Technical University of Denmark

Written examination: 26 May 2020, 10 AM - 2 PM.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

The exam is multiple choice. All questions have four possible answers marked by the letters A, B, C, and D as well as the answer "Don't know" marked by the letter E. Correct answer gives 3 points, wrong answer gives -1 point, and "Don't know" (E) gives 0 points.

When you hand in your answers you have to upload two files:

1. Your answers to the multiple choice exam using the "answers.txt" file.

2. Your written full explanations of how you found the answer to each question not marked as "E" (Don't know) either as a .zip (with bmp, png, tiff, jpg as allowed file formats if you take photos of your handwritten explanations) or as a PDF file (if you write a document with your answers).

You have to hand in your answers to the exam in file 1 no later than 2 PM and file 2 no later than 2:30 PM. Guessing on an answer is for the online exam not allowed, as each answer has to include an accompanying argumentation in writing for the answer. Failing to timely upload both documents will count as not having handed in the exam. Questions where we find answers in the "answers.txt" (file 1) that is different from the explanation or where explanations are insufficient in the accompanying file explaining the answers (file 2) will be treated as "Don't know". Systematic discrepancy between the answers in the two hand-in files will potentially count as attempt of cheating the exam.

Answers:

1	2	3	4	5	6	7	8	9	10
B	D	A	A	D	D	A	B	A	B
11	12	13	14	15	16	17	18	19	20
D	A	D	C	C	B	C	B	C	C
21	22	23	24	25	26	27			
D	A	A	B	B	D	D			

No.	Attribute description	Abbrev.
x_1	Live birth rate per 1000 population	BirthRt
x_2	Death rate per 1000 population	DeathRt
x_3	Infant deaths per 1000 population under 1 year	InfMort
x_4	Life expectancy at births for males	LExpM
x_5	Life expectancy at births for females	LExpF
x_6	Region encoded as $1, 2, \ldots, 6$	Region
y	Gross National Product, per capita, US$	GNP

Table 1: Description of the features of the Poverty dataset used in this exam. The dataset consists of population statistics of countries provided by the 1990 United Nations statistical almanacs. $x_1, \ldots, x_5$ respectively provide statistics on birth rates, death rates, infant deaths, and life expectancy by gender and x_6 denotes location of each country in terms of regions such that 1 = Eastern Europe, 2 = South America/Mexico, 3 = Western Europe/US/Canada/Australia/NewZealand/Japan, 4 = Middle East, 5 = Asia and 6 = Africa. The data has been processed such that countries having missing values have been removed. We consider the goal as predicting the gross national product (GNP) pr. capita both as a regression and classification task. For regression tasks, y_r will refer to the continuous value of GNP. For classification tasks the attribute y_b is discrete formed by thresholding y_r at the median value and takes values $y_b = 0$ (corresponding to low GNP level) and $y_b = 1$ (corresponding to a high GNP level). The dataset used has $N = 91$ observations in total.

Question 1. We will consider the Poverty dataset[1] described in Table 1. The dataset consists of 91 countries (observations) and six input attributes $x_1, \ldots, x_6$ as well as the output y_r providing the gross national product pr. capita (denoted GNP). Which one of the following statements regarding the dataset is correct?

A. All the input attributes $x_1, \ldots, x_6$ are ratio.

B. One of the six input attributes is nominal.

C. All the input attributes $x_1, \ldots, x_6$ are interval.

D. The output attribute y_r is ordinal.

E. Don't know.

Solution 1. For the atributes $x_1, \ldots, x_5$ zero means absence of what is being measured and we can naturally talk about a quantity being say twice as large as another etc. thus these five input attributes are all ratio. x_6 is nominal as this variable categorizes which region each observation belongs to of the six different regions in the dataset. The output y_r is ratio as zero naturally indicates absence of GNP and we again can naturally apply subtraction and addition (required for an interval attribute) but also multiplication (the GNP of one country can be three times larger that of another etc.).

[1] Dataset obtained from `https://www2.stetson.edu/~jrasp/data/Poverty.xls`

	Mean	Std	$x_{p=25\%}$	$x_{p=50\%}$	$x_{p=75\%}$
BirthRt	29.46	13.62	14.6	29	42.575
DeathRt	10.73	4.66	7.7	9.5	12.4
InfMort	55.28	46.05	13.025	43	88.25
LExpM	61.38	9.67	55.2	63.4	68.55

Table 2: Summary statistics of the first four attributes of the Poverty dataset. The column $x_{p=25\%}$ refers to the 25'th percentile of the given attribute, $x_{p=50\%}$ to the median and $x_{p=75\%}$ to the 75'th percentile.

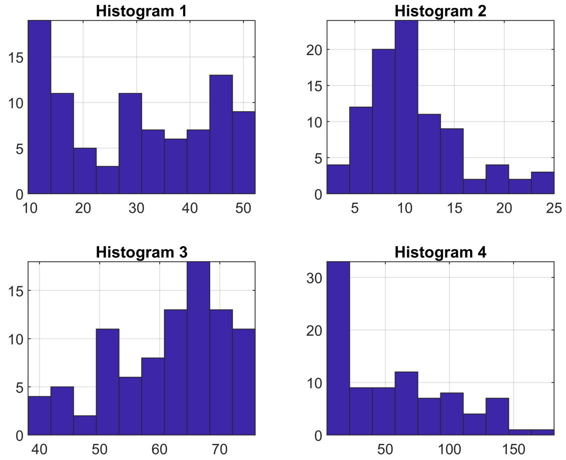

Figure 1: Four histograms corresponding to the variables with summary statistics given in Table 2 but not necessarily in that order.

Question 2.
Table 2 contains summary statistics of the first four attributes of the Poverty dataset. Which of the histograms in Figure 1 match which of the attributes according to their summary statistics?

A. *BirthRt* matches histogram 4, *DeathRt* matches histogram 2, *InfMort* matches histogram 1 and *LExpM* matches histogram 3.

B. *BirthRt* matches histogram 4, *DeathRt* matches histogram 1, *InfMort* matches histogram 3 and *LExpM* matches histogram 2.

C. *BirthRt* matches histogram 2, *DeathRt* matches histogram 3, *InfMort* matches histogram 1 and *LExpM* matches histogram 4.

D. *BirthRt* matches histogram 1, *DeathRt* matches histogram 2, *InfMort* matches histogram 4 and *LExpM* matches histogram 3.

E. Don't know.

Solution 2. To solve the problem, note that we can read of the median, 25'th, and 75'th percentiles from Table 2 as $q_{p=50\%}$, $q_{p=25\%}$, and $q_{p=75\%}$ respectively. These can be matched to the histograms in Figure 1 by observing histogram 2 does not have observations above 25 and thus must therefore be *DeathRt*. Histogram 4 is the only histogram having observations above 88.25 which only holds for *InfMort* (see 75th percentile). This only holds for answer option D.

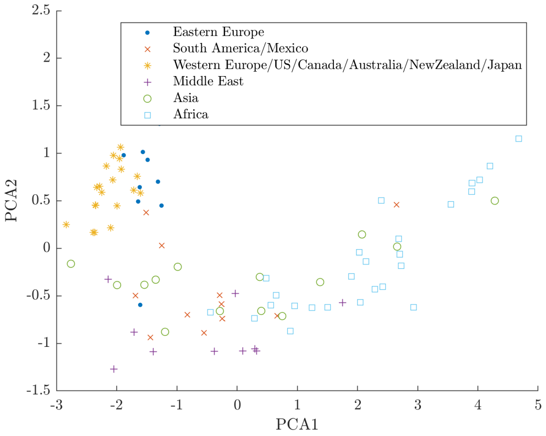

Figure 2: The Poverty data projected onto the first two principal component directions with each observation labelled according to the region it belongs to (given by x_6).

Question 3. A Principal Component Analysis (PCA) is carried out on the Poverty dataset in Table 1 based on the attributes x_1, x_2, x_3, x_4, x_5.

