Final Exam

You should answer precisely the questions below. The answer must be self-contained: you can introduce notations but you need to explain them. The following questions are general but require explanations and details.

1. Show that the solution of PCA when trying to represent the cloud of points of observations in 1 dimension is given by the first eigenvector associated with the first eigenvalue of the covariance matrix.

   Solution. For any dimension $1 \leqslant p \leqslant d$, let $\mathscr{F}_d^p$ be the set of all vector suspaces of $\mathbb{R}^d$ with dimension p. Principal Component Analysis computes a linear span V_p such as

   $$V_p \in \operatorname*{argmin}_{V \in \mathscr{F}_d^p} \sum_{i=1}^n \|X_i - \pi_V(X_i)\|^2 ,$$

   where π_V is the orthogonal projection onto the linear span V. Assume first that $p = 1$ and write $V_1 = \operatorname{span}\{v_1\}$ for $v_1 \in \mathbb{R}^d$ such that $\|v_1\| = 1$. Then,

   $$\begin{aligned}\sum_{i=1}^n \|X_i - \pi_{V_1}(X_i)\|^2 &= \sum_{i=1}^n \|x_i - \langle X_i; v_1\rangle v_1\|^2 ,\\ &= \sum_{i=1}^n \left(\|X_i\|^2 - 2\langle X_i; \langle x_i; v_1\rangle v_1\rangle + \|\langle X_i; v_1\rangle v_1\|^2\right) ,\\ &= \sum_{i=1}^n \left(\|X_i\|^2 - \sum_{j=1}^d \langle X_i; v_1\rangle^2\right).\end{aligned}$$

   Consequently, V_1 is a solution to the PCA optimization problem if and only if v_1 is solution to:

   $$v_1 \in \operatorname*{argmax}_{v \in \mathbb{R}^d ; \|v\|=1} \sum_{i=1}^n \langle X_i, v\rangle^2 .$$

   Note that for all $v \in \mathbb{R}^d$ such that $\|v\| = 1$,

   $$\frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 = \frac{1}{n}\sum_{i=1}^n (v' X_i)(X_i' v) = v' \Sigma_n v ,$$

   where $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ is the empirical covariance matrix. Let $(\vartheta_i)_{1\leqslant i\leqslant d}$ be the orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d \geqslant 0$ of Σ_n. Then,

   $$\frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 = v' \left(\sum_{i=1}^d \lambda_i \vartheta_i \vartheta_i'\right) v = \sum_{i=1}^d \lambda_i \langle v, \vartheta_i\rangle^2 \leqslant \lambda_1 \sum_{i=1}^d \langle v, \vartheta_i\rangle^2$$

   and, as $(\vartheta_i)_{1\leqslant i\leqslant d}$ is an orthonormal basis of $\mathbb{R}^d$, $\sum_{i=1}^d \langle v, \vartheta_i\rangle^2 = \|v\|^2 = 1$. Therefore,

   $$\frac{1}{n}\sum_{i=1}^n \langle X_i, v\rangle^2 \leqslant \lambda_1 .$$

On the other hand, for all $2 \leqslant i \leqslant d$, $\langle \vartheta_1, \vartheta_i \rangle = 0$ and $\langle \vartheta_1, \vartheta_1 \rangle = 1$ so that $\sum_{i=1}^{d} \lambda_i \langle \vartheta_1, \vartheta_i \rangle^2 = \lambda_1$ which proves that ϑ_1 is a solution.

2. Consider the following simulation. We generate a matrix of size 50 rows times 8 colums by drawing each row from a multivariate Gaussian distribution with a diagonal covariance matrix with 1 on its diagonal. Then, we perform a PCA and output the percentage of variability of the two first dimensions. We repeat this simulation 1000 times and take the 95% quantile of the percentages of variability. The obtained value is 41%. Comment the aim of such a simulation. Now you analyze a data set with the same size 50×8 and you obtained a percentage of variability of the two first dimensions of 60 %. Comment this number with respect to the number 41%.

