Final Exercise

Question 1:

The first exercise will be a collection of short questions, such as

- Recall that the cost-optimal Bayesian mapping for a new observation $\mathbf{x}$ is given by

$$\delta_C(\mathbf{x}) = r \;\leftrightarrow\; \sum_{k=1}^{K} c_{kr} P(k|\mathbf{x}) = \min_{j\in\{1,\ldots,K\}} \sum_{k=1}^{K} c_{kj} P(k|\mathbf{x}), \quad r = 1, \ldots, K\,.$$

  is given. How do the costs c_{kr} have to be chosen so that the Bayes assignment in each case corresponds to

  - of the maximum likelihood (ML) assignment and
  - of the standard Bayes assignment (classification)?

Question 2:

a) As part of a study, objects are to be grouped meaningfully according to similarity criteria. The following objects were observed:

   i. Twelve bytes (1 byte = 8 bits) e.g. [10001010] vs. [11001010] vs. [00101010] vs. ....

   ii. South African sharks (regarding strongly correlated measurements of weight and body length)

   iii. Berlin kebab stores (regarding location/coordinates)

   iv. Exams of two statistics students (plagiarism detection)

   v. Spanish first names

   vi. Two Weibull distributions (with different parameters) in the context of an analysis of lifetimes

   vii. Stores in Munich city center (regarding standardized, uncorrelated measurements of annual sales and time since opening)

   As an expert, you can suggest suitable similarity measures **for 3 cases of your choice**. Mention a possible distance or similarity measure and justify your decision for these three cases.

b) Give a function that is a metric (you may choose both known functions or your own trivial examples) and show that this functions fulfills all requirements of a metric.

Question 3:

a) Write out the model formula and assumptions for a logistic regression model.

b) Place the model from (a) in the context of (un-)supervised learning. Explain how it can it be used for both regression and classification.

Next, consider the following data as given, where "Yes"represents the outcome of interest:

predicted_prob_of_Yes	actual_outcome
0.31	No
0.79	Yes
0.51	No
0.14	No
0.67	Yes
0.42	No
0.50	Yes
0.43	No

You may assume that the probabilities were predicted by some logistic model.

c) Draw the *receiver operating characteristic* (ROC) for the following thresholds:

$$-\infty\,; \quad 0.465; \quad 0.505; \quad 0.590\,; \quad \infty$$

d) Calculate the *area under the curve* (AUC). What would you say about the model that produces the predicted probabilities based on the AUC value?

Question 4:

Consider a two-dimensional feature vector $\mathbf{X}$ that is normally distributed in three classes. As part of a discriminant analysis, the associated discriminant function was determined for each class using the Bayes assignment rule:

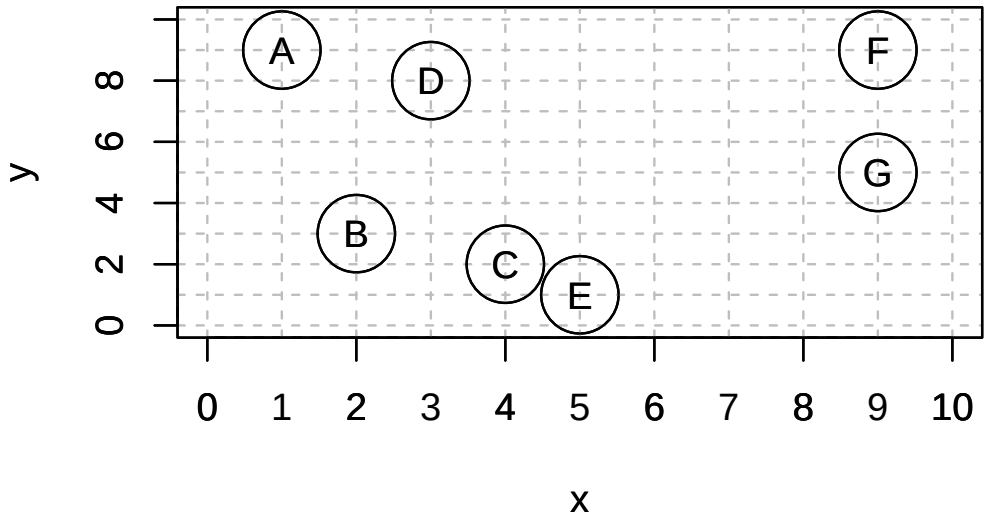
$$\begin{aligned} d_1(\mathbf{x}) &= x_1 + 2x_2 - 9 \\ d_2(\mathbf{x}) &= 2x_1 + x_2 - 3 \\ d_3(\mathbf{x}) &= -x_1 + x_2 - 3 \end{aligned}$$

a) Use the above discriminant functions to sketch the class areas. What is the nature of the separation? What is the assumption behind this form?

b) Could the same separation of a) occur within the framework of a QDA? Give a brief explanation.