The data is standardized by (i) substracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized data matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.43 & -0.5 & 0.7 & -0.25 & -0.07 \\ 0.38 & 0.85 & 0.3 & -0.2 & 0.03 \\ 0.46 & -0.13 & -0.61 & -0.61 & -0.15 \\ -0.48 & -0.0 & 0.13 & -0.63 & 0.6 \\ -0.48 & 0.1 & 0.16 & -0.36 & -0.78 \end{bmatrix} \quad (1)$$

$$\boldsymbol{S} = \begin{bmatrix} 19.64 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 6.87 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 3.26 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 2.30 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.12 \end{bmatrix}.$$

Which one of the following statements is true?

A. The variance explained by the first four principal components is greater than 99 %.

B. The variance explained by the last four principal components is greater than 15 %.

C. The variance explained by the first two principal components is greater than 97 %.

D. The variance explained by the first principal component is greater than 90 %.

E. Don't know.

Solution 3. The correct answer is A. To see this, recall that the variance explained by a given component k of the PCA is given by

$$\frac{\sigma_k^2}{\sum_{j=1}^M \sigma_j^2}$$

where M is the number of attributes in the dataset being analyzed. The values of σ_k can be read off as entry $\sigma_k = S_{kk}$ where $\boldsymbol{S}$ is the diagonal matrix of the SVD computed above. We therefore find the variance explained by the first four components is:

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2} =$$
$$\frac{19.64^2 + 6.87^2 + 3.26^2 + 2.30^2}{19.64^2 + 6.87^2 + 3.26^2 + 2.30^2 + 1.12^2} = 0.9972.$$

Question 4. Consider again the PCA analysis of the Poverty dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (1). In Figure 2 is given the data projected onto the first two principal components and each observation labelled according to the region it belongs to. Which one of the following statements is true?

A. An observation from Africa will typically have a relatively high value of BirthRt, a high value of DeathRt, a high value of InfMort, a low value of LExpM and a low value of LExpF as observed from the projection onto principal component number 1.

B. An observation from Western Europe/US/Canada/Australia/NewZealand/Japan will typically have a relatively high value of **BirthRt**, a low value of **DeathRt**, a high value of **InfMort**, and a low value of **LExpF** as observed from the projection onto principal component number 2.

C. As observed from the projection onto principal component number 1 observations from Eastern Europe will typically have a relatively low value of **BirthRt**, a high value of **DeathRt**, a low value of **InfMort**, a high value of **LExpM** whereas **LExpF** will have almost no influence (the coefficient is only -0.07).

D. As can be seen from the plot of the first and second principal components there is a negative correlation between the observations projected onto PC1 and PC2.

E. Don't know.

Solution 4. The correct answer is A. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_1$ (i.e. column one of $\boldsymbol{V}$) is

$$b_1 = \boldsymbol{x}^\top \boldsymbol{v}_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \begin{bmatrix} 0.43 \\ 0.38 \\ 0.46 \\ -0.48 \\ -0.48 \end{bmatrix}$$

for this projection to be (relatively large) and positive which is the case for observations coming from Afria, this occurs if x_1, x_2, x_3, x_4, x_5 has large magnitude and the sign convention given in option A. As the data projected onto PC1 and PC2 is given by $\tilde{\boldsymbol{X}}\boldsymbol{v}_1$ and $\tilde{\boldsymbol{X}}\boldsymbol{v}_2$ and the mean has been subtracted during standardization the mean values of the data projected onto $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ will be zero and thus the covariance between the observations projected onto PC1 and PC2 given by $\frac{1}{N-1}(\tilde{\boldsymbol{X}}\boldsymbol{v}_1)^T(\tilde{\boldsymbol{X}}\boldsymbol{v}_2) = (\boldsymbol{USV}^\top\boldsymbol{v}_1)^T(\boldsymbol{USV}^\top\boldsymbol{v}_2) = \boldsymbol{u}_1^\top s_{11}\boldsymbol{u}_2 s_{22} = s_{11}s_{22}\boldsymbol{u}_1^\top\boldsymbol{u}_2 = 0$ and there can therefore be no correlation between the observations projected onto PC1 and PC2.

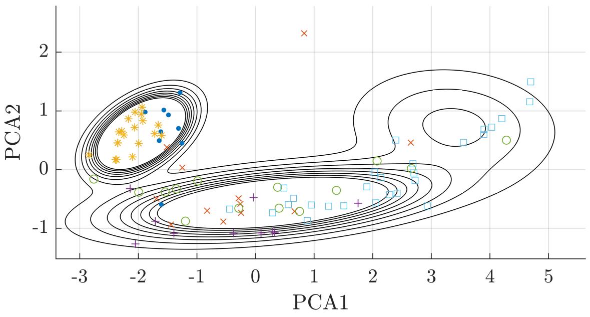

Figure 3: A GMM with K=3 clusters fitted to the poverty data projected onto the first two principal component directions. Each observation is again labelled according to the region it belongs to (given by x_6).

Question 5. In Figure 3 a Gaussian Mixture Model (GMM) is fitted to the standardized data projected onto the first two principal component directions using three mixture components (i.e., $K = 3$ clusters). Recall that the multivariate Gaussian distribution is given by: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Which one of the following GMM densities corresponds to the fitted density given in Figure 3?

A.

$$\begin{aligned} p(\boldsymbol{x}) = 0.1425 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ + 0.3235 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \\ + 0.5340 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \end{aligned}$$

B.

$$\begin{aligned} p(\boldsymbol{x}) = 0.1425 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ + 0.3235 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \\ + 0.5340 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \end{aligned}$$

C.

$$\begin{aligned} p(\boldsymbol{x}) = 0.3235 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ + 0.1425 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \\ + 0.5340 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \end{aligned}$$

D.

$$\begin{aligned} p(\boldsymbol{x}) = 0.3235 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}, \begin{bmatrix} 0.1695 & 0.0665 \\ 0.0665 & 0.1104 \end{bmatrix}) \\ + 0.1425 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}, \begin{bmatrix} 1.2137 & -0.0703 \\ -0.0703 & 0.3773 \end{bmatrix}) \\ + 0.5340 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}, \begin{bmatrix} 2.0700 & 0.1876 \\ 0.1876 & 0.1037 \end{bmatrix}) \end{aligned}$$

E. Don't know.

Solution 5. Inspecting the GMM density we observe that the cluster located at $\begin{bmatrix} 3.3884 \\ 0.7424 \end{bmatrix}$ will have the lowest mixing proportion as only few observations belong to this cluster. Furthermore, the cluster located at $\begin{bmatrix} -1.9482 \\ 0.6132 \end{bmatrix}$ clearly has positive covariance between PCA1 and PCA2 and much smaller variance (i.e., 0.1695) in the PCA1 direction when compared to the other cluster located at $\begin{bmatrix} 0.2756 \\ -0.5696 \end{bmatrix}$ having high variance (i.e., 2.0700) also having positive covariance. This only holds for answer option D.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}	o_{11}
o_1	0.0	1.7	1.4	0.4	2.2	3.7	5.2	0.2	4.3	6.8	6.0
o_2	1.7	0.0	1.0	2.0	1.3	2.6	4.5	1.8	3.2	5.9	5.2
o_3	1.4	1.0	0.0	1.7	0.9	2.4	4.1	1.5	3.0	5.5	4.8
o_4	0.4	2.0	1.7	0.0	2.6	4.0	5.5	0.3	4.6	7.1	6.3
o_5	2.2	1.3	0.9	2.6	0.0	1.7	3.4	2.4	2.1	4.8	4.1
o_6	3.7	2.6	2.4	4.0	1.7	0.0	2.0	3.8	1.6	3.3	2.7
o_7	5.2	4.5	4.1	5.5	3.4	2.0	0.0	5.4	2.5	1.6	0.9
o_8	0.2	1.8	1.5	0.3	2.4	3.8	5.4	0.0	4.4	6.9	6.1
o_9	4.3	3.2	3.0	4.6	2.1	1.6	2.5	4.4	0.0	3.4	2.9
o_{10}	6.8	5.9	5.5	7.1	4.8	3.3	1.6	6.9	3.4	0.0	1.0
o_{11}	6.0	5.2	4.8	6.3	4.1	2.7	0.9	6.1	2.9	1.0	0.0

Table 3: The pairwise Euclidian distances, $d(o_i, o_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 11 observations from the Poverty dataset based on $x_1, \ldots, x_5$. Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1 (excluding x_6). The colors indicate classes such that the red observations $\{o_1,\ o_2,\ o_3,\ o_4,\ o_5,\ o_6,\ o_7,\ o_8\}$ belong to class $y_b = 0$ (corresponding to a low GNP level), and the black observations $\{o_9,\ o_{10},\ o_{11}\}$ belongs to class $y_b = 1$ (corresponding to a relatively high GNP).

