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. In general, the following questions are general but require explanations and details.

1. What are the aims of PCA? Could you mathematically explain what is PCA?

   Solution. Principal Component Analysis is a dimension reduction techniques which aims at
   - studying similarities between observations (in the sense of the euclidian distance) from a multidimensional point of view
   - studying linear relationship between variable
   -relate both studies (describe groups of observations with the variables) PCA also finds "synthetic variables", that are the best linear combinations of input variables - they explain the most the data variance, i.e. they discriminate the most the observations. Such variables could be used as input for other machine learning method as linear regression (data preprocessing).

   The ultimate objective of PCA can be to visualize the observations and variables (by projecting the lcouds of points of variables and observations into the first two dimensions).

   Mathematically, a PCA consists in diagonalizing the correlation matrix and finding the largest eigenvalues/eigenvectors of this matrix. The eigenvector, sorted by decreasing order of their eigenvalues, will be the first principal component saxis. The associated eigenvalues correspond to the inertia explained by each dimension.

2. How are represented the quantitatives supplementary variables in PCA? **Solution.** The quantitative supplementary variables are represented on the correlation circle by an arrow. The coordinates are calculated by computing the correlation between the variables and the axis (v). It boils down to projecting these variables on the low-dimensional subspace.

3. In the correlation circle graph, explain how you can interpret the size of an arrow associated to a specific variable?

   Solution. The correlation circle is a representation of the variables. Each variable can be described by the vector of I entries, each entry corresponding to the value of the variable for a specific individual of the data set. The variables plotted in the correlation circle graph are the projections of the variables (belonging to $\mathbb{R}^I$) onto the space spanned by the first two principal components of PCA. Since, prior to any PCA, each variable is standardized, the variables in $\mathbb{R}^I$ are vectors of norm 1. Thus, we can see if the projection is close to the original vector by looking at the size of the arrow in the correlation circle: if the arrow is close to one, the projection is close to the original variable and we can interpret its correlation to the other variables that are well-represented in the correlation circle.

4. Explain the general principle of hierarchical clustering.

   Solution. Hierarchical clustering is a sequential clustering method. The algorithm starts with I clusters, where I is the number of observations, each cluster being composed of exactly one observation. Then the two closest clusters are merged according to some metric and the procedure is iterated until there is only one cluster left.

5. Give the definition of the total inertia, between inertia and within inertia. Give a relation between these three quantities. Prove this relation. You can denote Q the number of clusters, I the number of observations and I_q the number of obserbations in cluster q.

   Solution. Let Q be the number of clusters, I the number of observations, I_q the number of observations in cluster q. The total inertia is defined as

$$\text{Total inertia} = \sum_{i=1}^{I} ||x_i - \bar{x}||^2.$$

The following relation holds

$$\text{Total inertia} = \text{Between inertia} + \text{within inertia}$$

where

$$\text{Between inertia} = \sum_{q=1}^{Q} I_q ||\bar{x}_q - \bar{x}||^2.$$

and

$$\text{Within inertia} = \sum_{q=1}^{Q} \sum_{i=1}^{I} ||x_i - \bar{x}_q||^2 \mathbb{1}_{x_i \in C_q}.$$

Indeed,

$$\begin{aligned}
\text{Total inertia} &= \sum_{i=1}^{I} ||x_i - \bar{x}||^2 \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I} ||x_i - \bar{x}||^2 \mathbb{1}_{x_i \in C_q} \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I} ||x_i - \bar{x}_q + \bar{x}_q - \bar{x}||^2 \mathbb{1}_{x_i \in C_q} \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I} ||x_i - \bar{x}_q||^2 \mathbb{1}_{x_i \in C_q} + \sum_{q=1}^{Q} \sum_{i=1}^{I} ||\bar{x}_q - \bar{x}||^2 \mathbb{1}_{x_i \in C_q} \\
&\quad + 2 \sum_{q=1}^{Q} \sum_{i=1}^{I} \langle x_i - \bar{x}_q, \bar{x}_q - \bar{x} \rangle \mathbb{1}_{x_i \in C_q} \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I} ||x_i - \bar{x}_q||^2 \mathbb{1}_{x_i \in C_q} + \sum_{q=1}^{Q} I_q ||\bar{x}_q - \bar{x}||^2 \mathbb{1}_{x_i \in C_q},
\end{aligned}$$

since

$$\sum_{i=1}^{I} \langle x_i - \bar{x}_q, \bar{x}_q - \bar{x} \rangle \mathbb{1}_{x_i \in C_q} = \langle \sum_{i=1}^{I} x_i \mathbb{1}_{x_i \in C_q} - I_q \bar{x}_q, \bar{x}_q - \bar{x} \rangle$$
$$= 0.$$

6. What criterion can we use to assess the performance of a clustering method? What value of this criterion corresponds to a good clustering? Comment on this.