Question 5:

Consider the following scatter plot of data in $\mathbb{R}^2$.



a) Sketch a dendrogram for the data given in the scatter plot using the agglomerative single-linkage clustering method and the Manhattan distance, i.e. $d_{\text{Manhattan}}(x, y) := \sum_{i=1}^{m} |x_i - y_i|$, (a freehand drawing is sufficient, but with labeling).

b) Instead of using the single-linkage method, the scatter plot is now to be clustered using the centroid method (still using the Manhattan distance). The intermediate result is the following clustering

$$\mathbb{C} = \{\{A, D\}, \{B, C, E\}, \{F\}, \{G\}\}.$$

Determine the distance matrix of this clustering and specify the clustering in the next iteration after merging the corresponding clusters.

c) The k-Means clustering method was now carried out with the Euclidean distance as the distance measure. One iteration step ν results in the following clusters: $C_1 = \{A, B, D\}, C_2 = \{C, E\}, C_3 = \{F, G\}$. The distances are given by :

$$D^{(\nu)} = \begin{pmatrix} 2.54 & 3.67 & 5.08 & 1.67 & 6.41 & 7.38 & 7.20 \\ 8.25 & 8.06 & 7.07 & 6.08 & 7.21 & 2.0 & 2.0 \\ 8.28 & 2.92 & 0.71 & 6.67 & 0.71 & 8.75 & 5.70 \end{pmatrix}$$

Mark the assignment of the observations to the clusters in the matrix $D^{(\nu)}$ and enter the new clustering $\mathbb{C}^{(\nu+1)}$, as well as the new cluster centroids $x_i^{(\nu+1)}$ for $i = 1, ..., 3$.

Question 6:

In this task, we consider a data set 'trees' that specifies the variables `Volume`, `Height` and `Girth` for 31 trees.

a) A scaled PCA in R using the `prcomp()` function produces the following output:

```
Standard deviations (1, .., p=3):
[1] 1.5525141 0.7494113 0.1675789

Rotation (n x k) = (3 x 3):
              PC1         PC2         PC3
Girth  0.6085705  0.4099013  0.67942837
Height 0.4891267 -0.8680065  0.08555556
Volume 0.6248176  0.2802600 -0.72873681
```

Interpret this and specifically address the *Standard deviations*, and the columns and rows of the *Rotation matrix*.

If you were unable to solve the previous subtask or were only able to solve it partially, you can assume for partial tasks (f) and (g), you can assume that the empirical correlation matrix has the following eigenvalues: $\lambda_1 = 1.5;\ \lambda_2 = 1;\ \lambda_3 = 0.5$.

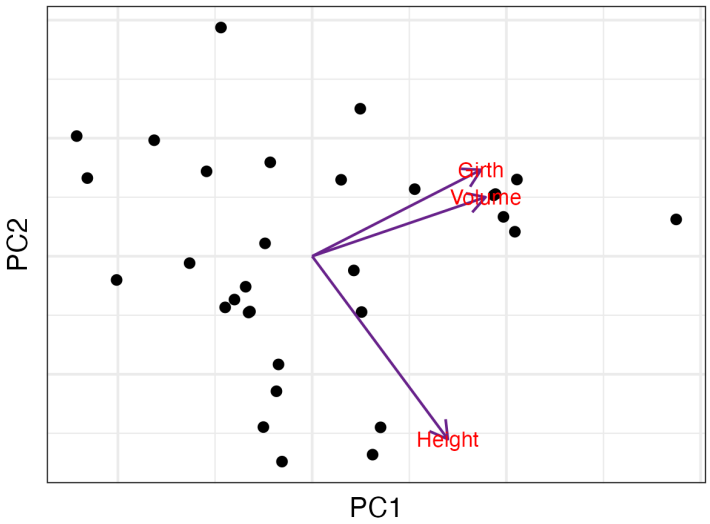
b) Draw the scree plot for this PCA.

c) Name two criteria that can be used to decide on the number of principal components (PCs) for dimensional reduction and apply both to the situation at hand.

d) Explain,

- what a biplot is and
- what it is used for in the context of PCA.

Also name (at least) 2 things that you can read from the biplot below for the analysis above. (*On the conceptual level only - the plot axes are intentionally not labeled*)



Question 7:

Consider the matrix

$$\begin{pmatrix} 1 & 0.75 \\ 0.75 & 1 \end{pmatrix}.$$

a) Perform an eigendecomposition of the above matrix.

b) Assume that this matrix is the sample correlation matrix of a data set that contains the observation $(10, 5)^\top$.
What would the new corresponding observation after dimension reduction via PCA to

(i) dimension $m = 1$

(ii) nonsensically, the same dimension, i.e. $m = 2$

be?

c) Next, assume that the above matrix equals the matrix B in the MDS decomposition $B = YY^\top$, with Y denoting the new representation.
Calculate the points in this MDS representation (you can assume that the centering step was skipped on purpose.)