Question 6. To examine if observation o_3 may be an outlier we will calculate the average relative density using the Euclidean distance based on the observations given in Table 3 only. We recall that the KNN density and average relative density (ard) for the observation $\boldsymbol{x}_i$ are given by:

$$\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K) = \frac{1}{\frac{1}{K}\sum_{\boldsymbol{x}' \in N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)} d(\boldsymbol{x}_i, \boldsymbol{x}')},$$

$$\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K) = \frac{\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)}{\frac{1}{K}\sum_{\boldsymbol{x}_j \in N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)} \text{density}_{\boldsymbol{X}_{\setminus j}}(\boldsymbol{x}_j, K)},$$

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors. What is the average relative density for observation o_3 for $K = 2$ nearest neighbors?

A. 0.59

B. 1.00

C. 1.05

D. 1.18

E. Don't know.

Solution 6.
To solve the problem, first observe the $k = 2$ neighborhood of o_3 and density is:

$$N_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = \{o_5, o_2\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = 1.053$$

For each element in the above neighborhood we can then compute their $K = 2$-neighborhoods and densities to be:

$$N_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = \{o_3, o_2\}, \quad N_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = \{o_3, o_5\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = 0.909, \text{density}_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = 0.870.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem and we obtain $1.053/(0.5 \cdot (0.909 + 0.870))$.

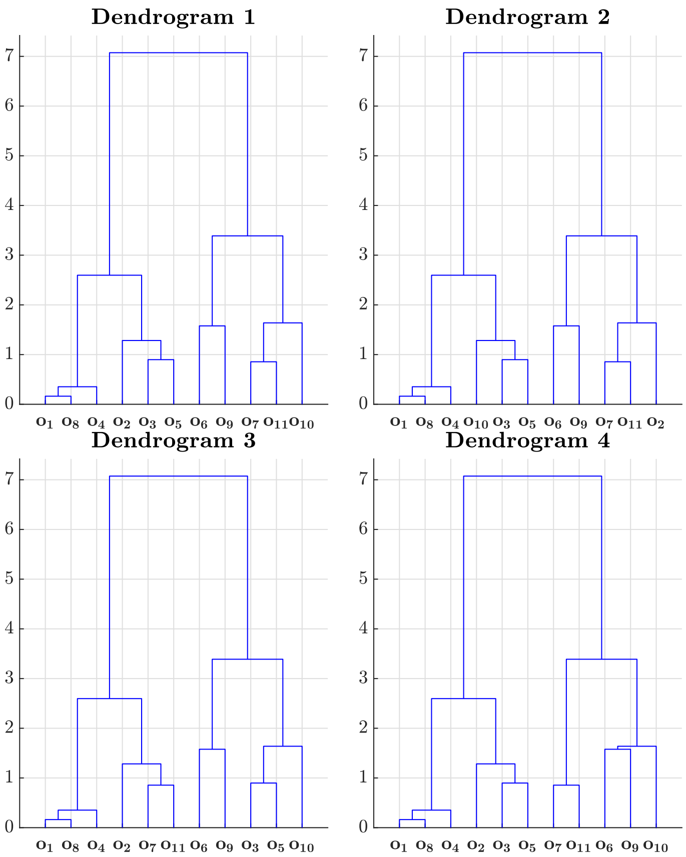

Figure 4: Four dendrograms for which one of the dendrograms corresponds to hierarchical clustering using maximum linkage of the 11 observations in Table 3.

Question 7. A hierarchical clustering is applied to the 11 observations in Table 3 using *maximum* linkage. Which one of the dendrograms shown in Figure 4 corresponds to the distances given in Table 3?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

Solution 7. The correct solution is A. We can rule out the other solutions by observing the first merge operation at which they diverge from the correct solution.

- In dendrogram 2, merge operation number 5 should have been between the sets $\{o_2\}$ and $\{o_3, o_5\}$ at a height of 1.28, however in dendrogram 2 merge number 5 is between the sets $\{o_{10}\}$ and $\{o_3, o_5\}$.

- In dendrogram 3, merge operation number 3 should have been between the sets $\{o_7\}$ and $\{o_{11}\}$ at a height of 0.86, however in dendrogram 3 merge number 3 is between the sets $\{o_3\}$ and $\{o_5\}$.
- In dendrogram 4, merge operation number 3 should have been between the sets $\{o_7\}$ and $\{o_{11}\}$ at a height of 0.86, however in dendrogram 4 merge number 3 is between the sets $\{o_6\}$ and $\{o_9\}$.

Question 8. Consider again the 11 observations in Table 3. We will use a one-nearest neighbor classifier to classify the observations. What will be the error rate of the KNN classifier when considering a leave-one-out cross-validation strategy to quantify performance?

A. 3/11

B. 4/11

C. 5/11

D. 6/11

E. Don't know.

Solution 8. observation o_1 has o_8 as nearest neighbor and correctly classified.
observation o_2 has o_3 as nearest neighbor and correctly classified.
observation o_3 has o_5 as nearest neighbor and correctly classified.
observation o_4 has o_8 as nearest neighbo and correctly classified.
observation o_5 has o_3 as nearest neighbor and correctly classified.
observation o_6 has o_9 as nearest neighbor and incorrectly classified.
observation o_7 has o_{11} as nearest neighbor and incorrectly classified.
observation o_8 has o_1 as nearest neighbor and correctly classified.
observation o_9 has o_6 as nearest neighbor and incorrectly classified.
observation o_{10} has o_{11} as nearest neighbor and correctly classified.
observation o_{11} has o_7 as nearest neighbor and incorrectly classified.
4 out of 11 observations are thus incorrectly classified.

Question 9. A logistic regression model is trained to distinguish between the two classes $y_b \in \{0, 1\}$, i.e., relatively low GNP (negative class) vs. relative high GNP (positive class). The model is trained using all observations except the 11 observations given in Table 3 that are used for testing the model (i.e., using the hold-out method). The features $x_1, \ldots, x_5$ are standardized (mean subtracted and each feature divided by its standard deviation). The feature x_6 is transformed using one-out-of-K coding and the last region removed to generate the new features c_1, c_2, c_3, c_4, c_5

that are included in the regression to produce the class-probability $\hat{y}$ defined by the trained model:

$$\hat{y} = \sigma(1.41 + 0.76x_1 + 1.76x_2 - 0.32x_3 - 0.96x_4 + 6.64x_5 - 5.13c_1 - 2.06c_2 + 96.73c_3 + 1.03c_4 - 2.74c_5).$$

We will predict the estimated output of the sixth of the eleven test observations given by:

$$\boldsymbol{x}_6 = [-0.06 \quad -0.28 \quad 0.43 \quad -0.30 \quad -0.36 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1$$

Which one of the following statements is correct?

A. According to the estimated model an increase in a country's birth rate will increase the probability that the country is rich.

