Bayesian Classification

- Select, given a certain pattern **x**, the class c_i which most likely predicts the pattern:

$$P(C = c_i | \mathbf{x}) > P(C = c_j | \mathbf{x}) \quad \forall j \neq i$$

- **Bayes Theorem (BT):**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Provides a method to calculate the probability of a random event A, while a priori knowing the occurrence of an event B

Bayesian Classification

- In classification:

$$P(c_j \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_j)P(c_j)}{P(\mathbf{x})}$$

- Since we want to know the class c_j that maximizes P ($c_j|\mathbf{x}$), we need to find the class c_j that maximizes
- P ($\mathbf{x}| c_j$) P (c_j)

Example

- Instances described by one attribute
- Two classes:
  - the patient has hepatitis;
  - the patient has not hepatitis.
- Attribute: result of a exam to the liver
- Two values: + (hepatitis) and - (no hepatitis).
- From the lab results we know:

P(hepatitis) = 0.008; P(no hepatitis) = 0.992

P(+ | hepatitis) = 0.98; P(- | hepatitis) = 0.02

P(+ | no hepatitis) = 0.03; P(- | no hepatitis) = 0.97

Example

- We observe that, according to a lab exam, a new patient's result is + and we want to classify it (has he hepatitis or not?).

- Since for BT (leaving out the denominator) we have:
  - P (+ | hepatitis) * P (hepatitis) = 0.078
  - P (+ | no hepatitis) * P (no hepatitis) = 0.298
- we conclude that it is more likely that the patient has not the hepatitis.

Bayesian classification

- In case where the instance **x** is described by more than one attributes, we find the class C_j such that

  $$P\ (c_j|X_1=x_1,X_2=x_2,\ \ldots,X_n=x_n)$$

  is maximum, i.e, we want to make a **belief revision**

  For brevity we write

  $$P\ (c_i|\ x_1,\ x_2,\ \ldots,\ x_n)$$

- By applying BT, this is equal to find the class that maximizes

  $$P\ (x_1,\ x_2,\ \ldots,\ x_n\ |\ c_i) \times P\ (c_i)$$

Bayesian Classification

- In the case of n attributes, the problem is to estimate $P(x_1, x_2, \ldots, x_n \mid c_i)$ from the available data
- Direct approach: $P(x_1, x_2, \ldots, x_n \mid c_i)$ is given by the number of instances of the class c_i equal to $x_1, x_2, \ldots, x_n$ that appear in the data divided by the number of instances of the class c_i.
- Problems:
  - In order to compute $P(x_1, x_2, \ldots, x_n \mid c_i)$, the instance $x_1, x_2, \ldots, x_n$ shoud appear more times in the data => we need a lot of data
  - we cannot classify instances not seen or not present in the training data => **no generalization**!

Bayesian Classification

- To overcome these problems we have to use the simplifying assumptions that the observed attributes are independent (**Naive Bayes assumptions**) given the class.
- If a and b are independent given the class, the likelihood that a and b occurr simultaneously given c P (a, b | c) is

  $P(a \mid c) \times P(b \mid a, c) = P(a \mid c) \times P(b \mid c)$

- Naive Bayes Method: using the Bayes theorem making the assumption of independence of the attributes

Classification Naive Bayes

- In our case

$$P(x_1, x_2, \ldots, x_n \mid c_i) = P(x_1 \mid c_i) \times P(x_2 \mid c_i) \times \ldots \times P(x_n \mid c_i)$$

- So with the Naive Bayes method the highest probability is assigned to the class c obtained by the following formula

$$c = \underset{c_j}{\operatorname{argmax}}\, P(c_j) \prod_{k=1}^{n} P(x_k \mid c_j)$$

Classification naive Bayes

- The parameters are calculated by the relative frequency
- $P(c_i)$ = proportion of examples of the training set that belong to c_i
- $P(x_k | c_i)$ = number of examples in the training set belonging to class c_i and that have the k-th attribute equal to x_k divided by the number of examples in the training set that belong to the class c_i

Classification naive Bayes

- The learning phase is to build a table of this form

	c_1	c_2	….	c_n
-		$P(c_2)$		
$X_1 = x_{1.1}$				
….				
$X_1 = x_{1,\ k1}$		$P(X_1 = x_{1,\ k1} \mid C = c_2)$		
….				
$X_n = x_{n,\ 1}$				
….				
$X_n = x_{n,\ kn}$				

Example

No	Outlook	Temp	Humid	Windy	Class
D1	sunny	mild	normal	T	P
D2	sunny	hot	high	T	N
D3	sunny	hot	high	F	N
D4	sunny	mild	high	F	N
D5	sunny	cool	normal	F	P
D6	overcast	mild	high	T	P
D7	overcast	hot	high	F	P
D8	overcast	cool	normal	T	P
D9	overcast	hot	normal	F	P
D10	rain	mild	high	T	N
D11	rain	cool	normal	T	N
D12	rain	mild	normal	F	P
D13	rain	cool	normal	F	P
D14	rain	mild	high	F	P

- Given a day with the following characteristics:

  G =<Outlook=sunny, Temp =cool,Humid =high,Windy=T >

we want to know whether or not to play tennis.

Example

- We do not calculate the entire matrix, we calculate only the probabilities that we need

Example (cont.)

P (Class=P) = 9/14 = 0.64

P (Class=N) = 5/14 = 0.36

P (Outlook=sunny | Class=P) = 2/9 = 0,222

P (Outlook=sunny | Class=N) = 3/5 = 0.6

P (Temp =cool | Class=P) = 3/9 = 0.333

P (Temp =cool | Class=N) = 1/5 = 0.2

P (Humid =high | Class=P) = 3/9 = 0.333

P (Humid =high | Class=N) = 4/5 = 0.8

P (Windy T = | Class=P) = 3/9 = 0.33

P (Windy T = | Class=N) = 3/5 = 0.6

P (Class=P) *P (Outlook=sunny|P)*P (Temp = cool |P) * P (Humid =high |P)*P (Windy T =|P) = **0.0052**

P (Class=N) *P (Outlook=sunny | No) * P (temp =cool | No) * P (Humid =high | No) * P (Windy T =| No) = **0.0207**

Example (cont.)

- The more 'probable class' is N
- We calculate the probability of classes:

P (**Class=P**| G) = 0.0053/ (0.0053+ 0.0206) =
=0.205

P (**Class=N**| G) = 0.0206/ (0.0053+ 0.0206) =
= 1-0.205 =0.795