TÉCNICO LISBOA

Aprendizagem 2025/26

Homework II

Deadline 08/10/2025 (Wednesday) 23:59 via Fenix as PDF

- Submit Gxxx.zip where xxx is your group number. The zip should contain two files: Gxxx.pdf with your report and Gxxx.ipynb with your notebook. 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]

Consider the following dataset

D	y_1	y_2	y_3	y_4	y_5	y_6
x_1	0.52	0.80	0	1	1	N
x_2	0.53	0.92	0	0	0	N
x_3	0.42	0.48	0	1	1	N
x_4	0.49	0.58	1	0	1	P
x_5	0.62	0.31	0	0	1	P
x_6	0.44	0.38	0	0	1	P
x_7	0.45	0.80	0	0	1	P
x_8	0.50	0.70	0	1	1	N

and the problem of classifying observations as positive (P) or negative (N), according to the target variable y_6.

1) Consider $x_1 - x_6$ to be training observations, $x_7 - x_8$ to be testing observations, and $y_1 - y_5$ to be input variables.

   a. [4.0v] Learn a Bayesian classifier using the training observations assuming: (i) $\{y_1, y_2\}$, $\{y_3, y_4\}$ and $\{y_5\}$ sets of independent variables (e.g., $y_1 \perp\!\!\!\perp y_3$ yet $y_1 \not\perp\!\!\!\perp y_2$), and (ii) $y_1 \times y_2 \in R^2$ is normally distributed. Show all parameters (distributions and priors for subsequent testing).

   b. [2.0v] Under a MAP assumption, classify each testing observation, showing all your calculus.

2) [4.0v] Let $y_3 - y_5$ be the input variables. Compute the accuracy of a kNN with $k = 3$ and Hamming distance using a leave-one-out evaluation schema. Show all your calculus.

   Note: For a given instance, if there is no radius containing only 3 neighbours, consider $k = 2$.

It was presented in class that, as the number of data points available tends to infinity, the error rate of the 1-nearest-neighbor classifier is at most twice the Bayesian error. This bound can be improved to $2 \times E_{Bayes} \times (1 - E_{Bayes})$. In this exercise, we prove that, under certain circumstances, this last bound is strict.

3) Let X be the observations and Θ their true category in a binary task. Let $n \in N$ and $\{x_i, \theta_i\}_{i=1,\ldots,n}$ be a collection of data points independent and identically distributed in $X \times \Theta$. Let $p \in (1/2, 1]$, $P(\Theta = 0) = p$, and, for all x in X's sample space, $P(X = x|\Theta = 0) = P(X = x|\Theta = 1)$ and $P(X = x) > 0$.
    a. [1.0v] Under a MAP assumption, show that the error rate of a Bayesian classifier, $E_{Bayes} = P(\Theta \neq \Theta_{Bayes})$, is given by $1 - p$.
    b. [1.0v] As n tends to infinity, the error rate of the 1-nearest-neighbor classifier, $E_{1NN} = P(\Theta \neq \Theta_{1NN})$, is given by $2 \times P(\Theta = 0|X = x) \times P(\Theta = 1|X = x)$ [1]. Show that $2 \times P(\Theta = 0|X = x) \times P(\Theta = 1|X = x) = 2p(1 - p)$.
    c. [1.0v] Conclude that $E_{1NN} = 2 \times E_{Bayes} \times (1 - E_{Bayes})$.

[1] - T. Cover and P. Hart, "Nearest neighbor pattern classification", in IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, January 1967, doi: 10.1109/TIT.1967.1053964.

II. Programming and critical analysis [7v]

Use the `Breast_cancer_dataset.csv` dataset available at the course's webpage and, using `sklearn`, apply a 5-fold stratified cross-validation with shuffling 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:
    a. [1.0v] Compute the accuracies for each classifier. Which is more stable than the other regarding performance, and why?
    b. [1.0v] Provide the accuracy of the *kNN* model, this time preprocessing the data with a Min-Max scaler before training the model. Explain the impact that this step has on the performance of the model, providing an explanation for the results.
    c. [1.0v] Using `scipy`, determine whether the scaled kNN model is statistically superior to Naïve Bayes (also scaled) when it comes to accuracy, and justify your result.

2) Using a 70-30 train-test split, vary the number of neighbors of a kNN classifier using $k = \{1, 5, 10, 15, 20, 25\}$. Additionally, for each k, train one classifier using uniform weights and distance weights.
    a. [1.0v] Plot the train and test accuracies for each model.
    b. [1.5v] Explain the impact of increasing the number of neighbors on the generalization ability of the models. Elaborate on the *trade-offs* between small and large values of *k* and suggest its optimal value.

3) [1.5v] Suppose you must deploy either kNN or Naïve Bayes in a clinical setting for breast cancer diagnosis. Discuss at least two factors that would influence your choice, referencing insights from your experiments. Comment on the models' performance with medical datasets and the overall models' characteristics from a more technical context (e.g., interpretability, computational cost, or scalability).

Note: Use all optional parameters as default or justify any changes in your answers.