B. The probability observation $\boldsymbol{x}_6$ belongs to class $y = 1$ is less than 1 %.

C. The attribute *Region* has very little influence on whether a country is poor or rich.

D. As the weight for x_1 and x_3 have opposing signs we can conclude the two features are negatively correlated.

E. Don't know.

Solution 9. As the coefficient infront of x_1 is positive this implies that increasing x_1 will according to the model increase the probability of being in the positive class (i.e., a rich country), thus this is a correct statement. The estimated output for $\boldsymbol{x}_6$ is

$$\begin{aligned}\hat{y} &= \sigma(1.41 + (0.76 \cdot -0.06) + (1.76 \cdot -0.28) \\ &- (0.32 \cdot 0.43) - (0.96 \cdot -0.30) + (6.64 \cdot -0.36) \\ &- (2.74 \cdot 1.00)) \\ &= \frac{1}{1+\exp(-(1.41+(0.76\cdot -0.06)+(1.76\cdot -0.28)-(0.32\cdot 0.43)-(0.96\cdot -0.30)+(6.64\cdot -0.36)-(2.74\cdot 1.00)))} \\ &= 1.62\%.\end{aligned}$$

From the model it is further observed that Region has a very strong influence on the estimated output - in particular $c_3 = 1$ corresponding to the country being in the Western Europe/US/Canada/Australia/NewZealand/Japan region strongly influences that the country will be given a high probabiltiy of being in the positive class (i.e., rich), i.e. the coefficient in front of c_3 is positive and very large with magnitude of 96.73. Notably, we can not use the sign of the estimated weights to deduce anything about feature correlation.

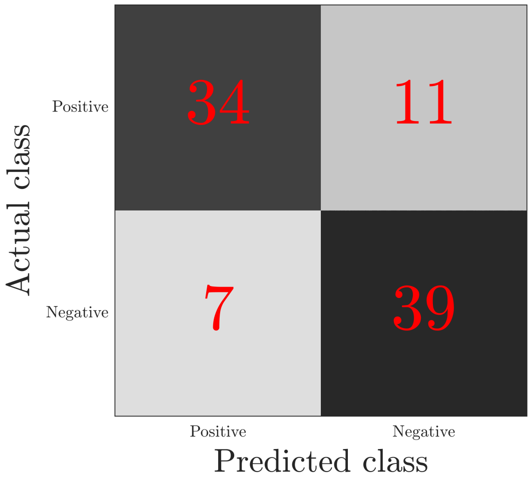

Figure 5: 5-fold cross validation applied to the entire dataset to evaluate logistic regression as an approach to predict low GNP ($y_b = 0$, negative class) versus high GNP ($y_b = 1$, positive class).

Question 10. Based on the entire dataset in Table 1, we use 5-fold cross-validation to estimate the performance of logistic regression. In Figure 5 is given the confusion matrix obtained using the cross-validation procedure. We will quantify the performance of the results using the F-measure given by $F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$.

Which one of the following statements is correct?

A. $F_1 = 0.7556$

B. $F_1 = 0.7907$

C. $F_1 = 0.7990$

D. $F_1 = 0.8293$

E. Don't know.

Solution 10. The Precision is 34/41 and the Recall 34/45. Thus, $F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot 34/41 \cdot 34/45}{34/41 + 34/45} = 0.7907$.

Question 11. Four different logistic regression models are trained to distinguish between the two classes $y_b \in \{0, 1\}$, (i.e., low GNP (negative class given as red plusses) vs. high GNP (positive class given as black crosses)) and evaluated on the 11 observations also considered in Table 3 presently used as a test set. In Figure 6 is in the top panel given the four classifiers' predictions on the 11 test observations and in the bottom panel a *reciever operator characteristic* (ROC) curve. Which classifier's performance corresponds to the shown ROC curve?

A. Classifier 1

B. Classifier 2

C. Classifier 3

D. Classifier 4

E. Don't know.

Solution 11. The correct answer is D. To see this, recall that the ROC curve is computed from the false positive rate (FPR) and true positive rate (TPR) for particular choices of threshold value $\hat{y}$. To compute e.g. the TPR, one assumes every observation predicted to belong to class 1 with a probability higher than $\hat{y}$ is actually assigned to class one. We then divide the total number of observations belonging to class one *and which are predicted to belong to class 1* with the number of observations in the *positive* class.

Similarly for the FPR, where we now count the number of observations that are assigned to class one *but in fact belongs to class 0*, divided by the total number of observations in the *negative* class.

We start by a very high threshold value say at > 1 in which TPR=0 and FPR=0. Subsequently we have that as we lower the threshold we first observe a positive observation is above the threshold giving FPR=0, TPR=1/3, then another positive observation is above the threshold as we lower it again giving FPR=0, TPR=2/3. Subsequently, a negative observation becomes above the threshold such that FPR=1/8, TPR=2/3 and the last positive observation become above the threshold as we next lower the threshold such that FPR=1/8, TPR=1. Finally, we traverse by lowering the threshold all the way down to 0 all the negative observations with no more positive observations added, i.e. FPR=1, TPR=1 when the threshold is lowered to 0 or greater.

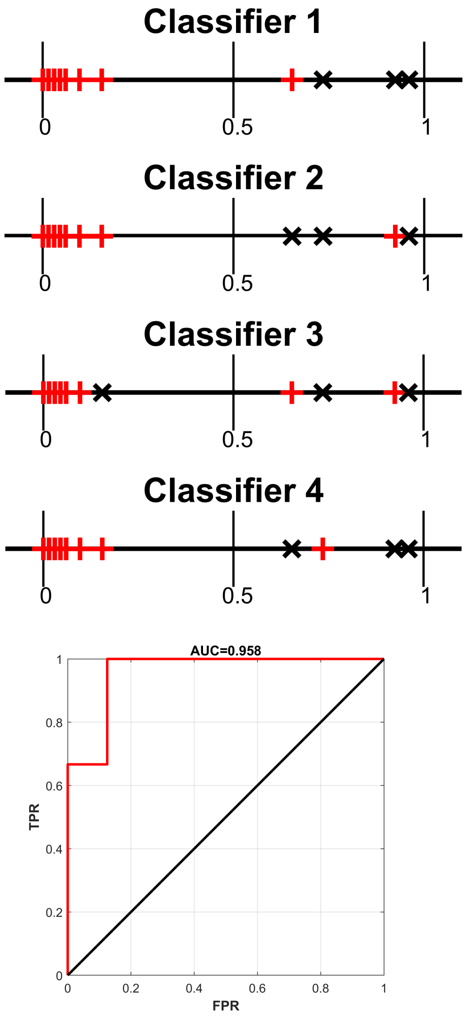

Figure 6: Top panel: Four different logistic regression models used to predict low GNP ($y_b = 0$, marked by red plusses) from high GNP ($y_b = 1$, marked by black crosses). Bottom panel: The ROC curve corresponding to one of the four classifiers in the top panel.

Question 12. Consider again the Poverty dataset in Table 1. We would like to predict GNP using a least squares linear regression model, and since we would like the model to be as interpretable as possible we will use variable selection to obtain a parsimonious model. We limit ourselves to the five features x_1, x_2, x_3, x_4 and x_5. In Table 4 we have pre-computed the estimated training and test errors for all combinations of the five attributes. Which one of the following statements is correct?

A. Forward selection will select attributes x_3.

B. Forward selection will select attributes x_1, x_3, x_4, x_5.

C. Forward selection will select attributes x_1, x_2, x_4.

D. Backward selection will select attributes x_1, x_4.

E. Don't know.

Solution 12. The correct answer is A. To solve this problem, it suffices to show which variables will be selected by forward or backward selection. First note that in variable selection, we only need to concern ourselves with the *test* error, as the training error should trivially drop when more variables are introduced and is furthermore not what we ultimately care about.

Forward selection: The method is initialized with the empty set $\{\}$ having an error of 2.02.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$. Since the lowest error of the available sets is 1.628, which is lower than 2.02, we update the current selected variables to $\{x_3\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_3\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1, x_4\}$, $\{x_2, x_4\}$, $\{x_3, x_4\}$, $\{x_1, x_5\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$, $\{x_4, x_5\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates at $\{x_3\}$.

Backward selection: The method is initialized with the set $\{x_1, x_2, x_3, x_4, x_5\}$ having an error of 2.03.

