Bayes's rule

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

A priori probability of H

Probability of event before evidence has been seen

Pr [H]

A posteriori probability of H

Probability of event after evidence has been seen

Pr [H|E]

Naïve Bayes for Classification

What's the probability of the class given an instance?

Evidence E = instance

Event H = class value for instance

Naïve Bayes assumption: evidence can be split into independent parts

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

Evidence:

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

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

$$\textbf{Probabilities for class YES} = \frac{\frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} * \frac{9}{14}}{Pr[E]}$$

Naïve Bayes Summary

Naïve Bayes works amazingly well

- Violated independence assumption

Because much of classification doesn’t require accurate probability estimates as long as maximum probability is assigned to correct class

Problem: Adding too many redundant attributes

- Example: identical attributes