Technical University of Denmark

Written examination: December 14th 2021, 9 AM — 1 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.

This exam only allows for electronic hand-in.
You hand in your answers at `https://eksamen.dtu.dk/`. To hand in your answers, write them in the file `answers.txt` (this file is available from the same place you downloaded this file). When you are done, upload the `answers.txt` file (and nothing else). Double-check that you uploaded the correct version of the file from your computer.

Do not change the format of `answers.txt`
The file is automatically parsed after hand-in. Do not change the file format of `answers.txt` to any other format such as `rtf`, `docx`, or `pdf`. Do not change the file structure. Only edit the portions of the file indicated by question marks.

Answers:

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

No.	Attribute description	Abbrev.
x_1	palmitic fatty acid content	palmitic
x_2	palmitoleic fatty acid content	palmitoleic
x_3	stearic fatty acid content	stearic
x_4	oleic fatty acid content	oleic
x_5	linoleic fatty acid content	linoleic
x_6	arachidic fatty acid content	arachidic
x_7	linolenic fatty acid content	linolenic
x_8	eicosenoic fatty acid content	eicosenoic
y	Region of origin in Italy	region

Table 1: Description of the features of the Olive Oil dataset used in this exam. The dataset consists of eight fatty acids measurements for olive oils from nine different regions of Italy. The content of each fatty acid is measured in percentages, i.e. in the interval $[0; 100]$. The dataset used here consists of $N = 572$ observations and the attribute y is discrete so that $y = 1$ (corresponding to North Apulia), $y = 2$ (corresponding to Calabria), $y = 3$ (corresponding to South Apulia), $y = 4$ (corresponding to Sicily), $y = 5$ (corresponding to Inner Sardinia), $y = 6$ (corresponding to Coastal Sardinia), $y = 7$ (corresponding to East Liguria), $y = 8$ (corresponding to West Liguria), and $y = 9$ (corresponding to Umbria).

Question 1. The main dataset used in this exam is the Olive Oil dataset[1] described in Table 1. In Figure 1 and Figure 2 are shown respectively histogram plots and boxplots of the attributes x_1 (palmitic), x_3 (stearic), x_5 (linoleic), and x_8 (eicosenoic) from the Olive Oil dataset described in Table 1. Which histogram plots

[1]Dataset obtained from `https://www2.chemie.uni-erlangen.de/publications/ANN-book/datasets/`

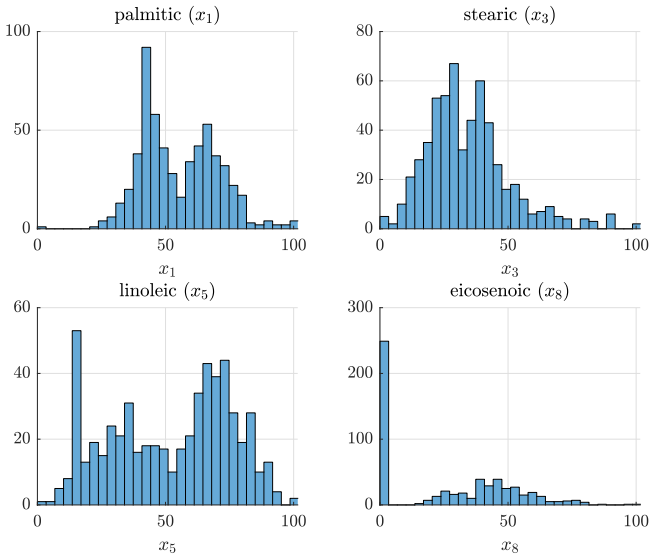

Figure 1: Plot of the observations of attributes x_1, x_3, x_5 and x_8 from the Olive Oil dataset of Table 1 as histogram plots.

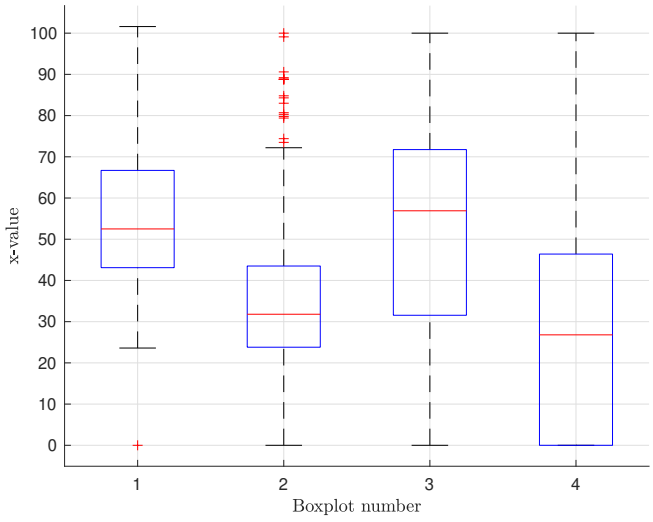

Figure 2: Boxplots corresponding to the variables plotted in Figure 1 but not necessarily in that order.

match which boxplots?

A. Boxplot 1 is x_5 (linoleic), boxplot 2 is x_1 (palmitic), boxplot 3 is x_3 (stearic) and boxplot 4 is x_8 (eicosenoic)

B. Boxplot 1 is x_1 (palmitic), boxplot 2 is x_8 (eicosenoic), boxplot 3 is x_5 (linoleic) and boxplot 4 is x_3 (stearic)

C. Boxplot 1 is x_1 (palmitic), boxplot 2 is x_3 (stearic), boxplot 3 is x_5 (linoleic) and boxplot 4 is x_8 (eicosenoic)

D. Boxplot 1 is x_1 (palmitic), boxplot 2 is x_5 (linoleic), boxplot 3 is x_8 (eicosenoic) and boxplot 4 is x_3 (stearic)

E. Don't know.

Solution 1. From the histograms, we see that x_3 (stearic) has a long right tail. For x_8 (eicosenoic) more than a quarter of the observations are close to 0, which means that the first quartile must also be close to 0. Using this knowledge boxplot 2 is matched to x_3 (stearic), and boxplot 4 is matched to x_8 (eicosenoic). Therefore ption C is the only correct.

Question 2. In this question we will only consider the first five attributes x_1, x_2, x_3, x_4 and x_5 of the Olive Oil dataset in Table 1. A scatter plot matrix for these attributes is shown in Figure 3. We also calculate the empirical covariance matrix, $\hat{\mathbf{\Sigma}}$, for the first five attributes. Which one of the following matrices is the correct empirical covariance matrix for these attributes?

A.
$$\begin{bmatrix} 564.3 & -77.5 & 292.5 & -388.5 & 164.0 \\ -77.5 & 271.5 & -72.5 & 36.0 & -42.0 \\ 292.5 & -72.5 & 392.4 & -324.8 & 248.1 \\ -388.5 & 36.0 & -324.8 & 369.9 & -241.4 \\ 164.0 & -42.0 & 248.1 & -241.4 & 224.6 \end{bmatrix}$$

B.
$$\begin{bmatrix} -564.3 & -77.5 & 292.5 & -388.5 & 164.0 \\ -77.5 & -271.5 & -72.5 & 36.0 & -42.0 \\ 292.5 & -72.5 & -392.4 & -324.8 & 248.1 \\ -388.5 & 36.0 & -324.8 & -369.9 & -241.4 \\ 164.0 & -42.0 & 248.1 & -241.4 & -224.6 \end{bmatrix}$$

C.
$$\begin{bmatrix} 224.6 & 248.1 & -42.0 & -241.4 & 164.0 \\ 248.1 & 392.4 & -72.5 & -324.8 & 292.5 \\ -42.0 & -72.5 & 271.5 & 36.0 & -77.5 \\ -241.4 & -324.8 & 36.0 & 369.9 & -388.5 \\ 164.0 & 292.5 & -77.5 & -388.5 & 564.3 \end{bmatrix}$$

D.
$$\begin{bmatrix} -224.6 & 248.1 & -42.0 & -241.4 & 164.0 \\ 248.1 & -392.4 & -72.5 & -324.8 & 292.5 \\ -42.0 & -72.5 & -271.5 & 36.0 & -77.5 \\ -241.4 & -324.8 & 36.0 & -369.9 & -388.5 \\ 164.0 & 292.5 & -77.5 & -388.5 & -564.3 \end{bmatrix}$$

E. Don't know.

Solution 2. Recall that the structure of the covariance matrix is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} & \sigma_{1,5} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} & \sigma_{2,4} & \sigma_{2,5} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} & \sigma_{3,4} & \sigma_{3,5} \\ \sigma_{4,1} & \sigma_{4,2} & \sigma_{4,3} & \sigma_{4,4} & \sigma_{4,5} \\ \sigma_{5,1} & \sigma_{5,2} & \sigma_{5,3} & \sigma_{5,4} & \sigma_{5,5} \end{bmatrix}$$

where $\sigma_{i,j} = \text{cov}(x_i, x_j)$.

We can rule out answer B and D, as they are not valid covariance matrices, since the diagonal is negative and $\sigma_{i,i} = \text{Var}(x_i) \geq 0$. For the scatter plots, we for instance see that $\sigma_{1,2} > 0$ and $\sigma_{1,4} < 0$. Of answer A and C, only the matrix in answer C satisfy this. Therefore the correct answer is C.

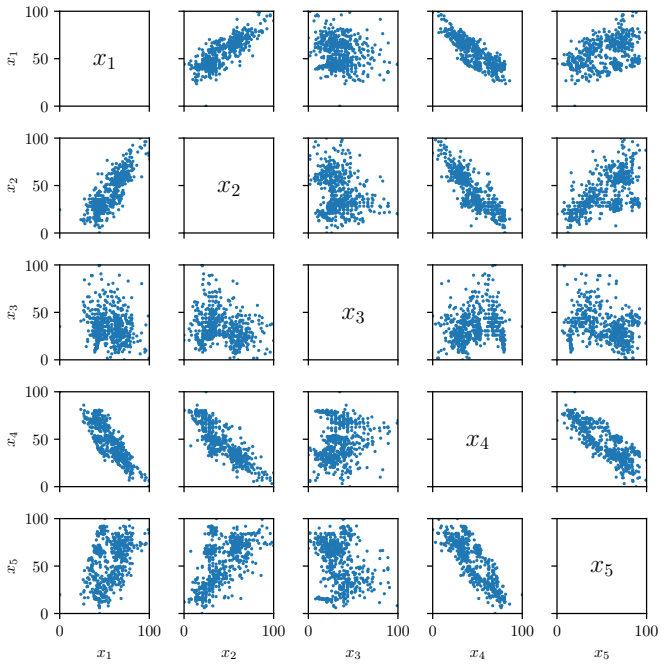

Figure 3: Scatter plot matrix for the attributes x_1, x_2, x_3, x_4, x_5 of the Olive Oil dataset of Table 1.

Question 3. A Principal Component Analysis (PCA) is carried out on the Olive Oil dataset in Table 1 based on the attributes x_1, x_2, x_3, x_4 and 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 $\boldsymbol{USV}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.48 & 0.09 & -0.57 & 0.52 & 0.42 \\ 0.51 & 0.03 & -0.27 & -0.82 & 0.05 \\ -0.15 & 0.98 & 0.03 & -0.07 & 0.08 \\ -0.54 & -0.16 & -0.14 & -0.25 & 0.78 \\ 0.45 & 0.01 & 0.77 & 0.05 & 0.46 \end{bmatrix} \tag{1}$$

$$\boldsymbol{S} = \begin{bmatrix} 43.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 23.39 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 18.26 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 9.34 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 2.14 \end{bmatrix}.$$

Which one of the following statements is true?

A. The variance explained by the last four principal components is less than 0.3 of the total variance.

B. The variance explained by the first three principal components is greater than 0.9 of the total variance.

C. The variance explained by the first four principal components is less than 0.95 of the total variance.

D. The variance explained by the first principal component is greater than 0.715 of the total variance.

E. Don't know.

Solution 3. The correct answer is B. To see this, recall 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 components x_1, x_2, x_3 is:

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2} = 0.9679.$$

Question 4. Consider again the PCA analysis for the Olive Oil dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (1). Which one of the following statements is true?

A. An observation with a low value of x_1 (palmitic), a low value of x_2 (palmitoleic), a high value of x_4 (oleic), and a low value of x_5 (linoleic) will typically have a negative value of the projection onto principal component number 1.

B. An observation with a high value of x_3 **(stearic)** will typically have a negative value of the projection onto principal component number 2.

C. An observation with a low value of x_1 **(palmitic)**, a high value of x_2 **(palmitoleic)**, and a high value of x_4 **(oleic)** will typically have a positive value of the projection onto principal component number 4.

D. An observation with a low value of x_1 **(palmitic)**, a low value of x_2 **(palmitoleic)**, and a high value of x_5 **(linoleic)** will typically have a negative value of the projection onto principal component number 3.

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

$$b_1 = \boldsymbol{x}^\top \boldsymbol{v}_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \begin{bmatrix} 0.48 \\ 0.51 \\ -0.15 \\ -0.54 \\ 0.45 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and negative, this occurs if x_1, x_2, x_4, x_5 has large magnitude and the sign convention given in option A.

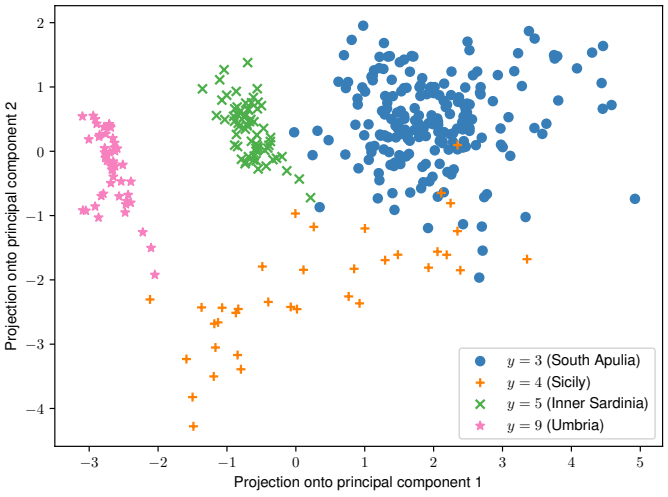

Figure 4: Scatter plot of the projection of observations belonging four classes from the Olive Oil dataset in Table 1 onto the first two principal components.

Question 5. A Principal Component Analysis (PCA) is carried out on all the eight attributes of the Olive Oil dataset in Table 1. All the objects from four regions of origin are projected onto the first two principal components and visualised as a scatter plot in Figure 4. Which one of the following statements is true?

A. There exists a logistic regression classifier that takes the observations projected onto the first two principal components as input, which can binary classify the observations in the two regions South Apulia ($y = 3$) and Sicily ($y = 4$) with 0 error.

B. Any classifications tree using axis-aligned splits that takes the observation projected onto the first two principal components as input and binary classify the observations in the two regions South Apulia ($y = 3$) and Umbria ($y = 9$) has an error strictly greater than 0

C. Any classification tree using axis-aligned splits that takes all eighth attributes as input and binary classify the observations in the two regions South Apulia ($y = 3$) and Inner Sardinia ($y = 5$) has an error strictly greater than 0.

D. There exists a logistic regression classifier that takes all eighth attributes as input, which can binary classify the observations in the two regions South Apulia ($y = 3$) and Umbria ($y = 9$) with 0 error.

E. Don't know.

Solution 5.

- Answer A is incorrect, since the points of the two classes South Apulia ($y = 3$) and Sicily ($y = 4$) are not linearly separable in Figure 4.
- Answer B is incorrect, since a tree with two leafs (splitting e.g. around -1 in the projection onto the first principal component) will be able to perfectly classify the objects.
- Answer C is incorrect, since a classification tree is always able to obtain an error of 0 when there is no identical training object in the two classes (unless the three complexity is limited).
- Answer D is correct, since the two classes South Apulia ($y = 3$) and Umbria ($y = 9$) are linearly separable in the PCA plot. Furthermore, if points are linearly separable in the projection onto the first two principal components, then the are also linearly separable in the original attribute space.

	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	53.8	87.0	67.4	67.5	71.2	65.2	117.9	56.1	90.3	109.8
o_2	53.8	0.0	69.9	75.5	62.9	58.0	63.0	135.0	84.1	107.9	131.5
o_3	87.0	69.9	0.0	49.7	38.5	19.3	35.5	91.8	76.9	78.7	89.1
o_4	67.4	75.5	49.7	0.0	24.2	47.2	47.0	62.3	33.4	37.2	60.0
o_5	67.5	62.9	38.5	24.2	0.0	37.7	41.7	79.5	52.4	60.2	78.9
o_6	71.2	58.0	19.3	47.2	37.7	0.0	21.5	95.6	68.3	78.4	91.0
o_7	65.2	63.0	35.5	47.0	41.7	21.5	0.0	96.0	64.3	75.5	89.4
o_8	117.9	135.0	91.8	62.3	79.5	95.6	96.0	0.0	66.9	44.3	24.2
o_9	56.1	84.1	76.9	33.4	52.4	68.3	64.3	66.9	0.0	39.2	60.7
o_{10}	90.3	107.9	78.7	37.2	60.2	78.4	75.5	44.3	39.2	0.0	39.4
o_{11}	109.8	131.5	89.1	60.0	78.9	91.0	89.4	24.2	60.7	39.4	0.0

Table 2: 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 Olive Oil dataset (recall that $M = 8$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1. The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belong to class C_1 (corresponding to North Apulia), the red observations $\{o_3,\ o_4,\ o_5,\ o_6,\ o_7\}$ belong to class C_2 (corresponding to Calabria), and the blue observations $\{o_8,\ o_9,\ o_{10},\ o_{11}\}$ belong to class C_3 (corresponding to South Apulia). To avoid single features to dominate, the dataset was standardized by subtracting the mean and dividing by the standard deviation.

Question 6. Consider the distances in Table 2 based on 11 observations from the Olive Oil dataset. The class labels $C_1,\ C_2,\ C_3$ (see table caption for details) will be predicted using a K-nearest neighbour classifier based on the distances given in Table 2 (ties are broken in the usual manner by considering the nearest observation from the tied classes). We will apply a 3-nearest neighbour classifier (i.e., $K = 3$) and *hold-out cross-validation*, in which the 11 observations is split into a training and test set. The training and test set is given by the observations:

$$\mathcal{D}^{\text{train}} = \{o_1, o_2, o_3, o_6, o_7, o_8, o_9, o_{11}\}$$
$$\mathcal{D}^{\text{test}} = \{o_4, o_5, o_{10}\}$$

If we train the model on the training set, what is the accuracy as computed on the test set?

A. accuracy $= 0$

B. accuracy $= \frac{1}{3}$

C. accuracy $= \frac{2}{3}$

D. accuracy $= 1$

E. Don't know.

Solution 6. The correct answer is D. To compute the accuracy for a particular observation o_i in the test set $\mathcal{D}^{\text{test}}$, we train a model on the observations in the training set and use it to predict the class of observation o_i. Doing this is simply a matter of finding the observations in the training set closest to o_i according to Table 2 and predict o_i as belonging to the majority class.

We find that the 3-nearest neighbours for the observations in the test set are

- $N(o_4, K = 3) = \{o_9, o_7, o_6\}$
- $N(o_5, K = 3) = \{o_6, o_3, o_7\}$
- $N(o_{10}, K = 3) = \{o_9, o_{11}, o_8\}$

So

- o_4 is predicted to belong to C_2 (which is correct).
- o_5 is predicted to belong to C_2 (which is correct).
- o_{10} is predicted to belong to C_3 (which is correct).

The accuracy is then found by observing how often the class label of the observation in the neighborhood agrees with the true class label. As none of the observations are predicted to have the correct class label, the accuracy is 1.

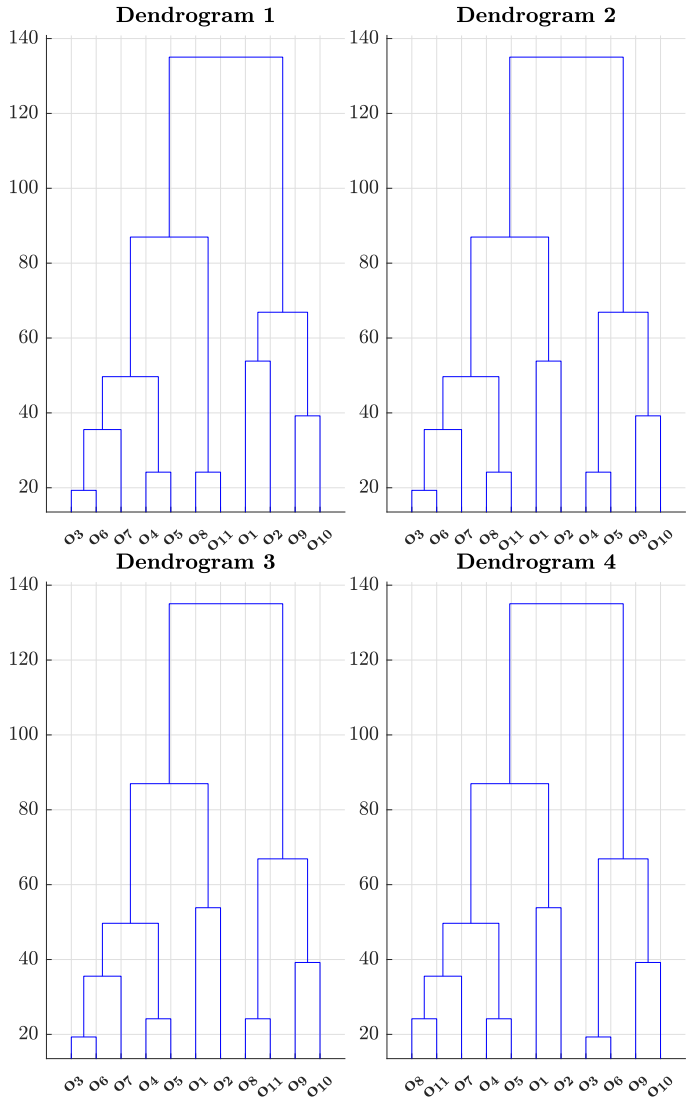

Figure 5: Proposed hierarchical clustering of the 11 observations in Table 2.

Question 7. A hierarchical clustering is applied to the 11 observations in Table 2 using *maximum* linkage. Which one of the dendrograms shown in Figure 5 corresponds to the distances given in Table 2?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

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

- In dendrogram 1, merge operation number 8 should have been between the sets $\{o_8, o_{11}\}$ and $\{o_9, o_{10}\}$, however in dendrogram 1 merge number 8 is between the sets $\{o_1, o_2\}$ and $\{o_9, o_{10}\}$.
- In dendrogram 2, merge operation number 6 should have been between the sets $\{o_3, o_6, o_7\}$ and $\{o_4, o_5\}$, however in dendrogram 2 merge number 6 is between the sets $\{o_3, o_6, o_7\}$ and $\{o_8, o_{11}\}$.
- In dendrogram 4, merge operation number 8 should have been between the sets $\{o_8, o_{11}\}$ and $\{o_9, o_{10}\}$, however in dendrogram 4 merge number 8 is between the sets $\{o_3, o_6\}$ and $\{o_9, o_{10}\}$.

Question 8. To examine if observation o_5 may be an outlier, we will calculate the K-nearest neighborhood density using only the observations and distances in Table 2. For an observation o_i, recall the density is computed using the set of K nearest neighbors of observation o_i excluding the i'th observation itself, $N_{\boldsymbol{X}_{\setminus i}}(o_i, K)$, and is denoted by $\text{density}_{\boldsymbol{X}_{\setminus i}}(o_i, K)$. What is the density for observation o_5 for $K = 3$ nearest neighbors?

A. 0.034

B. 0.030

C. 0.041

D. 0.879

E. Don't know.

Solution 8. The density is given as:

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

So to solve the problem, we only need to plug in the values. We find that the $k = 3$ neighborhood of o_5 and density is:

$$N_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = \{o_4, o_6, o_3\}$$
$$\text{density}_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = \frac{3}{24.2 + 37.7 + 37.7} \approx 0.030$$

Therefore option B is correct.

Question 9. Consider again the distances in Table 2 calculated from the Olive Oil dataset in Table 1 with $M = 8$ features. We wish to apply kernel density estimation for observations in the data-set. Apply kernel density estimation for the observation o_{11}, where *only* the closest two observations are used to estimate the kernel density and excluding o_{11}. Set the kernel width $\lambda = 20$. What is the estimated density at o_{11} using these assumptions?

A.
$$p_\lambda(o_{11}) \approx \frac{1}{2}\frac{1}{\sqrt{(2\pi \cdot 20^2)^8}} \cdot 0.6246$$

B.
$$p_\lambda(o_{11}) \approx \frac{1}{2}\frac{1}{\sqrt{(2\pi \cdot 20^2)^8}} \cdot 1.922$$

C.
$$p_\lambda(o_{11}) \approx \frac{1}{\sqrt{(2\pi \cdot 20^2)^8}} \cdot 0.6246$$

D.
$$p_\lambda(o_{11}) \approx \frac{1}{\sqrt{(2\pi \cdot 20^2)^8}} \cdot 1.922$$

E. Don't know.

Solution 9. The formula for kernel density estimation is given

$$p_\lambda(o_{11}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_i, \lambda^2 \boldsymbol{I}).$$

For the covariance matrix $\lambda^2 \boldsymbol{I}$ we can express the k-dimensional multivariate normal as

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{x}_i, \lambda^2 \boldsymbol{I}) = \frac{1}{\sqrt{(2\pi\lambda^2)^k}} \exp\left(\frac{-\|\boldsymbol{x} - \boldsymbol{x}_i\|_2^2}{2\lambda^2}\right).$$

We see that the distances reported in Table 2 can be used as $\|\boldsymbol{x} - \boldsymbol{x}_i\|_2$. If we use only the two closest observations, we have that $N = 2$ and $k = M = 8$, and thus we get

$$p_\lambda(o_{11}) = \frac{1}{2}\frac{1}{\sqrt{(2\pi \cdot 20^2)^8}}\left(\exp\left(\frac{-24.2^2}{2 \cdot 20^2}\right) + \exp\left(\frac{-39.4^2}{2 \cdot 20^2}\right)\right) \approx \frac{1}{2}\frac{1}{\sqrt{(2\pi \cdot 20^2)^8}} \cdot 0.6246$$

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
o_1	0	0	0	1	0	0	0	1
o_2	0	0	1	0	0	1	0	1
o_3	0	0	1	0	0	1	0	1
o_4	0	1	0	0	0	1	0	1
o_5	0	0	0	0	0	1	0	1
o_6	0	0	1	0	1	1	0	1
o_7	0	0	1	0	0	1	0	1
o_8	1	1	0	0	0	0	1	1
o_9	0	1	0	0	0	0	0	1
o_{10}	0	1	0	0	0	1	0	1
o_{11}	1	1	0	0	0	0	0	0

Table 3: Binarized version of the Olive Oil dataset. Each of the features f_i are obtained by taking a feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belong to class C_1 (corresponding to North Apulia), the red observations $\{o_3,\ o_4,\ o_5,\ o_6,\ o_7\}$ belong to class C_2 (corresponding to Calabria), and the blue observations $\{o_8,\ o_9,\ o_{10},\ o_{11}\}$ belong to class C_3 (corresponding to South Apulia).

Question 10. Now, we consider the binarized version of the Olive Oil dataset in Table 3. According to this dataset, what is the probability that a sample comes from the region Calabria given that we in that sample observe that the palmitic content is below the median and that the arachidic content is above the median?

A. $p(C_2|f_1 = 0, f_6 = 1) = \frac{5}{11}$

B. $p(C_2|f_1 = 0, f_6 = 1) = \frac{4}{7}$

C. $p(C_2|f_1 = 0, f_6 = 1) = \frac{5}{7}$

D. $p(C_2|f_1 = 0, f_6 = 1) = 1$

Solution 10. Using Bayes rule we have that

$$\begin{aligned} &p(C_2|f_1 = 0, f_6 = 1) \\ &\quad = \frac{p(f_1 = 0, f_6 = 1|C_2)p(C_2)}{\sum_{j=1}^{3} p(f_1 = 0, f_6 = 1|C_j)p(C_j)} \\ &= \frac{\frac{5}{5} \cdot \frac{5}{11}}{\frac{1}{2} \cdot \frac{2}{11} + \frac{5}{5} \cdot \frac{5}{11} + \frac{1}{4} \cdot \frac{4}{11}} = \frac{\frac{5}{11}}{\frac{1}{11} + \frac{5}{11} + \frac{1}{11}} = \frac{5}{7} \end{aligned}$$

Question 11. Consider the observations in Table 3. We consider these as 8-dimensional binary vectors and wish to compute the pairwise similarity. Which one of the following statements is true?

A. $\text{SMC}(o_2, o_4) \approx 0.626$

B. $\textbf{Cos}(o_1, o_2) \approx 0.408$

C. $\text{SMC}(o_3, o_4) \approx 0.263$

D. $\text{J}(o_2, o_4) \approx 0.843$

E. Don't know.

Solution 11. The problem is solved by simply using the definition of SMC, Jaccard similarity and cosine similarity as found in the lecture notes. The true values are:

$$\begin{aligned} \text{Cos}(o_1, o_2) &\approx 0.408 \\ \text{J}(o_2, o_4) &\approx 0.5 \\ \text{SMC}(o_3, o_4) &\approx 0.75 \\ \text{SMC}(o_2, o_4) &\approx 0.75 \end{aligned}$$

and therefore option B is correct.

Question 12. Consider again the binary data presented in Table 3 with three classes. We will use Hunt's algorithm to construct a classification tree using the Gini impurity measure. Suppose that the data in Table 3 is at the root node, and a binary split is made based on two different values of f_2. What is the impurity gain of this split?