   Solution. The idea here is reminiscent of what you can see if you compute a correlation coefficient between two independent variables but on a small sample size. You can have large values even if the variables are orthogonal. Here, the 95% quantile of the distribution of the percentage of inertia on the two first dimensions for a data set of size 50*8 is 41% when all the variables are independent (under the null). We can use this procedure as a test. When you have a data set of size 50*8, you can compare your percentage of variability to 41%. If this percentage is larger than 41%, we can reject the null hypothesis of abscence of relationship between variables. It means that you have more in your data set than noise, so you can continue and proceed to the analysis.

3. The composition of bottled water is studied using a PCA. For this purpose, 8 chemical concentrations (expressed in mg/l) of 34 waters were recorded.

   Among the following statements, check the ones that are true.

   (a) Waters that contain a lot of magnesium contain a lot of calcium.

   (b) Salvetat water has a lot of nitrates.

   (c) Hepar contains a lot of calcium.

   (d) Wattwiller and St. Georges waters are generally close in composition because they are close in the first factorial plane (plane composed of axes 1 and 2)?

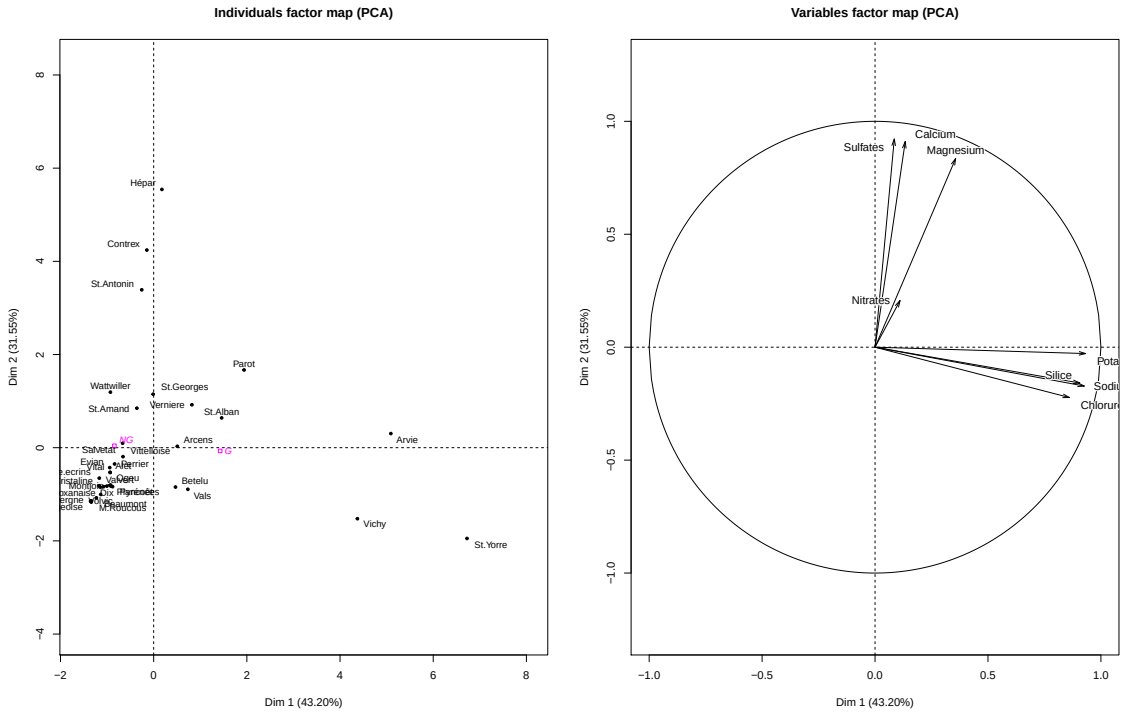
   (e) The waters of Contrex and St. Antonin are generally close in terms of composition because they are close in the first factorial plane (plane composed of axes 1 and 2)?

   Solution. 1-3-5

4. Total inertia, within inertia and between inertia. Tick the good answer(s).

   (a) If all clusters have the same average then the within-class inertia is zero.

   (b) The between inertia is equal to 0 if the averages of all clusters are identical.



(c) If in a clusters all individuals take the same values on all variables, then the within inertia of this class is zero.

