

Data Mining (Mineria de Dades)

Lluís A. Belanche
belanche@lsi.upc.edu

Soft Computing Research Group

Dept. de Llenguatges i Sistemes Informàtics (Software department)

Universitat Politècnica de Catalunya

2011-2012

Data Mining (Part II)

Contents

1. Association Rules
2. Bayesian decision theory
3. Bayesian classifiers: LDA/QDA
4. Bayesian classifiers: Naïve Bayes
5. Bayesian classifiers: Nearest Neighbours
6. Neural Networks: Multilayer Perceptrons
7. Support Vector Machines
8. Ensemble Learning

The Naïve-Bayes classifier

- We showed that the 0/1 loss Bayes rule minimizing the probability of error could be formulated in terms of a family of discriminant functions:

$$g_k(\mathbf{x}) = P(\omega_k)P(\mathbf{x}|\omega_k), k = 1, \dots, c.$$

- We can expand the conditional probability:

$$\begin{aligned} P(\omega_k)P(\mathbf{x}|\omega_k) &= P(\omega_k)P(X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_n = x_n \mid \omega_k) \\ &= P(\omega_k)P(X_1 = x_1|\omega_k) \prod_{j=2}^n P(X_j = x_j \mid \omega_k, X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_{j-1} = x_{j-1}) \end{aligned}$$

Assuming that $X_1, \dots, X_n$ (in pairs) are statistically independent *given the class*:

$$= P(\omega_k)P(X_1 = x_1|\omega_k) \prod_{j=2}^n P(X_j = x_j \mid \omega_k)$$

$$= P(\omega_k) \prod_{j=1}^n P(X_j = x_j \mid \omega_k)$$

$$\equiv \text{NB}_k(\mathbf{x})$$

The Naïve-Bayes classifier

Example

Outlook	Temperature	Humidity	Wind	PlayTennis?
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

The Naïve-Bayes classifier

Example

$$P(\omega_k) \approx \hat{P}(\omega_k) = \frac{|S_k|}{|S|}, \quad P(X_j = x_j \mid \omega_k) \approx \hat{P}(X_j = x_j \mid \omega_k) = \frac{|\{\mathbf{x} \in S_k \wedge X_j = x_j\}|}{|\{\mathbf{x} \in S_k\}|}$$

The prediction for the test example $\mathbf{x}^* = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{Weak})^t$ is $\mathbf{x}^* \in \text{Yes}$, given that:

- $\hat{P}(\text{No}) \cdot \hat{P}(\text{Sunny}|\text{No}) \cdot \hat{P}(\text{Hot}|\text{No}) \cdot \hat{P}(\text{Normal}|\text{No}) \cdot \hat{P}(\text{Weak}|\text{No})$
 $= \frac{5}{14} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \approx 0,0069$
- $\hat{P}(\text{Yes}|\mathbf{x}^*) = \hat{P}(\text{Yes}) \cdot \hat{P}(\text{Sunny}|\text{Yes}) \cdot \hat{P}(\text{Hot}|\text{Yes}) \cdot \hat{P}(\text{Normal}|\text{Yes}) \cdot \hat{P}(\text{Weak}|\text{Yes})$
 $= \frac{9}{14} \cdot \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} \approx 0,0141$

Note that the posteriors are:

$$\hat{P}(\text{No}|\mathbf{x}^*) \approx \frac{0,0069}{0,0069+0,0141} \approx 0,329, \quad \hat{P}(\text{Yes}|\mathbf{x}^*) \approx \frac{0,0141}{0,0069+0,0141} \approx 0,671$$

The Naïve-Bayes classifier

Extensions

1. Missing values
2. Null empirical probabilities
3. Continuous r.v.

The Naïve-Bayes classifier

Missing values

These values can be handled in two ways:

1. Ignoring the values in the probability counts in which they are involved
2. Adding a further modality named “Missing”

The second possibility is indicated when their number is significant and/or they can appear in future examples

The Naïve-Bayes classifier

Null empirical probabilities

In future examples $\mathbf{x}^*$, it may happen that some variable X_j has a value x_j^* not present in the sample used to create the classifier. In this case, $\hat{P}(X_j = x_j^* \mid \omega_k) = 0$ and therefore all the discriminants are null.

