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.

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.

	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.

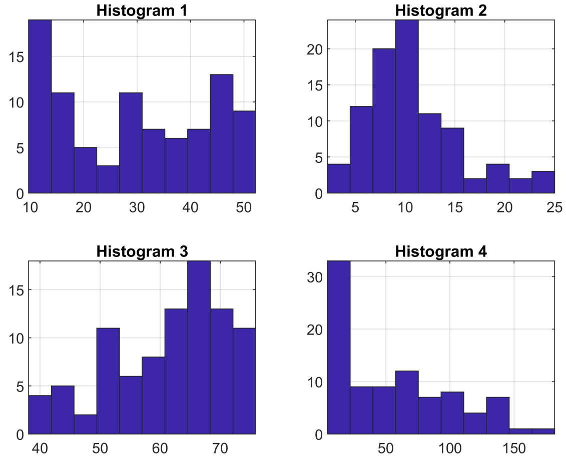

Figure 1: Four histograms corresponding to the variables with summary statistics given in Table 2 but not necessarily in that order.

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.

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

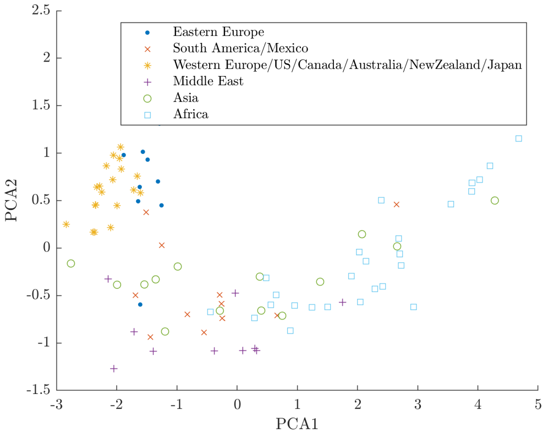

Figure 2: The Poverty data projected onto the first two principal component directions with each observation labelled according to the region it belongs to (given by x_6).

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.

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.

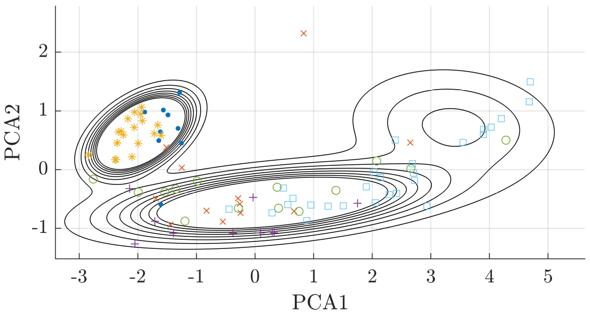

Figure 3: A GMM with K=3 clusters fitted to the poverty data projected onto the first two principal component directions. Each observation is again labelled according to the region it belongs to (given by x_6).

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.

	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:

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

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

where $N_{\boldsymbol{X}_{\backslash i}}(\boldsymbol{x}_i, K)$ 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.

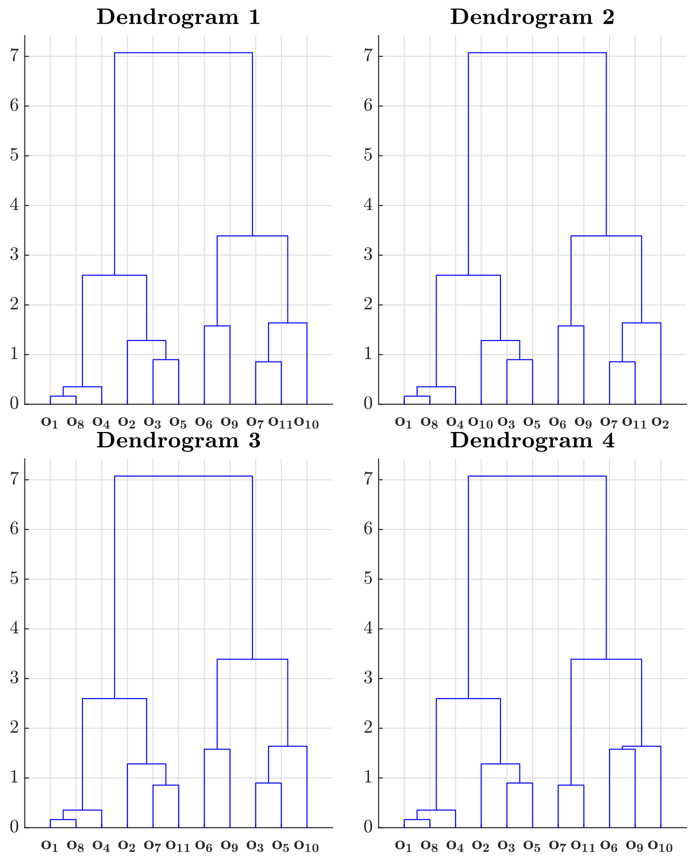

Figure 4: Four dendrograms for which one of the dendrograms corresponds to hierarchical clustering using maximum linkage of the 11 observations in Table 3.

