Naive Bayes Classifier

1. Bayes' law

Among all patients in the haematology section 1% is infected with the HIV virus, 97% of which test positive. Among non-infected patients this is 4%. Calculate the probability that a patient who is tested HIV positive really is infected with the HIV virus.

C_1 = "HIV+ patients"
C_2 = "HIV- patients"
X = "test result"

$P(C_1) = 0.01$ $\quad P(X = + \mid C_1) = 0.97$
$P(C_2) = 0.99$ $\quad P(X = + \mid C_2) = 0.04$

$$\boxed{P(C_1 \mid X = +) = \frac{P(C_1) \cdot P(X = + \mid C_1)}{P(C_1) \cdot P(X = + \mid C_1) + P(C_2) \cdot P(X = + \mid C_2)}}$$

Terminology

- $P(C_1)$, $P(C_2)$: prior probability

- $P(C_1 \mid X = +)$, $P(C_2 \mid X = +)$: posterior probability

In general

- n classes: $C_1, C_2, \ldots, C_n$

- k predictor variables $X_1, X_2, \ldots, X_k$

Question

If $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$, than what is the most probable class?

$$P(C_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)\ ?$$
$$= \quad P(C_i \mid x_1, x_2, \ldots, x_k)\ ?$$

Bayes' law:

Wet van Bayes:

$$P(C_i \mid x_1, x_2, \ldots, x_k) = \frac{P(C_i) \cdot P(x_1, x_2, \ldots, x_k \mid C_i)}{P(x_1, x_2, \ldots x_k)}$$

Example

- Titanic dataset

- 4 variables, 2201 observations

- X_1: class
  ('0' = personnel, '1'= most expensive class, '2' = middle class, '3' = cheapest class)

- X_2: age
  ('1' = adult, '0' = child)

- X_3: gender
  ('1' = male, '0' = female)

- C: rescued or not
  ('1' = rescued, '0' = not rescued)

Example question: Will a girl in the cheapest class probably be rescued or not?

P(rescued | cheapest, child, female) = ?
P(not rescued | cheapest, child, female) = ?

$$P(\text{rescued} \mid \text{cheapest, child, female}) = \frac{P(\text{rescued}) \cdot P(\text{cheapest, child, female} \mid \text{rescued})}{P(\text{cheapest, child, female})}$$

$$P(\text{not rescued} \mid \text{cheapest, child, female}) = \frac{P(\text{not rescued}) \cdot P(\text{cheapest, child, female} \mid \text{not rescued})}{P(\text{cheapest, child, female})}$$

Remarks:

- denominator P(cheapest, child, female) no need to compare.

- To answer all questions, the following needs to be known:

  P(rescued) and P(not rescued)
  $P(x_1, x_2, x_3 \mid \text{rescued})$ and
  $P(x_1, x_2, x_3 \mid \text{not rescued})$
  for all combinations of x_1, x_2, x_3
  $\rightarrow$ $2 \cdot 4 \cdot 2 \cdot 2 = 32$ possibilities
  Estimate probabilities via frequencies in training set

Problem:

- large amount of combinations x_1, x_2, x_3

- certain combinations may not occur in a class C_i in the training set

  → estimate: $P(x_1, x_2, x_3 \mid C_i) = 0$.
  → $P(C_i \mid x_1, x_2, x_3) = 0$

2. Solution: Naive Bayes

- Assumption: X_1, X_2, X_3 not interdependent in every class C_i

→ $P(x_1; x_2; x_3 \mid C_i)$
$= P(x_1 \mid C_i) \bullet P(x_2 \mid C_i) \bullet P(x_3 \mid C_i)$

- $P(x_1 \mid C_i)$, $P(x_2 \mid C_i)$, $P(x_3 \mid C_i)$
  Estimate via frequencies in training set

- less quickly zero; fewer chances: $2 \bullet (4+2+2) = 16$