Ensimag – Grenoble INP – UGA
Statistical Analysis and Document Mining
Pedro L. C. Rodrigues
Jacopo Iollo

Year 2022-2023

pedro.rodrigues@inria.fr
jacopo.iollo@inria.fr

Final **practical** exam – Duration: 3h

This exam is composed of four parts.

The maximum final grade is 20.0 points.

Be sure to **read through all of the questions** before starting to solve the exam.

All handwritten documents are allowed, as well as the electronic documents on your computer. No calculator or mobile phone allowed.

No internet connection is available from your computer.

To do this practical exam, you have to use `RStudio`. Open each `rmd` file located in the `$HOME/exam/` directory. Answer the questions directly in each `rmd` file and ensure that you can knit it into an `html` file, and save the `rmd` file. The different `rmd` files correspond to each of the parts of the exam.

Any answer not contained in the `rmd` files will not be taken into account, unless instructed explicitly by you or the teacher into the Rmd file (you can state into a `rmd` file: Please see figure `myfig.png` for example).

The contents of the `$HOME/exam/` directory are

- `exam.pdf` (this document)
- `Part1.rmd`, `Part2.rmd`, `Part3.rmd`, `Part4.rmd` (the files you have to complete and save)
- `/references` (lecture notes, slides, books, etc)

You should regularly **save the current state** of your whole home directory using the icons on your desktop. To do that: (1) Save the files currently open in your editor (e.g. `Rstudio`) then (2) Save your current work and continue the exam. **Before leaving the room** save your current work and stop the session. Please note that the backup operation is based on `rsync`, and does not store every intermediate saved version (in this order!). If your PC crashes, your latest backup can be recovered with the help of the teaching staff.

Good luck!

▶ Part 1: Multiple choice questions (6.0 points)

For each question, write the letter in your `Part1.rmd` file corresponding to the correct answer.

Each question has **exactly one correct** answer.

– Question 1: Robustness to outliers (credits to EPFL CS-433) (1.0 points)

We consider a classification problem on linearly separable data. Our dataset had an outlier – a point that is very far from the other datapoints in distance. We trained the linear discriminant analysis (LDA), logistic regression and 1-nearest-neighbour classifiers on this dataset. We tested the trained models on a test set that comes from the same distribution as the training set, but doesn't have any outlier points. After that, we removed the outlier and retrained our models.

After retraining, which classifier will **not change** its decision boundary around the test points.

(A) Logistic regression.

(B) 1-nearest-neighbors classifier.

(C) LDA.

(D) None of them.

– Question 2: Bias-variance decomposition (credits to EPFL CS-433) (1.0 points)

Consider a regression model where data (x, y) is generated by input $x \in \mathbf{R}$ uniformly sampled between $[0, 1]$ and $y = x + \varepsilon$, where ε is random noise with mean 0 and variance 1. Two models are carried out for regression: model $\mathcal{A}$ is a trained quadratic function $g_{\mathcal{A}}(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2$ and model $\mathcal{B}$ is a constant function $g_{\mathcal{B}}(x) = \frac{1}{2}$.

Compared to model $\mathcal{B}$, model $\mathcal{A}$ has

(A) Higher bias, higher variance.

(B) Lower bias, higher variance.

(C) Higher bias, lower variance.

(D) Lower bias, lower variance.

– Question 3: Linear regression (credits to EPFL CS-433) (1.0 points)

Assume we are doing linear regression with mean-squared loss and L2-regularization on four one-dimensional data points. Our prediction model can be written as $f(x) = ax + b$ and the optimization problem can be written as

$$a^\star, b^\star = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{4} \Big(y_i - f(x_i)\Big)^2 + \lambda a^2$$

Assume that our data points (x_i, y_i) are $\{(-2, 1), (-1, 3), (0, 2), (3, 4)\}$.

What is the optimal value for the bias, $b^\star$?

(A) Depends on the value of λ.

(B) 3

(C) 2.5

(D) None of the above answers.

– Question 4: PCA (credits to EPFL CS-433) (1.0 points)

Which of the following transformations to a data matrix $\mathbf{X}$ will affect the principal components obtained through PCA?

(A) Adding a constant value to all elements of $\mathbf{X}$.

(B) Multiplying one of the features of $\mathbf{X}$ by a constant.

(C) Adding an extra feature to $\mathbf{X}$ (i.e. an extra column) that is constant across all data points.

(D) None of the above answers.

– Question 5: PCA (credits to EPFL CS-433) (1.0 points)

In principal component analysis, the left singular vectors $\mathbf{U}$ of a data matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ are used to create a new data matrix $\mathbf{X}' = \mathbf{U}^\top \mathbf{X}$, where N is the number of data points and p is the number of features. Which property always holds for the matrix $\mathbf{X}'$?

(A) $\mathbf{X}'$ is a square matrix.