A. $\Delta = \frac{136}{1815}$

B. $\Delta = \frac{436}{1815}$

C. $\Delta = \frac{3}{11}$

D. $\Delta = \frac{1379}{1815}$

E. Don't know.

Solution 12. At the root node, we have 11 observations in total, and the class probabilities are

$$p(C_1|r) = 2/11,\ p(C_2|r) = 5/11,\ p(C_3|r) = 4/11.$$

The proposed split will yield two nodes with the following class probabilities

$$p(C_1|v_1) = \frac{2}{6},\ p(C_2|v_1) = \frac{4}{6},\ p(C_3|v_1) = \frac{0}{6}$$
$$p(C_1|v_2) = \frac{0}{5},\ p(C_2|v_2) = \frac{1}{/}5,\ p(C_3|v_2) = \frac{4}{5}$$

Using the Gini impurity function, we find that

$$I(r) = 1 - \frac{2}{11}^2 - \frac{5}{11}^2 - \frac{4}{11}^2 = \frac{76}{121}$$
$$I(v_1) = 1 - \frac{2}{6}^2 - \frac{4}{6}^2 - \frac{0}{6}^2 = \frac{4}{9}$$
$$I(v_2) = 1 - \frac{0}{5}^2 - \frac{1}{5}^2 - \frac{4}{5}^2 = \frac{8}{25}$$

Using the formula for impurity gain then yields

$$\Delta = I(r) - \frac{6}{11}I(v_1) - \frac{5}{11}I(v_2) = \frac{436}{1815}$$

Question 13. We consider the binary matrix from Table 3 as a market basket problem consisting of $N = 11$ transactions $o_1, \ldots, o_{11}$ and $M = 8$ items $f_1, \ldots, f_8$. What is the *confidence* of the rule $\{f_6, f_8\} \to \{f_3, f_5\}$?

A. The confidence is $\frac{1}{11}$

B. The confidence is $\frac{1}{7}$

C. The confidence is $\frac{4}{11}$

D. The confidence is 1

E. Don't know.

Solution 13. The confidence of the rule is computed as

$$\frac{\text{support}(\{f_6, f_8\} \cup \{f_3, f_5\})}{\text{support}(\{f_6, f_8\})} = \frac{\frac{1}{11}}{\frac{7}{11}} = \frac{1}{7}.$$

Therefore, answer B is correct.

Question 14. Again, we consider the binarized version of the Olive Oil dataset in Table 3 as a market basket problem consisting. We want to apply the Apriori algorithm (the specific variant described in Chapter 21 of the lecture notes) to find all itemsets with support greater than $\varepsilon = 0.3$.

What is the content of L_3 when the Apriori algorithm is completed?

A. $L_3 = \{\}$

B. $L_3 = \{\{f_3, f_6, f_8\}\}$

C. $L_3 = \{\{f_2, f_6, f_8\}, \{f_3, f_6, f_8\}\}$

D. $L_3 = \{\{f_2\}, \{f_3\}, \{f_6\}, \{f_8\}\}$

E. Don't know.

Solution 14. L_3 will contain all the itemsets with three items that has support grater than $\varepsilon = 0.3$. Since there are $N = 11$ transactions, an itemset needs to be contain in at least $\lceil \varepsilon N \rceil = \lceil 0.3 \cdot 11 \rceil = 4$ transactions to have support greater than ε.

By looking in Table 3, we see that $\{f_3, f_6, f_8\}$ is contained in four transactions (o_1, o_3, o_6 and o_7) and it is the only itemset of size three that is contained in at least four transactions.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
o_1	38.0	15.1	27.4	77.9	18.1	33.3	48.5	50.0
o_2	26.8	12.8	52.0	77.0	22.5	68.1	66.0	75.0
o_3	64.5	39.6	74.4	37.1	45.7	66.7	66.0	64.3
o_4	63.2	45.7	29.1	41.4	49.1	56.9	59.2	50.0
o_5	66.3	34.3	37.7	43.1	40.9	63.9	70.9	60.7
o_6	56.7	34.7	72.2	47.3	38.4	61.1	62.1	55.4
o_7	63.4	30.6	66.4	49.8	30.2	62.5	50.5	42.9
o_8	87.1	85.3	19.3	19.2	68.6	34.7	64.1	33.9
o_9	51.3	46.8	14.8	53.4	49.3	37.5	52.4	35.7
o_{10}	67.5	62.3	13.0	33.2	66.7	51.4	41.7	39.3
o_{11}	86.0	71.3	25.1	20.5	71.9	25.0	48.5	32.1

Table 4: A small subset of 11 observations for the Olive Oil dataset. The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belong to class C_1 (corresponding to North Apulia), the red observations $\{o_3,\ o_4,\ o_5,\ o_6,\ o_7\}$ belong to class C_2 (corresponding to Calabria), and the blue observations $\{o_8,\ o_9,\ o_{10},\ o_{11}\}$ belong to class C_3 (corresponding to South Apulia).

Question 15. Consider the small subset of the Olive Oil dataset shown in Table 4. Suppose we train a naïve-Bayes classifier on this subset to predict the class label y from only the attributes x_1 and x_2. In this naïve-Bayes classifier, we assume that the conditional density of each attributed is a 1D Gaussian,

$$p(x_i|C_j) = \mathcal{N}(x_i|\mu_{j,i}, \sigma^2),$$

where $\mu_{j,i}$ is the mean of the i't feature for class j. We will assume that $\sigma^2 = 400$ for all attributes and all classes. For a test Olive Oil sample, we observe that

$$x_1 = 32.0,\ x_2 = 14.0$$

Furthermore, you can assume that the value of denominator in the calculation of the class-probabilities using the naïve-bayes classifier is

$$p_{\text{NB}}(x_1 = 15.0, x_2 = 14.0) = 0.00010141$$

What is then the probability that the oil comes from the region North Apulia (C_1) according to the naïve-Bayes classifier?

A. $p_{\text{NB}}(C_1|x_1 = 32.0, x_2 = 14.0) \approx 59\%$

B. $p_{\text{NB}}(C_1|x_1 = 32.0, x_2 = 14.0) \approx 71\%$

C. $p_{\text{NB}}(C_1|x_1 = 32.0, x_2 = 14.0) \approx 84\%$

D. $p_{\text{NB}}(C_1|x_1 = 32.0, x_2 = 14.0) \approx 96\%$

E. Don't know.

Solution 15. First we calculate the mean of the two attributes for the class C_1 from Table 4:

$$\mu_{1,1} = \frac{38.0 + 26.8}{2} = 32.4$$
$$\mu_{1,2} = \frac{15.1 + 12.8}{2} = 13.95$$

From Table 4 we can also calculate the class probability

$$p(C_1) = \frac{2}{11}.$$

We then use the naïve-Bayes assumption, which is

$$p_{\text{NB}}(C_1|x_1 = 32.0, x_2 = 14.0) = \frac{p(x_1 = 32.0|C_1)p(x_2 = 14.0|C_1)p(C_1)}{\sum_{j=1}^{3} p(x_1 = 32.0|C_j)p(x_2 = 14.0|C_j)p(y = j)} = \frac{p(x_1 = 32.0|C_1)p(x_2 = 14.0|C_1)p(C_1)}{p_{\text{NB}}(x_1 = 32.0, x_2 = 14.0)}$$

The numerator evaluates to

$$\begin{aligned} &p(x_1 = 32.0|C_1)p(x_2 = 14.0|C_1)p(C_1) \\ &= \mathcal{N}(x_1 = 32.0|\mu_{1,1} = 32.4, \sigma^2 = 400) \\ &\quad \mathcal{N}(x_2 = 14.0|\mu_{1,2} = 13.95, \sigma^2 = 400) \cdot \frac{2}{11} \\ &= 0.019943 \cdot 0.019947 \cdot \frac{2}{11} = 0.000072328 \end{aligned}$$

And therefore

$$p_{\text{NB}}(C_1|x_1 = 32.0, x_2 = 14.0) = \frac{0.000072328}{0.00010141} = 71\%.$$

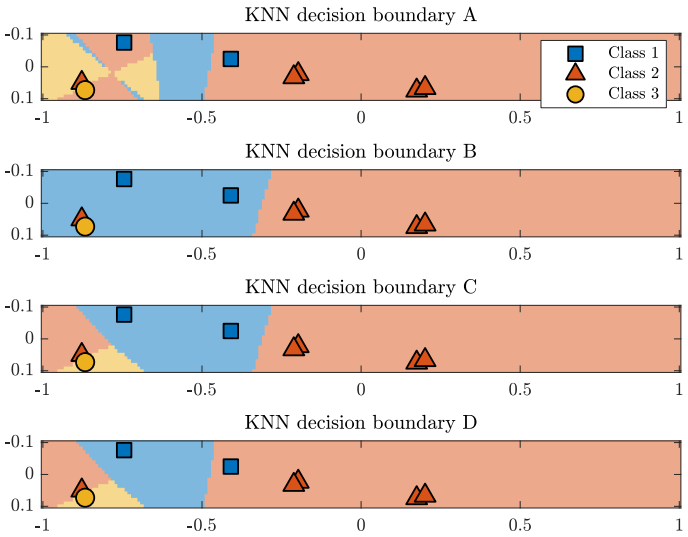

Figure 6: Decision boundaries for four KNN classifiers.

Question 16. Consider a two-dimensional data set comprised of $N = 8$ observations shown in Figure 6. The dataset consists of three classes indicated by the blue squares (class 1), red triangles (class 2) and yellow circles (class 3). In the figure, the decision boundaries for four K-nearest neighbor classifiers (KNN) are shown. Which one of the plots correspond to the $K = 3$ nearest-neighbour classifier assuming ties are broken by assigning to the *nearest* neighbour's class?

A. KNN decision boundary A

B. KNN decision boundary B

C. KNN decision boundary C

D. KNN decision boundary D

E. Don't know.

Solution 16. The point $(-1, 0)$ must be assigned to class 2, because there is a tie between all three classes and the nearest neighbour belongs to class 2. This rules out options A and B. Points close to the rightmost blue square must be assigned to class 2, since the two nearest neighbours belong to class 2. This rules out option C. Therefore D is the correct answer.

Question 17. An artificial neural network (ANN) trained on the Olive Oil dataset described in Table 1 will be used to predict the region of origin in Italy y as a multi-class classification problem based on all of the attributes $x_1, \ldots, x_8$. The neural network has a single hidden layer containing $n_h = 50$ units that uses a sigmoid non-linear activation function. The output layer uses a softmax activation function as described in the lecture notes, Section 15.3.2. How many parameters has to be trained to fit the neural network?

A. Network contains 501 parameters

B. Network contains 858 parameters

C. The network has 909 parameters

D. The network has 959 parameters

E. Don't know.

Solution 17. Each hidden unit has as many input weights as there are features in the dataset (i.e. $M = 8$) plus one (the bias), therefore they contribute with

$$(M + 1)n_h$$

weights. The multi-class output consists of $C = 9$ neurons (one for each class) which each also has a bias term and therefore contribute with:

$$(n_h + 1)C$$

weights. Adding these two numbers together gives the correct answer.

Question 18. Consider a two layer neural network $f : \mathbb{R}^2 \to \mathbb{R}$ for regression with one hidden unit and that can be written on the form

$$z^{(1)} = h^{(1)}(\tilde{\boldsymbol{x}}^\intercal \boldsymbol{w}^{(1)}),$$
$$f_{(\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)})}(\boldsymbol{x}) = \tilde{\boldsymbol{z}}^{(1)\intercal} \boldsymbol{w}^{(2)},$$

where $\tilde{\boldsymbol{x}} = \begin{bmatrix}1 & x_1 & x_2\end{bmatrix}^\intercal$, $z^{(1)} \in \mathbb{R}$, $\tilde{\boldsymbol{z}}^{(1)} = \begin{bmatrix}1 & z^{(1)}\end{bmatrix}$, and $h^{(1)}(x) = \max(0, x)$ is the activation function for the hidden layer (rectified linear unit). Assume that the weights of the first layer is fixed and given by

$$\boldsymbol{w}^{(1)\intercal} = \begin{bmatrix}-2 & 4 & 2\end{bmatrix}$$

Given N observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_N$ and corresponding targets $y_1, y_2, \dots y_N$, our learning objective is to find the value of the weight for the second layer $\boldsymbol{w}^{(2)}$ that minimizes the mean squared error,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}^{(2)}} \frac{1}{N} \sum_{i=1}^{N} \left\| \boldsymbol{f}_{(\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)})}(\boldsymbol{x}_i) - y_i \right\|^2, \quad (2)$$

where $\boldsymbol{w}^* = \begin{bmatrix}w_1^* & w_2^*\end{bmatrix}^\intercal \in \mathbb{R}^2$.

Consider the following dataset with $N = 4$ observations in $\boldsymbol{X}$ and the corresponding 4 targets in $\boldsymbol{y}$:

$$\boldsymbol{X} = \begin{bmatrix}0 & 0\\ 1 & 0\\ 1 & 1\\ 1 & 2\end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix}1\\ 3\\ 5\\ 7\end{bmatrix}$$

Which one of the following values of $\boldsymbol{w}^*$ minimizes mean squared error?

A. $\boldsymbol{w}^* = \begin{bmatrix}1 & 1\end{bmatrix}^\intercal$

B. $\boldsymbol{w}^* = \begin{bmatrix}1 & 2\end{bmatrix}^\intercal$

C. $\boldsymbol{w}^* = \begin{bmatrix}1 & 3\end{bmatrix}^\intercal$

D. $\boldsymbol{w}^* = \begin{bmatrix}1 & 4\end{bmatrix}^\intercal$

E. Don't know.

Solution 18. Since we know the weights of the first layer, we calculate the output of the first layer $z_i^{(1)}$ for each observations $\boldsymbol{x}_i$. We find that

$$\boldsymbol{Z}^{(1)} = \begin{bmatrix}z_1^{(1)}\\ z_2^{(1)}\\ z_3^{(1)}\\ z_4^{(1)}\end{bmatrix} = h^{(1)}(\tilde{\boldsymbol{X}}\boldsymbol{w}^{(1)})$$
$$= h^{(1)}\left(\begin{bmatrix}1 & 0 & 0\\ 1 & 1 & 0\\ 1 & 1 & 1\\ 1 & 1 & 2\end{bmatrix} \cdot \begin{bmatrix}-2\\ 4\\ 2\end{bmatrix}\right)$$
$$= h^{(1)}\left(\begin{bmatrix}-2\\ 2\\ 4\\ 6\end{bmatrix}\right) = \begin{bmatrix}0\\ 2\\ 4\\ 6\end{bmatrix}.$$

Now, to find $\boldsymbol{w}^*$ we can use regular linear regression with $\tilde{\boldsymbol{Z}}^{(1)}$ as the observations and $\tilde{\boldsymbol{y}}$ as the targets.

We observe that there is a linear relationship between $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{y}$, such that $y_i = Z_i^{(1)} + 1$. Expressed in vector notation that is

$$\tilde{\boldsymbol{Z}}^{(1)}\begin{bmatrix}1\\ 1\end{bmatrix} = \begin{bmatrix}1 & 0\\ 1 & 2\\ 1 & 4\\ 1 & 6\end{bmatrix}\begin{bmatrix}1\\ 1\end{bmatrix} = \begin{bmatrix}1\\ 3\\ 5\\ 7\end{bmatrix} = \boldsymbol{y}.$$

This means that $\boldsymbol{w}^* = \begin{bmatrix}1 & 1\end{bmatrix}^\intercal$ will give mean squared error of 0, and therefore A is the correct solution.

Question 19. Consider again the Olive Oil dataset of Table 1. Suppose we wish to predict the class label y using a decision tree model, and to improve performance we wish to apply AdaBoost. We apply AdaBoost to the full Olive Oil dataset. Recall the first steps of AdaBoost consists of: (i) Initialize weights, (ii) select a subset for training using sampling with replacement, and (iii) fit a model to the training set. Suppose the first fitted model has an accuracy of $\frac{3}{4}$ on the full dataset, what is the value of the weight of a correctly classified observation i after the first round of boosting?

A. $w_i(2) = \frac{2}{3} \cdot \frac{1}{572}$

B. $w_i(2) = \frac{3}{4} \cdot \frac{1}{572}$

C. $w_i(2) = \frac{4}{5} \cdot \frac{1}{572}$

D. $w_i(2) = \frac{5}{6} \cdot \frac{1}{572}$

E. Don't know.

Solution 19. For we note that the weight in AdaBoost are initialized as $w_i(1) = \frac{1}{N}$ for all $i = 1, \ldots, N$. Since the classifier has an accuracy of $\frac{3}{4}$, we see that $\epsilon_1 = \frac{\frac{1}{4}N}{N} = \frac{1}{4}$, since $N = 572$ divisible by 4, and therefor $\alpha_1 = \frac{1}{2}\log\frac{1-\epsilon_0}{\epsilon_0} = \frac{1}{2}\log 3$.

Using the weight update rule of AdaBoost, we find that the weight for a correctly classified observation is

$$w_i(2) = \frac{w_i(1)e^{-\alpha_1}}{\frac{3}{4}Nw_i(1)e^{-\alpha_1} + \frac{1}{4}Nw_i(1)e^{\alpha_1}} = \frac{N^{-1}e^{-\alpha_1}}{\frac{3}{4}e^{-\alpha_1} + \frac{1}{4}e^{\alpha_1}} = \frac{N^{-1}\frac{\sqrt{3}}{3}}{\frac{3}{4}\frac{\sqrt{3}}{3} + \frac{1}{4}\sqrt{3}} = \frac{2}{3} \cdot \frac{1}{572}$$

Question 20. Consider a small dataset comprised of $N = 4$ observations

$$x = \begin{bmatrix} 0.4 & 1.7 & 3.7 & 4.6 \end{bmatrix}^\mathsf{T}.$$

We wish to apply the k-means algorithm to the dataset using $K = 3$ and the farthest-first initialization method described in Section 18.2.2. Suppose the first selected centroid is $\mu_1 = 1.7$, what are the locations of the next two centroids?

A. $\mu_2 = 4.6$, $\mu_3 = 0.4$

B. $\mu_2 = 4.6$, $\mu_3 = 3.7$

C. $\mu_2 = 3.7$, $\mu_3 = 0.4$

D. $\mu_2 = 3.7$, $\mu_3 = 4.6$

E. Don't know.

Solution 20. According to the farthest-first initialization method, the second cluster is initialized at the location of the observation which is the most distant from μ_1, i.e. $\mu_2 = 4.6$. This rules out all options except A.

	$i=1$	$i=2$	$i=3$	$i=4$
x_i	1	2	3	4
y_i	6	2	3	4

Table 5: Simple 1D regression dataset

Question 21. Consider the small 1D dataset shown in Table 5 comprised of $N = 4$ observations and where the goal is to predict y_i given x_i. We apply linear regression to this datasets, where we transform the features using the transformation $\phi(x) = \begin{bmatrix}\cos(\frac{\pi}{2}x) & \sin(\frac{\pi}{2}x)\end{bmatrix}^\intercal$. Find the weights $\boldsymbol{w}^* = \begin{bmatrix}w_1^* & w_2^*\end{bmatrix}^\intercal$ that minimize the mean squared error. What is the value of w_2^*?

A. $w_2^* = \frac{1}{2}$

B. $w_2^* = 1$

C. $w_2^* = \frac{3}{2}$

D. $w_2^* = 2$

E. Don't know.

Solution 21. The solution to the least squares problem is given by

$$\boldsymbol{w}^* = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y},$$

where the transformed dataset is given by

$$\tilde{\boldsymbol{X}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \\ -0 & -1 \\ 1 & -0 \end{bmatrix},$$

and $\boldsymbol{y} = \begin{bmatrix}6 & 2 & 3 & 4\end{bmatrix}^T$. We find that

$$\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1} = \frac{1}{2}I_2 \text{ and } \tilde{\boldsymbol{X}}^T\boldsymbol{y} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

Therefore the optimal weights are

$$\boldsymbol{w}^* = \frac{1}{2}I_2\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{3}{2} \end{bmatrix}$$

and we see that $w_2^* = \frac{3}{2}$.

Question 22. Consider again the small 1D dataset shown in Table 5 comprised of $N = 4$ observations and where the goal is to predict y_i given x_i. Suppose we apply ridge regression to the problem in the form described in the lecture notes, Section 14.1, and find that the optimal weight and constant term are

$$\boldsymbol{w} = \begin{bmatrix} -\sqrt{\frac{3}{20}} \end{bmatrix} \qquad w_0 = \frac{15}{4}.$$

If the ridge regression cost function is $E_\lambda(\boldsymbol{w}, w_0) = 8$, what is the value of the regularization constant?

A. $\lambda = 1$

B. $\lambda = 2$

C. $\lambda = 4$

D. $\lambda = 8$

E. Don't know.

Solution 22. To calculate the cost function, we first standardized the feature matrix

$$\hat{\boldsymbol{X}} = \sqrt{\frac{3}{5}}\begin{bmatrix} -\frac{3}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{3}{2} \end{bmatrix}$$

The cost function is

$$E_\lambda(\boldsymbol{w}, w_0) = \left\|\boldsymbol{y} - w_0\boldsymbol{1} - \hat{\boldsymbol{X}}\boldsymbol{w}\right\|^2 + \lambda\|\boldsymbol{w}\|^2.$$

Solving for λ and setting in the values, we find that

$$\lambda = \frac{E_\lambda(\boldsymbol{w}, w_0) - \left\|\boldsymbol{y} - w_0\boldsymbol{1} - \hat{\boldsymbol{X}}\boldsymbol{w}\right\|^2}{\|\boldsymbol{w}\|^2} = 2.$$

Question 23. Consider again the Olive Oil dataset in Table 1. Using a neural network, Alice and Bob apply sequential feature selection to find a subset of the $M = 8$ attributes to predict the region y. They both choose the subsets based on the test error as determined by 5-fold cross-validation for *any subset of the attributes*. Alice does forward selection and Bob does backward selection.

Suppose that both Alice and Bob end up selecting the attributes x_1, x_2, x_4, x_5, x_7, and x_8. Let N_{forward} denote the minimal number of models that Alice trained during forward selection, and let N_{backward} denote the minimal number of models that Bob trained during backward selection. How many more models did Alice train in forward selection than Bob trained in backward selection?

A. $N_{\text{forward}} - N_{\text{backward}} = 14$

B. $N_{\text{forward}} - N_{\text{backward}} = 18$

C. $N_{\text{forward}} - N_{\text{backward}} = 70$

D. $N_{\text{forward}} - N_{\text{backward}} = 90$

E. Don't know.

Solution 23. First we see that both methods have 6 out of 8 attribute, and for any selection of attributes, we have to train $K = 5$ models.

In forward selection, we first train K model with no attributes. Then we train KM models with a single attribute, and select one attribute to proceed to the next level. Then we train $K(M-1)$ on two attributes, and choose one pair of attributes to proceed to the next level. We continue to do this, until we train $K(M-6)$ models on seven attributes, where we observe that for all seven selections of attributes, the error is higher than with six attributes, and the method terminates. So in total, we train $N_{\text{forward}} = K(1 + M + (M-1) + (M-2) + (M-3) + (M-4) + (M-5) + (M-6))$ models in forward selection.

In backward selection, we first train K model with all attributes. Then we train KM models with a seven attribute, and choose one selection of seven attributes to proceed to the next level. Then we train $K(M-1)$ on six attributes, and a choose one selection of six attributes to proceed to the next level. Finally, we train $K(M-2)$ models on five attributes, where we observe that for all selections of five attributes the error is higher than with six attributes and the method terminates. So in total, we train $N_{\text{backward}} = K(1 + M + (M-1) + (M-2))$ models in forward selection.

Therefore, the additional number of more models we train in forward selection than in backward section is

$$\begin{aligned} &N_{\text{forward}} - N_{\text{backward}} \\ &= K((M-3) + (M-4) + (M-5) + (M-6)) \\ &= 5 \cdot ((8-3) + (8-4) + (8-5) + (8-6)) = 70. \end{aligned}$$

Question 24. We want to estimate a confidence interval on the generalization error for a regression tree model using the procedure described in the lecture notes, Section 11.3.5. Using a small dataset, we perform $K = 3$ fold cross validation and evaluate the per-observation L_1 losses to be

$$z_1 = 1,\ z_2 = 3,\ z_3 = 3,\ z_4 = 1,\ z_5 = 2,\ z_6 = 3,\ z_7 = 1,$$

where z_i is the loss for the i'th observation. Assuming that the losses are normally distributed, the $1-\alpha$ confidence interval for the generalization error is obtained using the inverse cumulative distribution function of the student's t-distribution, $\text{cdf}_{\mathcal{T}}^{-1}(\cdot \mid \nu, \mu, \sigma)$. For the losses listed above, which one of the following combination of values should be use for ν, μ and σ?

A. $\nu = 6$, $\mu = 2$, $\sigma = \frac{1}{\sqrt{7}}$

B. $\nu = 6$, $\mu = 2$, $\sigma = 1$

C. $\nu = 7$, $\mu = 2$, $\sigma = \frac{1}{\sqrt{7}}$

D. $\nu = 7$, $\mu = 2$, $\sigma = 1$

E. Don't know.

Solution 24. Following Section 11.3.5, we have that $\nu = n - 1 = 6$ where n is the number of observations. μ is given by the empirical mean of $z_1, \ldots, z_7$, which is given by

$$\mu = \frac{1+3+3+1+2+3+1}{7} = 2$$

Finally, σ is the empirical standard deviation of the mean for $z_1, \ldots, z_7$. We find the empirical variance of the mean as

$$\begin{aligned} \sigma^2 = \frac{1}{7(7-1)}\Big((1-2)^2 + (3-2)^2 + (3-2)^2 + \\ (1-2)^2 + (2-2)^2 + (3-2)^2 + (1-2)^2\Big) = \frac{1}{7}. \end{aligned}$$

So we find that $\sigma = \frac{1}{\sqrt{7}}$

Question 25. We consider a regularized regression model for a dataset comprised of $N = 1000$ observations, and wishes to both select the optimal regularization strength and estimate the generalization error of the model. We consider three different values of the regularization strength.

We use a strategy where the hold-out method is used to estimate the generalization error and K-fold cross-validation is used to select the optimal regularization strength, i.e. the dataset is first divided into a test set $\mathcal{D}^{\text{test}}$, comprised of 20% of the full dataset, and the remainder $\mathcal{D}^{\text{train}}$ is used for cross-validation.

Suppose for any fixed value of the regularization strength, the time taken to train the regression model on a dataset of size n is $n \log_2 n$ units of time (note that $\log_2$ is the logarithm with base 2), and the time taken to test a trained model using a test dataset of size m is m units of time. Suppose the duration of all other tasks is negligible. You have a computational budget of 200 000 units of time.

What is the maximum number of folds K you can carry out in the cross-validation loop within your computational budget?

A. $K = 7$

B. $K = 8$

C. $K = 9$

D. $K = 10$

E. Don't know.

Solution 25. We have $N = 1000$ total points. For the hold-out outer loop we have $n_{\text{o}} = 800$ observations for training and $m_{\text{o}} = 200$ for testing. This means that the time used for in the out loop is

$$t_{\text{o}} = n_{\text{o}} \log_2 n_{\text{o}} + m_{\text{o}} = 800 \log_2 800 + 200$$

For the inner cross-validation loop, we have a total of 800 observations. So in each fold, we have $m_{\text{i}} = \frac{800}{K}$ observations for testing and $n_{\text{i}} = 800 \cdot \frac{K-1}{K}$ for training. In the cross-validation algorithm we train and test $K \cdot L$ times, where $L = 3$ is the different values of the regularization strength. So the time for the inner cross-validation loop is

$$\begin{aligned} t_{\text{i}} &= L \cdot K(n_{\text{i}} \log_2 n_{\text{i}} + m_{\text{i}}) \\ &= 3 \cdot K \left(\frac{800 \cdot (K-1)}{K} \log_2 \frac{800 \cdot (K-1)}{K} + \frac{800}{K} \right) \\ &= 2400 \cdot (K-1) \log_2 \frac{800 \cdot (K-1)}{K} + 2400 \end{aligned}$$

The total time is then given by

$$t_{\text{total}} = t_{\text{o}} + t_{\text{i}}$$

At this point we can try the different value of K, and $K = 9$ gives $1.92 \cdot 10^5$ units of time, whereas $K = 10$ gives $2.15 \cdot 10^5$ units of time.

Question 26. Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In Figure 7 is given 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Each observation is colored and marked in terms of which cluster it came from in the Gaussian Mixture model.

Which one of the following GMM densities was used to generate the data?

