An example classification problem

- Given
  - Outlook ∈ {Sunny, Overcast, Rainy}
  - Temperature ∈ {Hot, Mild, Cool}
  - Humidity ∈ {High, Normal}
  - Wind ∈ {True, False}
- Answer
  - Play: Yes or No

ZeroR

- Just answer the majority class of train data all the time

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

  - Play=No: 5/14 tuples
  - Play=Yes: 9/14 tuples
- Answer Play=Yes for every test data

Ian H. Witten's slide

OneR: One attribute does all the work

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

Attribute	Rules	Errors	Total errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Wind	False → Yes	2/8	5/14
	True → No*	3/6	

* indicates a tie

Ian H. Witten's slide

Naïve Bayes Classifier

Slides from Kyuseok Shim and Ian H. Witten

Review of Probability Theory

- Random variables
  - $V_1, V_2, \ldots, V_k$
- Joint probability
  - $P(V_1=v_1, V_2=v_2, \ldots, V_k=v_k)$
- Conditional probability

$$P(V_i \mid V_j) = \frac{P(V_i, V_j)}{P(V_j)}$$

Review of Probability Theory

- Chain rule

$$P(V_1, V_2, \ldots, V_k) = \prod_{i=1}^{k} P(V_i \mid V_{i-1}, \ldots, V_1)$$

  - e.g., *P*(A=a, B=b, C=c)
    - P(abc) = P(a)P(b|a)P(c|ab)

- Bayes theorem

$$P(V_i \mid V_j) = \frac{P(V_j \mid V_i) P(V_i)}{P(V_j)}$$

  - e.g., *P*(A=a|B=b)
    - P(a|b) = P(b|a) P(a) / P(b)

Review of Probability Theory

- Unconditional independence

$$P(V_1, V_2, \ldots V_k) = \prod_{i=1}^{k} P(V_i \mid V_{i-1}, \ldots, V_1) = \prod_{i=1}^{k} P(V_i)$$

  - e.g., *P* (A=a, B=b)
    - P(a,b)=P(ab)=P(a)P(b)

- Conditional independence

$$P(V_1, V_2, \ldots, V_k \mid V) = \prod_{i=1}^{k} P(V_i \mid V_{i-1}, \ldots V_1, V) = \prod_{i=1}^{k} P(V_i \mid V)$$

  - e.g., P(A=a,B=b|C=c)
    - P(ab|c)=P(a|c)P(b|c)

"Naïve Bayes" method

- Opposite strategy: use all the attributes
  - OneR: One attribute does all the work
- 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

Ian H. Witten's slide

Probability of event H given evidence E

- Thomas Bayes, British mathematician, 1702 –1761

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

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

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

Ian H. Witten's slide

Towards Naïve Bayesian Classifier

- Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector **X** = (x_1, x_2, …, x_n)
- Suppose there are *m* classes C_1, C_2, …, C_m.
- Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$
- This can be derived from Bayes' theorem

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

- Since P(X) is constant for all classes, only needs to be maximized

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$$

"Naïve Bayes" method

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	?

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

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

Ian H. Witten's slide

"Naïve Bayes" method

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"

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

Ian H. Witten's slide

Bayesian Classification: Why?

- A statistical classifier: performs probabilistic prediction, i.e., predicts class membership probabilities
- Foundation: Based on Bayes' Theorem.
- Performance: A simple Bayesian classifier, naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct — prior knowledge can be combined with observed data