(d) The total inertia of a data set does not depend on the number of classes.

Solution. 2-3-4

5. What is the criterion that we want to optimize using the K-means algorithm?

Solution. The K-means algorithm is a procedure which aims at partitioning a data set into K distinct, non-overlapping clusters. Consider $n \geqslant 1$ observations $(X_1, \ldots, X_n)$ taking values in $\mathbb{R}^p$. The K-means algorithm seeks to minimize over all partitions $C = (C_1, \ldots, C_K)$ of $\{1, \ldots, n\}$ the following criterion

$$\mathrm{crit}(C) = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \|X_a - X_b\|^2,$$

where for all $1 \leqslant i \leqslant n$, $1 \leqslant k \leqslant K$, $i \in C_k$ if and only if X_i is in the k-th cluster.

We may also use

$$2 \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \langle X_a, X_a - X_b \rangle = 2 \sum_{k=1}^{K} \sum_{a \in C_k} \|X_a - \bar{X}_{C_k}\|^2,$$

where

$$\bar{X}_{C_k} = \frac{1}{|C_k|} \sum_{b \in C_k} X_b$$

and

$$\begin{aligned}
\text{crit}(C) &= \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in G_k} \|X_a - X_b\|^2 , \\
&= \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \langle X_a - X_b, X_a - X_b \rangle , \\
&= \sum_{k=1}^{K} \frac{1}{|C_k|} \left\{ \sum_{a,b \in C_k} \langle X_a - X_b, X_a \rangle + \langle X_b - X_a, X_b \rangle \right\} , \\
&= 2 \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{a,b \in C_k} \langle X_a - X_b, X_a \rangle .
\end{aligned}$$

6. After applying a clustering algorithm, each observation is assigned to a cluster. Describe precisely how you can describe a cluster using the quantitative variable.

   Solution. We can perform a test by comparing the average of a variable for the individuals in the cluster q, to the average of the variable for all of the individuals. The null hypothesis of the test is: the values of the variables for the individuals in the cluster q are selected at random from all of the possible values of the variable. The null hypothesis of the test is the average of the variable for a clsuter q is equal to the general average, or in other words, the variable does not characterise the clsuter q. This test is only an indication and valid for supplementary variables. Indeed, active variables have been used to create the clsuster.

7. When describing the clusters obtained using qualitative variables:

   (a) a chi-square test helps to see whether overall, a qualitative variable characterizes well the clustering

   (b) a certain category characterizes a cluster well if it is over-represented in that cluster

   (c) if a category is very common, it will characterize many clusters well

   Solution. 1-2. A category characterizes a cluster well if the proportion of individuals in that category is very large or very small in this cluster compared with the overall proportion of individuals in that category in the whole data set. Therefore, being common does not imply that a category will characterize a cluster or clusters well.

8. High-dimensional data and clustering. Tick the good answer(s).

   (a) When there are many individuals (observations n), we can run a hierarchical clustering before doing K-means.

   (b) When there are many individuals, we can first group individuals together using K-means, then run hierarchical clustering.

   (c) When there are many variables, we can run a principal component method and retain a smaller number of dimensions, with which we can then run a clustering algorithm.

   Solution. 2-3

9. Let us consider a data set with two variables X_1 and X_2. There are missing values in X_1. The aim is to estimate the mean of the variable X_1. The missing values are MCAR. If you delete the observation with missing values and compute the mean of the variable X_1 on the observed values, what are the properties of your estimator in terms of biais and variance in comparison to the one that we would have obtained on the completed data. If the missing values are MNAR such that values are missing when X_1 is larger than a threshold, what are the properties of your estimator?

   Solution. MCAR: the estimator would be unbiased, we have a sample of a smaller size but representative of our sample, but the variance will be larger since we have less data. MNAR: in this setting, the estimator will be biases (downward) as the sample is not representative of the data.

10. In comparison to mean imputation, what can be the advantages of imputation with PCA?

    Solution. In comparison to the mean imputation, the imputation by PCA is based on the scores and the loadings and consequently take into account the similaties between the observations and the relationship between variables, so it can be seen as a mixed between imputation by nearest neightbor and imputation based on regression, so it will improve the quality of prediction of the value but it is also a single imputation method so it does not reflect the uncertainty associated to the prediction of a value.