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.

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.

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.

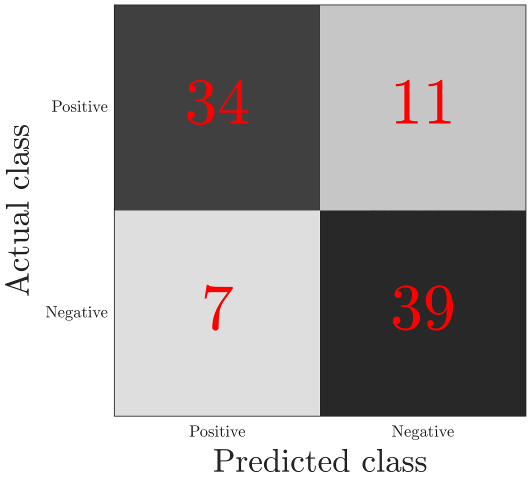

Figure 5: 5-fold cross validation applied to the entire dataset to evaluate logistic regression as an approach to predict low GNP ($y_b = 0$, negative class) versus high GNP ($y_b = 1$, positive class).

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.

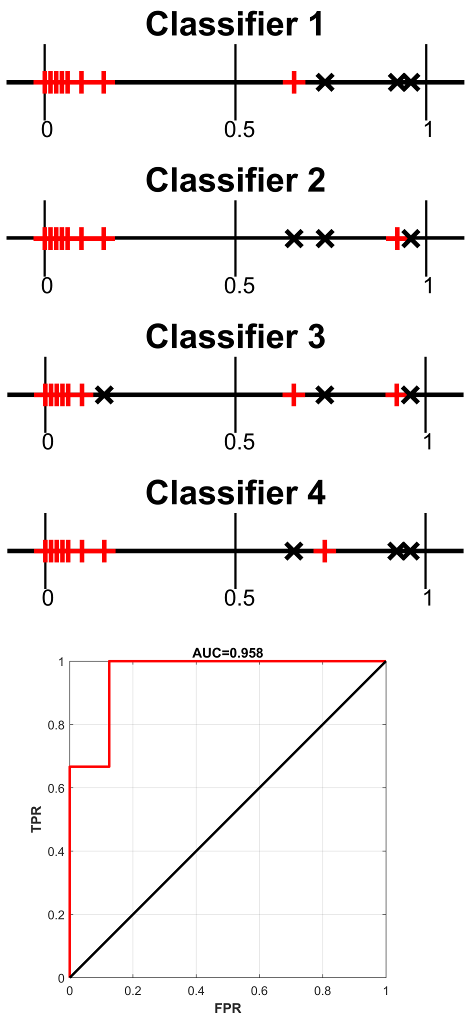

Figure 6: Top panel: Four different logistic regression models used to predict low GNP ($y_b = 0$, marked by red plusses) from high GNP ($y_b = 1$, marked by black crosses). Bottom panel: The ROC curve corresponding to one of the four classifiers in the top panel.

Question 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.

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.

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.

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.

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.

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.

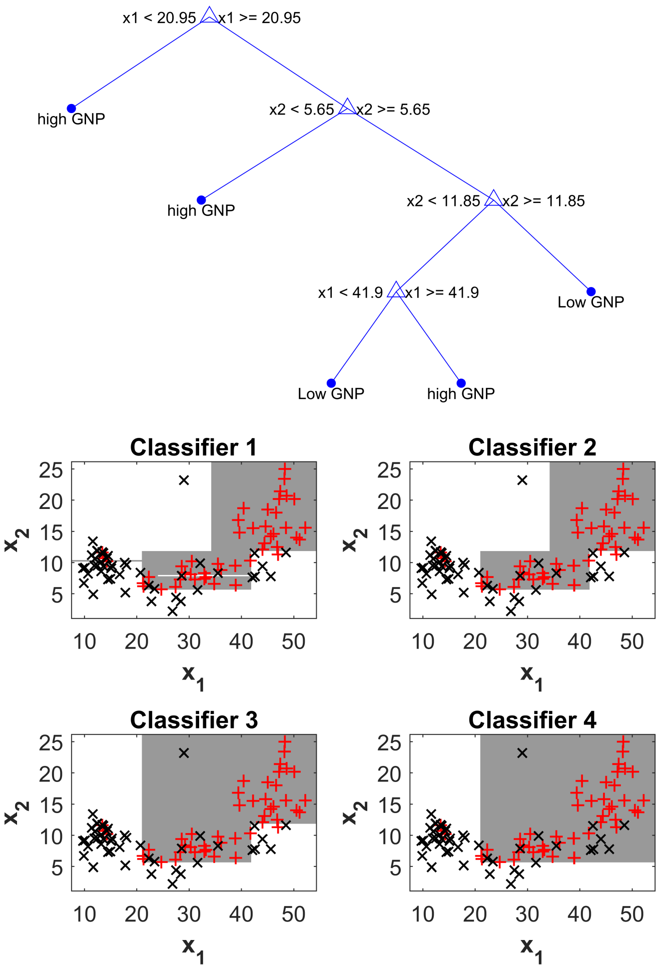

Figure 7: Top panel, a decision tree fitted to x_1 and x_2 of the Poverty data in order to predict wheter a country has relatively low or high GNP. Bottom panel, decision boundaries for four different decision trees in which gray regions correspond to regions predicted having low GNP ($y_b = 0$) and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

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

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

E. Don't know.

[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.