A.

$$p(\boldsymbol{x}) = \frac{1}{2}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-0.5\\-4.6\end{bmatrix}, \begin{bmatrix}1.7 & -1.3\\-1.3 & 2.1\end{bmatrix}\right) + \frac{1}{10}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-5.8\\13.1\end{bmatrix}, \begin{bmatrix}2.7 & 1.0\\1.0 & 1.4\end{bmatrix}\right) + \frac{2}{5}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.5\\1.0\end{bmatrix}, \begin{bmatrix}2.1 & -1.6\\-1.6 & 2.4\end{bmatrix}\right)$$

B.

$$p(\boldsymbol{x}) = \frac{2}{5}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-0.5\\-4.6\end{bmatrix}, \begin{bmatrix}1.7 & -1.3\\-1.3 & 2.1\end{bmatrix}\right) + \frac{1}{10}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-5.8\\13.1\end{bmatrix}, \begin{bmatrix}2.1 & -1.6\\-1.6 & 2.4\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.5\\1.0\end{bmatrix}, \begin{bmatrix}2.7 & 1.0\\1.0 & 1.4\end{bmatrix}\right)$$

C.

$$p(\boldsymbol{x}) = \frac{2}{5}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-0.5\\-4.6\end{bmatrix}, \begin{bmatrix}2.7 & 1.0\\1.0 & 1.4\end{bmatrix}\right) + \frac{1}{10}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-5.8\\13.1\end{bmatrix}, \begin{bmatrix}2.1 & -1.6\\-1.6 & 2.4\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.5\\1.0\end{bmatrix}, \begin{bmatrix}1.7 & -1.3\\-1.3 & 2.1\end{bmatrix}\right)$$

D.

$$p(\boldsymbol{x}) = \frac{1}{10}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-0.5\\-4.6\end{bmatrix}, \begin{bmatrix}1.7 & -1.3\\-1.3 & 2.1\end{bmatrix}\right) + \frac{2}{5}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-5.8\\13.1\end{bmatrix}, \begin{bmatrix}2.1 & -1.6\\-1.6 & 2.4\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.5\\1.0\end{bmatrix}, \begin{bmatrix}2.7 & 1.0\\1.0 & 1.4\end{bmatrix}\right)$$

E. Don't know.

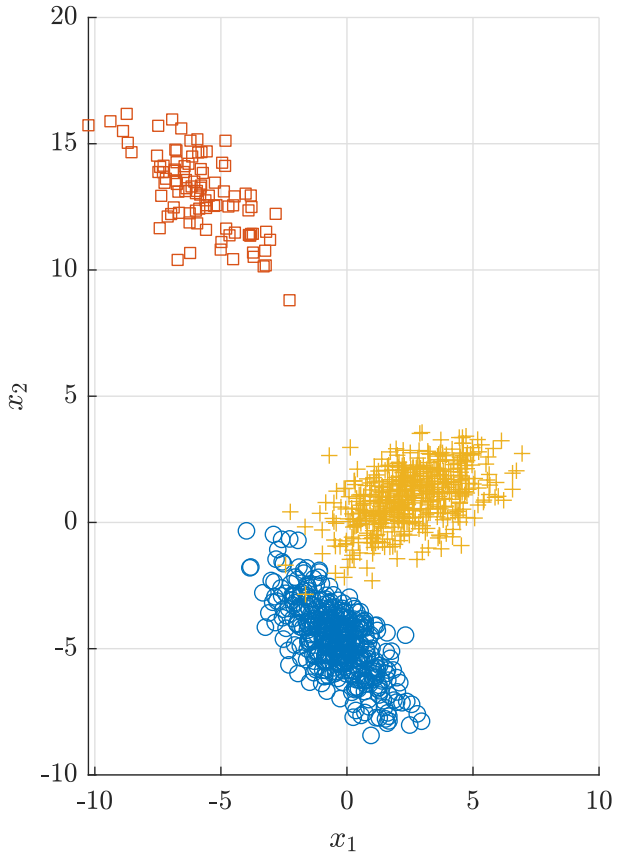

Figure 7: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

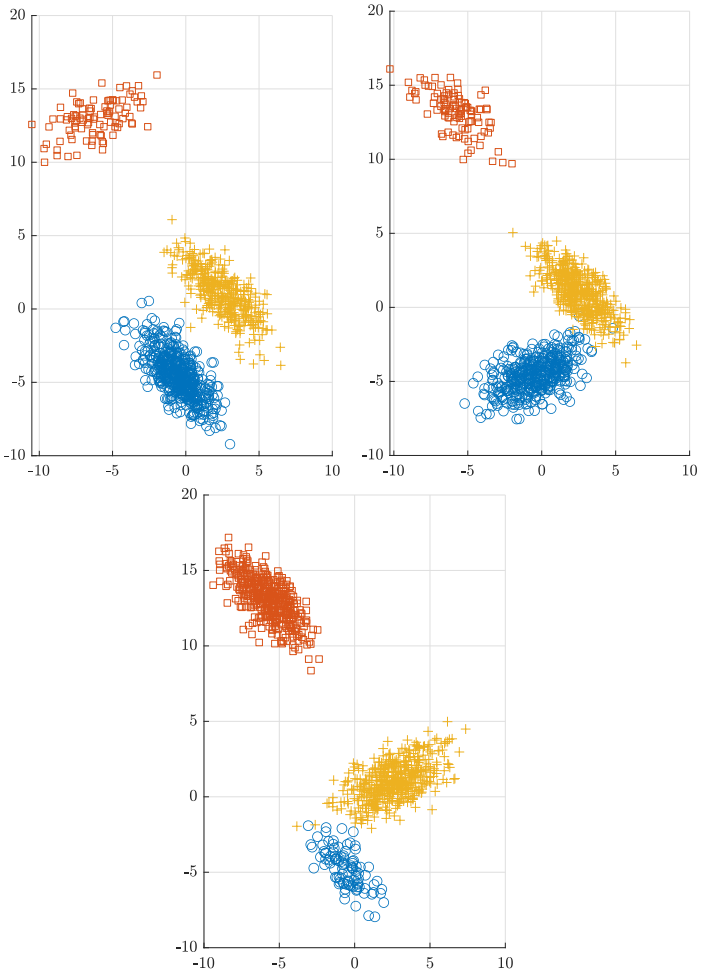

Figure 8: GMM mixtures corresponding to alternative options.

Solution 26. The three components in the candidate GMM densities can be matched to the colored observations by their mean values. Then, by considering the basic properties of the covariance matrices, we can easily rule out all options except B. Alternatively, in Figure 8 is shown the densities for densities corresponding to option A (upper left), C (upper right) and D (bottom center).

Question 27. A neural network classifier is trained to distinguish between two classes $y \in \{0, 1\}$ and produce class-probability $\hat{y}$ and the *reciever operator characteristic* (ROC) curve of the network when evaluated on a test set with $N = 9$ observations is shown in Figure 10. Suppose we plot the predictions on the $N = 9$ test observations by their $\hat{y}$ value along the x-axis and indicate the class labels by either a black circle (class $y = 0$) or red cross ($y = 1$), which one of the subplots in Figure 10 then corresponds to the ROC curve in Figure 9?

A. Prediction A

B. Prediction B

C. Prediction C

D. Prediction D

E. Don't know.

Solution 27. The correct answer is C. 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.

This procedure is then repeated for different threshold values to obtain the curves shown in Figure 11. The ROC curve is then obtained by plotting these two curves against each other. I.e. for each threshold value, the point

$$(x, y) = (\text{FPR}, \text{TPR})$$

is on the AUC curve. This rules out all options except C. For completeness, we have included the ROC curves for all options in Figure 12.

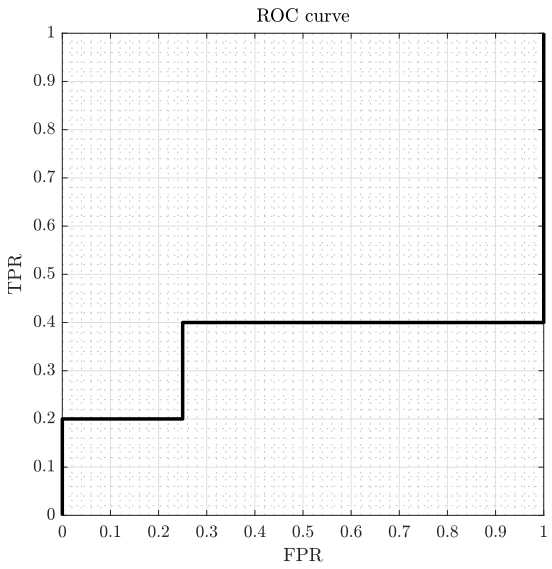

Figure 9: ROC curve for a neural network classifier, where the predictions and true class labels are one of the options in Figure 10.

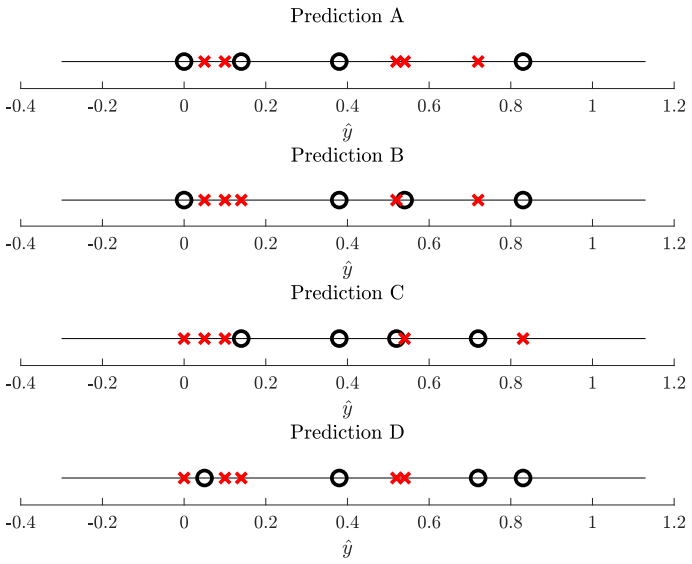

Figure 10: Four candidate predictions for the ROC curve in Figure 9. The observations are plotted horizontally, such that the position on the x-axis indicate the predicted value $\hat{y}_i$, and the marker/color indicate the class membership, such that the black circles indicate the observation belongs to class $y_i = 0$ and red crosses to $y_i = 1$.

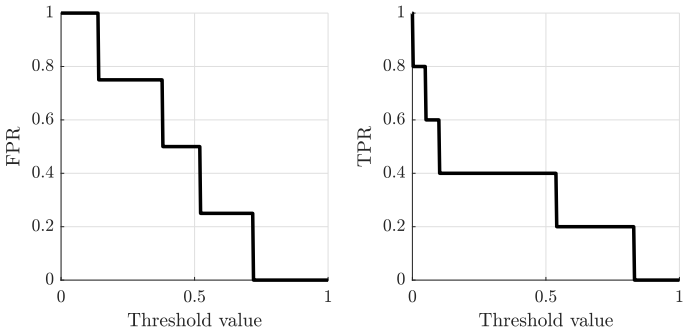

Figure 11: TPR, FPR curves for the classifier.

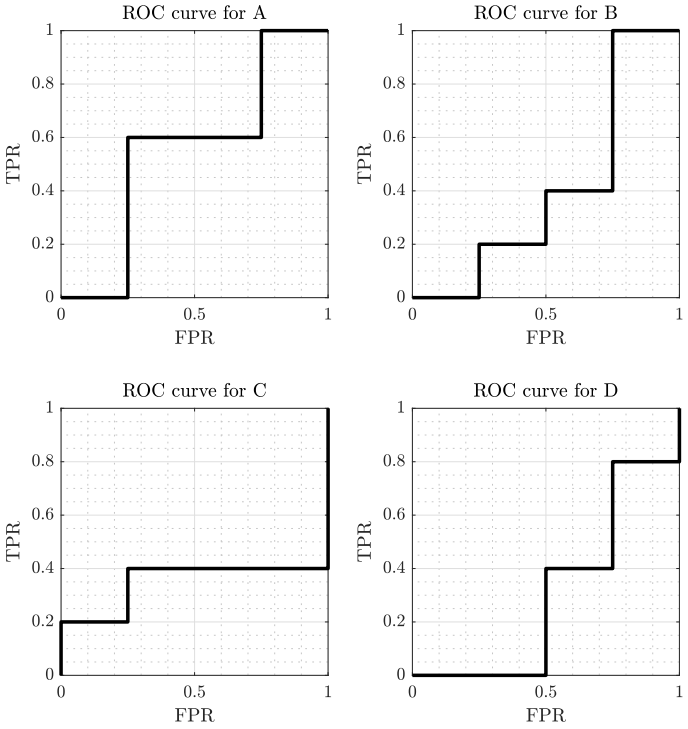

Figure 12: ROC curves for all options.

Technical University of Denmark

Written examination: 24 May 2018, 9 AM - 1 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.

Your answers to the questions are to be handed in using the electronic file. Use only this page for hand in if you are unable to hand in digitally. In case you have to hand in the answers using the form on this sheet, please print your name and student number clearly.

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.

The individual questions are answered by filling in the answer fields with one of the letters A, B, C, D, or E.

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

USE ONLY THIS PAGE FOR HAND IN IF YOU ARE UNABLE TO HAND IN DIGITALLY.

No.	Attribute description	Abbrev.
x_1	Number of seats times kilometers pr. week	S*KM/Week
x_2	Incidences 1985-1999	Inc. 85-99
x_3	Fatal accidents 1985-1999	FA 85-99
x_4	Fatalities 1985-1999	Fat. 85-99
x_5	Incidences 2000-2014	Inc. 00-14
x_6	Fatal accidents 2000-2014	FA 00-14
y	Fatalities 2000-2014	Fat. 00-14

Table 1: The attributes of the airline safety dataset that contains 56 observations of different airline companies and their properties in terms of number of seats times number of kilometers per week, incidences, fatal accidents, and fatalities accumulated over the period of 1985-1999 and 2000-2014 respectively. We presently consider as output y the number of fatalities from 2000-2014.

Question 1. We will consider the airline safety dataset consisting of 56 airline companies and their number of flights as quantified by number of seats times kilometers per week as well as incidences, fatal accidents, and fatalities quantified for the period of 1985-1999 and 2000-2014 respectively[1]. For brevity this dataset will be denoted the airline safety dataset. In Table 1 is given the attributes of the data as well as the output attribute y defined by the number of fatalities from 2000-2014. In Figure 1 is shown a matrix plot of the six attributes x_1–x_6.

Considering the attributes described in Table 1 and the matrix plot in Figure 1 which one of the following statements regarding the attributes x_1–x_6 is correct?

A. At least one of the attributes appears to be normal distributed.

B. x_2 corresponding to Inc. 85-99 and x_3 corresponding to FA 85-99 are negatively correlated.

C. All the six attributes are ratio.

D. All the attributes are continuous.

E. Don't know.

Solution 1. Inspecting the histograms along the diagonal of the matrix plot it is clearly seen that they all have a mode around zero/low values and decrease with no negative observations on the left side of the mode. This is not corresponding to the bell shape of a normal distribution and thus none of the attributes apear normally distributed. Inspecting, the plot of x_2 vs. x_3 we observe that airline with more incidences in 85-99 also have more fatal accidents in 85-99, there is thus a positive correlation between these two attributes. As for all attributes zero means absence of what is being measured and it makes sense to talk about a company having twice as many seats times kilometers per week, or incidents, or accidents, or fatalities, thus, all these attributes are ratio. As incidents, accidents and fatalities are counts they are discrete integer variables and not continuous.

[1]The dataset is taken from `https://github.com/fivethirtyeight/data/tree/master/airline-safety`

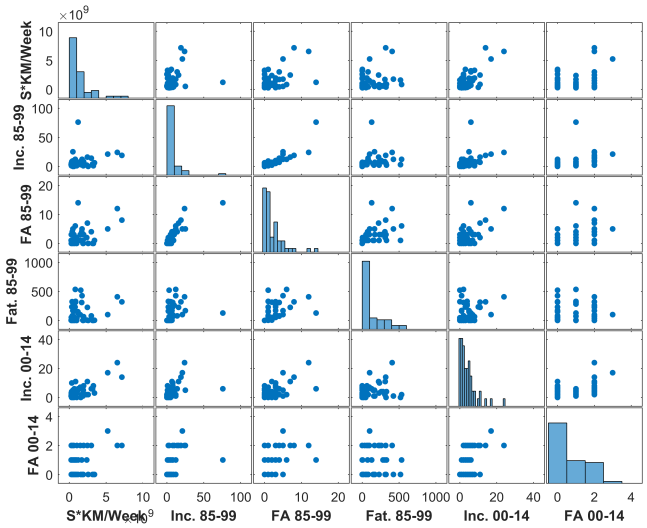

Figure 1: Matrix plot of the six attributes x_1–x_6. Along the diagonal of the matrix plot is given the histogram of each attribute.

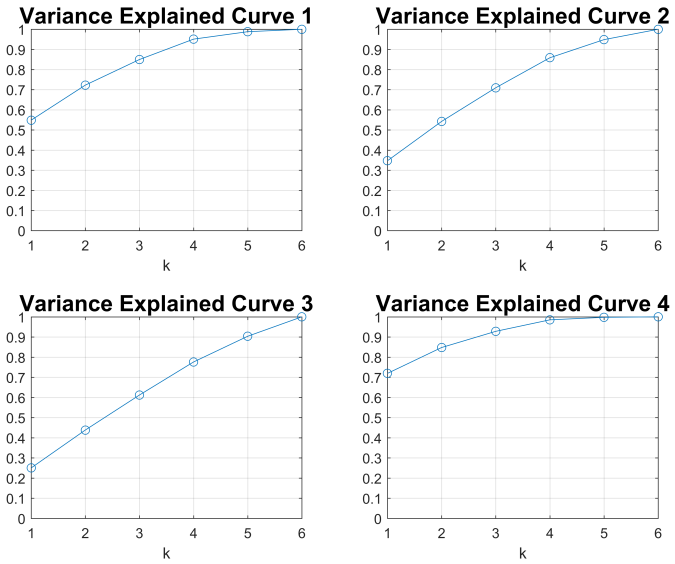

Figure 2: Four different curves of the variance explained as function of keeping the first k principal components when performing a PCA analysis. One of the four curves correctly corresponds to the PCA of the standardized airline safety data.

Question 2. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_6, forming the standardized matrix $\tilde{\boldsymbol{X}}$ (i.e., each attribute has been subtracted its mean and divided by its standard deviation). A singular value decomposition is applied to the standardized data matrix, i.e. $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ and we find the following solution in terms of the $\boldsymbol{S}$ and $\boldsymbol{V}$ matrices:

$$\boldsymbol{S} = \begin{bmatrix} 13.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 7.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.0 \end{bmatrix}.$$

$$\boldsymbol{V} = \begin{bmatrix} 0.38 & -0.51 & 0.23 & 0.47 & -0.55 & 0.11 \\ 0.41 & 0.41 & -0.53 & 0.24 & 0.00 & 0.58 \\ 0.50 & 0.34 & -0.13 & 0.15 & -0.05 & -0.77 \\ 0.29 & 0.48 & 0.78 & -0.17 & 0.00 & 0.23 \\ 0.45 & -0.42 & 0.09 & 0.03 & 0.78 & 0.04 \\ 0.39 & -0.23 & -0.20 & -0.82 & -0.30 & 0.04 \end{bmatrix}.$$

In Figure 2 is given the pct. of variance explained by retaining the first k principal components as a function of k. Which one of the four curves corresponds to the correct curve of variance explained as function of the number of principal components retained?

A. Variance Explained Curve 1.

B. Variance Explained Curve 2.

C. Variance Explained Curve 3.

D. Variance Explained Curve 4.

E. Don't know.

Solution 2. The variance explained by the first k principal components is given by $\frac{\sum_1^k \sigma_k^2}{\sum_{i'=1}^6 \sigma_{i'}^2}$. We thereby get that the curve should have the following values:

k=1: $\frac{13.5^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.5487$

k=2: $\frac{13.5^2+7.6^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.7226$

k=3: $\frac{13.5^2+7.6^2+6.5^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.8498$

k=4: $\frac{13.5^2+7.6^2+6.5^2+5.8^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.9511$

k=5: $\frac{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.9880$

k=6: $\frac{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+.02^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 1$

Only Variance Explained Curve 1 has this property.

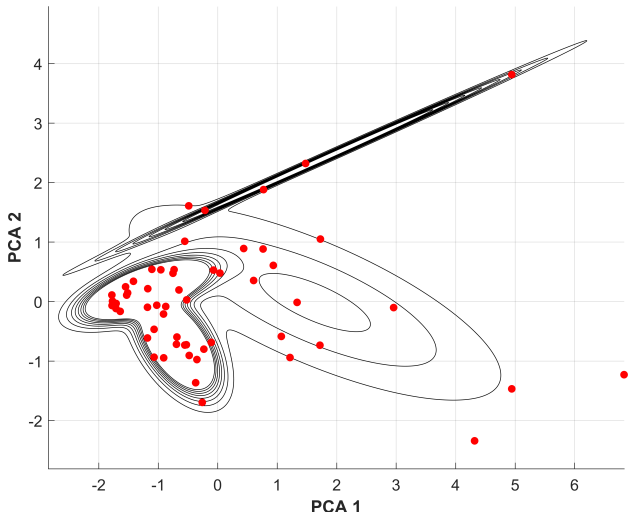

Figure 3: A Gaussian Mixture Model (GMM) using four clusters fitted to the standardized airline safety data projected onto the first two principal components.

Question 3. According to the extracted PCA directions given by the matrix $\boldsymbol{V}$ in the above what will be the coordinate of the standardized observation $\tilde{\boldsymbol{x}}^* = [-0.1 \;\; 0.2 \;\; 0.1 \;\; -0.3 \;\; 1 \;\; 0.5]$ when projected onto the first two principal components?

A. (-0.753, 0.206)

B. (0.652, -0.512)

C. (0.680, 0.019)

D. (0.671, -0.139)

E. Don't know.

Solution 3. The observation $\tilde{\boldsymbol{x}}^* = [-0.1 \;\; 0.2 \;\; 0.1 \;\; -0.3 \;\; 1 \;\; 0.5]$ will have the projection onto the two first principal components given by

$$[-0.1 \;\; 0.2 \;\; 0.1 \;\; -0.3 \;\; 1 \;\; 0.5] \begin{bmatrix} 0.38 & -0.51 \\ 0.41 & 0.41 \\ 0.50 & 0.34 \\ 0.29 & 0.48 \\ 0.45 & -0.42 \\ 0.39 & -0.23 \end{bmatrix} = [0.652 \;\; -0.512].$$

Thus, the observation will in the projection be located at (0.652, -0.512).

Question 4. Which one of the following statements regarding Gaussian Mixture Modeling (GMM) is correct?

A. The number of clusters used in the GMM can be determined by selecting the number of clusters that provides the best likelihood of the training data used for training the density.

B. For high-dimensional data, i.e. where the number of features M is large, it can be beneficial to constrain the covariance of each cluster to be diagonal, i.e. enforcing off-diagonal terms of the covariance matrices to be zero, in order to reduce the number of parameters in the GMM model.

C. The GMM is guaranteed to find the optimal clustering for a given dataset.

D. Similar to the k-means algorithm that assigns observations to the cluster in closest proximity, the EM-algorithm used to estimate the parameters of the GMM considers only the cluster each observation is the most likely to belong to when estimating the parameters in the M-step.

E. Don't know.

Solution 4. For the GMM we can use cross-validation to determine the number of clusters, however, this selection of the number of clusters must be based on the likelihood of the test data and not the training data, as this otherwise would result in overfitting of the density to the data. For high-dimensional data the covariance matrix can be ill-determined and it can therefore be beneficial to reduce the covariance matrix to a diagonal matrix for which only the variance of each features need to be estimated, which substantially reduces the number of free parameters in the GMM. The GMM is prone to local minima solutions and therefore it is recommended to fit the model using many restarts taking the best of these randomly initialized models. In the E-step of the GMM it is quantified how likely it is for the observations to belong to each cluster and thereby the observations are "soft" assigned to each cluster. Thus, in the subsequent M-step this soft-assignment is used taking into account how likely it is for the observations to belong to each of the clusters and contributing in the update of the cluster according to this, such that each clusters mean and covariance is a weighted average of the observations contribution weighted according to this probability.

Question 5. We fit a Gaussian Mixture Model (GMM) to the standardized data projected onto the first two principal component directions using four mixture components (i.e., 4 clusters). We 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.

$$p(\boldsymbol{x}) = 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix}) + 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix})$$

B.

$$p(\boldsymbol{x}) = 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix}) + 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix})$$

C.

$$p(\boldsymbol{x}) = 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix}) + 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix})$$

D.

$$p(\boldsymbol{x}) = 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix}) + 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix})$$

E. Don't know.

Solution 5. Inspecting the GMM density we observe that the cluster located at $\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}$ will have the lowest mixing proportion as only few observations belong to this cluster. Furthermore, it must have a large positive correlation and variance as given by $\begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}$. Only aswer option 2 and 4 have this property. The cluster located at $\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}$ has negative covariance but the variance of the cluster is also much larger than the variance of the other cluster having negative covariance. Thus, this cluster must have the covariance $\begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix}$. Only answer option 4 has this property.

Question 6. We would like to predict the safety of a given airline company. However, in order to take into account the volume of flights of the company when evaluating its safety we define a new output variable given by $\tilde{y} = y/x_1$. By defining $\tilde{y}$ as the number of fatalities divided by the number of seats times kilometers per week of the company, fatalities are quantified relative to the volume of flights that has been catered by the company. A least squares linear regression model is trained using different combinations of the five attributes x_2, x_3, x_4, x_5, and x_6 in order to predict $\tilde{y}$. Table 2 provides the training and test root-mean-square error (RMSE=$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{y}_i)^2}$) performance of the least squares linear regression model when trained using different combinations of the five attributes. Which one of the following statements is correct?

A. Forward and backward selection will result in the same features being selected.

B. Forward selection will terminate with four features in the feature set.

C. Backward selection will not remove any features.

D. x_4 will be among the features selected by forward selection.

E. Don't know.

Solution 6. Forward selection will result in x_6 being selected with performance 0.18749 and subsequently x_3 with performance 0.17624 and then x_5 with performance 0.17082 as no improvement can be achieved by adding additional features it will terminate at the set x_3, x_5, and x_6. Backward selection will result in removing x_4 to have x_2, x_3, x_5, x_6 with performance 0.17299 and then remove x_2 to attain the performance 0.17082 by se the feature set x_3, x_5, and x_6 upon which no further improvements can be achieved by removing features.

Feature(s)	Training RMSE	Test RMSE
none	0.11279	0.20677
x_2	0.10930	0.22301
x_3	0.10974	0.21773
x_4	0.10911	0.21362
x_5	0.11254	0.20729
x_6	0.09301	0.18749
x_2, x_3	0.10914	0.22247
x_2, x_4	0.10756	0.22145
x_2, x_5	0.10909	0.22513
x_2, x_6	0.09108	0.18555
x_3, x_4	0.10837	0.21768
x_3, x_5	0.10961	0.21800
x_3, x_6	0.09108	0.17624
x_4, x_5	0.10910	0.21368
x_4, x_6	0.09234	0.19121
x_5, x_6	0.08993	0.17657
x_2, x_3, x_4	0.10753	0.22138
x_2, x_3, x_5	0.10887	0.22435
x_2, x_3, x_6	0.09071	0.18029
x_2, x_4, x_5	0.10731	0.22315
x_2, x_4, x_6	0.08947	0.19339
x_2, x_5, x_6	0.08900	0.17610
x_3, x_4, x_5	0.10828	0.21795
x_3, x_4, x_6	0.08805	0.17900
x_3, x_5, x_6	0.08896	0.17082
x_4, x_5, x_6	0.08891	0.18062
x_2, x_3, x_4, x_5	0.10730	0.22314
x_2, x_3, x_4, x_6	0.08782	0.18371
x_2, x_3, x_5, x_6	0.08878	0.17299
x_2, x_4, x_5, x_6	0.08727	0.18336
x_3, x_4, x_5, x_6	0.08603	0.17440
x_2, x_3, x_4, x_5, x_6	0.08595	0.17685

Table 2: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict $\tilde{y}$ using different combinations of the five attributes (x_2–x_6).

Question 7. We would again like to predict $\tilde{y}$ based on x_2, x_3, x_4, x_5, and x_6. For this purpose, we will use the regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function: $E(\boldsymbol{w}) = \sum_n (\tilde{y}_n - [1\ x_{n2}\ x_{n3}\ x_{n4}\ x_{n5}\ x_{n6}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$, We will consider 20 different values of λ and use 10-fold cross-validation to select for the optimal value of λ. Which one of the following statement regarding the described regularized least squares regression procedure is correct?

A. Increasing λ will result in an increase in the 2-norm of the trained $\boldsymbol{w}$, i.e. in an increase of the quantity $\|\boldsymbol{w}\|_2$.

B. 10-fold cross-validation will require the fitting of 10 models in total to quantify the best value of λ.

C. The test error obtained for the optimal value of λ is a biased estimate of the generalization error.

D. When using regularization in least squares regression the model becomes more prone to overfitting.

E. Don't know.

Solution 7. When increasing λ the quantity $\boldsymbol{w}^\top \boldsymbol{w} = \|\boldsymbol{w}\|_2^2$ will be penalized more and therefore reduced (and not increased in terms of the quantity $\|\boldsymbol{w}\|_2$). To quantify the best value of λ using 10-fold cross-validation we need to carry out the cross-validation for each of the 20 considered values of λ resulting in $10 \times 20 = 200$ models to be fitted. The test error obtained for the optimal value of λ is biased as it is the best among 20 selected test-performances. To get an unbiased estimate thus requires two-level cross-validation. Regularization reduce overfitting as the model can less well fit the training data due to the regularization.

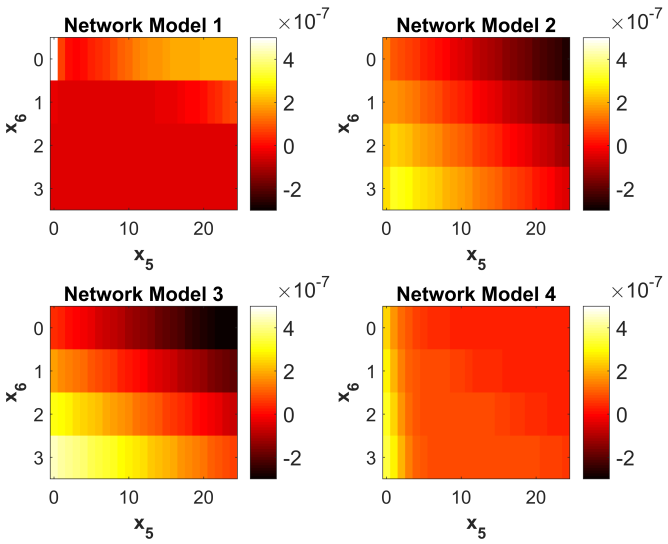

Figure 4: Four different artifical neural network models trained on the airline data to predict $\tilde{y}$ based on the features x_5 and x_6. In each image plot is given the output of one of the four networks using different combinations of x_5 and x_6. As such, $x_5 = 0$ and $x_6 = 0$ is given at the upper left corner whereas $x_5 = 24$ and $x_6 = 3$ is given at the lower right corner of each image.

Question 8. We will consider an artificial neural network (ANN) trained to predict $\tilde{y}$ based only on using x_5 and x_6 as inputs, i.e., incidences and fatal accidents during 2000-2014. The trained model is given by $f(\boldsymbol{x}, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x_5\ x_6]\boldsymbol{w}_j^{(1)})$, where $h^{(1)}(z) = 1/(1+exp(-z))$ is the logistic function used as activation function in the hidden layer (i.e., values are mapped to be between 0 and 1). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} 0.0189 \\ 0.9159 \\ -0.4256 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} 3.7336 \\ -0.8003 \\ 5.0741 \end{bmatrix},$$

and $w_0^{(2)} = 0.3799 \cdot 10^{-6}$, $w_1^{(2)} = -0.3440 \cdot 10^{-6}$, and $w_2^{(2)} = 0.0429 \cdot 10^{-6}$. Which one of the resulting outputs as function of x_5 and x_6 given in Figure 4 corresponds to the trained network?