11. When you suggest methods to deal with missing values to users, the recurrent question is ”What is the percentage of missing values that I can have in my data set, is 50% too much but 20% OK?” What is your answer to this question?

    Solution. The percentage of missing values is not the only thing which is important. If the variables are highly correlated, we can predict the missing values precisely even with a high fraction of missing values. On the contrary, if the data set is very noisy to begin with, even a small fraction of missing values can be troublesome. Multiple imputation can always be performed and enables to measure precisely the variability of the predictions, which evaluates how much we can trust the results obtained from a (very) incomplete dataset.

12. The logistic regression model assumes that the random variables $(\mathbf{X}, Y) \in \mathbb{R}^p \times \{0,1\}$ are such that

$$\mathbb{P}(Y=1|\mathbf{X}) = \frac{\exp(\mathbf{X}^t\beta)}{1+\exp(\mathbf{X}^t\beta)}, \tag{1}$$

with $\beta \in \mathbb{R}^d$. Give the expression of the conditional loglikelihood ℓ_n of n observations $\{(\mathbf{X}_i, Y_i)\}_{1\leqslant i\leqslant n}$.

Solution. Since the observations are assumed to be independent, the likelihood writes, for all $\beta \in \mathbb{R}^p$,

$$L_n(\beta) = \prod_{i=1}^n \left(\frac{e^{\mathbf{X}_i^t\beta}}{1+e^{\mathbf{X}_i^t\beta}}\right)^{Y_i} \left(\frac{1}{1+e^{\mathbf{X}_i^t\beta}}\right)^{1-Y_i}.$$

Therefore, the loglikelihood is, for all $\beta \in \mathbb{R}^p$,

$$\ell_n(\beta) = \log(L_n(\beta)) = \sum_{i=1}^n \left\{ Y_i \mathbf{X}_i^t\beta - \log(1+e^{\mathbf{X}_i^t\beta}) \right\}.$$

13. Compute the gradient of the function $\beta \mapsto -\ell_n(\beta)$ and prove that this function is convex.
Solution. The gradient of the negative loglikelihood is given, for all $\beta \in \mathbb{R}^p$,

$$-\frac{1}{n}\nabla\ell_n(\beta) = -\frac{1}{n}\sum_{i=1}^n Y_i\mathbf{X}_i + \frac{1}{n}\sum_{i=1}^n \frac{\exp(\mathbf{X}_i'\beta)}{1+\exp(\mathbf{X}_i'\beta)}\mathbf{X}_i = -\frac{1}{n}\sum_{i=1}^n \nabla f_i(\beta),$$

where, for all $\beta \in \mathbb{R}^p$,

$$f_i(\beta) = Y_i\mathbf{X}_i^t\beta - \log(1+e^{\mathbf{X}_i^t\beta}).$$

The function is convex because its Hessian matrix is positive as it is given by

$$H = \frac{1}{n}\sum_{i=1}^n \frac{\exp(\mathbf{X}_i^t\beta)}{(1+\exp(\mathbf{X}_i^t\beta))^2}\mathbf{X}_i\mathbf{X}_i^t.$$

14. Provide a stochastic gradient descent algorithm to minimize $\beta \mapsto -\ell_n(\beta)$.

Solution. A stochastic gradient descent algorithm to minimize $\beta \mapsto -\ell_n(\beta)$ proceeds for instance as follows.

(a) Choose $\beta_0 \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}_+^*$.

(b) At each iteration $k \geqslant 1$, choose randomly $i_k \in \{1,\dots,n\}$ and set

$$\beta_k = \beta_{k-1} + \frac{\gamma}{\sqrt{k}}\nabla f_{i_k}(\beta_{k-1}).$$

15. What is the difference between a gradient method to estimate the parameters of a logistic regression, a Newton method and a stochastic gradient method?