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.

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.

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.

[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

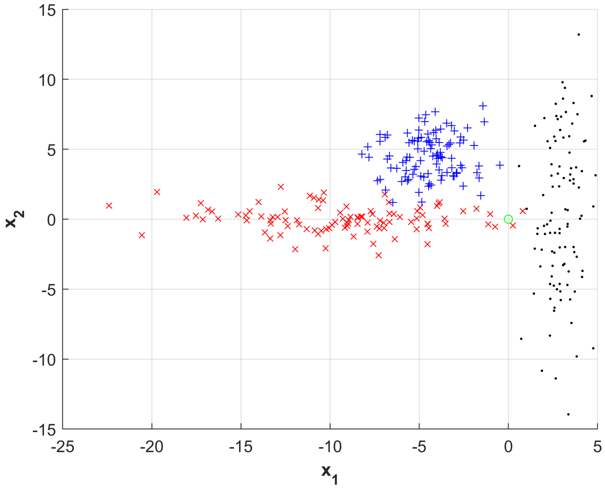

Figure 8: A dataset is separated into three clusters each having 100 observations given by blue plusses, red crosses and black dots. We would like to assign a new observation given by the green circle to one of the three clusters.

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.

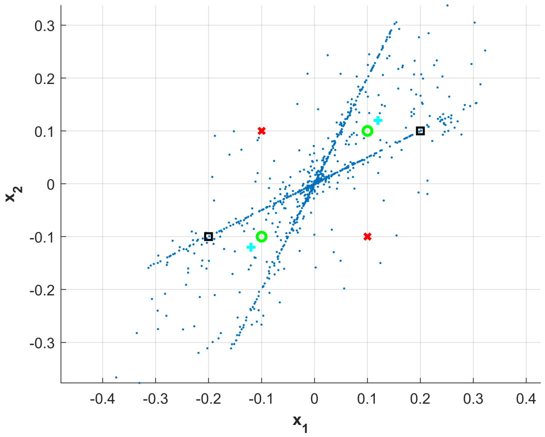

Figure 9: A dataset of 1000 observations given by the blue dots. In the plot is also given the location of two red crosses, two green circles, two cyan plusses and two black squares.

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

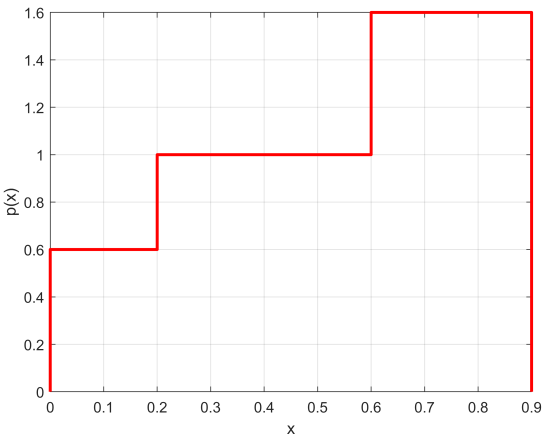

Figure 10: Probability density function for a random variable x. Outside the region from 0 to 0.9 the density function is zero.

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.

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.

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

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.

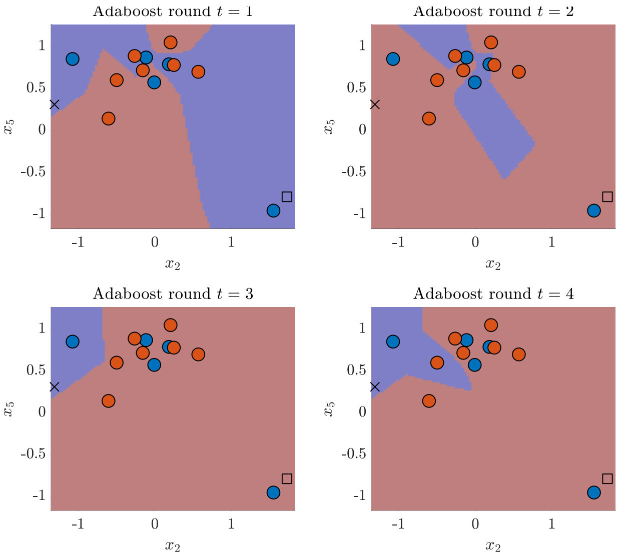

Figure 11: Decision boundaries for a KNN classifier for K=1 enhanced using $T = 4$ rounds of boosting. Notice, in addition to the training data the plot also includes two test points marked respectively by a black cross ($\boldsymbol{x}_1^{test}$) and square ($\boldsymbol{x}_2^{test}$). Observations 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.