A. Network Model 1.

B. Network Model 2.

C. Network Model 3.

D. Network Model 4.

E. Don't know.

Solution 8. Consider for instance the two corners located at $x_5 = 0$, $x_6 = 3$ and $x_5 = 24$, $x_6 = 0$ at these two locations we have that the outputs are respectively given by:

$$\begin{aligned} f([03], \boldsymbol{w}) &= 0.3799 \cdot 10^{-6} \\ &\quad - 0.3440 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 0\ 3] \cdot \begin{bmatrix} 0.0189 \\ 0.9159 \\ -0.4256 \end{bmatrix}))} \\ &\quad + 0.0429 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 0\ 3] \cdot \begin{bmatrix} 3.7336 \\ -0.8003 \\ 5.0741 \end{bmatrix}))} \\ &= 0.3799 \cdot 10^{-6} - 0.3440 \cdot 10^{-6} \cdot 0.2213 \\ &\quad + 0.0429 \cdot 10^{-6} \cdot 1.0000 \\ &= 3.4667 \cdot 10^{-7} \end{aligned}$$

and

$$\begin{aligned} f([240], \boldsymbol{w}) &= 0.3799 \cdot 10^{-6} \\ &\quad - 0.3440 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 24\ 0] \cdot \begin{bmatrix} 0.0189 \\ 0.9159 \\ -0.4256 \end{bmatrix}))} \\ &\quad + 0.0429 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 24\ 0] \cdot \begin{bmatrix} 3.7336 \\ -0.8003 \\ 5.0741 \end{bmatrix}))} \\ &= 0.3799 \cdot 10^{-6} - 0.3440 \cdot 10^{-6} \cdot 1 \\ &\quad + 0.0429 \cdot 10^{-6} \cdot 4.2410e \cdot 10^{-7} \\ &= 3.5900 \cdot 10^{-8}. \end{aligned}$$

The only network that has this property is Network Model 4.

Question 9. We would like to use two level cross-validation to select for the optimal number of hidden units in an artificial neural network (ANN) with one hidden layer as well as quantify the generalization of the selected model. For this purpose, we will use two-level cross-validation in which we in the outer fold use 5-fold cross-validation and in the inner fold (i.e., the fold in which we quantify the optimal number of hidden units in the hidden layer) use 10-fold cross-validation. As ANNs are prone to local minima issues we will train three models for each specification of the number of hidden unit based on three different random initializations and use the model out of these three with best training error. We have a computational budget of training 1000 models and would like to evaluate in steps of 1 from 1 to H hidden units (i.e., if H=3 we will evaluate ANNs with 1, 2, and 3 hidden units). What is the largest value of H for which no more than 1000 models will be trained?

A. 6

B. 19

C. 20

D. 66

E. Don't know.

Solution 9. In two level cross-validation we have for each inner fold $10 \cdot H$ different models. For the optimal of these selected models, we will have to train an additional model on the full dataset used for the inner fold to predict the test data of the outer fold. This thus requires the training of $5 \cdot (10 \cdot H + 1)$ models. As we for each trained model use three random initializations we obtain a total of $3 \cdot 5 \cdot (10 \cdot H + 1)$ models to be trained. We thereby obtain $3 \cdot 5 \cdot (10 \cdot H + 1) = 1000 \Rightarrow (10 \cdot H + 1) = 1000/15 \Rightarrow H = (1000/15 - 1)/10 = 6.5667$. We can thus maximally evaluate for H=6.

Question 10. Some people are afraid of flying and thus prefer to take for instance the car or bus. According to the economist Ian Savage[2] the probability of dying travelling 600 km is approximately:
- Chance dying travelling by car is 0.000271 %.
- Chance dying travelling by bus is 0.000004 %.
- Chance dying travelling by plane is 0.000003 %.

The distance travelling from Copenhagen to Oslo is 600 km regardless of the trip being based on car, bus, or plane. We will assume when travelling from Copenhagen to Oslo 30 % of people take the car, 10 % of people take the bus, and 60 % of people take the plane. Given a person died travelling between Copenhagen and Oslo what is the probability it was from travelling by plane?

A. $1.80 \cdot 10^{-4}$ %

B. 1.08 %

C. 2.16 %

D. 10.0 %

E. Don't know.

Solution 10. Let D denote the event dying travelling between Copenhagen and Oslo. Let C denote the event car, B the event Buss and F the event plane (i.e., flight). According to the numbers given we have P(D—C)=0.000271 %, P(D—B)=0.000004 %, and P(D—F)=0.000003 %. Using Bayes theorem we have:

$$\begin{aligned} P(F|D) &= \frac{P(D|F)P(F)}{P(D|F)P(F) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.000003\% \cdot 60\%}{0.000003\% \cdot 60\% + 0.000004\% \cdot 10\% + 0.000271\% \cdot 30\%} \\ &= 2.16\% \end{aligned}$$

[2] http://faculty.wcas.northwestern.edu/~ipsavage/MosesLecture.pdf

Question 11. We will predict whether an airline company is relatively safe (considered the positive class) or unsafe (considered the negative class) based on thresholding $\tilde{y}$ at its median value, i.e. if $\tilde{y} < \text{median}(\tilde{y})$ it is considered safe otherwise it is considered unsafe.

A decision tree is subsequently fitted to the data. At the root of the tree it is considered to split according to the median value of the number of incidences in 2000-2014 (i.e., x_5). For impurity we will use the classification error given by $I(v) = 1 - \max_c p(c|v)$. Before the split, we have 32 safe and 24 unsafe airline companies, and after the split we have

- 23 safe and 8 unsafe airline companies with relatively few incidences.
- 9 safe and 16 unsafe airline companies with relatively many incidences.

Which statement regarding the purity gain Δ of the split is correct?

A. $\Delta = -0.2679$

B. $\Delta = 0.1250$

C. $\Delta = 0.2500$

D. $\Delta = 0.4286$

E. Don't know.

Solution 11. The purity gain is given by

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

where

$$I(v) = 1 - \max_c p(c|v).$$

Evaluating the purity gain for the split we have:

$$\begin{aligned}\Delta =& (1 - 32/56) \\ & -[\tfrac{31}{56}(1 - (\tfrac{23}{31})) \\ & +\tfrac{25}{56}(1 - (\tfrac{16}{25}))] \\ =& \, 7/56\end{aligned}$$

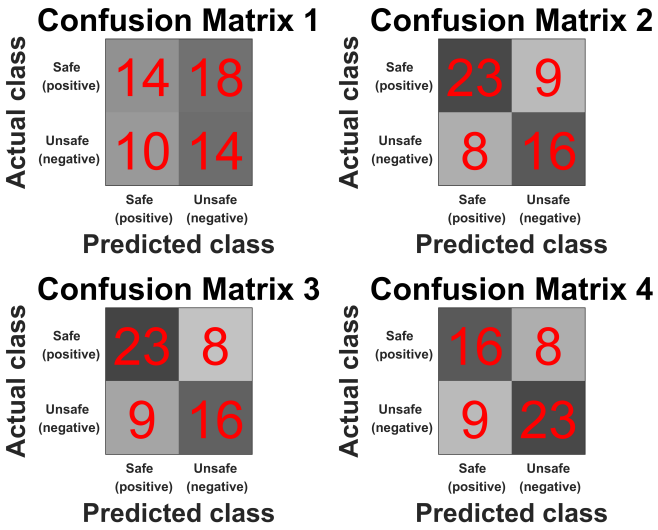

Figure 5: Four different confusion matrices where one corresponds to the confusion matrix of the decision tree with one split according to the median value of the number of incidences in 2000-2014 (i.e., x_5).

Question 12. We will consider the decision tree given by having only the above split (defined in question 11) as a decision and classifying according to this split using the largest class (i.e., using majority voting). In Figure 5 is given four different confusion matrices. Which one of the four confusion matrices corresponds to the decision tree's classification of the 56 observations?

A. Confusion Matrix 1.

B. Confusion Matrix 2.

C. Confusion Matrix 3.

D. Confusion Matrix 4.

E. Don't know.

Solution 12. The decision tree will use majority voting at each leaf in order to classify the 56 observations. For the left branch we have that the majority is safe and thus the 23 safe observations will be correctly classified whereas the 8 unsafe classification will be misclassified as safe. Likewise for the right branch, the majority is unsafe and thus the 16 unsafe observations will be classified as unsafe whereas the 8 safe observations will be misclassified as unsafe. This corresponds to confusion matrix 2.

Question 13. Which statement regarding classification is correct?

A. In classification the output value is continuous.

B. Logistic regression is not a classification approach but a regression method.

C. The k-means algorithm is a supervised classification method.

D. The softmax function is used to provide the probability that an observation is assigned to each class.

E. Don't know.

Solution 13. We make the distinction between a classification and a regression problem based on the property of the output variable being either categorical or continuous. Thus, for continuous outputs we use regression methods and not classification methods. Logistic regression is designed for binary outputs and thus a classification method. The k-means algorithm is used for unsupervised learning and thus not a supervised classification method as it only relies on the input data $\boldsymbol{X}$ and not on output values $\boldsymbol{y}$. The softmax function is used in multinomial regression and artificial neural networks for multi-class classification problems in order to provide outputs interpreted as the probability an observation is assigned to each class similar to the role of the logistic function in logistic regression.

Question 14. We will consider Confusion Matrix 1 given in Figure 5 and we regard the class safe as the *positive* class and unsafe as the *negative* class. Which statement regarding a classifier having performance given by the performance indicated by Confusion Matrix 1 is correct?

A. The classifier's precision is 7/12.

B. The classifier's recall is 1/2.

C. The false positive rate (FPR) of the classifier is 5/14.

D. The accuracy of the classifier is better than guessing everything to be the largest class.

E. Don't know.

Solution 14. The precision is 14/(14+10)=7/12. The recall is 14/(14+18)=7/16. The FPR is 10/(10+14)=5/12. Guessing everything to be the largest class would correspond to guessing everything as unsafe with accuracy of 32/56 whereas the accuracy of the classifier is 28/56.

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0	8.55	0.43	1.25	1.14	3.73	2.72	1.63	1.68	1.28
O2	8.55	0	8.23	8.13	8.49	6.84	8.23	8.28	8.13	7.66
O3	0.43	8.23	0	1.09	1.10	3.55	2.68	1.50	1.52	1.05
O4	1.25	8.13	1.09	0	1.23	3.21	2.17	1.29	1.33	0.56
O5	1.14	8.49	1.10	1.23	0	3.20	2.68	1.56	1.50	1.28
O6	3.73	6.84	3.55	3.21	3.20	0	2.98	2.66	2.50	3.00
O7	2.72	8.23	2.68	2.17	2.68	2.98	0	2.28	2.30	2.31
O8	1.63	8.28	1.50	1.29	1.56	2.66	2.28	0	0.25	1.46
O9	1.68	8.13	1.52	1.33	1.50	2.50	2.30	0.25	0	1.44
O10	1.28	7.66	1.05	0.56	1.28	3.00	2.31	1.46	1.44	0

Table 3: Pairwise Euclidean distances between the first ten observations of the standardized airline safety data. Black observations (i.e., O1, O3, O4, O5, O10) are observations corresponding to relatively safe airline companies, red observations (i.e., O2, O6, O7, O8, O9) are observations corresponding to relatively unsafe airline companies.

Question 15. To determine whether an airline company is relatively safe or unsafe we will use a k-nearest neighbor (KNN) classifier to predict each of the ten observations based on the Euclidean distances between the observations given in Table 3. We will use leave-one-out cross-validation for the KNN in order to classify the ten considered observations and use $K = 1$, i.e., a one nearest neighbor classifier. The analysis will be based only on the data given in Table 3. What will be the error rate of the classifier?

A. 0 %

B. 10 %

C. 20 %

D. 30 %

E. Don't know.

Solution 15. $N(O1, 1) = \{O3\}$ as O3 is closest it will be correctly classified as safe.
$N(O2, 1) = \{O6\}$ as O6 is closest it will be correctly classified as unsafe.
$N(O3, 1) = \{O1\}$ as O1 is closest it will be correctly classified as safe.
$N(O4, 1) = \{O10\}$ as O10 is closest it will be correctly classified as safe.
$N(O5, 1) = \{O3\}$ as O3 is closest it will be correctly classified as safe.
$N(O6, 1) = \{O9\}$ as O9 is closest it will be correctly classified as unsafe.
$N(O7, 1) = \{O4\}$ as O4 is closest it will be incorrectly classified as safe.
$N(O8, 1) = \{O9\}$ as O9 is closest it will be correctly classified as unsafe.
$N(O9, 1) = \{O8\}$ as O8 is closest it will be correctly classified as unsafe.
$N(O10, 1) = \{O4\}$ as O4 is closest it will be correctly classified as safe.
Thus, one out of the ten observations will be misclassified.

Question 16. We will again consider the Euclidean distances between the first ten observations given in Table 3. Agglomerative hierarchical clustering is used to cluster these ten observations based on their distances to each other using average linkage. Which one of the dendrograms given in Figure 6 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 16. As O2 merges last to the cluster containing all the remaining observations we can simply evaluate at what level O2 will merge which is given by O2's average distance to the observations {O1,O3,O4,O5,O6,O7,O8,O9,O10} which is given by $(8.55 + 8.23 + 8.13 + 8.49 + 6.84 + 8.24 + 8.28 + 8.13 + 7.66)/9 = 8.0611$. Only dendrogram 4 has this property.

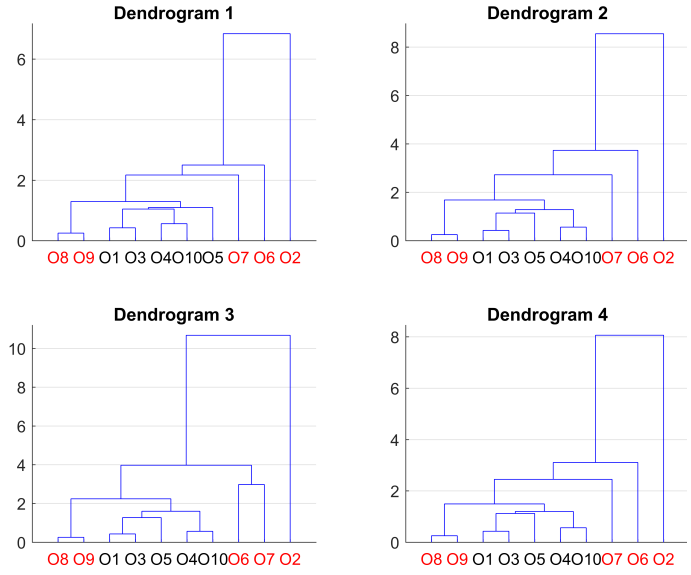

Figure 6: Four different dendrograms derived using the Euclidean distances between the first 10 observations in the airline safety data. Black observations correspond to relatively safe companies whereas red to relatively unsafe companies.

Question 17. We will cut dendrogram 1 at the level of three clusters and evaluate this clustering in terms of its correspondence with the class label information in which black observations, i.e., O1, O3, O4, O5, O10, are observations corresponding to relatively safe airline companies, and red observations, i.e., O2, O6, O7, O8, O9, are observations corresponding to relatively unsafe airline companies. We recall that the Rand index also denoted the simple matching coefficient (SMC) between the true labels and the extracted clusters is given by $R = \frac{f_{11}+f_{00}}{K}$, where f_{11} is the number of object pairs in same class assigned to same cluster, f_{00} is the number of object pairs in different class assigned to different clusters, and $K = N(N-1)/2$ is the total number of object pairs, where N is the number of observations considered. What is the value of R between the true labeling of the observations and the three extracted clusters?

A. 0.40

B. 0.47

C. 0.51

D. 0.60

E. Don't know.

Solution 17. The cluster indices are given by the vector: $[1211131111]^\top$, whereas the true class labels are given by the vector $[1211122221]^\top$. From this, we obtain:
$K = 10(10-1)/2 = 45$
$f_{00} = 5 \cdot 1 + 5 \cdot 1 + 1 \cdot 0 = 10$
$f_{11} = 5{\cdot}(5{-}1)/2{+}3{\cdot}(3{-}1)/2{+}1{\cdot}(1{-}1)/2{+}1{\cdot}(1{-}1)/2 = 13$
$R = \frac{f_{11}+f_{00}}{K} = \frac{13+10}{45} = 23/45.$

Question 18. We suspect that observation O2 may be an outlier. In order to quantify if this could be the case we will calculate the average relative KNN density based on Euclidean distance and 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 O2 for $K = 2$ nearest neighbors?

A. 0.085

B. 0.138

C. 0.169

D. 0.356

E. Don't know.

Solution 18.

$$\text{density}(\boldsymbol{x}_{O2}, 2) = (\frac{1}{2}(6.84 + 7.66))^{-1} = 0.1379$$
$$\text{density}(\boldsymbol{x}_{O6}, 2) = (\frac{1}{2}(2.50 + 2.66))^{-1} = 0.3876$$
$$\text{density}(\boldsymbol{x}_{O10}, 2) = (\frac{1}{2}(0.56 + 1.05))^{-1} = 1.2422$$
$$\text{a.r.d.}(\boldsymbol{x}_{O2}, 2)) = \frac{\text{density}(\boldsymbol{x}_{O2}, 2)}{\frac{1}{2}(\text{density}(\boldsymbol{x}_{O6}, 2) + \text{density}(\boldsymbol{x}_{O10}, 2))} = \frac{0.1379}{\frac{1}{2}(0.3876 + 1.2422)} = 0.169$$

	x_1^L	x_1^H	x_2^L	x_2^H	x_3^L	x_3^H	x_4^L	x_4^H	x_5^L	x_5^H	x_6^L	x_6^H
O1	1	0	1	0	1	0	1	0	1	0	1	0
O2	0	1	0	1	0	1	0	1	0	1	0	1
O3	1	0	0	1	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1	0	1	1	0
O5	0	1	1	0	1	0	1	0	1	0	1	0
O6	0	1	0	1	0	1	0	1	0	1	0	1
O7	0	1	1	0	1	0	0	1	0	1	0	1
O8	1	0	1	0	1	0	1	0	0	1	0	1
O9	0	1	0	1	1	0	1	0	0	1	0	1
O10	1	0	0	1	0	1	0	1	0	1	1	0

Table 4: The ten first observations of the airline safety dataset binarized considering the attribute x_1–x_6. The attributes are all binarized according to being below or equal (denoted L) or above the median value (denoted H). The ten observations are color coded in terms of relatively safe {O1, O3, O4, O5, O10} or unsafe {O2, O6, O7, O8, O9} airline companies.

Question 19. We will binarize each feature in the airline safety data according to the median value of the feature denoting below or equal the median value using the superscript L and above the median value using the superscript H. In Table 4 is given the first 10 observations after this binarization and we will consider these 10 observations as a dataset used for market basket analysis with observation O1–O10 corresponding to customers. What is the support for the association rule $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

Solution 19. The support of $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$ is given by the number of times out of the total number of customers that customers have relatively high values of both x_2, x_3, x_4, x_5, and x_6, i.e., given by the support of the itemset $\{x_2^H, x_3^H, x_4^H, x_5^H, x_6^H\}$. Only customer O2 and O6 have this property out of the 10 customers, thus the support is 2/10.

Question 20. We consider again the data in Table 4 as a market basket problem. What is the confidence of the association rule $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

Solution 20. The confidence is given as

$$P(x_6^H = 1 | x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1) = \frac{P(x_6^H = 1, x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1)}{P(x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1)} = \frac{2/10}{3/10} = 2/3 = 66.7\%$$

Question 21. We would like to predict whether an airline company is relatively safe or unsafe considering only the data given in Table 4. We will apply a Naïve Bayes classifier that assumes independence between the attributes given the class label (i.e., the class label is given by safe airlines in black, i.e., O1, O3, O4, O5, and O10, and unsafe airlines in red, i.e., O2, 06, O7, O8, and O9). Given that an airline company has $x_2^H = 1$, $x_3^H = 1$, $x_4^H = 1$, $x_5^H = 1$ what is the probability that the airline company is considered safe according to the Naïve Bayes classifier derived from the data in Table 4?

A. 0

B. 4/625

C. 4/49

D. 1/3

E. Don't know.

Solution 21. Let $\tilde{y} = 1$ denote that the airline is safe.

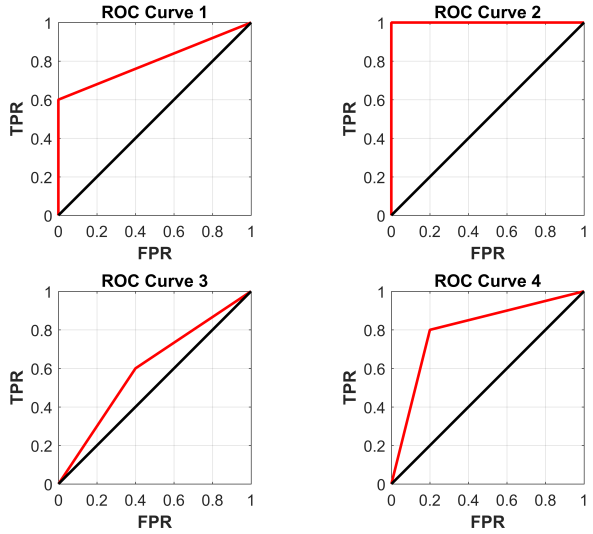

Figure 7: Four different receiver operator characteristic (ROC) curves.

According to the Naïve Bayes classifier we have

$$P(\tilde{y} = 1 | x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1) = \frac{\begin{pmatrix} P(x_2^H = 1|\tilde{y} = 1)\times \\ P(x_3^H = 1|\tilde{y} = 1)\times \\ P(x_4^H = 1|\tilde{y} = 1)\times \\ P(x_5^H = 1|\tilde{y} = 1)\times \\ P(\tilde{y} = 1) \end{pmatrix}}{\begin{pmatrix} P(x_2^H = 1|\tilde{y} = 1)\times \\ P(x_3^H = 1|\tilde{y} = 1)\times \\ P(x_4^H = 1|\tilde{y} = 1)\times \\ P(x_5^H = 1|\tilde{y} = 1)\times \\ P(\tilde{y} = 1) \\ +P(x_2^H = 1|\tilde{y} = 0)\times \\ P(x_3^H = 1|\tilde{y} = 0)\times \\ P(x_4^H = 1|\tilde{y} = 0)\times \\ P(x_5^H = 1|\tilde{y} = 0)\times \\ P(\tilde{y} = 0) \end{pmatrix}}$$

$$= \frac{2/5 \cdot 1/5 \cdot 2/5 \cdot 2/5 \cdot 5/10}{2/5 \cdot 1/5 \cdot 2/5 \cdot 2/5 \cdot 5/10 + 3/5 \cdot 2/5 \cdot 3/5 \cdot 5/5 \cdot 5/10}$$

$$= \frac{40/(5^4 \cdot 10)}{40/(5^4 \cdot 10) + 450/(5^4 \cdot 10)} = 40/490 = 4/49$$

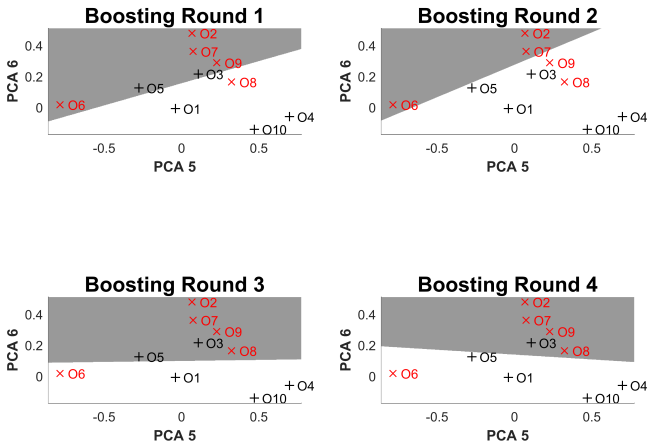

Figure 8: Decision boundaries for four rounds of boosting considering a logistic regression model using the fifth and sixth principal components as features and the first 10 observations of the airline safety data. Gray region indicates that the observation will be classified as unsafe (red crosses), white regions that the observation will be classified as safe (black plusses).

	Round 1	Round 2	Round 3	Round 4
O1	0.1000	0.0714	0.0469	0.0319
O2	0.1000	0.0714	0.0469	0.0319
O3	0.1000	0.1667	0.1094	0.2059
O4	0.1000	0.0714	0.0469	0.0319
O5	0.1000	0.1667	0.1094	0.2059
O6	0.1000	0.0714	0.0469	0.0882
O7	0.1000	0.0714	0.0469	0.0319
O8	0.1000	0.1667	0.3500	0.2383
O9	0.1000	0.0714	0.1500	0.1021
O10	0.1000	0.0714	0.0469	0.0319

Table 5: The weights for the first four rounds of AdaBoost.

Question 22. We will use x_5^L to determine if an airline is safe *considered the positive class* or unsafe *considered the negative class*. In Figure 7 are given four different receiver operator characteristic curves (ROC). Which one of the four ROC curves corresponds to using x_5^L to determine if an airline is safe (positive class) or unsafe (negative class)?

A. ROC curve 1.

B. ROC curve 2.

C. ROC curve 3.

D. ROC curve 4.

E. Don't know.

Solution 22. The ROC curve starts at (0,0) for which we threshold above 1. When tresholding at one we obtain that 3 out of 5 safe airlines (positive class) have $x_5^L = 1$ and 0 out of 5 unsafe (negative class) have $x_5^L = 1$ thus the ROC curve will be at the point (0,3/5). When we subsequently further lower the threshold to be at 0 all airline companies that are safe and unsafe will be at this threshold value or above, thus, the ROC curve will end here at (1,1). Only ROC curve 1 has this property.

Question 23. We would like to build a model for classifying whether and airline is safe or not based on the first 10 observations of the airline safety dataset. In order to do so, we will use the resulting classifier obtained by using the AdaBoost algorithm and a logistic regression classifier. The weights of the first four rounds of the AdaBoost procedure is given in Figure 8 and the associated sampling weights used for each round is given in Table 5. How will observation O5 and O6 be classified according to the ensemble classifier obtained by combining the four boosting rounds using the voting procedure defined by the AdaBoost algorithm?

A. Observation O5 and O6 will be tied between safe and unsafe by the AdaBoost classifier.

B. Both observation O5 and O6 will be correctly classified by the AdaBoost classifier.

C. Only one of the two observations O5 and O6 will be correctly classified by the AdaBoost classifier.

D. Neither of the two observations O5 and O6 will be correctly classified by the AdaBoost classifier.

E. Don't know.

Solution 23. The resulting boosting procedure weight each classifier according to their importance α_t for the t^{th} round, where $\alpha_t = 0.5 \log \frac{1-e_t}{e_t}$ such that $e_t = \sum_n w_n(t)(1 - \delta_{f_t(x_n),y_n})$ is the error rate weighted according to the weights of the t^{th} round. During the first round O3, O5 and O8 are misclassified and thus $e_1 = 0.1 + 0.1 + 0.1 = 0.3$ and consequently $\alpha_1 = 0.5 \log \frac{1-0.3}{0.3} = 0.4236$. In the second round O8 and O9 are mis-classified and thus $e_2 = 0.1667 + 0.0714 = 0.2381$ and therefore $\alpha_2 = 0.5 \log \frac{1-0.2381}{0.2381} = 0.5816$. In the third round O3, O5, and O6 are misclassified thus $e_3 = 0.1094 + 0.1094 + 0.0469 = 0.2657$ and therefore $\alpha_3 = 0.5 \log \frac{1-0.2657}{0.2657} = 0.5083$. Finally, in the fourth round O3 and O6 are misclassified and thus $e_4 = 0.2059 + 0.0882 = 0.2941$ and therefore $\alpha_4 = 0.5 \log \frac{1-0.2657}{0.2657} = 0.4378$. For observation O5 both classifier of round 2 and 4 vote for it being safe with the strength of vote given by $\alpha_2 + \alpha_4 = 0.5816 + 0.4378 = 1.0194$ whereas the strength of vote for unsafe is lower, i.e. $\alpha_1 + \alpha_3 = 0.4236 + 0.5083 = 0.9319$. Observation O6 will be classified as unsafe as classifier 1 and 2 classifies it as unsafe with strength $\alpha_1 + \alpha_2 = 0.4236 + 0.5816 = 1.0052$ and safe according to classifier 3 and 4 with the lower voting strength of $\alpha_3 + \alpha_4 = 0.5083 + 0.4378 = 0.9461$.

Question 24. Four different classifiers are trained on the airline safety data considering only the first 10 observations projected onto the fifth and sixth principal component to determine if an airline is relatively safe or unsafe. The decision boundary for each of the four classifiers is given in Figure 9 when only using as input the data projected onto the fifth (PCA 5) and sixth (PCA 6) principal component, i.e. each method has only these two inputs. Which one of the following statements is correct?

A. Classifier 1 corresponds to a logistic regression classifier, Classifier 2 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 3 is a decision tree classifier, and Classifier 4 corresponds to an artificial neural network (ANN).

B. Classifier 1 is a Naive Bayes classifier based on the use of univariate normal distributions, Classifier 2 is a 3-nearest neighbor classifier, Classifier 3 is a Decision Tree classifier, and Classifier 4 is a 1-nearest neighbor classifier using Euclidean distance.

C. Classifier 1 is a Naive Bayes classifier based on the use of univariate normal distributions, Classifier 2 is a 3-nearest neighbor classifier, Classifier 3 is a 1-nearest neighbor classifier, and Classifier 4 is a Decision Tree classifier.

D. Classifier 1 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 2 is a Naive Bayes classifier based on the use of univariate normal distributions, Classifier 3 is a Decision Tree classifier, and Classifier 4 is a 1-nearest neighbor classifier using Euclidean distance.

E. Don't know.

Solution 24. Classifier 1 is a 3-nearest neighbor classifier as such one red cross and one black plus are within the wrong decision boundary due to the majority voting of the three nearest neighbors. Classifier 2 has smooth decision boundaries which would correspond to the Naive Bayes classifier based on the use of univariate normal distributions. Classifier 3 is a Decision Tree classifier due to its vertical and horizontal decision boundaries. Classifier 4 is a 1-nearest neighbor classifier using Euclidean distance as the decision boundary clearly follows the most nearby observation.

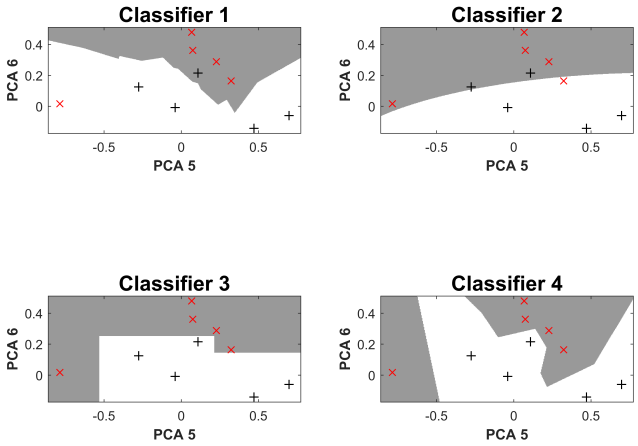

Figure 9: Decision boundaries for four different classifiers trained on the airline safety data using principal component 5 and 6 as input to the classifiers. Gray regions classify into red crosses whereas white regions into black plusses.

