7.3 Naïve Bayes Classifier

- ✓ Based on Bayes' theorem
- ✓ Statistical classifier
- ✓ Supervised learning
- ✓ Performance comparable to that of decision trees and neural networks
- ✓ High accuracy and speed when applied to large datasets
- ✓ Naïve: effect of one attribute is independent of the effect of other attributes (very simplified assumption)

❖ Basic ideas and major steps

1) Each data sample is represented by a feature vector,

$$V = (v_1, v_2, \ldots, v_n)$$, where n is the number of attributes.

2) Classifier predicts that unknown sample, $X = (x_1, x_2, \ldots, x_n)$ belongs to one of m classes, C_i with highest posterior probability

$P(C_i \mid X) > P(C_j \mid X), 1 \leq j \leq m \;\&\; j \neq i$. [Maximum posterior probability]

3) Posterior probabilitis are computed using Bayes' theorem as follows:

$$P(C_i \mid X) = (P(X \mid C_i) \times P(C_i)) / P(X)$$

4) $P(X)$ is constant for all classes, so, $P(X \mid C_i) \times P(C_i)$ needs to be maximized.

5) If classes are equally likely, $P(C_i)$ can also be dropped.

6) We take,

$P(C_i) = S_i / S$, where S_i – no. of samples of class C_i , S – total no. of samples.

7) Discarding attribute dependence,

$$P(X \mid C_i) = \Pi_{k=1:n} \; P(x_k \mid C_i).$$

8) For categorical attribute A_k, $P(x_k \mid C_i) = S_{ik} / S_i$, where S_i - no. of samples of class C_i and S_{ik} - those from S_i with attribute value x_k.

9) For continuous A_k , Gaussian distribution is typically assumed:

$P(x_k \mid C_i) = g(x_k, \mu_{Ci}, \sigma_{Ci})$ [Gaussian normal density function for A_k , while μ_{Ci} – mean and σ_{Ci} – standard deviation of samples with x_k of class C_i]

❖ **Example:** We take the same training dataset as for decision tree learning.

ID	**Age**	**Income**	**Student**	**Credit Rating**	*Decision/ Class/ Label*
1	≤ 30	high	no	fair	negative
2	≤ 30	high	no	excellent	negative
3	31…40	high	no	fair	positive
4	> 40	medium	no	fair	positive
5	> 40	low	yes	fair	positive
6	> 40	low	yes	excellent	negative
7	31…40	low	yes	excellent	positive
8	≤ 30	medium	no	fair	negative
9	≤ 30	low	yes	fair	positive
10	> 40	medium	yes	fair	positive
11	≤ 30	medium	yes	excellent	positive
12	31…40	medium	no	excellent	positive
13	31…40	high	yes	fair	positive
14	> 40	medium	no	excellent	negative

- C_1: 'Buys a computer' / 'positive
- C_2: 'Does not buy a computer' / 'negative'.

➢ Unknown sample: X = (age = 22, income = 'medium', student = 'yes', credit_rating = 'fair')

➢ We now compute P(X | C_i), for i = 1, 2 as follows:

P(age = '<=30' | C_1) = 2/9 = 0.222

P(age = '<=30' | C_2) = 3/5 = 0.600

…..

- Check that,

$$P(X \mid C_1) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$
$$P(X \mid C_2) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$$

- Also, $P(C_1) = 9/14 = 0.643$; $P(C_2) = 5/14 = 0.352$.
- Thus we have, $P(X \mid C_1)\, P(C_1) = 0.044 \times 0.643$ ***= 0.028***

$$P(X \mid C_2)\, P(C_2) = 0.019 \times 0.357 = 0.007$$

- That is, prediction for sample X is the same to that with decision tree: