

Méthodes d'apprentissage automatique pour la bioinformatique

BIOL-F-524

Gianluca Bontempi

Département d'Informatique
Boulevard de Triomphe - CP 212
<http://www.ulb.ac.be/di>

Naive Bayes classifier

- The Naive Bayes (NB) classifier has shown in some domains a performance comparable to that of neural networks and decision tree learning.
- Consider a classification problem with n inputs and a random output variable y that takes values in the set $\{c_1, \dots, c_K\}$.
- The Bayes optimal classifier for a 0-1 loss function should return

$$c^*(x) = \arg \max_{j=1, \dots, K} \text{Prob} \{y = c_j | x\}$$

- We can use the Bayes theorem to rewrite this expression as

$$\begin{aligned} c^*(x) &= \arg \max_{j=1, \dots, K} \frac{\text{Prob} \{x | y = c_j\} \text{Prob} \{y = c_j\}}{\text{Prob} \{x\}} = \\ &= \arg \max_{j=1, \dots, K} \text{Prob} \{x | y = c_j\} \text{Prob} \{y = c_j\} \end{aligned}$$

Naive Bayes classifier

- How to estimate these two terms on the basis of a finite set of data?
- It is easy to estimate each of the a priori probabilities $\text{Prob}\{y = c_j\}$ simply by counting the frequency with which each target class occurs in the training set. The estimation of $\text{Prob}\{x|y = c_j\}$ is much harder.
- The NB classifier is based on the simplifying assumption that the input values are conditionally independent given the target value:

$$\text{Prob}\{x|y = c_j\} = \text{Prob}\{x_1, \dots, x_n|y = c_j\} = \prod_{h=1}^n \text{Prob}\{x_h|y = c_j\}$$

- The NB classification is then

$$c_{NB}(x) = \arg \max_{j=1, \dots, K} \text{Prob}\{y = c_j\} \prod_{h=1}^n \text{Prob}\{x_h|y = c_j\}$$

- If the inputs x_h are discrete variables the estimation of $\text{Prob}\{x_h|y = c_j\}$ boils down to the counting of the frequencies of the occurrences of the different values of x_h for a given class c_j .

Example

Obs	G1	G2	G3	G
1	P.LOW	P.HIGH	N.HIGH	P.HIGH
2	N.LOW	P.HIGH	P.HIGH	N.HIGH
3	P.LOW	P.LOW	N.LOW	P.LOW
4	P.HIGH	P.HIGH	N.HIGH	P.HIGH
5	N.LOW	P.HIGH	N.LOW	P.LOW
6	N.HIGH	N.LOW	P.LOW	N.LOW
7	P.LOW	N.LOW	N.HIGH	P.LOW
8	P.LOW	N.HIGH	N.LOW	P.LOW
9	P.HIGH	P.LOW	P.LOW	N.LOW
10	P.HIGH	P.LOW	P.LOW	P.LOW

What is the NB classification for the query {G1=N.LOW G2= N.HIGH G3=N.LOW }?

Example

$$\text{Prob}\{y = P.HIGH\} = 2/10, \quad \text{Prob}\{y = P.LOW\} = 5/10$$

$$\text{Prob}\{y = N.HIGH\} = 1/10, \quad \text{Prob}\{y = N.LOW\} = 2/10$$

$$\text{Prob}\{G1 = N.LOW|y = P.HIGH\} = 0/2, \quad \text{Prob}\{G1 = N.LOW|y = P.LOW\} = 1/5$$

$$\text{Prob}\{G1 = N.LOW|y = N.HIGH\} = 1/1, \quad \text{Prob}\{G1 = N.LOW|y = N.LOW\} = 0/2$$

$$\text{Prob}\{G2 = N.HIGH|y = P.HIGH\} = 0/2, \quad \text{Prob}\{G2 = N.HIGH|y = P.LOW\} = 1/5$$

$$\text{Prob}\{G2 = N.HIGH|y = N.HIGH\} = 0/1, \quad \text{Prob}\{G2 = N.HIGH|y = N.LOW\} = 0/2$$

$$\text{Prob}\{G3 = N.LOW|y = P.HIGH\} = 0/2, \quad \text{Prob}\{G3 = N.LOW|y = P.LOW\} = 3/5$$

$$\text{Prob}\{G3 = N.LOW|y = N.HIGH\} = 0/1, \quad \text{Prob}\{G3 = N.LOW|y = N.LOW\} = 0/2$$

$$c_{NB}(x) = \arg \max_{P.H, P.L, N.H, N.L}$$

$$\{2/10 * 0 * 0 * 0, 5/10 * 1/5 * 1/5 * 3/5, 1/10 * 1 * 0 * 1, 2/10 * 0 * 0 * 0\} = P.LOW$$