Question 25. Consider a dataset with eight observations located at {1.0, 1.2, 1.5, 2.0, 2.2, 2.5, 3.0, 3.2}. We will cluster the dataset using the k-means algorithm using $k = 3$ clusters and initialize the clusters at the locations of the first three observations, i.e. cluster 1 will be initially located at 1.0, cluster 2 at 1.2 and cluster 3 at 1.5. What will be the converged clustering of the eight observations using the k-means procedure based on Euclidean distance as dissimilarity?

A. {1.0}, {1.2, 1.5}, {2.0, 2.2, 2.5, 3.0, 3.2}

B. {1.0, 1.2}, {1.5}, {2.0, 2.2, 2.5, 3.0, 3.2}

C. {1.0, 1.2, 1.5}, {2.0, 2.2, 2.5}, { 3.0, 3.2}

D. {1.0, 1.2, 1.5}, {2.0, 2.2}, {2.5, 3.0, 3.2}

E. Don't know.

Solution 25. The cluster located at 1.5 will be closest to the observations located at 2.0, 2.2, 2.5, 3.0, and 3.2 and will therefore be assigned these whereas cluster located at 1.0 and 1.2 will only be assigned respectively the observation located at 1.0 and 1.2. The location of the centroid for cluster 3 will thereby be changed such that cluster 3 is updated to be located at: $(1.5+2.0+2.2+2.5+3.0+3.2)/6 = 2.4$. Subsequently, observation 1.5 will be closer to cluster located at 1.2 than cluster located at 2.4 and will therefore as only observation change assignment, such that cluster 2 will

be updated to be located at $(1.2 + 1.5)/2 = 1.35$ whereas cluster 3 will be updated to be located at $(2.0+2.2+2.5+3.0+3.2)/5 = 2.58$. As no observation will change assignment based on the location of the updated clusters the algorithm will converge to the clustering given by:
{1.0}, {1.2, 1.5}, {2.0, 2.2, 2.5, 3.0, 3.2}.

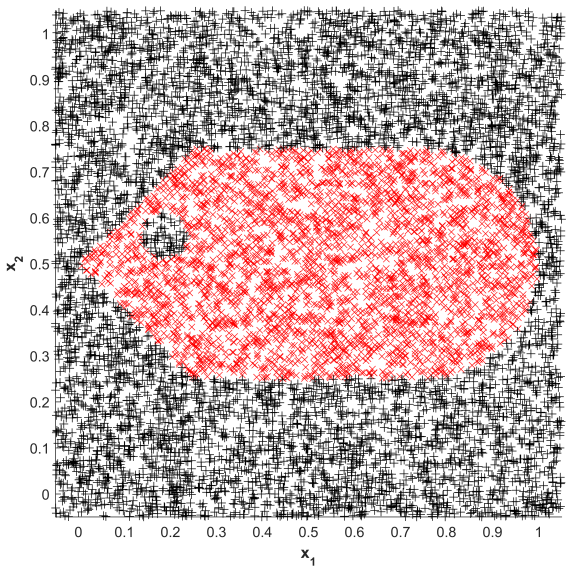

Figure 10: A two class classification problem with red crosses (i.e., x) and black plusses (i.e., +) constituting the two classes.

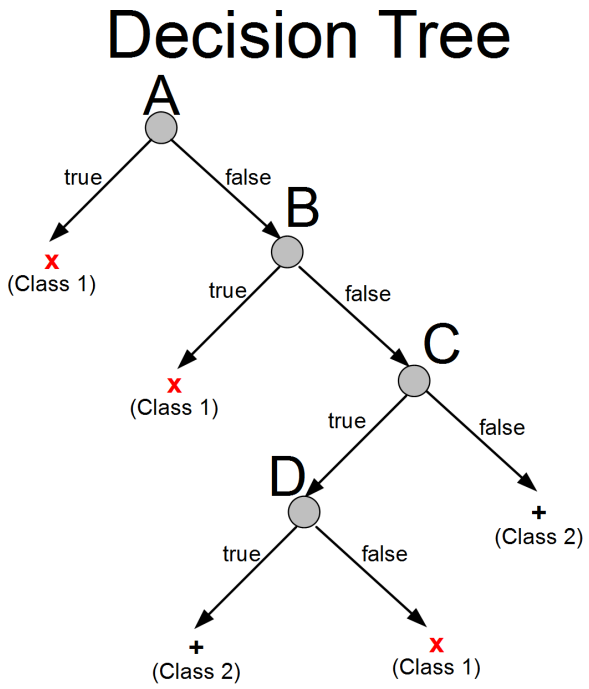

Figure 11: A decision tree with four decisions (A, B, C, and D) perfectly separating the black plusses from red crosses in Figure 10 if adequately defined.

Question 26. We will consider the two class classification problem given in Figure 10 in which the goal is to separate red crosses (i.e., x) from black plusses (i.e., +). Which one of the following procedures based on the decision tree given in Figure 11 will perfectly separate the two classes?

A. $\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathrm{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_2 \le 1/20,$
$\mathrm{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathrm{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/4.$

B. $\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathrm{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_2 \le 1/4,$
$\mathrm{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathrm{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/20.$

C. $\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathrm{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathrm{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_2 \le 1/4,$
$\mathrm{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/20.$

D. $\mathbf{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathbf{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_2 \le 1/4,$
$\mathbf{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathbf{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/20.$

E. Don't know.

Solution 26. The red crosses contains a circle located at (0.75,0.5) and a square located at (0.5, 0.5) and a diamond located at (0.25, 0.5) all with radius 0.25 corresponding to:
$\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_\infty \le 1/4,$

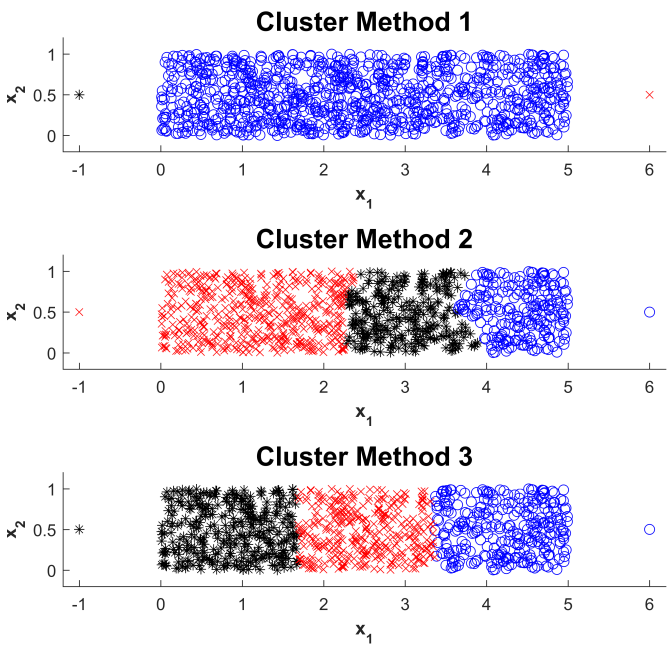

Figure 12: A dataset with two features x_1 and x_2 and 100 observations clustered using three different clutering approaches.

B=$\|\boldsymbol{x} - \begin{bmatrix} 3/4 \\ 1/2 \end{bmatrix}\|_2 \leq 1/4$,

C=$\|\boldsymbol{x} - \begin{bmatrix} 1/4 \\ 1/2 \end{bmatrix}\|_1 \leq 1/4$.

Finally, there are black plusses located at (0.1875, 0.5625) with a small radius of 0.05 corresponding to:

D=$\|\boldsymbol{x} - \begin{bmatrix} 3/16 \\ 9/16 \end{bmatrix}\|_2 \leq 1/20$.

This only holds for the last answer option.

Question 27. We will consider a dataset with two features x_1 and x_2 and 1000 observations that is clustered using three different clustering approaches all based on Euclidean distance as measure of distance. The clustering extracted by each of the three considered approaches are given in Figure 12. Which one of the following statements is correct?

A. Cluster Method 1 corresponds to k-means,
Cluster Method 2 corresponds to hierarchical clustering using single linkage,
Cluster method 3 corresponds to hierarchical clustering using complete linkage.

B. Cluster Method 1 corresponds to hierarchical clustering using single linkage,
Cluster Method 2 corresponds to k-means,
Cluster method 3 corresponds to hierarchical clustering using complete linkage.

**C. Cluster Method 1 corresponds to hierarchical clustering using single linkage,
Cluster method 2 corresponds to hierarchical clustering using complete linkage,
Cluster Method 3 corresponds to k-means.**

D. Cluster Method 1 corresponds to hierarchical clustering using complete linkage,
Cluster method 2 corresponds to hierarchical clustering using single linkage,
Cluster Method 3 corresponds to k-means.

E. Don't know.

Solution 27. Single linkage clusters consecutively by merging according to the two observations of each cluster that is closest to each other. As such, the clustering will be influenced by the gaps between the observations and therefore the two outlying observations at (-1,0.5) and (6,0.5) will be merged the latest in the dendrogram resulting in the three clusters given by Cluster Method 1. Complete linkage clusters consecutively by merging according to the two observations of each cluster that is the furthest apart. As such, the clustering will be influenced by how far the cluster extends as well as influenced by earlier merge decisions corresponding to the clustering given by Cluster Method 2. K-means will cluster observations according to their proximity to the center of the cluster which corresponds to a clustering given by Cluster Method 3. Neither Cluster Method 1

and Cluster Method 2 can be k-means as the distance to the centroid of each cluster would result in a different clustering configuration than the ones obtained. Furthermore, cluster method 2 cannot be single linkage as the major gaps to (-1,0.5) and (6,0.5) will make these merge latest in the dendrogram. Thus the only correct answer option is:
Cluster Method 1 corresponds to hierarchical clustering using single linkage,
Cluster method 2 corresponds to hierarchical clustering using complete linkage,
Cluster Method 3 corresponds to k-means.

Technical University of Denmark

Written examination: May 24th 2019, 9 AM - 1 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.

Please hand in your answers using the electronic file. Only use this page in the case where digital handin is unavailable. In case you have to hand in the answers using the form on this sheet, please follow these instructions:

Print name and study number clearly. 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.

The individual questions are answered by filling in the answer fields with one of the letters A, B, C, D, or E.

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

USE ONLY THIS PAGE FOR HAND IN IF YOU ARE UNABLE TO HAND IN DIGITALLY.

No.	Attribute description	Abbrev.
x_1	Average rating of art galleries	art galleries
x_2	Average rating of dance clubs	dance clubs
x_3	Average rating of juice bars	juice bars
x_4	Average rating of restaurants	restaurants
x_5	Average rating of museums	museums
x_6	Average rating of parks/picnic spots	parks
x_7	Average rating of beaches	beaches
x_8	Average rating of theaters	theaters
x_9	Average rating of religious institutions	religious
y	Rating of resort (poor, average, high)	Resort's rating

Table 1: Description of the features of the travel review dataset used in this exam. The dataset is obtained by crawling TripAdvisor.com and consists of reviews of destinations across East Asia in various categories. The scores in each category x_i is based on an average of reviews by travellers for a given resort where each traveler's rating is either Excellent (4), Very Good (3), Average (2), Poor (1), or Terrible (0). The overall score y also corresponds to an average of reviews but it has been discretized to obtain a classification problem. The dataset used here consists of $N = 980$ observations and the attribute y is discrete taking values $y = 1$ (corresponding to a poor rating),$y = 2$ (corresponding to an average rating), and$y = 3$ (corresponding to a high rating).

Question 1. The main dataset used in this exam is the travel review dataset[1] described in Table 1.

In Figure 1 is shown a scatter plot of the two attributes x_2 and x_9 from the travel review dataset and in Figure 2 boxplots of the attributes x_2, x_7, x_8, x_9 (not in that order). Which one of the following statements is true?

A. Attribute x_2 corresponds to boxplot 3 and x_9 corresponds to boxplot 2

B. Attribute x_2 corresponds to boxplot 2 and x_9 corresponds to boxplot 4

C. Attribute x_2 corresponds to boxplot 1 and x_9 corresponds to boxplot 4

D. Attribute x_2 corresponds to boxplot 2 and x_9 corresponds to boxplot 1

E. Don't know.

Solution 1.

[1]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Travel+Reviews`

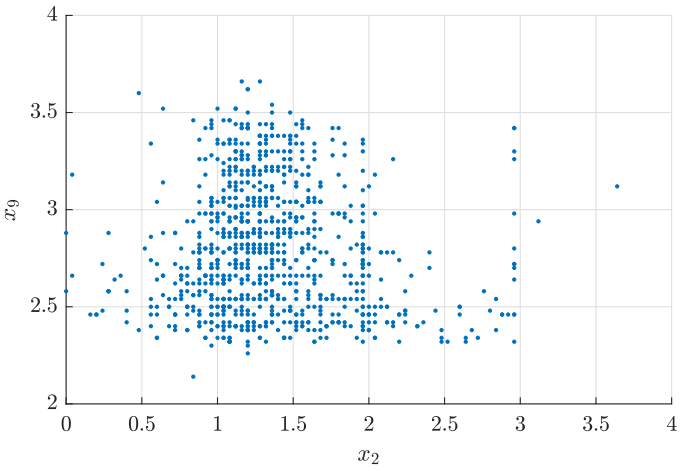

Figure 1: Scatter plot of observations x_2 and x_9 of the travel review dataset described in Table 1.

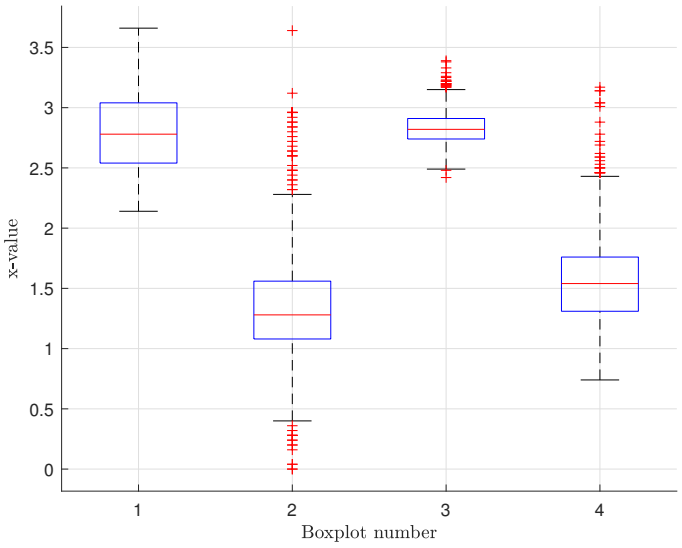

Figure 2: Four boxplots in which two of the boxplots correspond to the two variables plotted in Figure 1.

The correct answer is D. To see this, notice the red line in the boxplot agrees with the median of the attribute, and the median of the two attributes in Figure 1 can be derived by projecting onto either of the two axis and (visually estimate) the point such that half the mass of the data is above and below. For x_2 this is 1.3 and for x_9 this is 2.8, which rule out all but option D.

Question 2. A Principal Component Analysis (PCA) is carried out on the travel review dataset in Table 1 based on the attributes $x_5,\ x_6,\ x_7,\ x_8,\ x_9$.

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 $\boldsymbol{USV}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.94 & -0.12 & 0.32 & -0.0 & 0.0 \\ 0.01 & 0.0 & -0.02 & 0.0 & -1.0 \\ -0.01 & 0.07 & 0.07 & 0.99 & -0.0 \\ 0.11 & 0.99 & 0.06 & -0.08 & 0.0 \\ -0.33 & -0.02 & 0.94 & -0.07 & -0.02 \end{bmatrix} \tag{1}$$

$$\boldsymbol{S} = \begin{bmatrix} 14.14 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 11.41 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 9.46 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 4.19 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.17 \end{bmatrix}$$

Which one of the following statements is true?

A. The variance explained by the first two principal components is greater than 0.815

B. The variance explained by the first principal component is greater than 0.51

C. The variance explained by the last four principal components is less than 0.56

D. The variance explained by the first three principal components is less than 0.9

E. Don't know.

Solution 2. The correct answer is C. To see this, recall 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 components $x_2,\ x_3,\ x_4,\ x_5$ is:

$$\text{Var.Expl.} = \frac{\sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2} = 0.5427.$$

Question 3. Consider again the PCA analysis for the travel review dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (1). Which one of the following statements is true?

A. An observation with a low value of museums, and a high value of religious will typically have a negative value of the projection onto principal component number 1.

B. An observation with a low value of **museums**, and a low value of **religious** will typically have a positive value of the projection onto principal component number 3.

C. An observation with a low value of **museums**, and a high value of **religious** will typically have a positive value of the projection onto principal component number 1.

D. An observation with a high value of **parks** will typically have a positive value of the projection onto principal component number 5.

E. Don't know.

Solution 3. 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

$$b_1 = \boldsymbol{x}^\top \boldsymbol{v}_1 = \begin{bmatrix} x_5 & x_6 & x_7 & x_8 & x_9 \end{bmatrix} \begin{bmatrix} 0.94 \\ 0.01 \\ -0.01 \\ 0.11 \\ -0.33 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and negative, this occurs if x_5, x_9 has large magnitude and the sign convention given in option A.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	2.0	5.7	0.9	2.9	1.8	2.7	3.7	5.3	5.1
o_2	2.0	0.0	5.6	2.4	2.5	3.0	3.5	4.3	6.0	6.2
o_3	5.7	5.6	0.0	5.0	5.1	4.0	3.3	5.4	1.2	1.8
o_4	0.9	2.4	5.0	0.0	2.7	2.1	2.2	3.5	4.6	4.4
o_5	2.9	2.5	5.1	2.7	0.0	3.5	3.7	4.0	5.8	5.7
o_6	1.8	3.0	4.0	2.1	3.5	0.0	1.7	5.3	3.8	3.7
o_7	2.7	3.5	3.3	2.2	3.7	1.7	0.0	4.2	3.1	3.2
o_8	3.7	4.3	5.4	3.5	4.0	5.3	4.2	0.0	5.5	6.0
o_9	5.3	6.0	1.2	4.6	5.8	3.8	3.1	5.5	0.0	2.1
o_{10}	5.1	6.2	1.8	4.4	5.7	3.7	3.2	6.0	2.1	0.0

Table 2: The pairwise cityblock distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{p=1} = \sum_{k=1}^{M} |x_{ik} - x_{jk}|$ between 10 observations from the travel review dataset (recall $M = 9$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1. The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belongs to class C_1 (corresponding to a poor rating), the red observations $\{o_3,\ o_4,\ o_5\}$ belongs to class C_2 (corresponding to an average rating), and the blue observations $\{o_6,\ o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to a high rating).

Question 4. To examine if observation o_7 may be an outlier, we will calculate the average relative density using the cityblock distance and the observations given in Table 2 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_7 for $K = 2$ nearest neighbors?

A. 0.41

B. 1.0

C. 0.51

D. 0.83

E. Don't know.

Solution 4.

To solve the problem, first observe the $k = 2$ neighborhood of o_7 and density is:

$$N_{\boldsymbol{X}\setminus 7}(\boldsymbol{x}_7) = \{o_6, o_4\}, \quad \text{density}_{\boldsymbol{X}\setminus 7}(\boldsymbol{x}_7) = 0.513$$

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

$$N_{\boldsymbol{X}\setminus 6}(\boldsymbol{x}_6) = \{o_7, o_1\}, \quad N_{\boldsymbol{X}\setminus 4}(\boldsymbol{x}_4) = \{o_1, o_6\}$$

and

$$\text{density}_{\boldsymbol{X}\setminus 6}(\boldsymbol{x}_6) = 0.571, \text{density}_{\boldsymbol{X}\setminus 4}(\boldsymbol{x}_4) = 0.667.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem.

Question 5. Consider the distances in Table 2 based on 10 observations from the travel review dataset. The class labels C_1, C_2, C_3 (see table caption for details) will be predicted using a k-nearest neighbour classifier based on the distances given in Table 2 (ties are broken in the usual manner by considering the nearest observation from the tied classes). Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 3-nearest neighbour classifier (i.e. $k = 3$). What is the error rate computed for all $N = 10$ observations?

A. error rate $= \frac{3}{10}$

B. error rate $= \frac{5}{10}$

C. error rate $= \frac{6}{10}$

D. error rate $= \frac{7}{10}$

E. Don't know.

Solution 5.

The correct answer is C. To see this, recall that leave-one-out cross-validation means we train a total of $N = 10$ models, each model being tested on a single observation and trained on the remaining such that each observation is used for testing exactly once.

The model considered is KNN classifier with $k = 3$. To figure out the error for a particular observation i (i.e. the test set for this fold), we train a model on the other observations and predict on observation i. To do that, simply find the observation different than i closest to i according to Table 2 and predict i as belonging to it's class. Concretely, we find: $N(o_1, k) = \{o_4, o_6, o_2\}$, $N(o_2, k) = \{o_1, o_4, o_5\}$, $N(o_3, k) = \{o_9, o_{10}, o_7\}$, $N(o_4, k) = \{o_1, o_6, o_7\}$, $N(o_5, k) = \{o_2, o_4, o_1\}$, $N(o_6, k) = \{o_7, o_1, o_4\}$, $N(o_7, k) = \{o_6, o_4, o_1\}$, $N(o_8, k) = \{o_4, o_1, o_5\}$, $N(o_9, k) = \{o_3, o_{10}, o_7\}$, and $N(o_{10}, k) = \{o_3, o_9, o_7\}$.

The error is then found by observing how often the class label of the observation in the neighborhood agrees with the true class label. We find this happens for observations

$$\{o_6, o_7, o_9, o_{10}\}$$

and the remaining observations are therefore erroneously classified, in other words, the classification error is $\frac{6}{10}$.

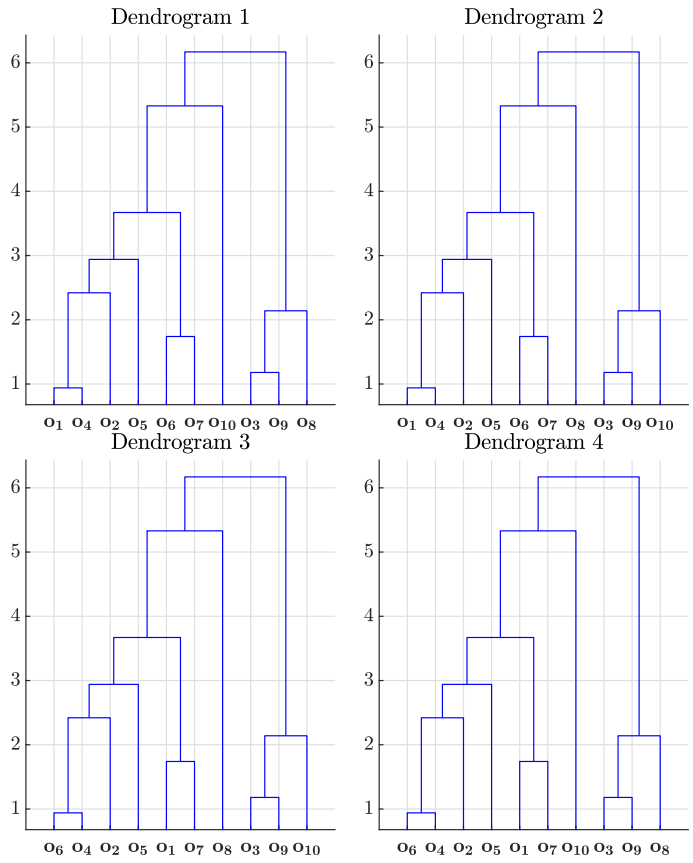

Figure 3: Proposed hierarchical clustering of the 10 observations in Table 2.

Question 6. A hierarchical clustering is applied to the 10 observations in Table 2 using *maximum* linkage. Which one of the dendrograms shown in Figure 3 corresponds to the distances given in Table 2?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

Solution 6. The correct solution is B. We can rule out the other solutions by observing the first merge operation at which they diverge from the correct solution.

- In dendrogram 1, merge operation number 4 should have been between the sets $\{f_{10}\}$ and $\{f_3, f_9\}$ at a height of 2.14, however in dendrogram 1 merge number 4 is between the sets $\{f_8\}$ and $\{f_3, f_9\}$.
- In dendrogram 3, merge operation number 1 should have been between the sets $\{f_1\}$ and $\{f_4\}$ at a height of 0.94, however in dendrogram 3 merge number 1 is between the sets $\{f_6\}$ and $\{f_4\}$.
- In dendrogram 4, merge operation number 1 should have been between the sets $\{f_1\}$ and $\{f_4\}$ at a height of 0.94, however in dendrogram 4 merge number 1 is between the sets $\{f_6\}$ and $\{f_4\}$.

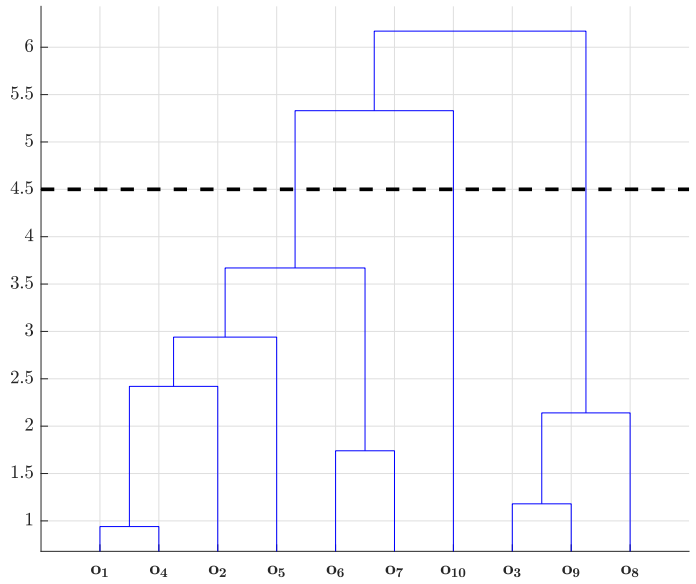

Figure 4: Dendrogram 1 from Figure 3 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

Question 7. Consider dendrogram 1 from Figure 3. Suppose we apply a cutoff (indicated by the black line) thereby generating three clusters. We wish to compare the quality of this clustering, Q, to the ground-truth clustering, Z, indicated by the colors in Table 2. Recall the *Jaccard similarity* of the two clusters is

$$\mathrm{J}[Z,Q] = \frac{S}{\frac{1}{2}N(N-1) - D}$$

in the notation of the lecture notes. What is the Jaccard similarity of the two clusterings?

A. $\mathrm{J}[Z,Q] \approx 0.104$

B. $\mathrm{J}[Z,Q] \approx 0.143$

C. $\mathrm{J}[Z,Q] \approx 0.174$

D. $\mathrm{J}[Z,Q] \approx 0.153$

E. Don't know.

Solution 7. To compute $\mathrm{J}[Z,Q]$, note Z is the clustering corresponding to the colors in Table 2 and Q the clustering obtained by cutting the dendrogram in Figure 4 given as:

$$\{10\},\ \{1,2,4,5,6,7\},\ \{3,8,9\}$$

From this information we can define the counting matrix $\boldsymbol{n}$ as

$$\boldsymbol{n} = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 2 & 1 \\ 1 & 2 & 2 \end{bmatrix}$$

It is then a simple matter of using the definitions in the lecture notes (see chapter 17.4) to compute

$$S = 4, D = 17$$

From this the answer by simply plugging the values into the formula given in the text and answer B is correct.

	$x_4 \leq 0.43$	$x_4 \leq 0.55$
$y = 1$	143	223
$y = 2$	137	251
$y = 3$	54	197

Table 3: Proposed split of the travel review dataset based on the attribute x_4. We consider a two-way split where for each interval we count how many observations belonging to that interval has the given class label.

Question 8. Suppose we wish to build a classification tree based on Hunt's algorithm where the goal is to predict Resort's rating which can belong to three classes, $y = 1,\ y = 2,\ y = 3$. The number of observations in each of the classes are:

$$n_{y=1} = 263,\ n_{y=2} = 359,\ n_{y=3} = 358.$$

We consider binary splits based on the value of x_4 of the form $x_4 < z$ for two different values of z. In Table 3 we have indicated the number of observations in each of the three classes for different values of z. Suppose we use the *classification error* as impurity measure, which one of the following statements is true?

A. The impurity gain of the split $x_4 \leq 0.43$ is $\Delta \approx 0.1045$

B. The impurity gain of the split $x_4 \leq 0.43$ is $\Delta \approx 0.0898$

C. The best split is $x_4 \leq 0.55$

D. The impurity gain of the split $x_4 \leq 0.55$ is $\Delta \approx 0.1589$

E. Don't know.

Solution 8. Recall the information gain Δ is given as:

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

These quantities are easiest computed by forming the matrix R_{ki}, defined as the number of observations in split k belonging to class i. This can in turn be obtained from the information given in the problem for $x_4 \leq 0.43$ as:

$$R = \begin{bmatrix} 143 & 120 \\ 137 & 222 \\ 54 & 304 \end{bmatrix}.$$

We obtain $N(r) = \sum_{ki} R_{ki} = 980$ as the total number of observations and the number of observations in each branch is simply:

$$N(v_k) = \sum_i R_{ki}.$$

Next, the impurities $I(v_k)$ is computed from the probabilities

$$p_i = \frac{R_{ki}}{N(v_k)}$$

and the impurity I_0 from

$$p_i = \frac{\sum_k R_{ki}}{N(r)}.$$

In particular we obtain:

$$I_0 = 0.634, I(v_1) = 0.626,\ I(v_2) = 0.479.$$

Combining these we see that $\Delta = 0.09$ and therefore option B is correct.

Question 9. Consider the splits in Table 3. Suppose we build a classification tree considering only the split $x_4 \leq 0.55$ and evaluate it on the same data it was trained upon. What is the accuracy?

A. The accuracy is: 0.42

B. The accuracy is: 0.685

C. The accuracy is: 0.338

D. The accuracy is: 0.097

E. Don't know.

Solution 9.

We will first form the matrix R_{ki}, defined as the number of observations in split k belonging to class i:

$$R = \begin{bmatrix} 223 & 40 \\ 251 & 108 \\ 197 & 161 \end{bmatrix}.$$

From this we obtain $N = \sum_{ki} R_{ki} = 980$ as the total number of observations. For each split, the number of observations in the largest classes, n_k, is:

$$n_1 = \max_i R_{ik} = 251,\ n_2 = \max_i R_{ik} = 161.$$

Therefore, the accuracy is:

$$\text{Accuracy:} \quad \frac{251 + 161}{980}$$

and answer A is correct.

