TÉCNICO LISBOA

Aprendizagem 2024/25

Homework II

Deadline 07/10/2024 (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]

We collected four positive (P) observations, $\{x_1 = (A, 0), x_2 = (B, 1), x_3 = (A, 1), x_4 = (A, 0)\}$, and four negative (N) observations, $\{x_5 = (B, 0), x_6 = (B, 0), x_7 = (A, 1), x_8 = (B, 1)\}$. Consider the problem of classifying observations as positive or negative.

1) [3.0v] Compute the F1-measure of a kNN with $k = 5$ and Hamming distance using a leave-one-out evaluation schema. Show all calculus.

2) [2.5v] Propose a new metric (distance and/or k) that improves the latter's performance (i.e., the F1-measure) by three fold.

An additional positive observation was acquired, $x_9 = (B, 0)$, and a third variable y_3 was independently monitored, yielding estimates,

$$y_3|P = \{1.1, 0.8, 0.5, 0.9, 0.8\} \text{ and } y_3|N = \{1, 0.9, 1.2, 0.9\}.$$

3) [2.5v] 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.

Consider now three testing observations,

$$\{(A, 1, 0.8), (B, 1, 1), (B, 0, 0.9)\}.$$

4) [2.5v] Under a MAP assumption, classify each testing observation showing all your calculus.

At last, consider only the following sentences and their respective connotations,

$$\{("Amazing\ run", P), ("I\ like\ it", P), ("Too\ tired", N), ("Bad\ run", N)\}.$$

5) [2.5v] Using a naïve Bayes under a ML assumption, classify the new sentence "$I\ like\ to\ run$". For the likelihoods calculation consider the following formula,

$$p(t_i|c) = (freq(t_i) + 1)/(N_c + V),$$

where t_i represents a certain term i, V the number of unique terms in the vocabulary, and N_c the total number of terms in class c. Show all calculus.

II. Programming and critical analysis [7v]

Consider the `heart-disease.csv` dataset available at the course webpage's homework tab. Using `sklearn`, apply a 5-fold stratified cross-validation with shuffling (`random_state=0`) for the assessment of predictive models along this section.

1) Compare the performance of a kNN with $k = 5$ and a naïve Bayes with Gaussian assumption (consider all remaining parameters as default):
    a. [1.0v] Plot two boxplots with the fold accuracies for each classifier. Is there one more stable than the other regarding performance? Why do you think that is the case? Explain.
    b. [1.0v] Report the accuracy of both models, this time scaling the data with a Min-Max scaler before training the models. Explain the impact that this preprocessing step has on the performance of each model, providing an explanation for the results.
    c. [1.0v] Using `scipy`, test the hypothesis "the kNN model is statistically superior to naïve Bayes regarding accuracy", asserting whether it is true.

2) Using a 80-20 train-test split, vary the number of neighbors of a kNN classifier using $k = \{1,\ 5,\ 10,\ 20,\ 30\}$. Additionally, for each k, train one classifier using uniform weights and distance weights.
    a. [1.0v] Plot the train and test accuracy for each model.
    b. [1.5v] Explain the impact of increasing the neighbors on the generalization ability of the models.

3) [1.5v] Considering the unique properties of the `heart-disease.csv` dataset, identify two possible difficulties of the naïve Bayes model used in the previous exercises when learning from the given dataset.