(B) The mean of any row $\mathbf{X}'_i$ is 0

(C) $\mathbf{X}'$ has only positive values.

(D) For any two rows i, j with $i \neq j$ from $\mathbf{X}'$, the dot product between the rows $\mathbf{X}'_i$ and $\mathbf{X}'_j$ is 0.

– Question 6: Ridge regularization (credits to EPFL CS-433) (1.0 points)

Assume we have N training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ where each $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.

For $\lambda \geq 0$, we consider the following loss function:

$$\mathcal{L}_\lambda(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2$$

and let $C_\lambda = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{L}_\lambda(\boldsymbol{\beta})$ denote the optimal loss value. Which of the following statements is **true**?

(A) C_λ is a non-increasing function of λ.

(B) For $\lambda = 0$, the loss $\mathcal{L}_0$ is non-convex and might have several minimizers.

(C) C_λ is a non-decreasing function of λ.

(D) None of the above statements are true.

▶ Part 2: Multiple linear regression (4.0 points)

In this part we will be considering the `mtcars` dataset available in `R`. In this dataset, car consumptions are measured in MPGs (miles per gallon): higher MPG values mean that the car is able to do more miles with one gallon of fuel. To determine which car features migth have a positive or negative effect on consumption, we estimate the following multiple linear regression model.

```
M <- lm(mpg~., data=mtcars)
summary(M)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

(a) Which predictors seem to have an effect on `mpg`? Explain why and which aspects of the summary above you have used to make your conclusion. **(1.0 points)**

(b) Provide a 95% confidence interval for the `carb` parameter. What can you deduce from the fact that zero is in the confidence interval? **(1.0 points)**

(c) Explain what the `p-value: 3.793e-07` in the last line of the summary above stands for. What can you conclude based on this value? Does it seem in discordance with your answer in (a)? In this case, what could explain such discordance? **(2.0 points)**

▶ Part 3: Classification (5.0 points)

Consider a simulated dataset which you will generate as follows:

(1) Set the seed of your `R` script with `set.seed(42)`.

(2) 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.

(3) 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:

$$y_i = 0 \quad \Rightarrow \quad \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

$$y_i = 1 \quad \Rightarrow \quad \mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

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

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

> 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 $\mathcal{D}(N \mid \varepsilon, p)$.

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

(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. **(1.0 points)**

> 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.

(b) What is the mathematical expression for the optimal Bayes classifier in this setting? And for its boundary region? **(1.0 points)**

> 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? (**1.0 points**)

(d) Given the structure of the model generating the datasets, which classifier presented in our lectures seems to be the most adequate? **(0.5 points)**

(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 (b)? **(0.5 points)**

(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? **(1.0 points)**

▶ Part 4: Community detection (5.0 points)

In this part you will be using the `igraph` package from TP4 which is already installed in your PC.

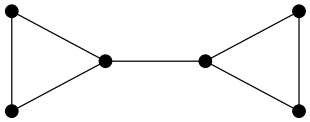
Remember to load it with `library(igraph)`.

The documentation for `igraph` is available at `/references/R/igraph.pdf`

Note that some questions can be easily answered if you've read the paper "Modularity and community structure in networks" (available to you at `/references/articles/article-newman.pdf`)

– Question 1: Modularity maximization (Credits to Mark Newman) (3.0 points)

Consider the small network displayed below



(a) Define in your own words the notion of modularity of a network and how it can be used to split a network into communities. Your description should include whether the modularity depends solely on the structure of the network or not. **(0.5 points)**

(b) Use `igraph` to build the adjacency matrix of the network and then calculate the modularity matrix $\mathbf{B}$ defined in our lectures. Calculate the leading eigenvector of $\mathbf{B}$ and use it to split the network into two groups of nodes. Provide the value of the modularity for the split. **(1.0 points)**

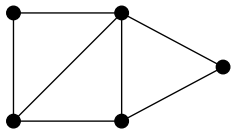
(c) Are there any nodes for which the split in (b) looks more ambiguous than others? Which aspect of the eigenvectors of $\mathbf{B}$ could be useful to check this information? **(0.5 points)**

(d) How would you proceed to split the network into more than two communities? Is your algorithm capable of splitting the network into three networks? **(0.5 points)**

(e) What would happen if all the eigenvalues of the modularity matrix $\mathbf{B}$ were smaller than zero? What would this indicate in terms of the structure of the network? **(0.5 points)**

– Question 2: Hierarchical agglomerative clustering (Credits to Mark Newman) (2.0 points)

Consider the small network with five nodes displayed below



(a) Calculate the cosine similarity for all pairs of nodes. **(1.0 points)**

(b) Using the values of the similarities construct the dendogram for the single-linkage hierarchical agglomerative clustering of the network according to cosine similarity. You will probably need function `hclust` for that. **(1.0 points)**