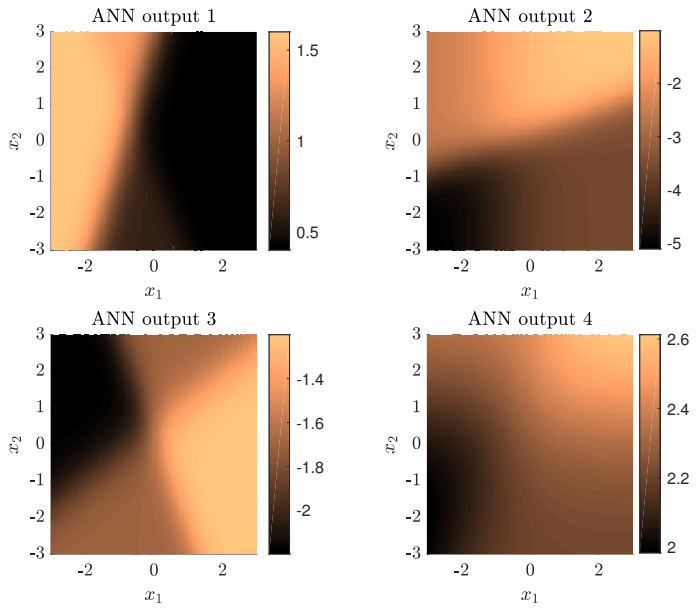

Figure 5: Suggested outputs of an ANN trained on the two attributes x_1 and x_2 from the travel review dataset to predict y.

Question 10. We will consider an artificial neural network (ANN) trained on the travel review dataset described in Table 1 to predict y from the two attributes x_1 and x_2. Suppose the neural network takes the form:

$$f(x, \boldsymbol{w}) = h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x_1\ x_2]\boldsymbol{w}_j^{(1)})\right).$$

where the activation functions are selected as $h^{(1)}(x) = \sigma(x)$ (the sigmoid activation function) and $h^{(2)}(x) = x$ (the linear activation function) and the weights are given as:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -1.2 \\ -1.3 \\ 0.6 \end{bmatrix}, \quad \boldsymbol{w}_2^{(1)} = \begin{bmatrix} -1.0 \\ -0.0 \\ 0.9 \end{bmatrix},$$
$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -0.3 \\ 0.5 \end{bmatrix}, \quad w_0^{(2)} = 2.2.$$

Which one of the curves in Figure 5 will then correspond to the function f?

A. ANN output 1

B. ANN output 2

C. ANN output 3

D. ANN output 4

E. Don't know.

Solution 10.

It suffices to compute the activation of the neural network at $\begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 3 & 3 \end{bmatrix}$. The activation of each of the two hidden neurons is:

$$n_1 = h^{(1)}(\begin{bmatrix} 1 & 3 & 3 \end{bmatrix} \boldsymbol{w}_1^{(1)}) = 0.036$$
$$n_2 = h^{(1)}(\begin{bmatrix} 1 & 3 & 3 \end{bmatrix} \boldsymbol{w}_2^{(1)}) = 0.846.$$

The final output is then computed by a simple linear transformation:

$$f(x, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}(\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} \boldsymbol{w}_j^{(1)})$$
$$= w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} n_j = 2.612.$$

This rules out all options except D.

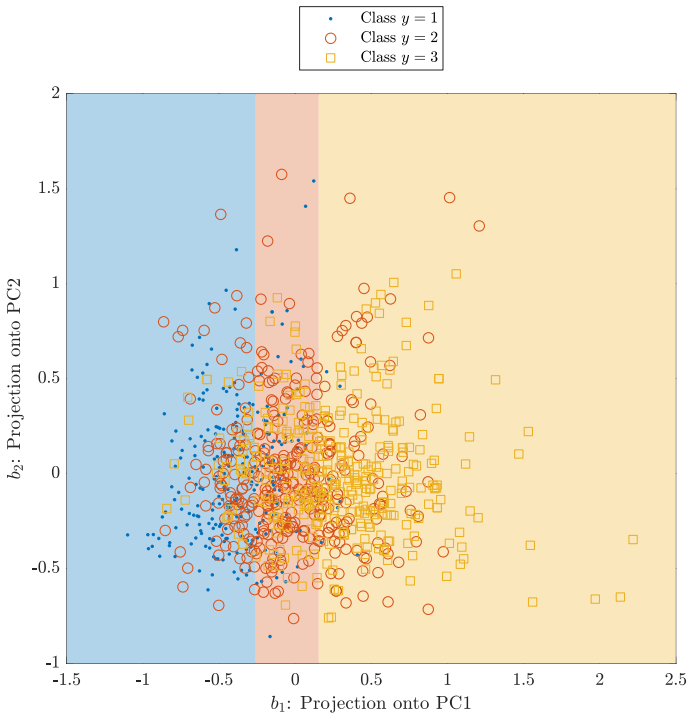

Figure 6: Output of a logistic regression classifier trained on observations from the travel review dataset.

Question 11. Consider again the travel review dataset. We consider a multinomial regression model applied to the dataset projected onto the first two principal directions, giving the two coordinates b_1 and b_2 for each observation. Multinomial regression then computes the per-class probability by first computing the numbers:

$$\hat{y}_1 = \begin{bmatrix} 1 \\ b_1 \\ b_2 \end{bmatrix}^\top \boldsymbol{w}_1, \quad \hat{y}_2 = \begin{bmatrix} 1 \\ b_1 \\ b_2 \end{bmatrix}^\top \boldsymbol{w}_2,$$

and then use the softmax transformation in the form:

$$P(y = k|\boldsymbol{x}) = \begin{cases} \frac{e^{\hat{y}_k}}{1+\sum_{k'=1}^{2} e^{\hat{y}_{k'}}}. & \text{if } k \leq 2 \\ \frac{1}{1+\sum_{k'=1}^{2} e^{\hat{y}_{k'}}} & \text{if } k = 3. \end{cases}$$

Suppose the resulting decision boundary is as shown in Figure 6, what are the weights?

A. $\boldsymbol{w}_1 = \begin{bmatrix} -0.77 \\ -5.54 \\ 0.01 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} 0.26 \\ -2.09 \\ -0.03 \end{bmatrix}$

B. $\boldsymbol{w}_1 = \begin{bmatrix} 0.51 \\ 1.65 \\ 0.01 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} 0.1 \\ 3.8 \\ 0.04 \end{bmatrix}$

C. $\boldsymbol{w}_1 = \begin{bmatrix} -0.9 \\ -4.39 \\ -0.0 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} -0.09 \\ -2.45 \\ -0.04 \end{bmatrix}$

D. $\boldsymbol{w}_1 = \begin{bmatrix} -1.22 \\ -9.88 \\ -0.01 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} -0.28 \\ -2.9 \\ -0.01 \end{bmatrix}$

E. Don't know.

Solution 11. The solution is found by simply observing three of the weights will lead to misclassification. For instance, consider the point

$$\boldsymbol{b} = \begin{bmatrix} -0.0 \\ -1.0 \end{bmatrix}$$

The projections onto the four options are, in order,

- $\begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 \end{bmatrix} = \begin{bmatrix} -0.78 & 0.29 & 0.0 \end{bmatrix}$
- $\begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 0.5 & 0.06 & 0.0 \end{bmatrix}$
- $\begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 \end{bmatrix} = \begin{bmatrix} -0.9 & -0.05 & -0.0 \end{bmatrix}$
- $\begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 \end{bmatrix} = \begin{bmatrix} -1.21 & -0.27 & -0.0 \end{bmatrix}$

Since we select the maximal class, this means the four predicted classes for this point are: 2, 1, 3 and 3 and Inspecting the figure we see that the correct class is $y = 2$, which mean option A is correct.

Question 12. Consider a small dataset comprised of $N = 10$ observations

$$x = \begin{bmatrix} 1.0 & 1.2 & 1.8 & 2.3 & 2.6 & 3.4 & 4.0 & 4.1 & 4.2 & 4.6 \end{bmatrix}.$$

Suppose a k-means algorithm is applied to the dataset with $K = 3$ and using Euclidian distances. The algorithm is initialized with K cluster centers located at

$$\mu_1 = 1.8, \ \mu_2 = 3.3, \ \mu_3 = 3.6$$

What will the location of the cluster centers be after the k-means algorithm has converged?

A. $\mu_1 = 2.05,\ \mu_2 = 4,\ \mu_3 = 4.3$

B. $\mu_1 = 1.58,\ \mu_2 = 3.33,\ \mu_3 = 4.3$

C. $\mu_1 = 1.33,\ \mu_2 = 2.77,\ \mu_3 = 4.22$

D. $\mu_1 = 1.58,\ \mu_2 = 3.53,\ \mu_3 = 4.4$

E. Don't know.

Solution 12. Recall the K-means algorithm iterates between assigning the observations to their nearest centroids, and then updating the centroids to be equal to the average of the observations assigned to them. Given the initial centroids, the K-means algorithm assign observations to the nearest centroid resulting in the partition:

$$\{1, 1.2, 1.8, 2.3\},\ \{2.6, 3.4\},\ \{4, 4.1, 4.2, 4.6\}.$$

Therefore, the subsequent steps in the K-means algorithm are:

Step $t = 1$: The centroids are computed to be:

$$\mu_1 = 1.575,\ \mu_2 = 3,\ \mu_3 = 4.225.$$

And the updated assignment of observations to nearest centroids results in the clustering:

$$\{1, 1.2, 1.8\},\ \{2.3, 2.6, 3.4\},\ \{4, 4.1, 4.2, 4.6\}.$$

Step $t = 2$: The centroids are computed to be:

$$\mu_1 = 1.33333,\ \mu_2 = 2.76667,\ \mu_3 = 4.225.$$

And the updated assignment of observations to nearest centroids results in the clustering:

$$\{1, 1.2, 1.8\},\ \{2.3, 2.6, 3.4\},\ \{4, 4.1, 4.2, 4.6\}.$$

At this point, the centroids are no longer changing and the algorithm terminates. Hence, C is correct.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
o_1	0	0	0	1	0	0	0	0	0
o_2	0	0	0	0	0	0	0	0	1
o_3	0	1	1	1	1	1	0	0	0
o_4	1	0	0	0	0	0	0	0	0
o_5	1	0	0	1	0	0	0	0	0
o_6	0	0	1	1	0	0	0	1	0
o_7	0	0	1	1	1	0	0	0	0
o_8	0	0	0	0	1	0	0	0	0
o_9	0	1	1	0	1	0	0	0	0
o_{10}	0	0	1	1	1	0	1	0	0

Table 4: Binarized version of the travel review dataset. Each of the features f_i are obtained by taking a feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belongs to class C_1 (corresponding to a poor rating), the red observations $\{o_3,\ o_4,\ o_5\}$ belongs to class C_2 (corresponding to an average rating), and the blue observations $\{o_6,\ o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to a high rating).

Question 13. We again consider the travel review dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce 9 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 $N \times M = 10 \times 9$ binary matrix in Table 4. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_2,\ f_4,\ f_5$. If for an observations we observe

$$f_2 = 0,\ f_4 = 1,\ f_5 = 0$$

what is then the probability it has average rating ($y = 2$) according to the Naïve-Bayes classifier?

A. $p_{\mathbf{NB}}(y = 2 | f_2 = 0,\ f_4 = 1,\ f_5 = 0) = \frac{200}{533}$

B. $p_{\mathrm{NB}}(y = 2 | f_2 = 0,\ f_4 = 1,\ f_5 = 0) = \frac{25}{79}$

C. $p_{\mathrm{NB}}(y = 2 | f_2 = 0,\ f_4 = 1,\ f_5 = 0) = \frac{2000}{6023}$

D. $p_{\mathrm{NB}}(y = 2 | f_2 = 0,\ f_4 = 1,\ f_5 = 0) = \frac{125}{287}$

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

Solution 13. To solve this problem, we simply use the general form of the naïve-Bayes approximation and plug in the relevant numbers. We get:

$$p_{\text{NB}}(y=2|f_2=0,\ f_4=1,\ f_5=0) = \frac{p(f_2=0|y=2)p(f_4=1|y=2)p(f_5=0|y=2)p(y=2)}{\sum_{j=1}^{3} p(f_2=0|y=j)p(f_4=1|y=j)p(f_5=0|y=j)p(y=j)}$$
$$= \frac{\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{3}{10}}{\frac{1}{1}\frac{1}{2}\frac{1}{1}\frac{1}{5} + \frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{3}{10} + \frac{4}{5}\frac{3}{5}\frac{1}{5}\frac{1}{2}}$$
$$= \frac{200}{533}.$$

Therefore, answer A is correct.

Question 14. Consider the binarized version of the travel review dataset shown in Table 4.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 9$ items $f_1, f_2, \ldots, f_9$. Which of the following options represents all (non-empty) itemsets with support greater than 0.15 (and only itemsets with support greater than 0.15)?

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

B. $\{f_3\}$, $\{f_4\}$, $\{f_5\}$, $\{f_3, f_4\}$, $\{f_3, f_5\}$

C. $\{f_3\}$, $\{f_4\}$, $\{f_5\}$, $\{f_3, f_4\}$, $\{f_3, f_5\}$, $\{f_4, f_5\}$, $\{f_3, f_4, f_5\}$

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

E. Don't know.

Solution 14. 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 0.15, an itemset needs to be contained in 2 rows. It is easy to see this rules out all options except *A*.

Question 15. We again consider the binary matrix from Table 4 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 9$ items $f_1, \ldots, f_9$.

What is the *confidence* of the rule $\{f_2\} \rightarrow \{f_3, f_4, f_5, f_6\}$?

A. The confidence is $\frac{3}{20}$

B. The confidence is $\frac{1}{2}$

C. The confidence is 1

D. The confidence is $\frac{1}{10}$

E. Don't know.

Solution 15. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_2\} \cup \{f_3, f_4, f_5, f_6\})}{\text{support}(\{f_2\})} = \frac{\frac{1}{10}}{\frac{1}{5}} = \frac{1}{2}.$$

Therefore, answer B is correct.

Question 16. We will again consider the binarized version of the travel review dataset already encountered in Table 4, however, we will now only consider the first $M = 4$ features f_1, f_2, f_3, f_4. We wish to apply the a-priori algorithm (the specific variant encountered in chapter 19 of the lecture notes) to find all itemsets with support greater than $\varepsilon = 0.35$. Suppose at iteration $k = 2$ we know that:

$$L_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

What, in the notation of the lecture notes, is C_2?

A. $C_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

B. $C_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}$

C. $C_2 = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix}$

D. $C_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$

E. Don't know.

Solution 16. To compute C_2, we need to run the a-priori algorithm for 2 steps. We will therefore simply list the intermediate values which are computed entirely similar to those in the example in the lecture notes.

$t = 1$: Initially, let L_1 be all singleton itemsets with a support of at least $\varepsilon = 0.35$.

$$L_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$t = 2$: Define C'_2 by forming all itemsets that can be obtained by taking an element in L_1 and adding a single item not already contained within it:

$$C'_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Then, for each itemset in C'_2, check that all subsets of size $k - 1$ are in L_1. If so, keep them as C_2:

$$C_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}$$

Finally, for each itemset in the C_2, check it has support of at least $\varepsilon = 0.35$ and if so keep them as L_2:

$$L_2 = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}$$

Therefore, answer B is correct.

Question 17. Consider the observations in Table 4. We consider these as 9-dimensional binary vectors and wish to compute the pairwise similarity. Which of the following statements are true?

A. $\text{Cos}(o_1, o_3) \approx 0.132$

B. $\mathbf{J}(o_2, o_3) \approx 0.0$

C. $\text{SMC}(o_1, o_3) \approx 0.268$

D. $\text{SMC}(o_2, o_4) \approx 0.701$

E. Don't know.

Solution 17. The problem is solved by simply using the definition of SMC, Jaccard similarity and cosine similarity as found in the lecture notes. The true values are:

$$\begin{aligned} \text{J}(o_2, o_3) &\approx 0.0 \\ \text{SMC}(o_1, o_3) &\approx 0.556 \\ \text{Cos}(o_1, o_3) &\approx 0.447 \\ \text{SMC}(o_2, o_4) &\approx 0.778 \end{aligned}$$

and therefore option B is correct.

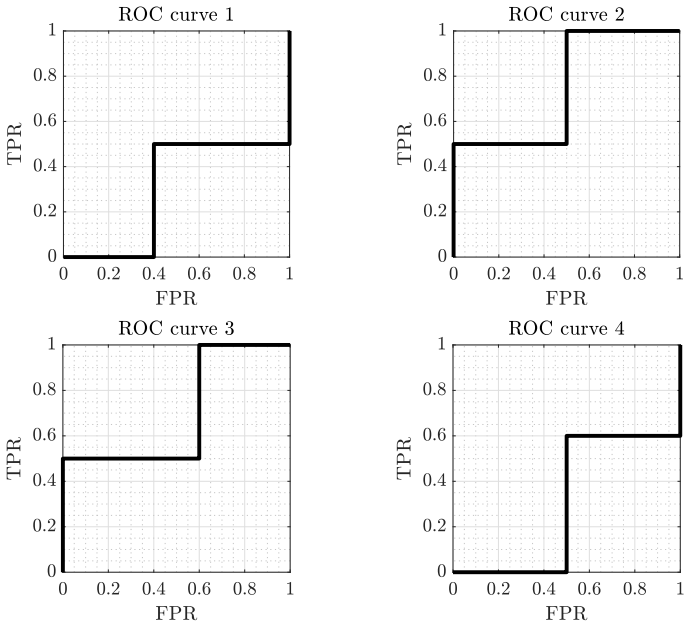

Figure 7: Proposed ROC curves for the neural network classifier with predictions/true class labels given in Table 5

y	1	1	0	1	1	1	0
$\hat{y}$	0.14	0.15	0.27	0.61	0.71	0.75	0.81

Table 5: Small binary classification dataset of $N = 7$ observations along with the predicted class probability $\hat{y}$.

Question 18. A neural network classifier is trained to distinguish between two classes $y \in \{0, 1\}$ in a small dataset consisting of $N = 7$ observations. Suppose the true class label y and predicted probability an observation belongs to class 1, $\hat{y}$, is as given in Table 5.

To evaluate the classifier, we will use the *area under curve* (AUC) of the *reciever operator characteristic* (ROC) curve. In Figure 7 is given four proposed ROC curves, which one of the curves corresponds to the classifier?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

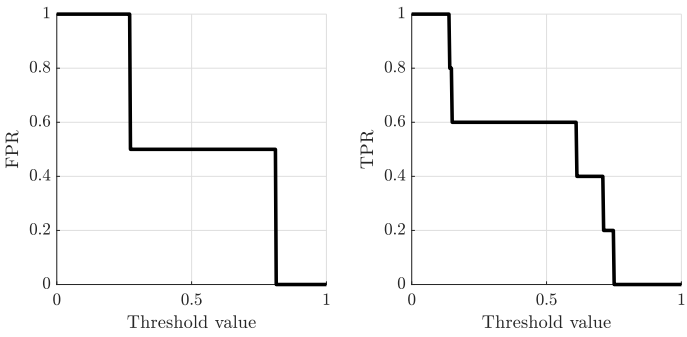

Figure 8: TPR, FPR curves for the classifier.

Solution 18. To compute the AUC, we need to compute 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.

This procedure is then repeated for different threshold values to obtain the curves shown in Figure 8. The ROC curve is then obtained by plotting these two curves against each other. I.e. for each threshold value, the point

$$(x, y) = (\text{FPR}, \text{TPR})$$

is on the AUC curve. This rules out all options except D.

Question 19. Consider again the travel review dataset in Table 1. We would like to predict a resort's rating using a linear regression, 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 $x_1,\ x_6,\ x_7,\ x_8,\ x_9$ and in Table 6 we have pre-computed the estimated training and test error for different variable combinations of the dataset. Which of the following statements is correct?

A. Forward selection will select attributes x_6

B. Forward selection will select attributes $x_1,\ x_6,\ x_7,\ x_8$

C. Backward selection will select attributes $x_1,\ x_6$

D. Forward selection will select attributes $x_1,\ x_6$

E. Don't know.

Solution 19.

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

Forward selection: The method is initialized with the set $\{\}$ having an error of 5.528.

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 $\{x_1\}$, $\{x_6\}$, $\{x_7\}$, $\{x_8\}$, $\{x_9\}$. Since the lowest error of the available sets is 4.57, which is lower than 5.528, we update the current selected variables to $\{x_6\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_6\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_6\}$, $\{x_1, x_7\}$, $\{x_6, x_7\}$, $\{x_1, x_8\}$, $\{x_6, x_8\}$, $\{x_7, x_8\}$, $\{x_1, x_9\}$, $\{x_6, x_9\}$, $\{x_7, x_9\}$, $\{x_8, x_9\}$. Since the lowest error of the available sets is 4.213, which is lower than 4.57, we update the current selected variables to $\{x_1, x_6\}$

Step $i = 3$ The available variable sets to choose between is obtained by taking the current variable

Feature(s)	Training RMSE	Test RMSE
none	5.25	5.528
x_1	4.794	5.566
x_6	4.563	4.57
x_7	5.246	5.52
x_8	5.245	5.475
x_9	4.683	5.185
$x_1,\ x_6$	3.344	4.213
$x_1,\ x_7$	4.794	5.565
$x_6,\ x_7$	4.561	4.591
$x_1,\ x_8$	4.742	5.481
$x_6,\ x_8$	4.559	4.614
$x_7,\ x_8$	5.242	5.473
$x_1,\ x_9$	3.945	4.967
$x_6,\ x_9$	4.552	4.643
$x_7,\ x_9$	4.679	5.223
$x_8,\ x_9$	4.674	5.284
$x_1,\ x_6,\ x_7$	3.338	4.165
$x_1,\ x_6,\ x_8$	3.325	4.161
$x_1,\ x_7,\ x_8$	4.741	5.494
$x_6,\ x_7,\ x_8$	4.557	4.648
$x_1,\ x_6,\ x_9$	3.314	4.258
$x_1,\ x_7,\ x_9$	3.945	4.958
$x_6,\ x_7,\ x_9$	4.55	4.67
$x_1,\ x_8,\ x_9$	3.942	4.93
$x_6,\ x_8,\ x_9$	4.546	4.717
$x_7,\ x_8,\ x_9$	4.667	5.354
$x_1,\ x_6,\ x_7,\ x_8$	3.315	4.098
$x_1,\ x_6,\ x_7,\ x_9$	3.307	4.218
$x_1,\ x_6,\ x_8,\ x_9$	3.282	4.234
$x_1,\ x_7,\ x_8,\ x_9$	3.942	4.911
$x_6,\ x_7,\ x_8,\ x_9$	4.542	4.767
$x_1,\ x_6,\ x_7,\ x_8,\ x_9$	3.266	4.195

Table 6: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y in the travel review dataset using different combinations of the features $x_1,\ x_6,\ x_7,\ x_8,\ x_9$.

set $\{x_1, x_6\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_6, x_7\}$, $\{x_1, x_6, x_8\}$, $\{x_1, x_7, x_8\}$, $\{x_6, x_7, x_8\}$, $\{x_1, x_6, x_9\}$, $\{x_1, x_7, x_9\}$, $\{x_6, x_7, x_9\}$, $\{x_1, x_8, x_9\}$, $\{x_6, x_8, x_9\}$, $\{x_7, x_8, x_9\}$. Since the lowest error of the available sets is 4.161, which is lower than 4.213, we update the current selected variables to $\{x_1, x_6, x_8\}$

Step $i = 4$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_6, x_8\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_6, x_7, x_8\}$, $\{x_1, x_6, x_7, x_9\}$, $\{x_1, x_6, x_8, x_9\}$, $\{x_1, x_7, x_8, x_9\}$, $\{x_6, x_7, x_8, x_9\}$. Since the lowest error of the available sets is 4.098, which is lower than 4.161, we update the current selected variables to $\{x_1, x_6, x_7, x_8\}$

Step $i = 5$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_6, x_7, x_8\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_6, x_7, x_8, x_9\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Backward selection: The method is initialized with the set $\{x_1, x_6, x_7, x_8, x_9\}$ having an error of 4.195.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_6, x_7, x_8, x_9\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_6, x_7, x_8\}$, $\{x_1, x_6, x_7, x_9\}$, $\{x_1, x_6, x_8, x_9\}$, $\{x_1, x_7, x_8, x_9\}$, $\{x_6, x_7, x_8, x_9\}$. Since the lowest error of the available sets is 4.098, which is lower than 4.195, we update the current selected variables to $\{x_1, x_6, x_7, x_8\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_6, x_7, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_6, x_7\}$, $\{x_1, x_6, x_8\}$, $\{x_1, x_7, x_8\}$, $\{x_6, x_7, x_8\}$, $\{x_1, x_6, x_9\}$, $\{x_1, x_7, x_9\}$, $\{x_6, x_7, x_9\}$, $\{x_1, x_8, x_9\}$, $\{x_6, x_8, x_9\}$, $\{x_7, x_8, x_9\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Question 20. Consider the travel review dataset from Table 1. We wish to predict the resort's rating based on the attributes *dance clubs* and *juice bars* using a Bayes classifier.

$p(\hat{x}_2, \hat{x}_3 \mid y)$	$y = 1$	$y = 2$	$y = 3$
$\hat{x}_2 = 0,\ \hat{x}_3 = 0$	0.41	0.28	0.15
$\hat{x}_2 = 0,\ \hat{x}_3 = 1$	0.17	0.28	0.33
$\hat{x}_2 = 1,\ \hat{x}_3 = 0$	0.33	0.25	0.15
$\hat{x}_2 = 1,\ \hat{x}_3 = 1$	0.09	0.19	0.37

Table 7: Probability of observing particular values of $\hat{x}_2$ and $\hat{x}_3$ conditional on y.

Therefore, suppose the attributes have been binarized such that $\hat{x}_2 = 0$ corresponds to $x_2 \leq 1.28$ (and otherwise $\hat{x}_2 = 1$) and $\hat{x}_3 = 0$ corresponds to $x_3 \leq 0.82$ (and otherwise $\hat{x}_3 = 1$). Suppose the probability for each of the configurations of $\hat{x}_2$ and $\hat{x}_3$ conditional on the resort's rating y are as given in Table 7. and the prior probability of the resort's ratings are

$$p(y = 1) = 0.268,\ p(y = 2) = 0.366,\ p(y = 3) = 0.365.$$

Using this, what is then the probability an observation had poor rating given that $\hat{x}_2 = 0$ and $\hat{x}_3 = 1$?

A. $p(y = 1 \mid \hat{x}_2 = 0, \hat{x}_3 = 1) = 0.17$

B. $p(y = 1 \mid \hat{x}_2 = 0, \hat{x}_3 = 1) = 0.411$

C. $p(y = 1 \mid \hat{x}_2 = 0, \hat{x}_3 = 1) = 0.218$

D. $p(y = 1 \mid \hat{x}_2 = 0, \hat{x}_3 = 1) = 0.046$

E. Don't know.

Solution 20. The problem is solved by a simple application of Bayes' theorem:

$$p(y = 1 \mid \tilde{x}_2 = 0, \tilde{x}_3 = 1) = \frac{p(\tilde{x}_2 = 0, \tilde{x}_3 = 1 \mid y = 1)p(y = 1)}{\sum_{k=1}^{3} p(\tilde{x}_2 = 0, \tilde{x}_3 = 1 \mid y = k)p(y = k)}$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_2 = 0, \tilde{x}_3 = 1 \mid y)$ in Table 7. Inserting the values we see option A is correct.

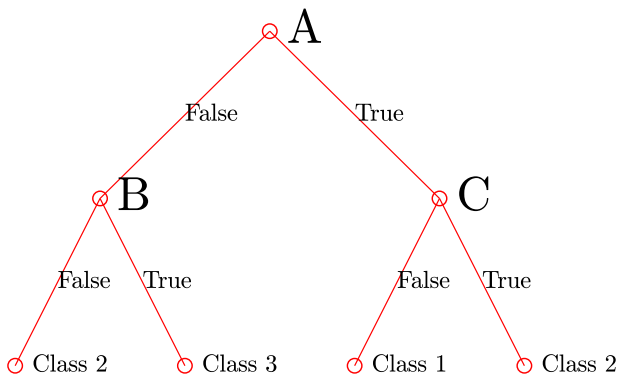

Figure 9: Example classification tree.

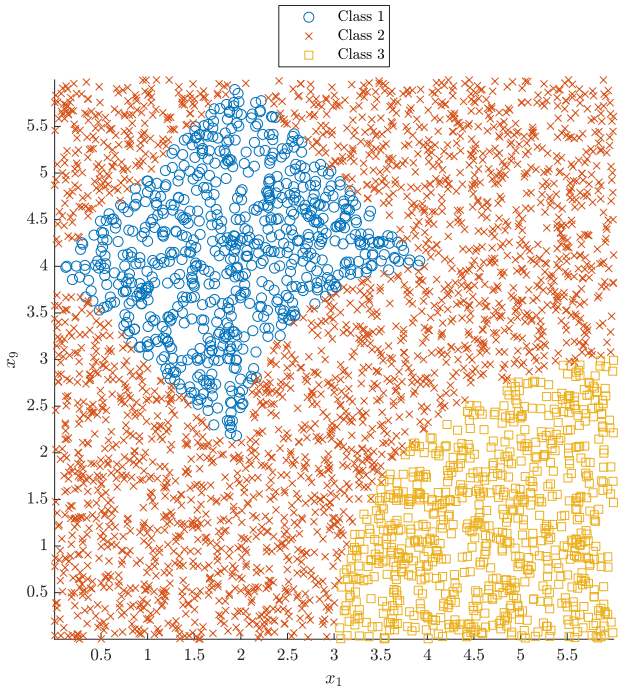

Figure 10: classification boundary.

Question 21. We consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 9 resulting in a partition into classes indicated by the colors/markers in Figure 10. What is the correct rule assignment to the nodes in the decision tree?

A. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 2$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 0 \end{bmatrix} \right\|_2 < 3$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_2 < 2$

B. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 2$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_2 < 2$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 0 \end{bmatrix} \right\|_2 < 3$

C. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_2 < 2$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 0 \end{bmatrix} \right\|_2 < 3$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 2$

D. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_2 < 2$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 2$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 0 \end{bmatrix} \right\|_2 < 3$

E. Don't know.

Solution 21.

This problem is solved by using the definition of a decision tree and observing what classification rule each of the assignment of features to node names in the decision tree will result in. I.e. beginning at the top of the tree, check if the condition assigned to the node is met and proceed along the true or false leg of the tree.