Feature(s)	Training RMSE	Test RMSE
none	1.429	2.02
x_1	0.755	1.662
x_2	1.421	1.977
x_3	0.636	1.628
x_4	0.847	1.636
x_5	0.773	1.702
x_1, x_2	0.640	1.706
x_1, x_3	0.636	1.638
x_2, x_3	0.401	1.912
x_1, x_4	0.745	1.602
x_2, x_4	0.565	1.799
x_3, x_4	0.587	1.890
x_1, x_5	0.728	1.647
x_2, x_5	0.449	1.767
x_3, x_5	0.613	1.824
x_4, x_5	0.733	2.155
x_1, x_2, x_3	0.380	2.135
x_1, x_2, x_4	0.541	1.696
x_1, x_3, x_4	0.586	1.914
x_2, x_3, x_4	0.399	1.954
x_1, x_2, x_5	0.448	1.779
x_1, x_3, x_5	0.613	1.831
x_2, x_3, x_5	0.396	1.828
x_1, x_4, x_5	0.702	2.022
x_2, x_4, x_5	0.407	2.087
x_3, x_4, x_5	0.582	1.901
x_1, x_2, x_3, x_4	0.379	2.168
x_1, x_2, x_3, x_5	0.369	1.988
x_1, x_2, x_4, x_5	0.400	2.138
x_1, x_3, x_4, x_5	0.580	1.927
x_2, x_3, x_4, x_5	0.359	1.935
x_1, x_2, x_3, x_4, x_5	0.315	2.030

Table 4: Root-mean-square error (RMSE) for the training and test set using least squares regression to predict GNP in the Poverty dataset using different combinations of the features x_1, x_2, x_3, x_4, and x_5.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_2, x_3, x_4, x_5\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_2, x_3, x_4\}$, $\{x_1, x_2, x_3, x_5\}$, $\{x_1, x_2, x_4, x_5\}$, $\{x_1, x_3, x_4, x_5\}$, $\{x_2, x_3, x_4, x_5\}$. Since the lowest error of the available sets is 1.927, which is lower than 2.03, we update the current selected variables to $\{x_1, x_3, x_4, x_5\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_3, x_4, x_5\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, $\{x_1, x_3, x_4\}$, $\{x_2, x_3, x_4\}$, $\{x_1, x_2, x_5\}$, $\{x_1, x_3, x_5\}$, $\{x_2, x_3, x_5\}$, $\{x_1, x_4, x_5\}$, $\{x_2, x_4, x_5\}$, $\{x_3, x_4, x_5\}$. Since the lowest error of the available sets is 1.831, which is lower than 1.927, we update the current selected variables to $\{x_1, x_3, x_5\}$

Step $i = 3$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_3, x_5\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1, x_4\}$, $\{x_2, x_4\}$, $\{x_3, x_4\}$, $\{x_1, x_5\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$, $\{x_4, x_5\}$. Since the lowest error of the available sets is 1.638, which is lower than 1.831, we update the current selected variables to $\{x_1, x_3\}$

Step $i = 4$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_3\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$. Since the lowest error of the available sets is 1.628, which is lower than 1.638, we update the current selected variables to $\{x_3\}$

Step $i = 5$ The available variable sets to choose between is obtained by taking the current variable set $\{x_3\}$ and removing each of the left-out variables thereby resulting in the sets $\{\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Question 13. Suppose a neural network is trained to predict GNP. As part of training the network, we wish to select between three different model architectures respectively with 5, 10 and 20 hidden units and estimate the generalization error of the optimal choice. In the outer loop we opt for $K_1 = 4$-fold cross-validation, and in the inner $K_2 = 7$-fold cross-validation. The time taken to *train* a single model is 20 seconds, and this can be assumed constant for each fold. If the time taken to test a model is 1 second what is then the total time required to complete the 2-level cross-validation procedure?

A. 1760 seconds

B. 1764 seconds

C. 1844 seconds

D. 1848 seconds

E. Don't know.

Solution 13. Let $S = 3$ denote the three different models considered. Going over the 2-level cross-validation algorithm we see the total number of models to be *trained* is:

$$K_1(K_2 S + 1) = 88$$

Multiplying by the time taken to train a single model we obtain a total training time of 1760 seconds.

As every model we use to train is also used for testing a dataset the number of times we test a model is:

$$K_1(K_2 S + 1) = 88$$

As each of these take 1 second we obtain in total 1760+88=1848 seconds.

Question 14. We will fit a decision tree in order to determine based on the features x_1 and x_2 if a country has a relatively low or high GNP. In the top panel of Figure 7 is given the fitted decision tree and in the bottom panel is given four different decision boundaries in which one of the four decision boundaries corresponds to the boundaries generated by the decision tree given in the top panel.

Which one of the the four decision boundaries corresponds to the decision boundaries of the illustrated classification tree?

A. Decision boundary of Classifier 1

B. Decision boundary of Classifier 2

C. Decision boundary of Classifier 3

D. Decision boundary of Classifier 4

E. Don't know.

Solution 14. The decision tree includes four decisions two based on x_1 and two based on x_2. As such the decision boundaries must have two horizontal and vertical lines which only holdes for Classifier 3.

Question 15. According to the poverty dataset we have that 15.4% of countries are from Africa. We are further told that if a country is from Africa the probability that the country has a GNP above 1000 US$ pr. capita is 28.6% whereas if a country is outside of Africa the probability that the GNP is above 1000 US$ pr. capita is 68.8%.

Given that a country's GNP is above 1000 US$ pr. capita what is the probabily it is in Africa?

A. 4.4 %

B. 6.4 %

C. 7.0 %

D. 7.6 %

E. Don't know.

Solution 15. According to Bayes' theorem we have:

$$P(Africa|GNP > 1000) = \frac{P(GNP>1000|Africa)P(Africa)}{P(GNP>1000)}$$

$$= \frac{P(GNP>1000|Africa)P(Africa)}{P(GNP>1000|Africa)P(Africa)+P(GNP>1000|not\ Africa)P(not\ Africa)}$$

$$= \frac{0.286 \cdot 0.154}{0.286 \cdot 0.154 + 0.688 \cdot (1-t0.154)} = 7.0\%$$

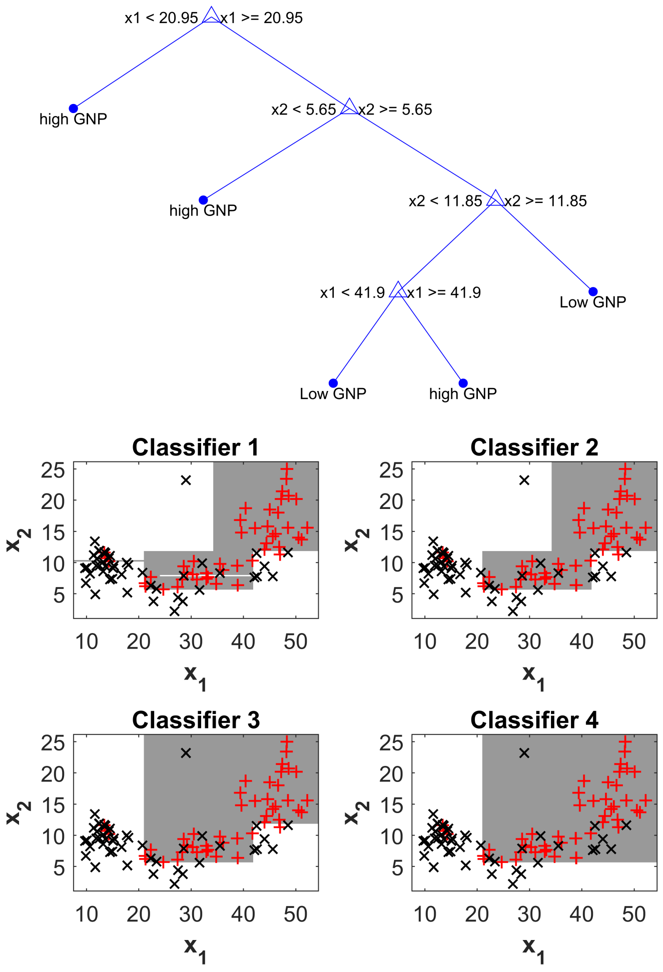

Figure 7: Top panel, a decision tree fitted to x_1 and x_2 of the Poverty data in order to predict wheter a country has relatively low or high GNP. Bottom panel, decision boundaries for four different decision trees in which gray regions correspond to regions predicted having low GNP ($y_b = 0$) and white regions to predictions having high GNP ($y_b = 1$). One of the four decision boundaries corresponds to the decision boundary of the classification tree given in the top panel.

