TÉCNICO LISBOA

Aprendizagem 2022/23

Homework II

Deadline 17/10/2022 (Monday) 23:59 via Fenix as PDF

- Submission Gxxx.PDF in Fenix where xxx is your group number. Please note that it is possible to submit several times on Fenix to prevent last-minute problems. Yet, only the last submission is considered valid
- Use the provided report template. Include your programming code as an Appendix
- Exchange of ideas is encouraged. Yet, if copy is detected after automatic or manual clearance, homework is nullified and IST guidelines apply for content sharers and consumers, irrespectively of the underlying intent
- Please consult the FAQ before posting questions to your faculty hosts

I. Pen-and-paper [13v]

Four positive observations, $\left\{\binom{A}{0}, \binom{B}{1}, \binom{A}{1}, \binom{A}{0}\right\}$, and four negative observations, $\left\{\binom{B}{0}, \binom{B}{0}, \binom{A}{1}, \binom{B}{1}\right\}$, were collected. Consider the problem of classifying observations as positive or negative.

1) [4v] Compute the recall of a distance-weighted kNN with $k = 5$ and distance $d(\mathbf{x}_1, \mathbf{x}_2) = Hamming(\mathbf{x}_1, \mathbf{x}_2) + \frac{1}{2}$ using leave-one-out evaluation schema (i.e., when classifying one observation, use all remaining ones).

An additional positive observation was acquired, $\binom{B}{0}$, and a third variable y_3 was independently monitored, yielding estimates $y_3|P = \{1.2, 0.8, 0.5, 0.9, 0.8\}$ and $y_3|N = \{1, 0.9, 1.2, 0.8\}$.

2) [4v] 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 ii) y_3 is normally distributed. Show all parameters.

Considering three testing observations, $\left\{\left(\begin{pmatrix} A \\ 1 \\ 0.8 \end{pmatrix}, \text{Positive}\right), \left(\begin{pmatrix} B \\ 1 \\ 1 \end{pmatrix}, \text{Positive}\right), \left(\begin{pmatrix} B \\ 0 \\ 0.9 \end{pmatrix}, \text{Negative}\right)\right\}$.

3) [3v] Under a MAP assumption, compute $P(\text{Positive}|\mathbf{x})$ of each testing observation.

4) [2v] Given a binary class variable, the default decision threshold of $\theta = 0.5$,

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

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

II. Programming and critical analysis [7v]

Considering the `pd_speech.arff` dataset available at the course webpage.

5) [3v] 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) [2v] Using `scipy`, test the hypothesis "kNN is statistically superior to Naïve Bayes regarding accuracy", asserting whether is true.

7) [2v] Enumerate three possible reasons that could underlie the observed differences in predictive accuracy between kNN and Naïve Bayes.

END