Solution. A gradient algorithm computes at each iteration the gradient of the loglikelihood: for all $k \geqslant 1$,

$$\beta_k = \beta_{k-1} + \frac{\gamma}{n} \nabla \ell_n(\beta_{k-1}) \,.$$

A more efficient approach consists in approximating this gradient by a noisy approximation and replacing this update by the stochastic gradient update given in the previous question. At each iteration the gradient of only one function f_i is required so that the complexity is of order p while the complexity of one iteration of the gradient algorithm is of order np.

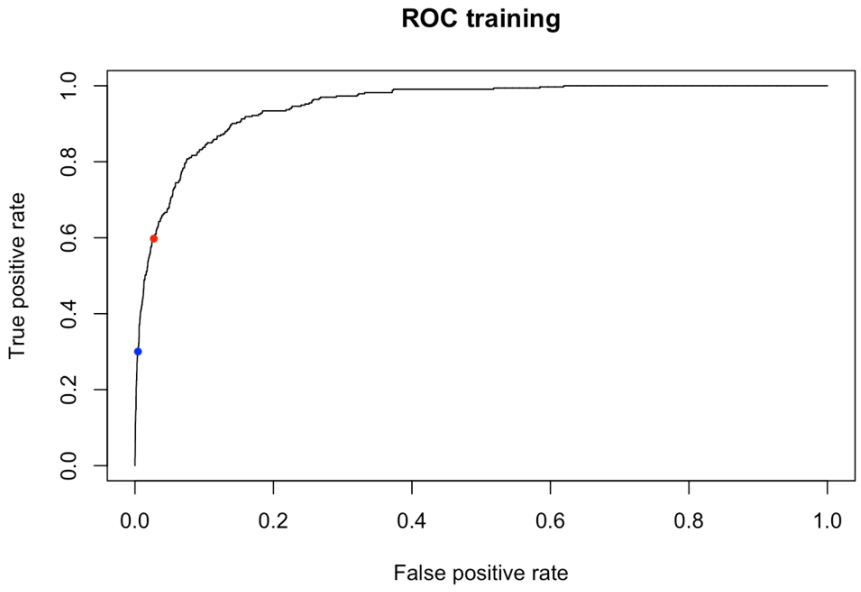
Newton's approcah minimizes the *best locally quadratic approximation* of $-\ell_n/n$. By Taylor's expansion, for $h \in \mathbb{R}^p$:

$$\ell_n(\beta_{k-1} + h) \simeq \ell_n(\beta_{k-1}) + \nabla \ell_n(\beta_{k-1})^T h + \frac{1}{2} h^T \nabla^2 \ell_n(\beta_{k-1}) h \,.$$

β_k is then obtained by minimizing $h \mapsto -\ell_n(\beta_{k-1}) - \nabla \ell_n(\beta_{k-1})^T h - \frac{1}{2} h^T \nabla^2 \ell_n(\beta_{k-1}) h$.

16. The usual logistic regression classifier is defined by $h_n : \mathbf{x} \mapsto 1$ is $\mathbf{x}' \hat{\beta}_n > 0$ and 0 otherwise, where $\hat{\beta}_n$ is the maximum likelihood estimator of β. Therefore $h_n(\mathbf{x}) = 1$ if and only if $\mathbb{P}(Y = 1 | X = \mathbf{x}) > 1/2$. Other classifiers can be defined by setting $h_n(\mathbf{x}) = 1$ if and only if $\mathbb{P}(Y = 1 | X = \mathbf{x}) > p_*$ for a chosen $p_* \in (0,1)$. Two classifiers were built with $p_* = 0.5$ and $p_* = 0.2$, associate each classifier with its point on ROC curve displayed above.

Solution. $p_* = 0.2$ corresponds to the red dot and $p_* = 0.5$ to the blue dot.



For American families, we have the annual income and age of the car available; we are trying to explain the purchase variable, which indicates whether the family plans to buy a new car in the coming year.

17. Let consider the output. What is the sample size of this data?

Solution. The sample size is $n = 33$.

```
> d = read.table("car_income.txt", header=TRUE)
> g = glm(purchase ~ income + age, data=d, family=binomial)
> summary(g)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6189  -0.8949  -0.5880   0.9653   2.0846

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73931    2.10195  -2.255   0.0242 *
income       0.06773    0.02806   2.414   0.0158 *
age          0.59863    0.39007   1.535   0.1249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 36.690  on 30  degrees of freedom
AIC: 42.69

Number of Fisher Scoring iterations: 4
```

18. How the value *z value* 2.414 is computed in the R output above ? Provide the statistical test associated with this statistic.