	f_1	f_2	f_3	f_4	f_5
o_1	1	1	1	0	0
o_2	1	1	1	0	0
o_3	1	1	1	0	0
o_4	1	1	1	0	0
o_5	1	1	1	0	0
o_6	0	1	1	0	0
o_7	0	1	0	1	1
o_8	1	1	1	0	0
o_9	1	0	1	0	0
o_{10}	0	0	0	1	1
o_{11}	0	1	0	1	1

Table 5: Binarized version of the Poverty dataset in which the features $x_1, \dots, x_5$ are binarized. Each of the binarized features f_i are obtained by taking the corresponding feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). As in Table 3 the colors indicate the two classes such that the red observations $\{o_1,\ o_2,\ o_3,\ o_4,\ o_5,\ o_6,\ o_7,\ o_8\}$ belong to class $y_b = 0$ (corresponding to a low GNP). and black observations $\{o_9,\ o_{10},\ o_{11}\}$ belongs to class $y_b = 1$ (corresponding to a high GNP)

Question 16. We again consider the Poverty dataset from Table 1 and the $N = 11$ observations we already encountered in Table 3. The first five features of the dataset is processed to produce five new, binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[2], and we thereby arrive at the $N \times M = 11 \times 5$ binary matrix in Table 5. We wish to apply a Bayesian classifier to the dataset and as part of this task we have to estimate the probability

$$p(f_2 = 1,\ f_3 = 1 | y_b = 1).$$

For better numerical stability, we will use robust estimation to obtain the probability by introducing a regularization factor of α such that:

$$p(A|B) = \frac{\{\text{Occurences matching } A \text{ and } B\} + \alpha}{\{\text{Occurences matching } B\} + 2\alpha}.$$

What is the probability estimate if $\alpha = 1$?

A. $p(f_2 = 1,\ f_3 = 1 | y_b = 1) = \frac{1}{9}$

B. $p(f_2 = 1,\ f_3 = 1 | y_b = 1) = \frac{1}{5}$

C. $p(f_2 = 1,\ f_3 = 1 | y_b = 1) = \frac{4}{11}$

D. $p(f_2 = 1,\ f_3 = 1 | y_b = 1) = \frac{2}{3}$

E. Don't know.

Solution 16. Of the observations in class $y_b = 1$ zero have simultaneously $f_2 = 1$ and $f_3 = 1$. As this class contains *three* observations, we see the answer is

$$\frac{0 + \alpha}{3 + 2\alpha} = \frac{1}{5}$$

Therefore, answer B is correct.

Question 17. Consider again the binarized version of the Poverty dataset given in Table 5. We will no longer use robust estimation (i.e., we set $\alpha = 0$) and train a naïve-Bayes classifier in order to predict the class label y_b using only the features f_2 and f_3. If for an observation we have

$$f_2 = 1,\ f_3 = 0$$

what is then the probability that the observation has high GNP (i.e., $y_b = 1$) according to a naïve-Bayes classifier trained using only the data in Table 5?

A. $p_{\text{NB}}(y_b = 1 | f_2 = 1,\ f_3 = 0) = \frac{2}{9}$

B. $p_{\text{NB}}(y_b = 1 | f_2 = 1,\ f_3 = 0) = \frac{1}{3}$

C. $p_{\textbf{NB}}(y_b = 1 | f_2 = 1,\ f_3 = 0) = \frac{2}{5}$

D. $p_{\text{NB}}(y_b = 1 | f_2 = 1,\ f_3 = 0) = \frac{16}{25}$

E. Don't know.

Solution 17. To solve this problem, we simply use the general form of the naïve-Bayes approximation and plug in the relevant numbers. We get:

$$\begin{aligned} &p_{\text{NB}}(y_b = 1 | f_2 = 1,\ f_3 = 0) = \\ &\frac{p(f_2 = 1 | y = 1) p(f_3 = 0 | y = 1) p(y_b = 1)}{\sum_{j=0}^{1} p(f_2 = 1 | y = j) p(f_3 = 0 | y = j) p(y_b = j)} \\ &= \frac{\frac{1}{3}\frac{2}{3}\frac{3}{11}}{\frac{8}{8}\frac{1}{8}\frac{8}{11} + \frac{1}{3}\frac{2}{3}\frac{3}{11}} \\ &= \frac{2}{5}. \end{aligned}$$

[2] Note that in association mining, we would normally also include features f_i such that $f_i = 1$ if the corresponding feature is less than the median; for brevity we will not consider features of this kind in this problem.

Question 18. We will develop a decision tree classifier in order to dermine wether a country is relatively poor ($y_b = 0$) or rich ($y_b = 1$) considering only the data in Table 5. During the training of the classifier the purity gain using feature f_1 corresponding to thresholding x_1 by the median value is evaluated by Hunt's algortihm as the first decision in the tree (i.e., as decision for the root of the tree). As impurity measure we will use Gini which is given by $I(v) = 1 - \sum_c p(c|v)^2$.

What is the purity gain Δ of this considered split?

A. $\Delta = 0.000$

B. $\Delta = 0.059$

C. $\Delta = 0.125$

D. $\Delta = 0.148$

E. Don't know.

Solution 18. The purity gain is given by

$$\Delta = I(r) - \sum_{k=1}^{K} \frac{N(v_k)}{N} I(v_k),$$

where

$$I(v) = 1 - \sum_c p(c|v)^2.$$

Evaluating the purity gain for the split we have:

$$\begin{aligned}\Delta =&(1 - ((8/11)^2 + (3/11)^2)\\ &-[\tfrac{4}{11}(1 - ((2/4)^2 + (2/4)^2)\\ &+\tfrac{7}{11}(1 - ((6/7)^2 + (1/7)^2)]\\ &= 0.059\end{aligned}$$

Question 19. We again consider the dataset in Table 5. This time it is decided to group the observations according to f_2 corresponding to having a relatively low or high death rate (DeathRt). We will thereby cluster the observations such that $f_2 = 0$ corresponds to observations in the first cluster and $f_2 = 1$ corresponds to observations in the second cluster[3]. We wish to compare this clustering to that corresponding to the true class labels $y_b = 0$ and $y_b = 1$ according to the Jaccard index. Recall that the Jaccard index is given by $J = \frac{S}{N(N-1)/2-D}$ where S denotes the number of pairs of observations assigned to the same cluster that are in the same class, and D denotes the number of pairs of observations assigned to different clusters that are also in different classes. What is the value of J between the true class labels given by $y_b = 0$ and $y_b = 1$ and the two extracted clusters given by $f_2 = 0$ and $f_2 = 1$?

A. $J = 0.0909$

B. $J = 0.5273$

C. $J = 0.7436$

D. $J = 0.7838$

E. Don't know.

Solution 19. for the elleven observations we have the true labels are $\boldsymbol{y}_b = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]^\top$ and the clustering is given as $\boldsymbol{z} = [2\ 2\ 2\ 2\ 2\ 2\ 2\ 2\ 1\ 1\ 2]^\top$. We now consider all object pairs in same cluster having same class, for the 9 observations in cluster 2 we have that 8 are in the same class giving 8(8-1)/2 pairs and for the two observations in cluster 1 we have that both are in the same class giving 2(2-1)/2 pairs. Thus, $S = 8(8-1)/2+2(2-1)/2 = 29$. For the 9 observations in cluster 2 we have that 8 have $y_b = 0$ whereas both the two observations in cluster 2 have $y_b = 1$. Thus, $D = 8 \cdot 2 = 16$. As a result, we have that $J = \frac{S}{N(N-1)/2-D} = 29/(11 \cdot (11-1)/2 - 16) = 0.7436$,

[3] This clustering would correspond to the optimally converged k-means solution for $k = 2$ clusters using only the binary feature f_2 as input to the k-means algorithm

Question 20. Consider the binarized version of the Poverty dataset shown in Table 5. The matrix can be considered as representing $N = 11$ transactions $o_1, o_2, \ldots, o_{11}$ and $M = 5$ items $f_1, f_2, \ldots, f_5$. Which one of the following options represents all (non-empty) itemsets with support greater than 0.3 (and only itemsets with support greater than 0.3)?

A. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_5\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_1, f_2\}$, $\{f_1, f_3\}$, $\{f_2, f_3\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_1, f_2\}$, $\{f_1, f_3\}$, $\{f_2, f_3\}$, $\{f_1, f_2, f_3\}$

D. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_5\}$, $\{f_1, f_2\}$, $\{f_1, f_3\}$, $\{f_2, f_3\}$, $\{f_4, f_5\}$, $\{f_1, f_2, f_3\}$

E. Don't know.

Solution 20. Recall the support of an itemset is the number of rows containing all items in the itemset divided by the total number of rows. Therefore, to have a support of 0.3, an itemset needs to be contained in 4 rows. Only option C has this property and the itemsets are found by first identifying one-itemsets $\{f_1\}$, $\{f_2\}$, $\{f_3\}$ then combining these one-itemsets to form two itemsets and keeping itemsets with support greater than 0.3 we obtain $\{f_1, f_2\}$, $\{f_1, f_3\}$, $\{f_2, f_3\}$. Finally, forming the candidate three itemsets we see that only $\{f_1, f_2, f_3\}$ have support greater than 0.3.

Question 21. We again consider the binary matrix from Table 5 as a market basket problem consisting of $N = 11$ transactions $o_1, \ldots, o_{11}$ and $M = 5$ items $f_1, \ldots, f_5$. What is the *confidence* of the rule $\{f_1, f_2\} \rightarrow \{f_3\}$?

A. The confidence is $\frac{6}{11}$

B. The confidence is $\frac{7}{11}$

C. The confidence is $\frac{3}{4}$

D. The confidence is 1

E. Don't know.

Solution 21. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_1, f_2\} \cup \{f_3\})}{\text{support}(\{f_1, f_2\})} = \frac{\frac{6}{11}}{\frac{6}{11}} = 1.$$

Therefore, answer D is correct.

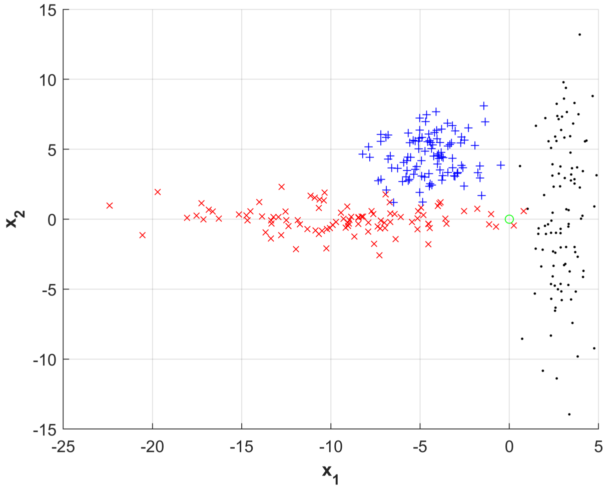

Figure 8: A dataset is separated into three clusters each having 100 observations given by blue plusses, red crosses and black dots. We would like to assign a new observation given by the green circle to one of the three clusters.

Question 22. Consider the data set given in Figure 8 in which three clusters have been extracted given by blue plusses, red crosses and black dots. We have a new observation given by the green circle located at $(0, 0)$. We assign the green observation to one of the three cluster by considering the proximity measure as computed based on Euclidean distance between the green point, and the points in the cluster.

Which one of the following statements is correct?

A. If we use *maximum* linkage the new observation will be assigned to the cluster given by blue plusses.

B. If we use *minimum* linkage the new observation will be assigned to the cluster given by black dots.

C. If we use *average* linkage the new observation will be assigned to the cluster given by red crosses.

D. If we use *minimum* linkage the new observation will be assigned to the cluster given by blue plusses.

E. Don't know.

Solution 22. The correct answer is A. Minimum linkage will result in the new observation assigned to red crosses as the closest most observation is a red cross. Maximum linkage will assign the new observation to the blue plusses according to the furthest most observation of each cluster having a blue plus as closest to the new observation. Average linkage corresponds to considering the avarege distance and can be considered a center based approach. Here the center of the red crosses are furthest away and therefore this cannot be the cluster the new observation is assigned to.

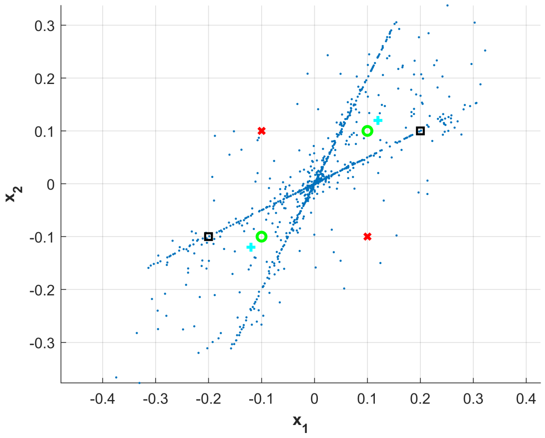

Figure 9: A dataset of 1000 observations given by the blue dots. In the plot is also given the location of two red crosses, two green circles, two cyan plusses and two black squares.

Question 23.

Consider the dataset given in Figure 9. We will consider the Mahanalobis distance using the empirical covariance matrix estimated based on the 1000 blue observations. Which one of the following statements is correct?

A. The Mahanalobis distance between the two green circles is smaller than the Mahanalobis distance between the two black squares.

B. The Mahanalobis distance between the two red crosses is the same as the Mahanalobis distance between the two green circles.

C. The Mahanalobis distance between the two black squares is smaller than the Mahanalobis distance bewteen the two cyan plusses.

D. The empirical covariance matrix estimated based on the blue observations has at least one element that is negative.

E. Don't know

Solution 23. As the correlation between x_1 and x_2 is positive the covariance matrix only has positive elements. The covariance matrix will have a shape in the direction of the green cirles and blue plusses and therefore these pairs of observations will have relatively short Mahanalobis distance between each other, when compared to the other pairs of observations. Thus, the

Mahanalobis distance between the two green circles is smaller than the Mahanalobis distance between the two black squares.

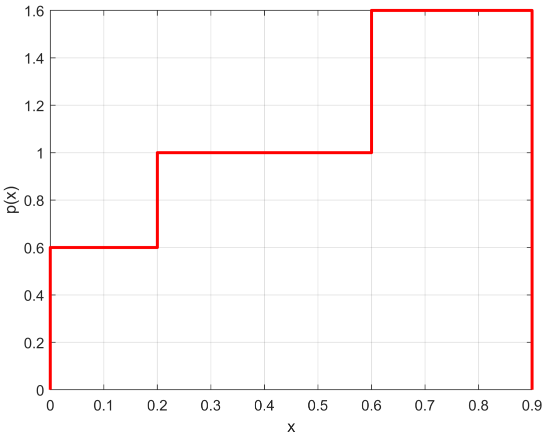

Figure 10: Probability density function for a random variable x. Outside the region from 0 to 0.9 the density function is zero.

Question 24. In Figure 10 is given the denstity function $p(x)$ of a random variable x. What is the expected value of x, i.e. $\mathbb{E}[x]$?