   Solution. We can use the ratio Between inertia/Total inertia to assess the quality of a clustering method. If this ratio is close to one then almost all inertia is explained by the clusters (if we summarize each point by the centroid of its cluster, we do not lose much variance). On the other hand, if this ratio is close to zero the clustering is almost non informative, almost all variance being carried out by the variability of points inside each cluster. Since there exists no ideal clustering, there exist no threshold value for this ratio. We should always take into account the interpretability of the clustering together with the value of this ratio in order to assess the quality of the clustering.

7. Computationally speaking, would you prefer to use Kmeans or hierarchical clustering? Can you think of a method to draw benefits from both methods?

   Solution. Since hierarchical clustering starts with one cluster per observation, hierarchical clustering is far more computationally intensive than Kmeans if the number of clusters K in Kmeans is small compared to I, which is usually the case. However, hierarchical clustering allows us to choose the number of clusters after running the algorithm. To draw benefits from both algorithms, one could first run Kmeans with a reasonable number of clusters and then run hierarchical clustering on the clusters output by Kmeans. In this way, hierarchical clustering would start with K clusters and not with I clusters.

8. 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.

9. Describe a single imputation method. What is the aim of single imputation? What is the drawback of single imputation? When applying a statistical analysis on a dataset that has been completed, is it possible to get an unbiased estimator?

   Solution. A possible single imputation method consists in replacing all missing values by the mean over the variable for which the entries are missing. The aim of single imputation is to provide a good estimation of the missing values. However, is does not take into account the variability of the data set: single imputation can lead to unbiased estimates but induces an underestimation of the variance.

10. How can you assess the performance of a single imputation method?

    Solution. You can use a cross-validation strategy: you remove some available entries, you predict them with the 3 imputation methods and you compute the errors of prediction. You repeat this procedure say K-times. You select the methods which minimizes the prediction error.

11. Your aim is to perform a PCA and you have missing values in the data set. What can you do? Describe it.

    Solution. You can start by imputing missing values using the iterative PCA algorithm and then apply a regular PCA to the completed data set.

12. Describe an application of unsupervised learning and an application of supervised learning.

    Solution. Clustering (unsupervised learning) can be used to create groups of customers in order to find pattern in their consumption behaviour. Predictive algorithm as logistic regression can be used to detect risk of developing such disease by measuring some relevant biological quantities of interest.

13. Explain what is a SVM?

    Solution. Support Vector Machine is a supervised learning algorithm that works on linearly separable data set. It finds the best hyperplane that separates the two classes in the data set by maximizing the margin defined as the distance between the hyperplane and the closest data points to the hyperplane. In other words, SVM outputs an hyperplane of maximal margin that correctly classifies the observations if they are linearly separable.

14. Explain why the optimization problem in SVM is computationally quick to solve.

    Solution. The optimization problem consists in minimizing a convex function on a convex constraint space. This problem can thus be solved by a gradient descent algorithm. Furthermore, the problem is quadratic which makes the optimization even quicker.

15. What are the main assumptions in Linear Discriminant Analysis?

Solution. Linear Discriminant Analysis assumes that the distribution of $\mathbf{X}|Y$ is Gaussian, that is

$$\mathbf{X}|Y = 0 \sim \mathcal{N}(\mu_0, \Sigma_0) \quad \text{and} \quad \mathbf{X}|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma_1)$$

and that the covariance matrices Σ_0 and Σ_1 are equal.

16. Give the definition of the Bayes classifier for the risk $\mathbb{P}[Y \neq f(X)]$. Give also its expression in terms of probability.

Solution. The Bayes classifier $f^\star$ is by definition the function that minimizes the risk, that is

$$f^\star \in \operatorname*{argmin}_f \mathbb{P}[Y \neq f(X)].$$

As seen in the course, we can prove that this is equivalent to

$$f^\star(\mathbf{x}) = \begin{cases} +1 & \text{if } \eta(\mathbf{x}) \geq 1/2 \\ -1 & \text{otherwise} \end{cases},$$

where $\eta(\mathbf{x}) = \mathbb{P}[Y = 1|\mathbf{X} = \mathbf{x}]$.