Solution. This value is the value of the Wald statistic ϖ for the variable income which is computed as the ratio of the estimaed parameter over its estimated standard deviation:

$$\varpi = \frac{0.06773}{0.02806} = 2.414\,.$$

This is the sample based value of a random variable with standard Gaussian distribution under the null hypothesis $\beta_1 = 0$.

19. Interpret the value 0.06773.

Solution. In this case, the unknown parameter β lies in $\mathbb{R}^3$ and the approximate maximum likelihood estimated by *glm* is $\widehat{\beta} = (-4.73931, 0.06773, 0.59863)$, where $\widehat{\beta}_1 = 0.06773$ is the estimated parameter associated with the variable *income*.

20. Linear discriminant analysis assumes that the random variables $(X,Y) \in \mathbb{R}^p \times \{0,1\}$ has the following distribution. For all $A \in \mathscr{B}(\mathbb{R}^p)$ and all $y \in \{0,1\}$,

$$\mathbb{P}(X \in A; Y = y) = \pi_y \int_A g_y(x)\mathrm{d}x\,,$$

where π_0 and π_1 are positive real numbers such that $\pi_0 + \pi_1 = 1$ and g_0 (resp. g_1) is the probability density of a Gaussian random variable with mean $\mu_0 \in \mathbb{R}^d$ (resp. μ_1) and positive definite covariance matrix Σ. Prove that the Bayes classifier $h_* : \mathbb{R}^p \to \{0,1\}$ is defined by

$$h_* : x \mapsto \mathbb{1}_{\{\pi_1 g_1(x) > \pi_0 g_0(x)\}}\,.$$

Solution. For all $A \in \mathscr{B}(\mathbb{R}^p)$,

$$\begin{aligned}\mathbb{P}(X \in A) &= \mathbb{P}(Y=0)\mathbb{P}(X \in A|Y=0) + \mathbb{P}(Y=1)\mathbb{P}(X \in A|Y=1)\,,\\ &= \pi_0 \int_A g_0(x)\mathrm{d}x + \pi_1 \int_A g_1(x)\mathrm{d}x\,.\end{aligned}$$

The probability density of the random variable X is given, for all $x \in \mathbb{R}^d$, by

$$g(x) = \pi_0 g_0(x) + \pi_1 g_1(x)\,.$$

Then, note that for all $x \in \mathbb{R}^d$,

$$\eta(x) = \mathbb{P}(Y=1|X)|_{X=x} = \frac{\mathbb{P}(X|Y=1)|_{X=x}\,\mathbb{P}(Y=1)}{g(x)} = \frac{\pi_1 g_1(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)}\,,$$

and the condition $\eta(x) \leqslant 1/2$ can be rewritten as

$$\frac{\pi_1 g_1(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)} \leqslant 1/2\,,$$

that is $\pi_1 g_1(x) \leqslant \pi_0 g_0(x)$.

21. Prove that the Bayes classifier of the previous question is given, for all $x \in \mathbb{R}^d$, by:

$$h_*(x) = 1 \Leftrightarrow (\mu_1 - \mu_0)^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_0}{2}\right) > \log(\pi_0/\pi_1)\,.$$

Solution. For all $x \in \mathbb{R}^d$,

$$\begin{aligned}&\pi_1 g_1(x) > \pi_0 g_0(x)\\ &\quad\Leftrightarrow \log(\pi_1 g_1(x)) > \log(\pi_0 g_0(x))\,,\\ &\quad\Leftrightarrow -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) > \log(\pi_0/\pi_1)\,,\\ &\quad\Leftrightarrow -\frac{1}{2}\Big(-\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}\mu_1 + \mu_0^T\Sigma^{-1}x - \mu_0^T\Sigma^{-1}\mu_0 + x^T\Sigma^{-1}\mu_0\Big) > \log(\pi_0/\pi_1)\,,\\ &\quad\Leftrightarrow x^T\Sigma^{-1}\mu_1 - x^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 > \log(\pi_0/\pi_1)\,,\\ &\quad\Leftrightarrow (\mu_1-\mu_0)^T\Sigma^{-1}\left(x - \frac{\mu_1+\mu_0}{2}\right) > \log(\pi_0/\pi_1)\,.\end{aligned}$$

Therefore, all $x \in \mathbb{R}^d$ is classified according to its position with respect to an affine hyperplane orthogonal to $\Sigma^{-1}(\mu_1 - \mu_{-1})$.