A possible workaround is the Laplace correction:

$$\hat{P}_L(X_j = x_j^* \mid \omega_k) = \frac{|\{\mathbf{x} \in S_k \wedge X_j = x_j\}| + p}{|\{\mathbf{x} \in S_k\}| + pV_k}, \quad p \in \mathbb{N}$$

where p is the “weight” assigned to the prior probability and V_k is the number of modalities of variable k

Example. Take $p = 1$ and $V_k = 2$. Assume that $|\{\mathbf{x} \in S_k \wedge X_j = x_j\}| = 0$. Then $\hat{P}_L(X_j = x_j^* \mid \omega_k) = \frac{1}{|\{\mathbf{x} \in S_k\}| + 2}$. Can you give an interpretation?

The Naïve-Bayes classifier

Continuous variables (I)

There are at least three ways of handling continuous variables in this classifier:

1. Assume a particular pdf for the variable and estimate its parameters
2. Discretize the variable and use the standard discrete classifier
3. Do not assume a particular pdf for the variable and use KDE

The Naïve-Bayes classifier

Continuous variables (II)

1. Since we have assumed statistically independent variables (*given the class*), we can model every variable of each class differently: if the variable is continuous, we can resort to a univariate pdf that we find convenient.

The parameters of the pdf can be found in the usual way, by maximum likelihood. For instance, if X_j is normally distributed for class ω_k :

$$p_{jk}(x) = \frac{1}{\sqrt{2\pi}\sigma_{jk}} \exp \left\{ -\frac{(x - \mu_{jk})^2}{2\sigma_{jk}^2} \right\}$$

The factor $\hat{P}(X_j = x_j^* \mid \omega_k)$ would be replaced by $p_{jk}(x_j^*)$

As an example, consider the variable Temperature in the previous exercise.

Note that there are (somewhat) standard ways to make a univariate pdf closer to that of a gaussian pdf (e.g. log is suitable in presence of a positive skewness).

The Naïve-Bayes classifier

Continuous variables (II)

2. Discretizing the variable is an attractive option, specially in presence of a limited number of examples:

1. It greatly reduces the amount of noise
2. It makes the probability estimations more reliable
3. However, it may entail a loss of information:
 - a) We may remove more variability than desired
 - b) We loose information on ordering

A simple method for doing this is to use a histogram

Another method is to use a special-purpose algorithm, such as CAIM

The Naïve-Bayes classifier

Continuous variables (III)

3. If the pdfs are not known they can be estimated using the data set non-parametric methods like *kernel density estimation* (KDE)

- A difficult task in the general case, in the univariate case the estimation of the pdf $p(x)$ by kernels is obtained as follows:

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^n \varphi\left(\frac{x - x_i}{h}\right) \quad (1)$$

where $x_1, \dots, x_n$ is an i.i.d. sample of a continuous r.v. X with unknown pdf $p(x)$.

- The weight function (also called Parzen window or *kernel*) φ is defined by

$$\varphi(z) = \begin{cases} \frac{1}{2} & \text{if } |z| \leq 1; \\ 0 & \text{otherwise} \end{cases}$$

- This choice for the kernel is called *uniform* and h is called the *bandwidth*, a parameter for the contribution of each data point x_i to the approximated density at x .

The Naïve-Bayes classifier

Further questions

1. How is the classifier affected by irrelevant variables?

→ Not much affected, given that often they are evenly distributed across the classes

2. How is the classifier affected by redundant variables?

→ Quite affected, given that, assumed independent, their contribution is multiplied

3. Can you show me an example of conditional independence given the class?