Ensimag – Grenoble INP – UGA Year 2023-2024
Introduction to Statistical Learning and Applications
Pedro L. C. Rodrigues pedro.rodrigues@inria.fr
Alexandre Wendling alexandre.wendling@univ-grenoble-alpes.fr

⚠ General guidelines for TPs

Each team shall upload its report on Teide before the deadline indicated at the course website. Please **include the name of all members of the team** on top of your report.

The report should contain graphical representations. For each graph, axis names should be provided as well as a legend when it is appropriate. Figures should be explained by a few sentences in the text. Answer to the **questions in order and refer to the question number in your report**. Computations and graphics have to be performed with `R`.

The report should be written using the `Rmarkdown` format. This is a file format that allows users to format documents containing text and `R` instructions. You should include all of the `R` instructions that you have used in the `rmd` document so that it may be possible to replicate your results. From your `rmd` file, you are asked to generate an `html` file for the final report. In Teide, you are asked to submit both the `rmd` and the `html` files. In the `html` file, you should limit the displayed `R` code to the most important instructions.

TP 3: Benchmarking classification methods

▶ Part 1

Consider a simulated dataset which you will generate as follows:

- Set the seed of your `R` script with `set.seed(42)`.
- For each data point i, sample its label from a Bernoulli distribution $y_i \sim \mathcal{B}(p)$, i.e. $y_i = 1$ with probability p and $y_i = 0$ with probability $1 - p$. To sample a random variable B from $\mathcal{B}(p)$ you can first sample U from an uniform distribution with function `runif` from the `stats` package and then $B = \mathbf{1}(U < p)$ where $\mathbf{1}(\cdot)$ is an indicator function.
- Then, depending on the label $y_i \in \{0, 1\}$ the associated data point $\mathbf{x}_i \in \mathbb{R}^2$ is sampled as follows:

$$\mathbf{x}_i \mid y_i = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{and} \quad \mathbf{x}_i \mid y_i = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ with pdf

$$p_{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}(x) = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

and

$$\boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\mu}_1 = \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}$$

Note that to sample a p-dimensional vector $\mathbf{x}$ from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, you can use the function `mvrnorm` from the `MASS` package.

We will denote a set of N data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ simulated with ε and p as $\mathcal{D}(N \mid \varepsilon, p)$. Define two datasets:

$$\mathcal{D}_{\text{train}} = \mathcal{D}(50 \mid 1, 0.2) \quad \text{and} \quad \mathcal{D}_{\text{test}} = \mathcal{D}(1000 \mid 1, 0.2) \ .$$

(a) Plot the data points in $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}$ using different colors to indicate the classes of each data point and different pointing symbols to indicate whether a point is from the train or test set. The keyword `pch` from the `plot` function allows you to change the symbol of the scatter plot. For instance, using `pch=1` will give circles and `pch=6` are triangles pointing down. Note that you can overlay scatter plots on an existing figure with the command `points`.

(b) What is the mathematical expression for the optimal Bayes classifier in this setting? And for its boundary region? Remember that the Bayes classifier can be written in terms of the ratio of $\text{Prob}(Y = 1 \mid \mathbf{x})$ over $\text{Prob}(Y = 0 \mid \mathbf{x})$ and that the values of $\mathbf{x} \in \mathbb{R}^2$ for which this ratio is 1 are those defining its boundary.

(c) Estimate the error of the Bayes classifier on the samples from $\mathcal{D}_{\text{test}}$. How you would expect it to change in terms of ε? Plot a curve showing how the Bayes error rate changes for different choices ε (note that you will have to generate new test datasets for this).

(d) Given the structure of the model generating the datasets, which classifier presented in our lectures seems to be the most adequate? Justify your answer in terms of the assumptions behind the construction of each classifier.

(e) Train a LDA, a QDA, and a logistic regression classifier on $\mathcal{D}_{\text{train}}$ and estimate their errors on the samples from $\mathcal{D}_{\text{test}}$. How do their errors compare to the value obtained in (c)? Can we expect the gap between the Bayes error rate and test error for each classifier change when the number of samples in $\mathcal{D}_{\text{train}}$ in change? Justify your answer.

(f) Consider a new test set defined as $\mathcal{D}'_{\text{test}} = \mathcal{D}(1000 \mid 1, 0.8)$. Use the same classifiers trained in (e) and estimate their new test errors. Do you observe any difference in the results? Can you explain what is happening?

▶ Part 2

In this part, we will consider a simulated benchmark similar to that from Section 4.5.2 in James et al presented and discussed in class. Our benchmark will compare the performance of four classifiers under three different scenarios.

– Scenario 1

The observations for this scenario are generated as per:

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{2N} = \{(\mathbf{x}_i, 0)\}_{i=1}^{N} \cup \{(\mathbf{x}_i, 1)\}_{i=1}^{N}$$

with

$$\mathbf{x}_i | y_i = 0 \sim \mathcal{N}(\mu_0, \mathbf{\Sigma}_0) \quad \text{with} \quad \mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

and

$$\mathbf{x}_i | y_i = 1 \sim \mathcal{N}(\mu_1, \mathbf{\Sigma}_1) \quad \text{with} \quad \mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

The training set always have $N = 20$ and the test set $N = 5000$.

- Compare the performances of LDA, logistic regression, Gaussian naive Bayes, and QDA in this scenario. For this, you should generate 100 pairs of training-test datasets and evaluate the test errors for each of the classifiers. Use the command `boxplot` to display the results for each of the classifiers along the different realizations. Explain the differences of the performances in terms of the assumptions of each classifier and the structure of the data generating mechanism.

– Scenario 2

The observations for this scenario are generated as per:

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{2N} = \{(\mathbf{x}_i, 0)\}_{i=1}^{N} \cup \{(\mathbf{x}_i, 1)\}_{i=1}^{N}$$

with

$$\mathbf{x}_i | y_i = 0 \sim \mathcal{N}(\mu_0, \mathbf{\Sigma}_0) \quad \text{with} \quad \mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_0 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 2 \end{bmatrix}$$

and

$$\mathbf{x}_i | y_i = 1 \sim \mathcal{N}(\mu_1, \mathbf{\Sigma}_1) \quad \text{with} \quad \mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_1 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 2 \end{bmatrix}.$$

The training set always have $N = 20$ and the test set $N = 5000$.

- Perform the same comparison as done for Scenario 1.

– **Scenario 3**

The observations for this scenario are generated as per:

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{2N} = \{(\mathbf{x}_i, 0)\}_{i=1}^{N} \cup \{(\mathbf{x}_i, 1)\}_{i=1}^{N}$$

with

$$\mathbf{x}_i | y_i = 0 \sim \mathcal{N}(\mu_0, \mathbf{\Sigma}_0) \quad \text{with} \quad \mu_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_0 = \begin{bmatrix} 1 & -0.7 \\ -0.7 & 2 \end{bmatrix}$$

and

$$\mathbf{x}_i | y_i = 1 \sim \mathcal{N}(\mu_1, \mathbf{\Sigma}_1) \quad \text{with} \quad \mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_1 = \begin{bmatrix} 1 & +0.7 \\ +0.7 & 2 \end{bmatrix}.$$

The training set always have $N = 20$ and the test set $N = 5000$.

- Perform the same comparison as done for Scenarios 1 and 2.