

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 kNN classifier

Let X be a set. A *metric* in X is a two-place function $d : X \rightarrow \mathbb{R}^+ \cup \{0\}$ satisfying, for all $x, y, z \in X$:

$$1. \ d(x, y) = 0 \iff x = y$$

$$2. \ d(x, y) = d(y, x)$$

$$3. \ d(x, y) \leq d(x, z) + d(z, y)$$

We say that the pair (X, d) is a metric space.

The kNN classifier

Given a class-labelled data sample $S \subset X$ and a metric d in X , the 1NN technique classifies an $x \in S \setminus X$ by looking at the class of the “nearest neighbour” of x in S , that is:

the class of x is the same as the class of $\arg \min_{x' \in S} d(x, x')$

(ties are broken randomly)

The kNN classifier

The kNN technique is a generalization that considers the $k \geq 1$ “nearest neighbours” of x in S and votes for the most represented class:

the class of x is the same as the majority class of those k elements of S that are closest to x

- ties may happen again and are broken randomly
- for two-class problems and odd k there can be no ties

The kNN classifier

Non-parametric density estimation

Suppose we have an i.i.d. set of (unlabelled) observations $X = \{x_1, \dots, x_N\}$ generated from an unknown pdf $p(x)$.

The goal is to determine an estimation $\hat{p}(x)$ of $p(x)$ based on X

One possible approach is the parametric way (used in LDA/QDA and Naïve Bayes, among others)

Define the probability $\mathcal{P}$ that a specific observation x^* “falls” within a region R as:

$$\mathcal{P} = \int_R p(x) dx$$

The kNN classifier

Non-parametric density estimation

If R is small and p is continuous, we can reasonably assume that p is approximately constant (and equal to $p(x^*)$ in R and hence:

$$\mathcal{P} = \int_R p(x) dx \approx \int_R p(x^*) dx = p(x^*) \int_R dx = p(x^*) V_R$$

where V_R is the *volume* of R .

Therefore we obtain the estimate: $p(x^*) \approx \frac{\mathcal{P}}{V_R}$.

The kNN classifier

Non-parametric density estimation

How do we find $\mathcal{P}$? The probability that k out of the N observations “fall” within R obeys the binomial distribution $Y \sim B(\mathcal{P}, N)$:

$$P(Y = k) = \frac{N!}{k!(N - k)!} \mathcal{P}^k (1 - \mathcal{P})^{N-k}, k \in \{0, \dots, N\}$$

We have $E[Y] = \mathcal{P}N$ and $Var[Y] = \mathcal{P}(1 - \mathcal{P})N$. Since the distribution is unimodal, the expected value is also the most probable value.

Therefore the most probable value for k is $\mathcal{P}N$ and then $\mathcal{P} \approx \frac{k}{N}$.

Therefore we obtain the estimate: $p(x^*) \approx \frac{\mathcal{P}}{V_R} \approx \frac{k}{N \cdot V_R}$.

The kNN classifier

Given x^* , d and X , the kNN technique builds the smallest d -ball centered at x^* that contains k observations in S .

Note that k is chosen beforehand and this choice leads to the region R (the ball)

It has been proved that the probability estimation provided by this method converges to $\mathcal{P}$ when $N \rightarrow \infty$ if:

1. R shrinks sufficiently fast (as a consequence of larger N)
2. k grows sufficiently fast (as a consequence of larger N)

The kNN classifier

This method is the basis of the kNN classifier. Suppose we have an i.i.d. sample S of *labelled* observations, known to belong to a class in $\{\omega_1, \dots, \omega_c\}$.

Call N_i the number of elements in X belonging to class ω_i . Given x^*, d and S , build the smallest d -ball R centered at x^* that contains k observations; this ball has volume V_R .

Call k_i the number of elements in R belonging to class ω_i . We have:

$$\hat{p}(x^*) = \frac{k}{N \cdot V_R}, \quad \hat{p}(x^*|\omega_i) = \frac{k_i}{N_i \cdot V_R}, \quad \hat{P}(\omega_i) = \frac{N_i}{N}$$

$$P(\omega_i|x^*) = \frac{p(x^*|\omega_i)P(\omega_i)}{p(x^*)} \approx \frac{\hat{p}(x^*|\omega_i)\hat{P}(\omega_i)}{\hat{p}(x^*)} = \frac{\frac{k_i}{N_i \cdot V_R} \cdot \frac{N_i}{N}}{\frac{k}{N \cdot V_R}} = \frac{k_i}{k}$$

And the Bayes rule prescribes choosing that ω_i for which $P(\omega_i|x^*)$ is largest, which (using estimated probabilities) is exactly what the kNN classifier does.

The kNN classifier

Cover-Hart theorem

Let ϵ_{1NN} be the probability of error of 1NN and ϵ_B be the Bayes error, then:

$$\epsilon_B \leq \epsilon_{1NN} \leq \epsilon_B \left(2 - \epsilon_B \frac{c}{c-1} \right) \leq 2\epsilon_B$$

in the limit $N \rightarrow \infty$.

In particular, for two-class problems, $\epsilon_B \leq \epsilon_{1NN} \leq 2\epsilon_B(1 - \epsilon_B)$