The resulting decision boundaries for each of the options are shown in Figure 11 and it follows answer A is correct.

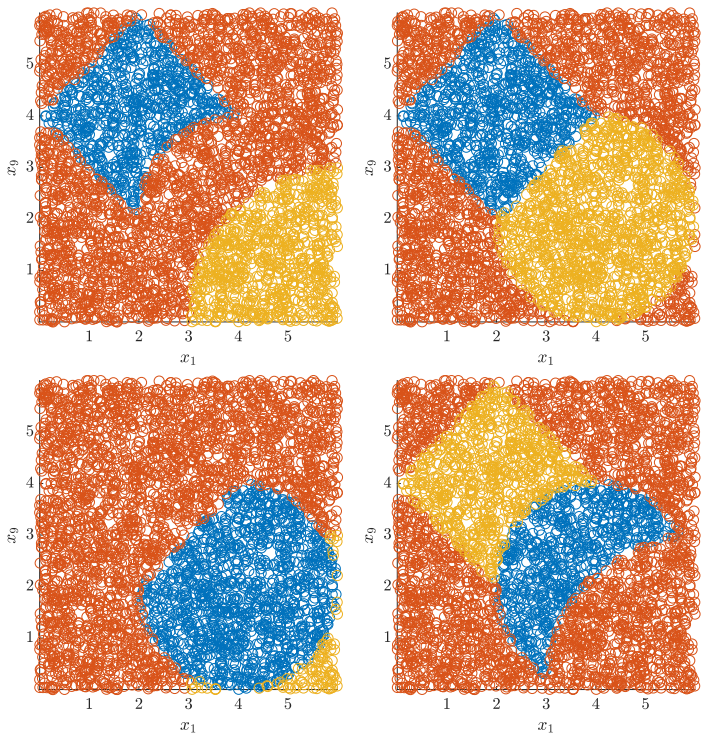

Figure 11: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

Question 22. Suppose we wish to compare a neural network model and a regularized logistic regression model on the travel review dataset. For the neural network, we wish to find the optimal number of hidden neurons n_h, and for the regression model the optimal value of λ. We therefore opt for a two-level cross-validation approach where for each outer fold, we train the model on the training split, and use the test split to find the optimal number of hidden units (or regularization strength) using cross-validation with $K_2 = 5$ folds. The tested values are:

$$\lambda : \{0.01, 0.1, 0.5, 1, 10\}$$
$$n_h : \{1, 2, 3, 4, 5\}.$$

Then, given this optimal number of hidden units n_h^* or regularization strength λ^*, the model is trained and evaluated on the current outer test split. This produces Table 8 which shows the optimal number of hidden units/lambda as well as the (outer) test classification errors E_1^{test} (neural network model) and E_2^{test} (logistic regression model). Note these errors are averaged over the number of observations in the the (outer) test splits.

	ANN		Log.reg.	
	n_h^*	E_1^{test}	λ^*	E_2^{test}
Outer fold 1	1	0.561	0.1	0.439
Outer fold 2	1	0.513	0.1	0.487
Outer fold 3	1	0.564	0.1	0.436
Outer fold 4	1	0.671	0.1	0.329

Table 8: Result of applying two-level cross-validation to a neural network model and a logistic regression model. The table contains the optimally selected parameters from each outer fold (n_h^*, hidden units and λ^*, regularization strength) and the corresponding test errors E_1^{test} and E_2^{test} when the models are evaluated on the current outer split.

How many models were *trained* to compose the table?

A. 208 models

B. 100 models

C. 200 models

D. 104 models

E. Don't know.

Solution 22. Going over the 2-level cross-validation algorithm we see the total number of models to be *trained* is:

$$K_1(K_2 S + 1) = 104$$

Since we have to do this for each model, and $S = 5$ in both cases, we need to train twice this number of models and therefore A is correct.

Question 23. We fit a GMM to a single feature x_6 from the travel review dataset. Recall the density of a 1D GMM is

$$p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \sigma_k^2)$$

and suppose that the identified values of the mixture weights are

$$w_1 = 0.19,\ w_2 = 0.34,\ w_3 = 0.48$$

and the parameters of the multivariate normal densities:

$$\mu_1 = 3.177,\ \mu_2 = 3.181,\ \mu_3 = 3.184$$
$$\sigma_1 = 0.0062,\ \sigma_2 = 0.0076,\ \sigma_3 = 0.0075.$$

According to the GMM, what is the probability an observation at $x_0 = 3.19$ is assigned to cluster $k = 2$?

A. 0.49

B. 0.31

C. 0.08

D. 0.68

E. Don't know.

Solution 23.

Recall γ_{ik} is the posterior probability that observation i is assigned to mixture component 2 which can easily be obtained using Bayes' theorem. We see that:

$$\gamma_{i,2} = \frac{p(x_i|z_{i,2} = 1)\pi_2}{\sum_{k=1}^{3} p(x_i|z_{ik} = 1)\pi_k}.$$

To use Bayes' theorem, we need to compute the probabilities using the normal density. These are:

$$p(x_i|z_{i1} = 1) = 7.142$$
$$p(x_i|z_{i2} = 1) = 26.036$$
$$p(x_i|z_{i3} = 1) = 38.626$$

Combining these with the class-assignment probabilities we obtain:

$$\gamma_{i,2} = 0.308$$

and conclude the solution is B.

Variable	y^{true}	$t = 1$
y_1	1	1
y_2	2	1
y_3	2	1
y_4	1	2
y_5	1	1
y_6	1	2
y_7	2	1

Table 9: For each of the $N = 7$ observations (first column), the table indicate the true class labels y^{true} (second column) and the predicted outputs of the AdaBoost classifier (third column) which is also shown in Figure 12.

Question 24. Consider again the travel review dataset of Table 1. Suppose we limit ourselves to $N = 7$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_4 and x_6. We use a KNN classification model ($K = 1$) to this dataset and apply AdaBoost to improve the performance. After the first $T = 1$ round of boosting, we obtain the decision boundaries shown in Figure 12 (the predictions of the $T = 1$ weaker classifiers and the true class labels is also tabulated in Table 9).

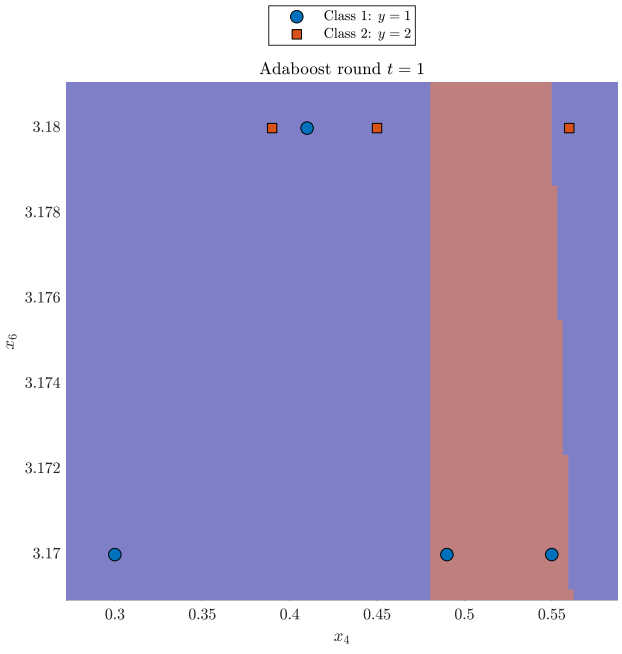

Figure 12: Decision boundaries for a KNN classifier for the first $T = 1$ rounds of boosting.

Given this information, how will the AdaBoost update the weights $\boldsymbol{w}$?

A. $[0.25 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.25 \quad 0.1 \quad 0.1]$

B. $[0.388 \quad 0.045 \quad 0.045 \quad 0.045 \quad 0.388 \quad 0.045 \quad 0.045]$

C. $[0.126 \quad 0.15 \quad 0.15 \quad 0.15 \quad 0.126 \quad 0.15 \quad 0.15]$

D. $[0.066 \quad 0.173 \quad 0.173 \quad 0.173 \quad 0.066 \quad 0.173 \quad 0.173]$

E. Don't know.

Solution 24.
We first observe the AdaBoost classifier at $t = 1$ mis-classify observations:

$$\{y_2, y_3, y_4, y_6, y_7\}$$

Since the weights are just $w_i = \frac{1}{N}$, we therefore get:

$$\epsilon_{t=1} = \sum_i w_i(t)(1 - \delta_{f_t(x_i),y_i}) = 0.714$$

From this, we compute α_t as

$$\alpha_t = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t} = -0.458$$

Scaling the observations corresponding to the mis-classified weights as $w_i e^{\alpha_t}$ and those corresponding to the correctly classified weights as $w_i e^{-\alpha_t}$ and normalizing the new weights to sum to one then give answer A.

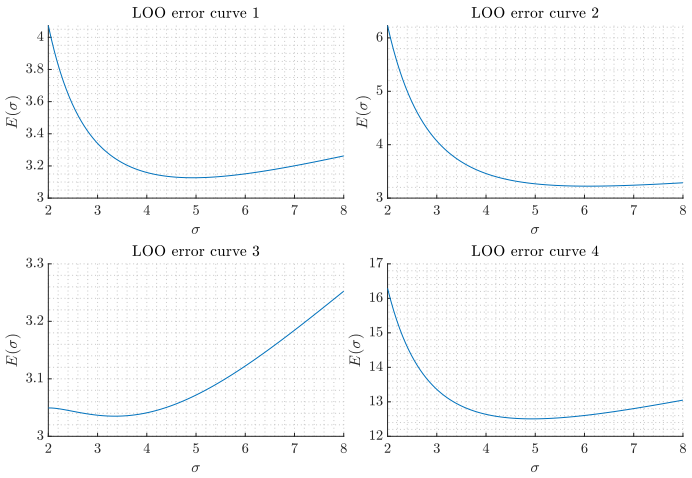

Figure 13: Estimated negative log-likelihood as obtained using LOO cross-validation on a small, $N = 4$ one-dimensional dataset as a function of kernel width σ.

Question 25. Consider the following $N = 4$ observations from a one-dimensional dataset:

$$\{3.918, -6.35, -2.677, -3.003\}.$$

Suppose we apply a Kernel Density Estimator (KDE) to the dataset with kernel width σ (i.e., σ is the standard deviation of the Gaussian kernels), and we wish to find σ by using leave-one-out (LOO) cross-validation using the average (per observation) negative log-likelihood

$$E(\sigma) = \frac{-1}{4} \sum_{i=1}^{4} \log p_\sigma(x_i).$$

Which of the curves in Figure 13 shows the LOO estimate of the generalization error $E(\sigma)$?

A. LOO curve 1

B. LOO curve 2

C. LOO curve 3

D. LOO curve 4

E. Don't know.

Solution 25. To solve the problem, we will compute the LOO cross-validation estimate of the generalization error at $\sigma = 2$. To do so, recall the density at each

observation i, when the KDE is fitted on the other $N-1$ observations, is:

$$p_\sigma(x_i) = \frac{1}{N-1}\sum_{j\neq i}\mathcal{N}(x_i|x_j,\sigma=2)$$

These values are approximately:

$p_\sigma(x_1) = 0,\ p_\sigma(x_2) = 0.029,\ p_\sigma(x_3) = 0.078,\ p_\sigma(x_4) = 0.($

The LOO error is then:

$$E(\sigma=2) = \frac{1}{N}\sum_{i=1}^{N} -\log p_\sigma(x_i) = 4.073$$

Therefore, the correct answer is A.

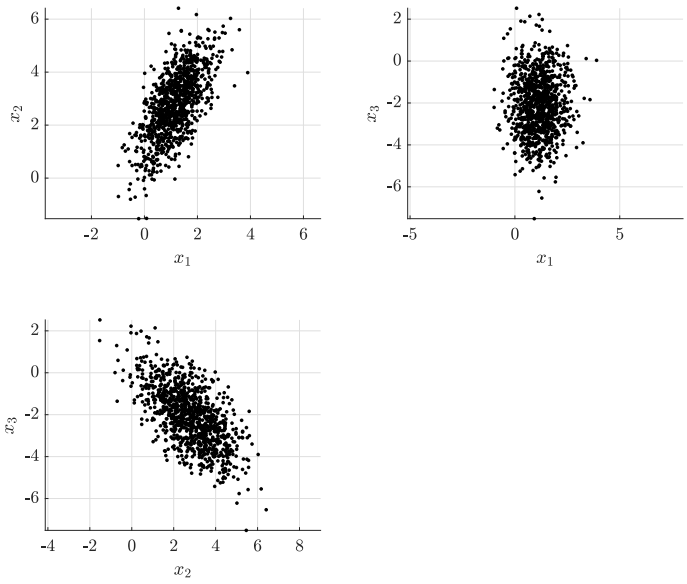

Figure 14: Scatter plots of all pairs of attributes of a vector $\boldsymbol{x}$ when $\boldsymbol{x}$ is a random vector distributed as a multivariate normal distribution of 3 dimensions.

Question 26. Consider a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ and mean μ and suppose we generate 1000 random samples from it:

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Plots of each pair of coordinates of the draws $\boldsymbol{x}$ is shown in Figure 14. One of the following covariance matrices was used to generate the data:

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.5 & 0.56 & 0.0 \\ 0.56 & 1.5 & -1.12 \\ 0.0 & -1.12 & 2.0 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.0 & -1.12 & 0.0 \\ -1.12 & 1.5 & 0.56 \\ 0.0 & 0.56 & 0.5 \end{bmatrix}$$

What is the *correlation* between variables x_1 and x_2?

A. The correlation between x_1 and x_2 is 0.647

B. The correlation between x_1 and x_2 is -0.611

C. The correlation between x_1 and x_2 is 0.747

D. The correlation between x_1 and x_2 is 0.56

E. Don't know.

Solution 26. To solve this problem, recall that the correlation between coordinates x_i, x_j of an observation drawn from a multivariate normal distribution is

positive if $\Sigma_{ij} > 0$, negative if $\Sigma_{ij} < 0$ and zero if $\Sigma_{ij} \approx 0$. Furthermore, recall positive correlation in a scatter plot means the points (x_i, x_j) tend to lie on a line sloping upwards, negative correlation means it is sloping downwards and zero means the data is axis-aligned.

We can therefore use the scatter plots of variables x_i, x_j to read of the sign off Σ_{ij} (or whether it is zero). We thereby find that $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ generated the data. We can now read off the *covariance* as $\text{Cov}[x_1, x_2] = \mathbf{\Sigma}_{1,2}$ and the variance of each variable as

$$\text{Var}[x_1] = \mathbf{\Sigma}_{1,1}, \quad \text{Var}[x_2] = \mathbf{\Sigma}_{2,2}.$$

The correlation is then given as:

$$\text{Corr}[x_1, x_2] = \frac{\text{Cov}[x_1, x_2]}{\sqrt{\text{Var}[x_1]\text{Var}[x_2]}} = 0.647$$

and therefore answer A is correct.

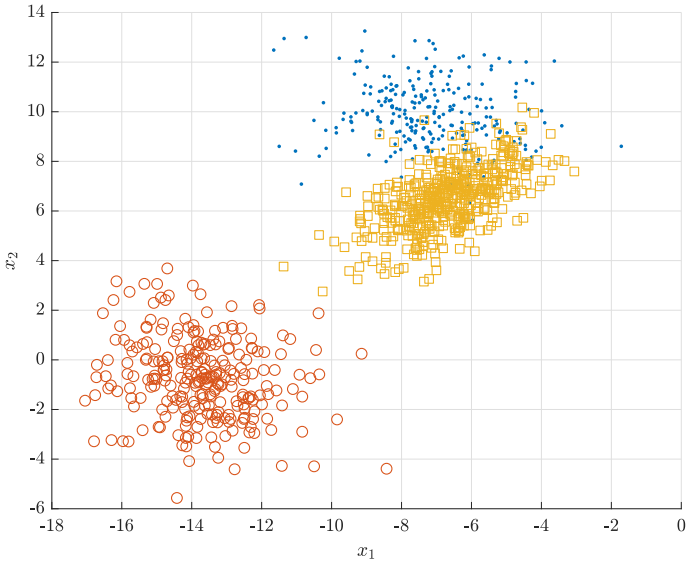

Figure 15: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

Question 27. Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. In Figure 15 is given 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Each observation is colored and marked in terms of which cluster it came from in the Gaussian Mixture.
Which one of the following GMM densities was used to

generate the data?

A.

$$p(\boldsymbol{x}) = \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-7.2\\10.0\end{bmatrix},\begin{bmatrix}2.4 & -0.4\\-0.4 & 1.7\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-13.8\\-0.8\end{bmatrix},\begin{bmatrix}1.7 & -0.3\\-0.3 & 2.3\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-6.8\\6.4\end{bmatrix},\begin{bmatrix}1.6 & 0.9\\0.9 & 1.5\end{bmatrix}\right)$$

B.

$$p(\boldsymbol{x}) = \frac{1}{2}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-7.2\\10.0\end{bmatrix},\begin{bmatrix}1.6 & 0.9\\0.9 & 1.5\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-13.8\\-0.8\end{bmatrix},\begin{bmatrix}1.7 & -0.3\\-0.3 & 2.3\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-6.8\\6.4\end{bmatrix},\begin{bmatrix}2.4 & -0.4\\-0.4 & 1.7\end{bmatrix}\right)$$

C.

$$p(\boldsymbol{x}) = \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-7.2\\10.0\end{bmatrix},\begin{bmatrix}1.6 & 0.9\\0.9 & 1.5\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-13.8\\-0.8\end{bmatrix},\begin{bmatrix}2.4 & -0.4\\-0.4 & 1.7\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-6.8\\6.4\end{bmatrix},\begin{bmatrix}1.7 & -0.3\\-0.3 & 2.3\end{bmatrix}\right)$$

D.

$$p(\boldsymbol{x}) = \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-7.2\\10.0\end{bmatrix},\begin{bmatrix}2.4 & -0.4\\-0.4 & 1.7\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-13.8\\-0.8\end{bmatrix},\begin{bmatrix}1.6 & 0.9\\0.9 & 1.5\end{bmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\boldsymbol{x}\middle|\begin{bmatrix}-6.8\\6.4\end{bmatrix},\begin{bmatrix}1.7 & -0.3\\-0.3 & 2.3\end{bmatrix}\right)$$

E. Don't know.

Solution 27.

The three components in the candidate GMM densities can be matched to the colored observations by their mean values. Then, by considering the basic properties of the covariance matrices, we can easily rule out all options except A. Alternatively, in Figure 16 is shown the densities for densities corresponding to option B (upper left), C (upper right) and D (bottom center).

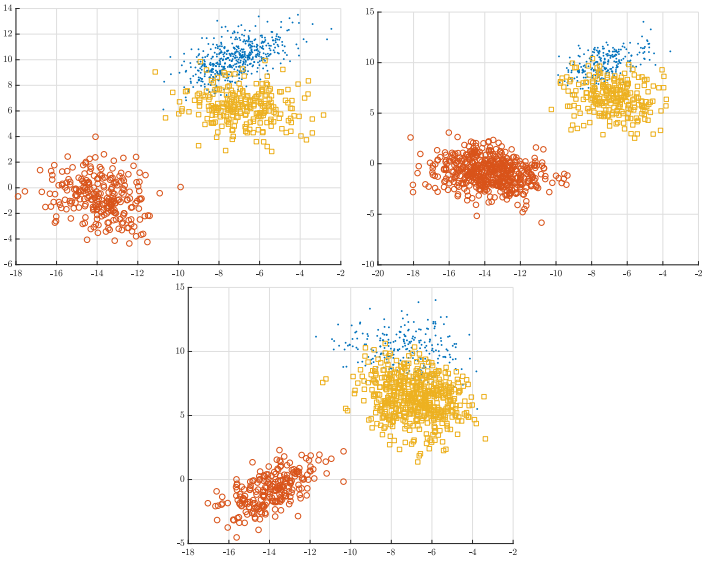

Figure 16: GMM mixtures corresponding to alternative options.

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.

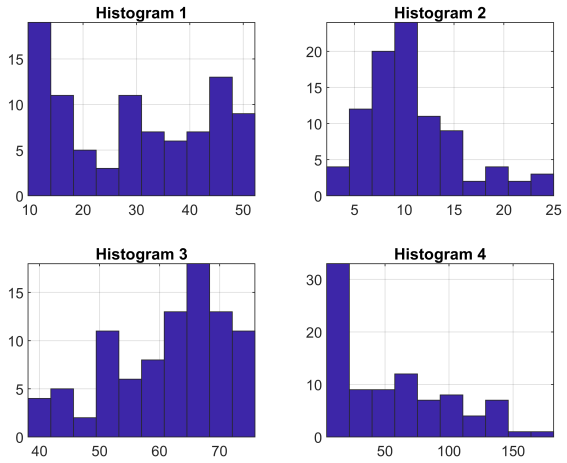

Figure 1: Four histograms corresponding to the variables with summary statistics given in Table 2 but not necessarily in that order.

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.

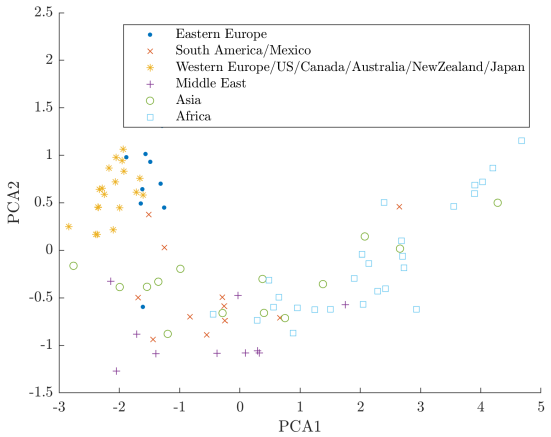

Figure 2: The Poverty data projected onto the first two principal component directions with each observation labelled according to the region it belongs to (given by x_6).

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 $\boldsymbol{USV}^T = \tilde{\boldsymbol{X}}$

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

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

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 $\tilde{\boldsymbol{X}}\boldsymbol{v}_2$

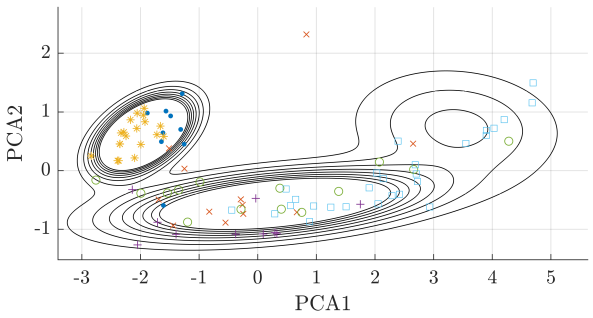

Figure 3: A GMM with K=3 clusters fitted to the poverty data projected onto the first two principal component directions. Each observation is again labelled according to the region it belongs to (given by x_6).

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 $\frac{1}{N-1}(\tilde{\boldsymbol{X}}\boldsymbol{v}_1)^T(\tilde{\boldsymbol{X}}\boldsymbol{v}_2) = (\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\boldsymbol{v}_1)^T(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\boldsymbol{v}_2) = \boldsymbol{u}_1^\top s_{11}\boldsymbol{u}_2 s_{22} = s_{11}s_{22}\boldsymbol{u}_1^\top\boldsymbol{u}_2 = 0$ and there can therefore be no correlation between the observations projected onto PC1 and PC2.

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.

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

B.

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

C.

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

D.

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

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 $1.053/(0.5 \cdot (0.909 + 0.870))$.

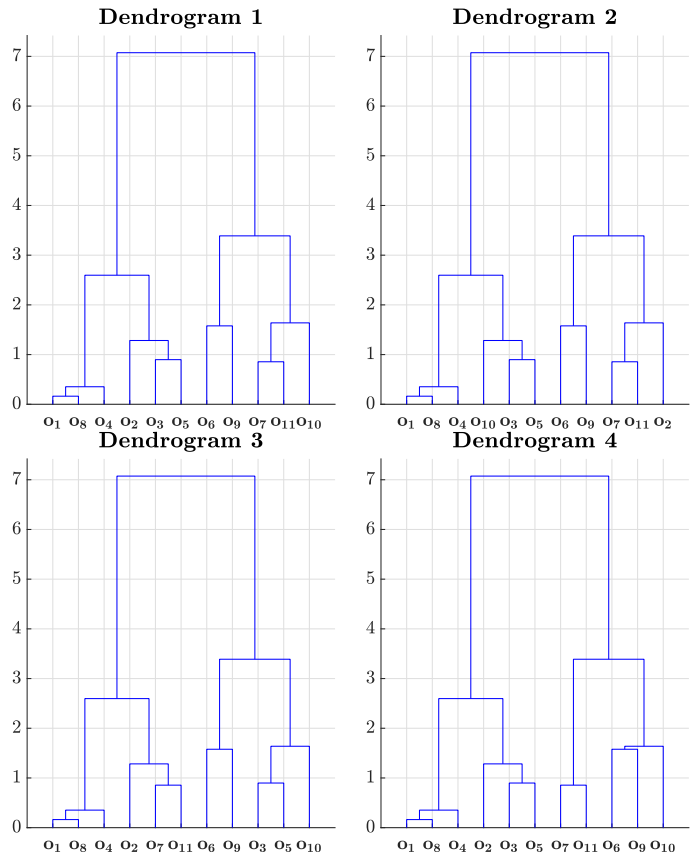

Figure 4: Four dendrograms for which one of the dendrograms corresponds to hierarchical clustering using maximum linkage of the 11 observations in Table 3.

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

$$\begin{aligned}\hat{y} &= \sigma(1.41 + (0.76 \cdot -0.06) + (1.76 \cdot -0.28) \\ &\quad - (0.32 \cdot 0.43) - (0.96 \cdot -0.30) + (6.64 \cdot -0.36) \\ &\quad - (2.74 \cdot 1.00)) \\ &= \frac{1}{1+\exp(-(1.41+(0.76\cdot-0.06)+(1.76\cdot-0.28)-(0.32\cdot0.43)-(0.96\cdot-0.30)+(6.64\cdot-0.36)-(2.74\cdot1.00)))} \\ &= 1.62\%.\end{aligned}$$

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.

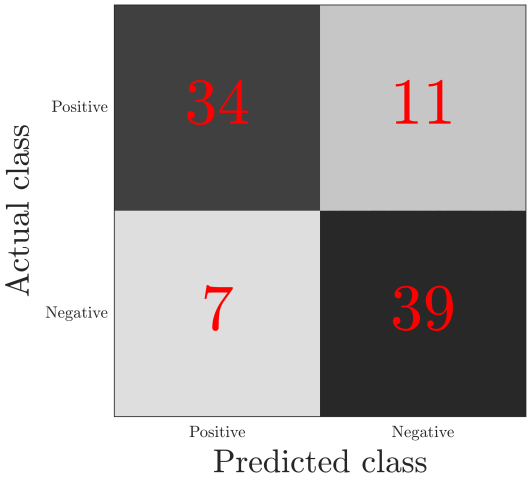

Figure 5: 5-fold cross validation applied to the entire dataset to evaluate logistic regression as an approach to predict low GNP ($y_b = 0$, negative class) versus high GNP ($y_b = 1$, positive class).

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.

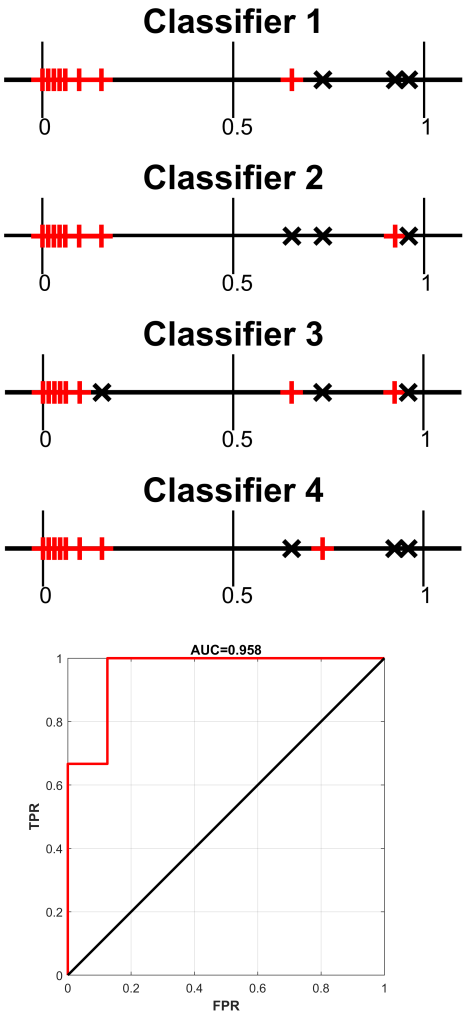

Figure 6: Top panel: Four different logistic regression models used to predict low GNP ($y_b = 0$, marked by red plusses) from high GNP ($y_b = 1$, marked by black crosses). Bottom panel: The ROC curve corresponding to one of the four classifiers in the top panel.