22. Explain the difference between linear discriminant analysis and quadratic discriminant analysis.

Solution. In the LDA setting both Gaussian distributions have the same covariance matrix Σ while in the QDA framework the covariance matrices are different in the two groups.

23. In the Hard Support Vector Machines (SVM) framework, explain what is a linearly separable dataset. Give one expression of the Soft SVM optimization problem when the dataset is not linearly separable.

Solution. The hard Support Vector Machines is a classification procedure which can be used when a dataset is linearly separable, i.e. if the training dataset $(X_i)_{1\leqslant i\leqslant n}$ can be correctly splitted into two groups using an hyperplane.

The hard Support Vector Machines can be reduced to a quadratic optimization problem with linear constraints, which may be solved in a reasonable computational time, when the training data sets is linearly separable. Restricting the problem to linearly separable training data sets is a somehow strong assumption. Soft Support Vector Machines algorithm introduces a relaxation of this constraint which can be applied with nonlinearly separable data sets. In this setting, the inequality constraints in the quadratic optimization problem can be relaxed by introducing nonnegative variables $(\xi_i)_{1\leqslant i\leqslant n}$ which quantify for each variable $1 \leqslant i \leqslant n$, the nonfeasability of the constraint $Y_i(\langle w\,;\,X_i\rangle + b) \geqslant 1$. Therefore, the soft Support Vector Machines algorithm simultaneously the margin of the linear classifier and the average value of these slack variables $(\xi_i)_{1\leqslant i\leqslant n}$:

$$(w_\star, b_\star, \xi_\star) \in \underset{\substack{(w,b,\xi)\in\mathbb{R}^d\times\mathbb{R}\times\mathbb{R}_+^d \\ \forall i\in\{1,\dots,n\},\, Y_i(\langle w\,;\,X_i\rangle+b)\geqslant 1-\xi_i}}{\operatorname{argmin}} \left\{ \lambda \|w\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i \right\},$$

where $\lambda > 0$.

24. What is a kernel base SVM ? What does 62 means in the R output below ?

Solution. Let $k : \mathsf{X}\times\mathsf{X} :\to \mathbb{R}$ be a positive definite kernel and $\mathscr{F}$ the RKHS with kernel k. Then, kernel based SVM amounts to solving the following optimization problem:

$$\widehat{f}^n_{\mathscr{F}} \in \underset{f\in\mathscr{F}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^n (1 - Y_i f(X_i))_+ + \lambda \|f\|^2 .$$

The solution is of the form $\widehat{f}^n_{\mathscr{F}} : x \mapsto \sum_{i=1}^n \widehat{\alpha}_i k(X_i, x)$, where, for all $1 \leqslant i \leqslant n$, $\widehat{\alpha}_i \in \mathbb{R}$. A vector X_i is said to be a suuport vector if and only if it is actually involved in the definition of $\widehat{f}^n_{\mathscr{F}}$, i.e. if $\widehat{\alpha}_i \neq 0$.

```
library(ISLR)
library(e1071)
khan_frame <- data.frame(x=Khan$xtrain, y=as.factor (Khan$ytrain ))
khan_svm   <- svm(y~., data = khan_frame, kernel = "radial")

summary(khan_svm)
```

```
##
## Call:
## svm(formula = y ~ ., data = khan_frame, kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.0004332756
##
## Number of Support Vectors:  62
##
##  ( 22 20 12 8 )
##
##
## Number of Classes:  4
```

The value 62 means that there are exactly 62 support vectors, i.e. 62 X_i involved in the definition of the kernel based classifier.