[BAYES CLASSIFACTION]

$$P(h/D) = \frac{P(D/h)\ P(h)}{P(D)}$$

maximum likelihood (ML) hypothesis

$$h_{ML} = \underset{h_i \in H}{\operatorname{argmax}}\ P(D/h_i)$$

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\ P(h/D)$$

$$= \underset{h \in H}{\operatorname{argmax}}\ \frac{P(D/h)\ P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}}\ P(D/h) \cdot P(h)$$

Does Patient have cancer or not ?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result (+) in only 98% of the cases in which the disease is actually present, and a correct negative result (−) in only 97% of the cases in which the disease is not present.

Furthermore 0.008 of the entire population have this cancer.

$P(\text{cancer}) = 0.008$ $\quad P(\neg\text{cancer}) = 0.992$

$P(+|\text{cancer}) = 0.98$ $\quad P(-|\text{cancer}) = 0.02$

$P(+|\neg\text{cancer}) = 0.03$ $\quad P(-|\neg\text{cancer}) = 0.97$

$P(+|\text{cancer}) \cdot P(\text{cancer}) = 0.98 * 0.008 = \underline{0.0078}$

$P(+|\neg\text{cancer}) \cdot P(\neg\text{cancer}) = 0.03 \cdot 0.992 = \underline{0.0298}$

$h_{MAP} = \neg\text{cancer}$

- Naive assumption: Attribute independence

$$P(x_1, \ldots, x_k|c) = P(x_1|c) \cdot \ldots \cdot P(x_k|c)$$

- If i-th attribute is categorical:
  $P(x_i|c)$ is estimated as the relative freq. of samples having value x_i as i-th attribute in class c.
- If i-th attribute is continuous
  $P(x_i|c)$ is estimated thru a Gaussian density function

Play-Tennis example: estimating $P(x_i|c)$

$P(\text{Play})$, $n(\text{don't play})$

$P(p) = 9/14$ $\quad$ $P(n) = 5/14$

Outlook	
$P(\text{Sunny}\mid p) = 2/9$	$P(\text{sunny}\mid n) = 3/5$
$P(\text{overcast}\mid p) = 4/9$	$P(\text{overcast}\mid n) = 0$
$P(\text{rain}\mid p) = 3/9$	$P(\text{rain}\mid n) = 2/5$
Temperature	
$P(\text{hot}\mid p) = 2/9$	$P(\text{hot}\mid n) = 2/5$
$P(\text{mild}\mid p) = 4/9$	$P(\text{mild}\mid n) = 2/5$
$P(\text{cool}\mid p) = 3/9$	$P(\text{cool}\mid n) = 1/5$
Humidity	
$P(\text{high}\mid p) = 3/9$	$P(\text{high}\mid n) = 4/5$
$P(\text{normal}\mid p) = 6/9$	$P(\text{normal}\mid n) = 2/5$
Windy	
$P(\text{true}\mid p) = 3/9$	$P(\text{true}\mid n) = 3/5$
$P(\text{false}\mid p) = 6/9$	$P(\text{false}\mid n) = 2/5$

Play-tennis example: classifying X

$X = <rain, hot, high, false>$

$P(X|p) \cdot P(p) = P(rain|p) \cdot P(hot|p) \cdot P(high|p) \cdot P(false|p) \cdot P(p)$

$= 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$

$P(X|n) \cdot P(n) = P(rain|n) \cdot P(hot|n) \cdot P(high|n) \cdot P(false|n) \cdot P(n)$

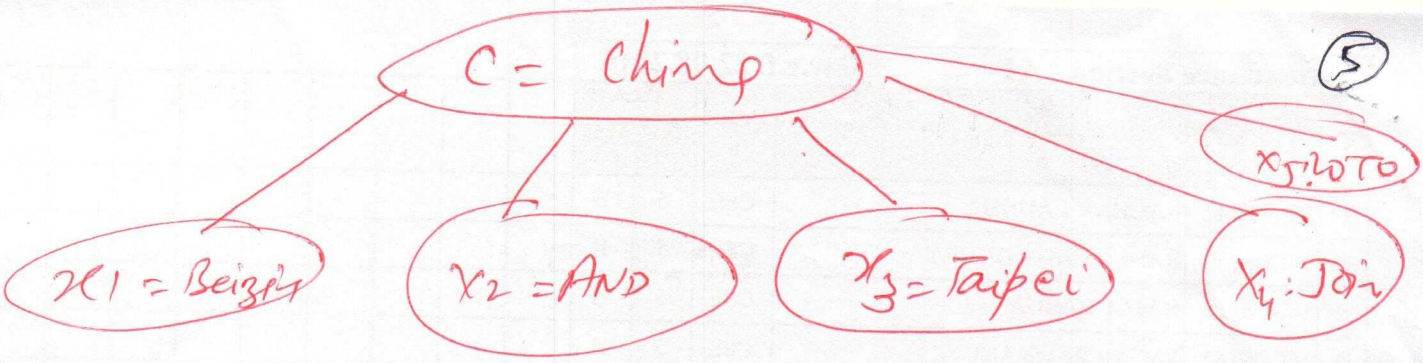
$= 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = 0.018286.$

• Sample X is classified in class n (don't play)

Training set

Doc ID			in c = China
1	Chinese	Beijing chines	yes
2	Chinese	Chinese Shanghai	yes
3	Chinese	Macao	yes
4	Tokyo	Japan Chinese	no

Test is: Chinese Chinese Chinese Tokyo Japan = ?



Beijing joins the WTO: China)

$P(\text{Yes}) = 3/4$ $\quad P(\text{No}) = 1/4$

$P(\text{chinese} | \text{yes}) = (5+1)/8+6 = 6/14 = 3/7$

$P(\text{tokyo} | \text{yes}) = (0+1)/14 =$

$P(\text{Japan} | \text{yes}) = (0+1)/14$

$P(\text{chinese} | \text{No}) = (1+1)/(3+6) = 2/9$

$P(\text{Japan} | \text{No}) = (1+1)/(3+6) = 2/9$

$P(\text{tokyo} | \text{No} = (1+1)/(3+6) = 2/9.$

Using Laplace smoothing

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\left(\sum_{t' \in V}(T_{ct'}+1)\right)}$$

$$= \frac{T_{ct}+1}{\left(\sum_{t' \in V} t_{ct'}\right) + B'}$$

$B' = |V|$ is the no of terms in the vocabulary

$t_{ct} \Rightarrow$ No of occurances of t in training doc from class c.

$$P(Yes \mid DS) = \frac{3}{4} \times \frac{3}{7} \times \frac{3}{7} \times \frac{3}{7} \cdot \frac{1}{14} \times \frac{1}{14}$$

$$= \underline{0.0003}$$

$$P(No \mid DS) = \frac{1}{4} \cdot \frac{2}{9} \times \frac{2}{9} \times \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{2}{9}$$

$$= \underline{0.0001}$$

So the DS belongs to class China.