Aprendizagem 2022

Homework II – Group 019

Diogo Gaspar 99207, Rafael Oliveira 99311

Part I: Pen and paper

1. **Compute the recall of a distance-weighted kNN with $k = 5$ and distance $d(x_1, x_2) = \text{Hamming}(\text{x}_1, \text{x}_2) + \frac{1}{2}$ using leave-one-out evaluation schema (i.e., when classifying one observation, use all remaining ones).**

2. **Considering the nine training observations, learn a Bayesian classifier assuming: i) y_1 and y_2 are dependent, ii) $\{y_1, y_2\}$ and $\{y_3\}$ variable sets are independent and equally important, and iii) y_3 is normally distributed. Show all parameters.**

3. **Under a MAP assumption, compute $P(Positive|x)$ of each testing observation.**

4. **Given a binary class variable, the default decision threshold of $\theta = 0.5$,**

$$f(x|\theta) = \begin{cases} Positive & \textbf{if } P(positive|x) > \theta \\ Negative & otherwise \end{cases}$$

can be adjusted. Which decision threshold – 0.3, 0.5 or 0.7 – optimizes testing accuracy?

Part II: Programming

5. **Using `sklearn`, considering a 10-fold stratified cross validation (`random=0`), plot the cumulative testing confusion matrices of kNN (uniform weights, $k = 5$, Euclidean distance) and Naïve Bayes (Gaussian assumption). Use all remaining classifier parameters as default.**

6. **Using `scipy`, test the hypothesis "kNN is statistically superior to Naïve Bayes regarding accuracy", asserting whether is true.**

7. **Enumerate three possible reasons that could underlie the observed differences in predictive accuracy between kNN and Naïve Bayes.**

Appendix