17. Give the expression of the logistic model with one continuous input variable $X \in \mathbb{R}$ and one binary variable $Y \in \{0, 1\}$. Give the values of X such that the probability that $Y = 1$ is larger than 0.5.

Solution. In this context, a logistic model is of the form

$$\log\left(\frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)}\right) = \beta_0 + \beta_1 X.$$

Let $p(X) = \mathbb{P}[Y = 1|X]$. Since the function $p \mapsto \log(p/(1-p))$ is non-decreasing, we have

$$\begin{aligned} & p(X) \geq 1/2 \\ \Leftrightarrow & \log\left(\frac{p(X)}{p(X)}\right) \geq 0 \\ \Leftrightarrow & \beta_0 + \beta_1 X \geq 0 \\ \Leftrightarrow & X \geq -\beta_0/\beta_1, \end{aligned}$$

if $\beta_1 > 0$ and $X < -\beta_0/\beta_1$ otherwise.

18. Give the definition of the odd ratio for the variable X in the previous model. Give the odd ratio expression as a function of the logistic regression coefficient.

    Solution. In the previous model, the odd ration associated to X is given by

$$\text{OR} = \left(\frac{\frac{\mathbb{P}(Y=1|X=x+1)}{\mathbb{P}(Y=0|X=x+1)}}{\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}} \right) = \exp(\beta_j).$$

19. Give the expression for the coefficients in logistic regression. How are these coefficients computed in practice?

    Solution. Contrary to the linear regression we do not have a close form for estimating the coefficients of logistic regression. In practice, the coefficients are estimated via maximum likelihood obtained by a gradient descent procedure as Iterative Reweighted Least Squares - Newton-Raphson.

20. Describe how the ROC curve is built. What is the ideal ROC curve? What is the corresponding value for the AUC?

    Solution. A point of the ROC curve corresponds to a particular threshold of the estimated probability output by logistic regression model. The coefficient in logistic regression are first computed, then the estimated probabilities are also computed for each point in the data set. Then, we need to threshold probabilities to obtain a $0-1$ prediction. For each choice of the threshold the performance of the resulting classifier are computed via the false positive rate (x-axis of the ROC curve) and the true positive rate (y-axis of the ROC curve). The ideal ROC curve is the broken line that goes by $(0,0),(0,1),(1,1)$ since the point $(0,1)$ corresponds to a true positive rate of one and a false positive rate of 0. The Area Under the Curve (AUC) is equal in that case to 1.

21. Give the equation to describe a regular gradient descent. What are the condition(s) for this procedure to converge to the global minimum?

    Solution.

    Assume we want to optimize a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ where we have access to the gradient. A gradient descent algorithm writes:

```
Initialize: $x_0$, k=0
while not stop
    compute descent direction: $d_k = -\nabla f(x_k)$
    choose a step-size $\sigma_k$ (via line search for instance)
    update: $x_{k+1} = x_k + \sigma_k d_k$
    $k = k+1$
```

Stopping criteria: can be based on $\|\nabla f(x_k)\| \leq \varepsilon$; $\|x_{k+1} - x_k\| \leq \varepsilon$ and $\frac{|f(x_{k+1})-f(x_k)|}{|f(x_k)|} \leq \varepsilon$
Line search: based on Armijo-rule or backtracking line search
To ensure the convergence of gradient descent, the function f has to be differentiable and convex.

22. Assume you optimize a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Define the Newton direction. When would you like to use this direction in an optimization algorithm?

    Solution. The Newton direction is given by $-\nabla^2 f(x_k)\nabla f(x_k)$ (minus the inverse of the Hessian matrix times the gradient). This direction points towards the global optimum on convex-quadratic functions. In the case of ill-conditioned problems the direction given by minus the gradient and by the Newton direction will differ a lot. A gradient descent algorithm will be slow and we will prefer to use Newton's methods (using Newton's direction).

23. Explain the motivation and the main ideas behind quasi-Newton's methods.

    Solution. Very often, we do not have access to the Hessian matrix and estimating it is expensive (at it requires estimating of the order of n^2 parameters). Quasi-Newton's methods approximate Newton's direction $-\nabla^2 f(x_k)\nabla f(x_k)$ using only gradient information. The main idea is that successive iterates x_k, x_{k+1} and gradients $\nabla f(x_k)$ yield second order information via:

    $$q_k \approx \nabla^2 f(x_{k+1}) p_k$$

    where $p_k = x_{k+1} - x_k$ and $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.

    One standard quasi-Newton's method is the BFGS algorithm.

Solution.