A. 0.450

B. 0.532

C. 0.600

D. 1.000

E. Don't know.

Solution 24.

$$\begin{aligned}\mathbb{E}[x] &= \int xp(x) = \int_0^{0.2} x \cdot 0.6 + \int_{0.2}^{0.6} x \cdot 1 + \int_{0.6}^{0.9} x \cdot 1.6 \\ &= 0.6 \cdot 0.5 \cdot (0.2^2 - 0^2) + 1 \cdot 0.5 \cdot (0.6^2 - 0.2^2) \\ &+ 1.6 \cdot 0.5 \cdot (0.9^2 - 0.6^2) = 0.532,\end{aligned}$$

where we have used that $\int_a^b x dx = 0.5 \cdot (b^2 - a^2)$

Question 25. Consider again the Poverty dataset from Table 1 and in particular the first three attributes x_1, x_2 and x_3 of the 35'th and 53'th observation

$$\boldsymbol{x}_{35} = \begin{bmatrix} -1.24 \\ -0.26 \\ -1.04 \end{bmatrix}, \quad \boldsymbol{x}_{53} = \begin{bmatrix} -0.60 \\ -0.86 \\ -0.50 \end{bmatrix}.$$

Let the p-norm distance be donoted $d_p(\cdot, \cdot)$ and the cosine similarity be denoted $cos(\cdot, \cdot)$. Which one of the following statements is correct?

A. $d_{p=1}(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.64$

B. $d_{p=4}(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.79$

C. $d_{p=\infty}(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.68$

D. $cos(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.67$

E. Don't know.

Solution 25.

Solving this problem simply consist of recalling the definition of the p-norm since $d_p(x, y) = \|x - y\|_p$. For $1 \leq p < \infty$ it is:

$$\|\boldsymbol{x}\|_p = \left(\sum_{j=1}^{3} |x_j|^p \right)^{\frac{1}{p}}$$

and for $p = \infty$:

$$\|\boldsymbol{x}\|_p = \max\{|x_1|, \ldots, |x_3|\}.$$

We see the correct values are:

- $d_{p=1}(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 1.78$
- $d_{p=4}(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.79$
- $d_{p=\infty}(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.64$
- $cos(\boldsymbol{x}_{35}, \boldsymbol{x}_{53}) = 0.78$

Therefore, answer B is correct.

Question 26. Which one of the following statements regarding machine learning and cross-validation is correct?

A. In machine learning we are mainly concerned about the training error as opposed to the test error.

B. As we get more training data the trained model becomes more prone to overfitting.

C. For a classifier the test error rate will in general be lower than the training error rate.

D. The number of observations used for testing is the same for five-fold and ten-fold cross-validation.

E. Don't know

Solution 26. In machine learning we are mainly concerned with model generalization and here the test-error using cross-validation is used to estimate the generalization. As we get more training data a model trained will be less prone to overfitting and not the reverse. In general, the training error will be lower than the test error due to overfitting. Only when we have a very large training set can we expect overfitting to be negligible and the training and test error to have same magnitude. The number of observations used for testing is the same when we use K-fold cross-validation for all values of K as all observations are used once for testing.

Variable	$t=1$	$t=2$	$t=3$	$t=4$
y_1	0	1	1	1
y_2	0	1	0	0
y_3	0	1	1	1
y_4	1	1	1	1
y_5	0	0	1	1
y_6	1	1	1	0
y_7	1	1	1	1
y_8	0	0	1	1
y_9	0	1	1	1
y_{10}	0	0	1	1
y_{11}	0	1	1	1
y_{12}	1	0	1	1
y_1^{test}	0	1	0	0
y_2^{test}	0	1	1	1
ϵ_t	0.417	0.243	0.307	0.534

Table 6: Tabulation of each of the predicted outputs of the AdaBoost classifiers, as well as the weighted error rate ϵ_t when evaluating the AdaBoost algorithm for $T = 4$ rounds. Note the table includes the prediction of the two test points in Figure 11.

Question 27.

Consider again the Poverty dataset of Table 1. Suppose we limit ourselves to $N = 12$ randomly selected observations from the original dataset and only consider the features x_2 and x_5. We apply a KNN classification model ($K = 1$) to this dataset and use AdaBoost in order to enhance the performance of the classifier. During the first $T = 4$ rounds of boosting, we obtain the decision boundaries shown in Figure 11. The figure also contains two test observations marked by a cross and a square located respectively at $\boldsymbol{x}_1^{test}$ and $\boldsymbol{x}_2^{test}$.

The prediction of the intermediate AdaBoost classifiers and ϵ_t are given in Table 6. Using this information, how will the AdaBoost classifier as obtained by combining the $T = 4$ weak KNN-classifiers classify the two test observations $\boldsymbol{x}_1^{test}$ and $\boldsymbol{x}_2^{test}$?

A. $\tilde{y}_1^{\text{test}} = 0$ and $\tilde{y}_2^{\text{test}} = 0$

B. $\tilde{y}_1^{\text{test}} = 1$ and $\tilde{y}_2^{\text{test}} = 0$

C. $\tilde{y}_1^{\text{test}} = 0$ and $\tilde{y}_2^{\text{test}} = 1$

D. $\tilde{y}_1^{\text{test}} = 1$ and $\tilde{y}_2^{\text{test}} = 1$

E. Don't know.

Solution 27.

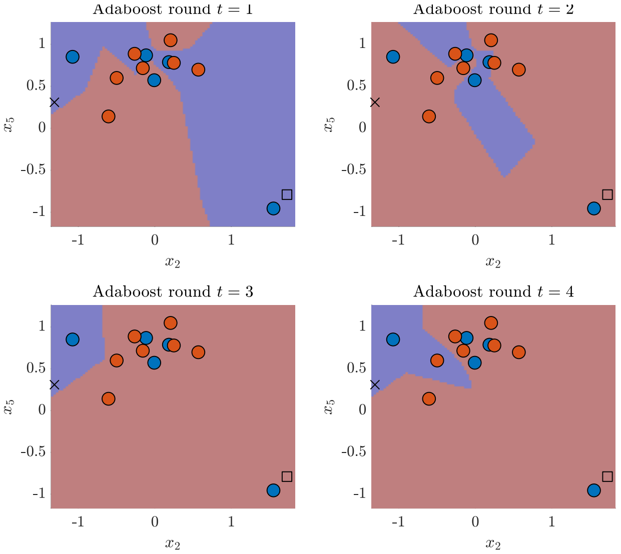

Figure 11: Decision boundaries for a KNN classifier for K=1 enhanced using $T = 4$ rounds of boosting. Notice, in addition to the training data the plot also includes two test points marked respectively by a black cross ($\boldsymbol{x}_1^{test}$) and square ($\boldsymbol{x}_2^{test}$). Observations in blue corresponds to low GNP ($y_b = 0$) whereas observations in red corresponds to high GNP ($y_b = 1$) and the associated class specific decision boundaries are respectively also given in blue and red.

According to the AdaBoost algorithm, the classification rule when combining T AdaBoost algorithms is:

$$f^*(\boldsymbol{x}) = \arg\max_{y=0,1} \sum_{t=1}^{T} \alpha_t \delta_{f_t(\boldsymbol{x}),y},$$

where $\alpha_t = 0.5log(\frac{1-\epsilon_t}{\epsilon_t})$. In other words, the classification rule is obtained by summing the α_t where $f_t(\boldsymbol{x}) = 0$ (as F_0) and those where $f_t(\boldsymbol{x}) = 1$ (as F_1) and then selecting the y corresponding to the largest value. For the four rounds we obtain $\alpha_1 = 0.168$, $\alpha_2 = 0.568$, $\alpha_3 = 0.407$ and $\alpha_4 = -0.068$ and we thereby have for the two test points:

$$F_0(\boldsymbol{x}_1^{\text{test}}) = \alpha_1 + \alpha_3 + \alpha_4 = 0.507$$
$$F_1(\boldsymbol{x}_1^{\text{test}}) = \alpha_2 = 0.568$$
$$F_0(\boldsymbol{x}_2^{\text{test}}) = \alpha_1 = 0.168$$
$$F_1(\boldsymbol{x}_2^{\text{test}}) = \alpha_2 + \alpha_3 + \alpha_4 = 0.907.$$

As a result, we have $y_1^{\text{test}} = 1$ and $y_2^{\text{test}} = 1$.