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:

$$P(Africa|GNP > 1000) = \frac{P(GNP>1000|Africa)P(Africa)}{P(GNP>1000)}$$
$$= \frac{P(GNP>1000|Africa)P(Africa)}{P(GNP>1000|Africa)P(Africa)+P(GNP>1000|not\ Africa)P(not\ Africa)}$$
$$= \frac{0.286 \cdot 0.154}{0.286 \cdot 0.154 + 0.688 \cdot (1-t0.154)} = 7.0\%$$

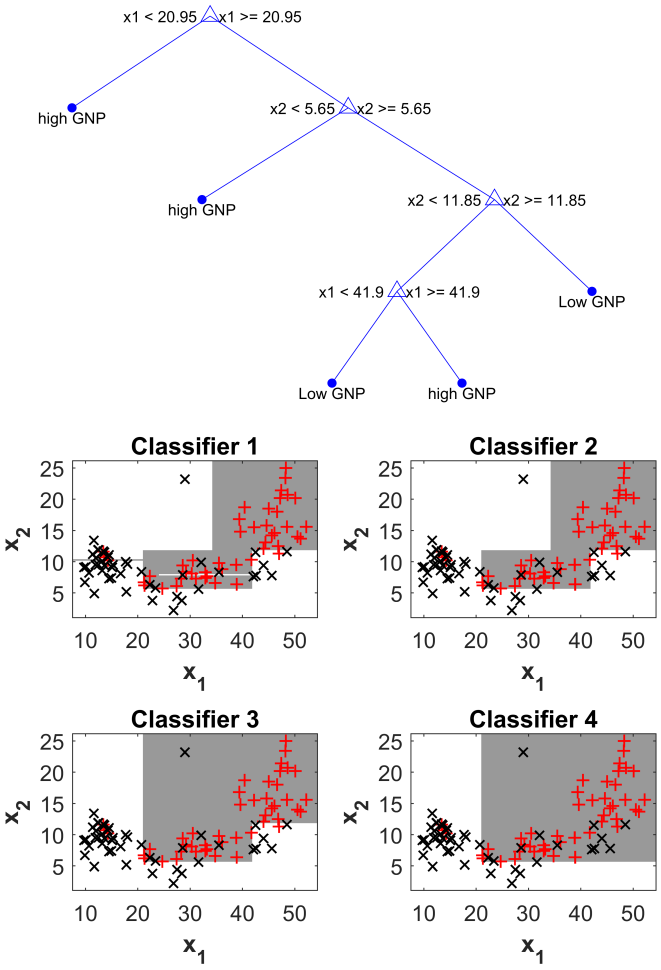

Figure 7: Top panel, a decision tree fitted to x_1 and x_2 of the Poverty data in order to predict wheter a country has relatively low or high GNP. Bottom panel, decision boundaries for four different decision trees in which gray regions correspond to regions predicted having low GNP ($y_b = 0$) and 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 $x_1, \ldots, x_5$ 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:

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

[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.

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

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:

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

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 $\{f_1, f_2\} \to \{f_3\}$?

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.

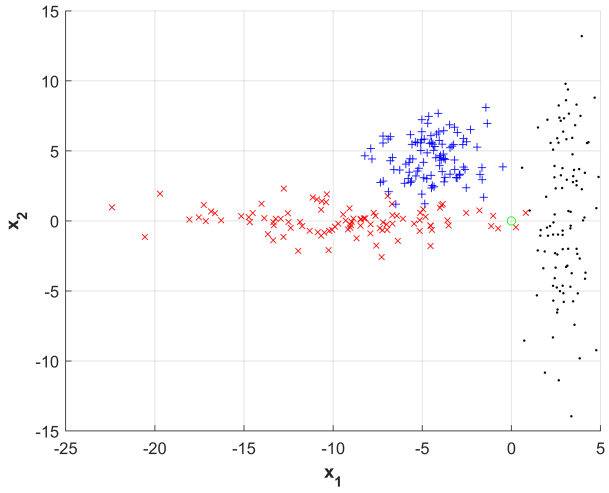

Figure 8: A dataset is separated into three clusters each having 100 observations given by blue plusses, red crosses and black dots. We would like to assign a new observation given by the green circle to one of the three clusters.

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.

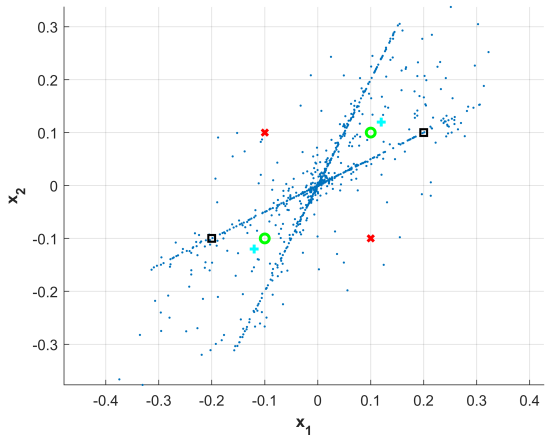

Figure 9: A dataset of 1000 observations given by the blue dots. In the plot is also given the location of two red crosses, two green circles, two cyan plusses and two black squares.

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.

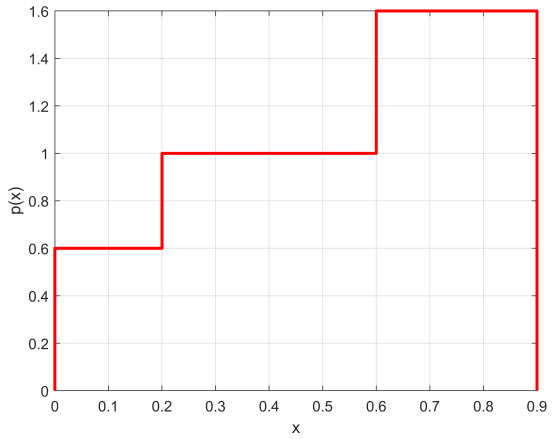

Figure 10: Probability density function for a random variable x. Outside the region from 0 to 0.9 the density function is zero.

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.

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

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.

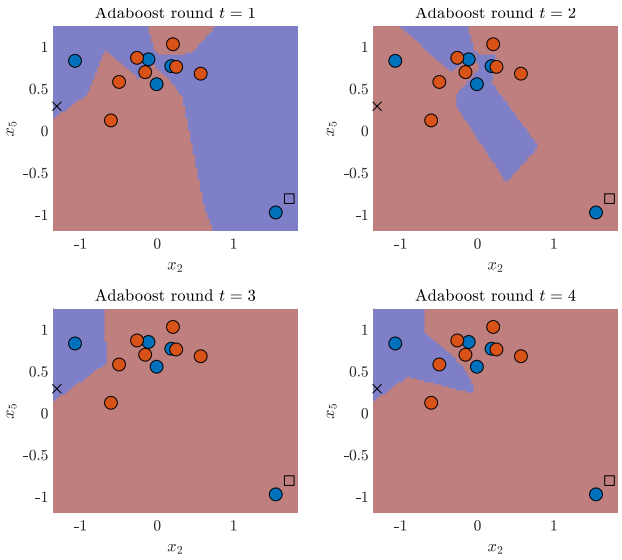

Figure 11: Decision boundaries for a KNN classifier for K=1 enhanced using $T = 4$ rounds of boosting. Notice, in addition to the training data the plot also includes two test points marked respectively by a black cross ($\boldsymbol{x}_1^{test}$) and square ($\boldsymbol{x}_2^{test}$). Observations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

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

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

Technical University of Denmark

Written examination: May 26th 2021, 9 AM - 1 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.

This exam only allows for electronic hand-in.
You hand in your answers at `https://eksamen.dtu.dk/`. To hand in your answers, write them in the file `answers.txt` (this file is available from the same place you downloaded this file). When you are done, upload the `answers.txt` file (and nothing else). Double-check that you uploaded the correct version of the file from your computer.

Do not change the format of `answers.txt`
The file is automatically parsed after hand-in. Do not change the file format of `answers.txt` to any other format such as `rtf`, `docx`, or `pdf`. Do not change the file structure. Only edit the portions of the file indicated by question marks.

No.	Attribute description	Abbrev.
x_1	Hour (0-23)	Hour
x_2	Temperature (Celcius)	Temperature
x_3	Humidity (percent)	Humidity
x_4	Wind speed (m/s)	Wind
x_5	Visibility (10m)	Visibility
x_6	Dew point temperature (Celcius)	Dewpoint
x_7	Solar Radiation (MJ/m2)	Solar
x_8	Rainfall (mm)	Rain
y_r	Bike rental/demand (bikes/hour)	Bike rental

Table 1: Description of the features of the Bicycle rental dataset used in this exam. Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. To ensure bikes are available at all times, it is important to forecast the number of bikes rented per hour y_r as a function of the time of day (measured by the hour attribute so that e.g. $x_1 = 15$ is 15:00-16:00) as well as other features. Visibility is the degree of visibility at 10m of distance (0 meaning no visibility at all) and humidity is measured in percentage of full water saturation (0 being completely dry air). The unit for solar radiation is mega joules per square meter. For classification, the attribute y_r is discretized to create the variable y, taking values $y = 1$ (corresponding to a low demand), $y = 2$ (corresponding to a medium demand), and $y = 3$ (corresponding to a high demand). There are $N = 8760$ observations in total.

Question 1. The main dataset used in this exam is the Bicycle rental dataset[1] described in Table 1. We will consider the type of an attribute as the highest level it obtains in the type-hierarchy (nominal, ordinal, interval, and ratio). Which of the following statements are true about the types of the attributes in the Bicycle rental dataset?

A. x_1 (*Hour*) is nominal, x_2 (*Temperature*) is ratio, x_4 (*Wind*) is ratio, and x_6 (*Dewpoint*) is interval

B. x_2 (*Temperature*) is nominal, x_4 (*Wind*) is nominal, x_7 (*Solar*) is ratio, and x_8 (*Rain*) is ratio

C. x_1 (*Hour*) is nominal, x_2 (*Temperature*) is interval, x_3 (*Humidity*) is ratio, and x_6 (*Dewpoint*) is interval

D. x_2 (*Temperature*) is interval, x_5 (*Visibility*) is ratio, x_6 (*Dewpoint*) is interval, and x_7 (*Solar*) is ratio

E. Don't know.

Solution 1. The problem is solved by simply thinking about what the attributes represent and comparing them to the definition in the different types. Recall that

- Nominal is a type that only allow comparison (equal or different)
- Ordinal allows ordering (but not differences)
- Interval allows differences but no (physically well-defined) zero
- Ratio is a type with a zero with a well-defined meaning

With these definitions, we see that

x_1 (***Hour***) is interval

x_2 (***Temperature***) is interval

x_3 (***Humidity***) is ratio

x_4 (***Wind***) is ratio

x_5 (***Visibility***) is ratio

x_6 (***Dewpoint***) is interval

x_7 (***Solar***) is ratio

x_8 (***Rain***) is ratio

y (***Bike rental***) is ratio

and therefore option D is correct.

[1]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand`

Question 2. A Principal Component Analysis (PCA) is carried out on the Bicycle rental dataset in Table 1 based on the attributes x_1 (HOUR), x_2 (TEMPERATURE), x_3 (HUMIDITY), x_6 (DEWPOINT), and x_7 (SOLAR).

The data is pre-processed by substracting the mean to obtain the centered data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out to obtain the decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.11 & -0.8 & 0.3 & -0.17 & -0.48 \\ -0.58 & -0.31 & 0.01 & -0.5 & 0.56 \\ 0.49 & 0.08 & -0.49 & -0.72 & -0.07 \\ 0.6 & -0.36 & 0.04 & 0.27 & 0.66 \\ -0.23 & -0.36 & -0.82 & 0.37 & -0.09 \end{bmatrix} \tag{1}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 126.15 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 104.44 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 92.19 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 75.07 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 53.48 \end{bmatrix}.$$

We let $\boldsymbol{u}_i$ denote the i'th column of $\boldsymbol{U}$ and $\boldsymbol{v}_i$ the i'th column of $\boldsymbol{V}$. Furthermore, suppose $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the first two unit vectors. The unit vectors are defined such that only coordinate 1 of $\boldsymbol{e}_1$ is 1 (and all other coordinates are zero) and only coordinate 2 of $\boldsymbol{e}_2$ is 1 and (and all other coordinates are zero), and it is assumed the dimensions of the unit vectors are such the matrix/vector multiplications below are possible. Finally, recall $\|\boldsymbol{X}\|_F$ is the Frobenius norm.

Which one of the following statements computes the variance explained by the first two principal components?

A. $\frac{(\boldsymbol{u}_1^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1)^2 + (\boldsymbol{u}_2^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_2)^2}{\|\boldsymbol{\Sigma}\|_F^2}$

B. $\frac{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}\boldsymbol{e}_1 + \boldsymbol{e}_2^\top \boldsymbol{\Sigma}\boldsymbol{e}_2}{\|\boldsymbol{\Sigma}\|_F}$

C. $\frac{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1 + \boldsymbol{e}_2^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_2}{\|\boldsymbol{\Sigma}\|_F}$

D. $\frac{(\boldsymbol{e}_1^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1)^2 + (\boldsymbol{e}_2^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_2)^2}{\|\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\|_F^2}$

E. Don't know.

Solution 2. The correct answer is A. To see this, recall 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 $\sigma_k = \boldsymbol{\Sigma}_{kk}$ where $\boldsymbol{\Sigma}$ is the diagonal matrix of the SVD computed above. The denominator is therefore equal to the sum of the squares of all elements in $\boldsymbol{\Sigma}$ i.e. $\sum_{j=1}^M \sigma_j^2 = \|\boldsymbol{\Sigma}\|_F^2$. To obtain the numerator, note that since $\boldsymbol{U}, \boldsymbol{V}$ are orthonormal it holds that

$$\boldsymbol{u}_1^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1 = \boldsymbol{e}_1^\top \boldsymbol{\Sigma}\boldsymbol{e}_1 = \Sigma_{11}.$$

Hence expressions of this form a suitable to compute the numerator, meaning that A are the correct answer.

Question 3. Consider again the PCA analysis for the Bicycle rental dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (1). Which one of the following statements is true?

A. An observation with a low value of Temperature, a high value of Humidity, a high value of Dewpoint, and a low value of Solar will typically have a positive value of the projection onto principal component number 1.

B. An observation with a high value of **Hour**, a low value of **Humidity**, and a low value of **Solar** will typically have a negative value of the projection onto principal component number 3.

C. An observation with a high value of **Hour**, a low value of **Temperature**, and a low value of **Dewpoint** will typically have a positive value of the projection onto principal component number 5.

D. An observation with a low value of **Hour**, a low value of **Temperature**, a low value of **Dewpoint**, and a low value of **Solar** will typically have a negative value of the projection onto principal component number 2.

E. Don't know.

Solution 3. 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

$$b_1 = \boldsymbol{x}^\top \boldsymbol{v}_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_6 & x_7 \end{bmatrix} \begin{bmatrix} 0.11 \\ -0.58 \\ 0.49 \\ 0.6 \\ -0.23 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and positive, this occurs if x_2, x_3, x_6, x_7 has large magnitude and the sign convention given in option A.

Question 4. Consider again the Bicycle rental dataset and the PCA decomposition described in Equation (1). Recall the PCA decomposition is obtained by first forming the centered data matrix $\tilde{\boldsymbol{X}}$ by subtracting the column-wise mean

$$\boldsymbol{\mu} = \begin{bmatrix} 12.9 \\ 58.2 \\ 1.7 \\ 1436.8 \\ 4.1 \end{bmatrix}$$

from the data matrix $\boldsymbol{X}$. Assume an observation has coordinates

$$\boldsymbol{x} = \begin{bmatrix} 15.5 \\ 59.2 \\ 1.4 \\ 1438.0 \\ 5.3 \end{bmatrix}.$$

Which coordinates in the coordinate system spanned by the principal component vectors corresponds to $\boldsymbol{x}$?

A. $\boldsymbol{b} = \begin{bmatrix} 0.0 & -3.2 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

B. $\boldsymbol{b} = \begin{bmatrix} 0.0 & 1.2 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

C. $\boldsymbol{b} = \begin{bmatrix} 0.0 & 1.5 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

D. $\boldsymbol{b} = \begin{bmatrix} 0.0 & -1.6 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

E. Don't know.

Solution 4. The simplest way to solve this problem is to begin with one of the possible values of $\boldsymbol{b}$ and check if it corresponds to $\boldsymbol{x}$ or not. Recall that this is done by computing:

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{b} + \boldsymbol{\mu}. \tag{2}$$

We will only compute the value of $\boldsymbol{x}$ at the first coordinate, i.e. x_1. Since most entries in $\boldsymbol{b}$ are zero this can be done as simply:

$$x_1 = V_{12} b_2 + \mu_1. \tag{3}$$

Knowing that $x_1 = 15.5$ we observe that A is the correct answer.

Question 5. Consider again the Bicycle rental dataset. The empirical covariance matrix of the first 5 attributes $x_1, \ldots, x_5$ is given by:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 143.0 & 39.0 & -0.0 & 253.0 & 142.0 \\ 39.0 & 415.0 & -7.0 & -6727.0 & 143.0 \\ -0.0 & -7.0 & 1.0 & 108.0 & -2.0 \\ 253.0 & -6727.0 & 108.0 & 370027.0 & -1403.0 \\ 142.0 & 143.0 & -2.0 & -1403.0 & 171.0 \end{bmatrix}.$$

What is the empirical correlation of x_2 (Temperature) and x_3 (Humidity)?

A. -0.12987

B. -0.01687

C. -0.34362

D. -2.64575

E. Don't know.

Solution 5. Recall the correlation is defined as

$$\text{cor}[x, y] = \frac{\text{cov}[x, y]}{\text{std}[x]\,\text{std}[y]}$$

Next, by definition the diagonal elements of the covariance matrix are estimates of the variance and the off-diagonal elements are estimates of the covariance, i.e. for $i \neq j$:

$$\hat{\Sigma}_{ii} = \text{Var}[x_i], \quad \hat{\Sigma}_{ij} = \text{cov}[x_i, x_j]$$

Therefore we get:

$$\text{cor}[x_i, y_j] = \frac{\hat{\Sigma}_{ij}}{\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}}}.$$

By simple insertion, we see option C is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	5.0	7.7	6.1	4.2	11.0	7.3	9.0	11.3	1.4
o_2	5.0	0.0	5.4	4.0	7.5	7.9	5.3	6.8	11.9	3.5
o_3	7.7	5.4	0.0	5.2	7.2	6.1	7.8	6.7	12.9	6.4
o_4	6.1	4.0	5.2	0.0	5.1	5.4	8.4	3.3	8.1	4.8
o_5	4.2	7.5	7.2	5.1	0.0	8.7	8.8	6.6	7.7	4.1
o_6	11.0	7.9	6.1	5.4	8.7	0.0	12.0	4.2	9.3	9.8
o_7	7.3	5.3	7.8	8.4	8.8	12.0	0.0	11.0	16.3	6.7
o_8	9.0	6.8	6.7	3.3	6.6	4.2	11.0	0.0	6.2	7.8
o_9	11.3	11.9	12.9	8.1	7.7	9.3	16.3	6.2	0.0	10.4
o_{10}	1.4	3.5	6.4	4.8	4.1	9.8	6.7	7.8	10.4	0.0

Table 2: The pairwise cityblock distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{p=1} = \sum_{k=1}^{M} |x_{ik} - x_{jk}|$ between 10 observations from the Bicycle rental dataset (recall that $M = 8$. Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1. The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belong to class C_1 (corresponding to a low demand), the red observations $\{o_3,\ o_4,\ o_5,\ o_6\}$ belong to class C_2 (corresponding to a medium demand), and the blue observations $\{o_7,\ o_8,\ o_9,\ o_{10}\}$ belong to class C_3 (corresponding to a high demand). To avoid single features to dominate, the dataset was standardized by subtracting the mean and dividing by the standard deviation.

Question 6. To examine if observation o_3 may be an outlier, we will calculate the average relative density using the cityblock distance based on the observations given in Table 2 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. 0.7

B. 0.4

C. 0.63

D. 0.19

E. Don't know.

Solution 6.

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

$$N_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = \{o_4, o_2\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = 0.189$$

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

$$N_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = \{o_8, o_2\}, \quad N_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = \{o_{10}, o_4\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = 0.274, \text{density}_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = 0.267.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem.

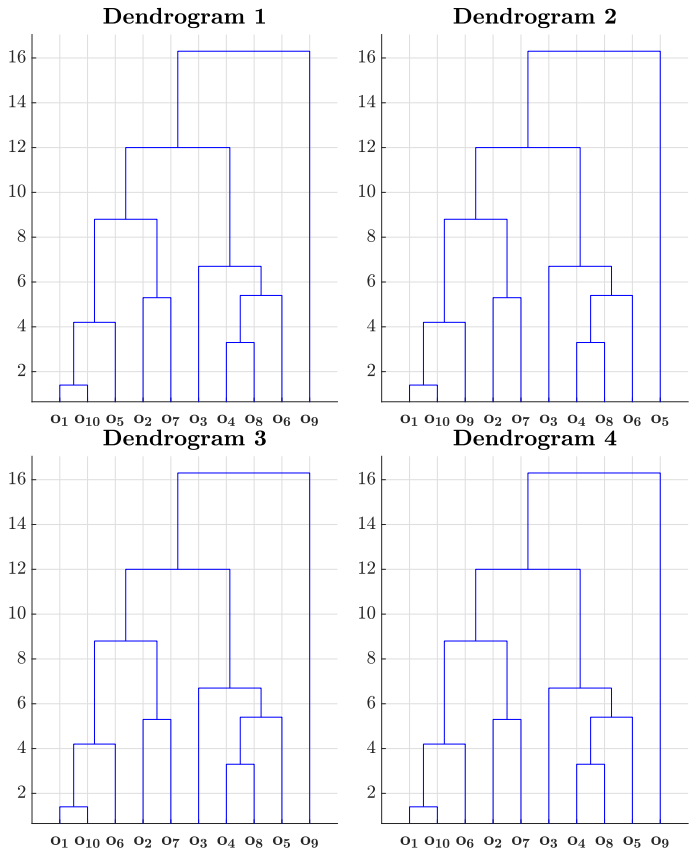

Figure 1: Proposed hierarchical clustering of the 10 observations in Table 2.

Question 7. A hierarchical clustering is applied to the 10 observations in Table 2 using *maximum* linkage. Which one of the dendrograms shown in Figure 1 corresponds to the distances given in Table 2?

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 3 should have been between the sets $\{f_5\}$ and $\{f_1, f_{10}\}$ at a height of 4.2, however in dendrogram 2 merge number 3 is between the sets $\{f_9\}$ and $\{f_1, f_{10}\}$.

- In dendrogram 3, merge operation number 3 should have been between the sets $\{f_5\}$ and $\{f_1, f_{10}\}$ at a height of 4.2, however in dendrogram 3 merge number 3 is between the sets $\{f_6\}$ and $\{f_1, f_{10}\}$.
- In dendrogram 4, merge operation number 3 should have been between the sets $\{f_5\}$ and $\{f_1, f_{10}\}$ at a height of 4.2, however in dendrogram 4 merge number 3 is between the sets $\{f_6\}$ and $\{f_1, f_{10}\}$.

Question 8. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two binary vectors of (even) dimension M such that the first two elements of $\boldsymbol{x}_1$ are 1 (and the rest are 0) and the first $\frac{M}{2}$ elements of $\boldsymbol{x}_2$ are 1 (and the rest are 0).

Which of the following expressions computes the Jaccard similarity of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ when $M \geq 4$?

A. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{4}{M}$

B. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\frac{1}{2}M}{\frac{1}{2}M+2}$

C. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{2}{M}$

D. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{2}{\frac{1}{2}M-2}$

E. Don't know.

Solution 8. The Jaccard similarity is given as

$$J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{n_{11}}{M - n_{00}}.$$

Given the information in the problem the number of 0-0 matches is $n_{00} = \frac{M}{2}$ while the number of 1-1 matches is $n_{11} = 2$. The solution is therefore A.

Question 9. Consider again the Bicycle rental dataset in Table 1. We apply backward selection to find an interpretable linear regression model which uses a subset of the $M = 8$ attributes to predict the bike rental y_r. Recall backward selection chooses models based on the test error as determined by cross-validation, and in our case we use the hold-out method to generate a single test/training split.

Suppose backward selection ends up selecting the attributes x_1, x_3, x_4, x_5, x_6, x_7, and x_8, what is the minimal number of models which were *tested* in order to obtain this result?

A. 15 models

B. 18 models

C. 16 models

D. 8 models

E. Don't know.

Solution 9.

Note the solution selected all variables except one. Since we use backward selection, we first have to evaluate a single model with all features. Then we evaluated all models with a single missing feature

giving an additional M models. One of these models were selected and variable selection proceeded at the next level where an additonal $M-1$ models were evaluated. However, since all had a higher cost, none were selected and the method terminated. This gives

$$1+M+M-1$$

evaluations and so C is correct.

Question 10. We wish to predict which of the three classes an observation $\boldsymbol{x}$ belong to in the Bicycle rental dataset described in Table 1. To accomplish this we apply a Naïve-Bayes classifier where we model each of the $M = 8$ features using a 1-dimensional normal distribution. The classifier will be used in an embedded setting where model prediction speed is paramount. Therefore, consider a single model evaluation:

$$p(y = \text{LOW DEMAND}|\boldsymbol{x}).$$

What is the minimum number of evaluations of the normal density function $\mathcal{N}(x|\mu, \sigma^2)$ we have to perform to compute this quantity?

A. 24

B. 27

C. 36

D. 32

E. Don't know.

Solution 10. Recall the formula for Naïve-Bayes is

$$p(y|\boldsymbol{x}) = \frac{p(y)\prod_{k=1}^{M} p(x_k|y)}{\sum_{y'} p(y')\prod_{k=1}^{M} p(x_k|y')}.$$

The total number of evaluations is equal to the total number of evaluations of terms $p(x_k|y)$. Note that once we have evaluated the denominator, we will also have evaluated the numerator since they share the same terms $p(x_k|y)$, cutting down on computations. The total number of evaluations is therefore simply CM where C is the number of classes and therefore answer A is correct.

Question 11. Consider the Bicycle rental dataset from Table 1 consisting of $N = 8760$ observations, and suppose the attribute Humidity has been binarized into low and high values. We still consider the goal to predict the bike rental and are given the following information

- Of the 3285 observations with low demand, 1327 had a low value of Humidity.
- Of the 2190 observations with medium demand, 1718 had a low value of Humidity.
- Of the 3285 observations with high demand, 2344 had a low value of Humidity.

Suppose a particular observation has a high value of Humidity, what is the probability of observing high demand?

A. 0.279

B. 0.286

C. 0.04

D. 0.487

E. Don't know.

Solution 11. The problem is solved by applying Bayes rule. Introducing the binary variable x such that $x = 1$ if an observation has a high value of Humidity (and otherwise $x = 0$) the question asked is equivalent to computing $p(y = 3|x = 1)$. Applying Bayes' theorem we get:

$$p(y = 3|x = 1) = \frac{p(x = 1|y = 3)p(y = 3)}{\sum_{k=1}^{3} p(x = 1|y = k)p(y = k)}$$

Recall that $p(x = 1|y) = 1 - p(x = 0|y)$, we can obtain the required probabilities from each of the three bullet points above. We obtain:

- $p(y = 1) = \frac{3285}{N}$ and $p(x = 0|y = 1) = \frac{1327}{3285}$.
- $p(y = 2) = \frac{2190}{N}$ and $p(x = 0|y = 2) = \frac{1718}{2190}$.
- $p(y = 3) = \frac{3285}{N}$ and $p(x = 0|y = 3) = \frac{2344}{3285}$.

Plugging these into Bayes theorem, and using that $p(x = 0|y) = 1 - p(x = 1|y)$ because x is binary, we see $p(y = 3|x = 1) = 0.279$ and hence that option A is correct.

Question 12. Consider the Bicycle rental dataset described in Table 1. Suppose we apply a market basket analysis to the dataset in the usual fashion: We first binarize each of the attributes, thereby obtaining $M = 8$ items, and consider each of the $N = 8760$ observations as a transaction containing a (subset) of the binarized attributes. We will let $C(\{I_1, \ldots, I_k\})$ be the number of the $N = 8760$ transactions containing the itemset $\{I_1, \ldots, I_k\}$. For this problem we focus on just three items and are given the information:

- $C(\{\text{VISIBILITY}\}) = 4091$.
- $C(\{\text{HUMIDITY}\}) = 3637$.
- $C(\{\text{DEWPOINT}\}) = 3459$.

Finally, consider the itemset:

$$I: \{\text{VISIBILITY}, \text{HUMIDITY}\}.$$

Which of the following options indicate the *highest possible* support of the itemset I which is consistent (i.e., obtainable) given the information in the bullet list above?

A. $\textbf{supp}(I) = 0.415$

B. $\text{supp}(I) = 0.441$

C. $\text{supp}(I) = 0.217$

D. $\text{supp}(I) = 0.467$

E. Don't know.

Solution 12. For a transaction to include an itemset, it must (by the downwards closure property) include all subsets, and hence the maximal number of transactions which contain $\{\text{VISIBILITY}, \text{HUMIDITY}\}$ is upper-bounded by the number of transactions which contain either of the items, i.e. $\min\{C(\{\text{VISIBILITY}\}), C(\{\text{HUMIDITY}\})\}$:

$$C(\{\text{VISIBILITY}, \text{HUMIDITY}\}) \leq \min\{C(\{\text{VISIBILITY}\}), C(\{\text{HUMIDITY}\})\} = 3637.$$

Dividing by N we see answer A is correct.

	1	2	3	4	5	6	7	8
x_1	-1.1	-0.8	0.08	0.18	0.34	0.6	1.42	1.68
y_r	12	5	10	23	6	17	14	13

Table 3: Values of x_1 and the corresponding value of y_r.

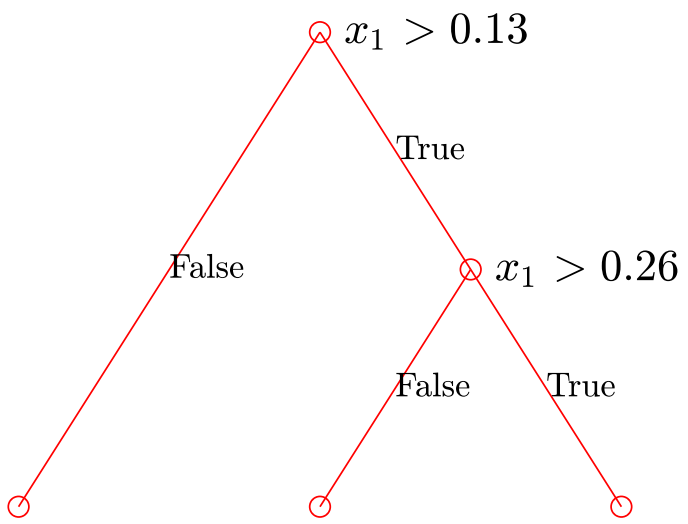

Figure 2: Structure of a regression tree. The nodes show the decision rules which determine how the observations are propagated towards the leafs of the tree.

Question 13. We will consider the first 8 observations of the Bicycle rental dataset shown in Table 2. Table 3 shows their corresponding value of x_1 and y_r. We fit a small regression tree to this dataset. The structure (and binary splitting rules) is depicted in Figure 2. Which one of the prediction rules (i.e., the model output $\hat{y}_r$ as a function of x_1) shown in Figure 3 corresponds to the tree?

A. Prediction rule 1

B. Prediction rule 2

C. Prediction rule 3

D. Prediction rule 4

E. Don't know.

Solution 13.

The problem is easiest solved by selecting a lucky x-value and using it to rule out the wrong plot. In our case, we select $x_1 = 0.4$. The predicted value for a given input is computed as the average y-value of those

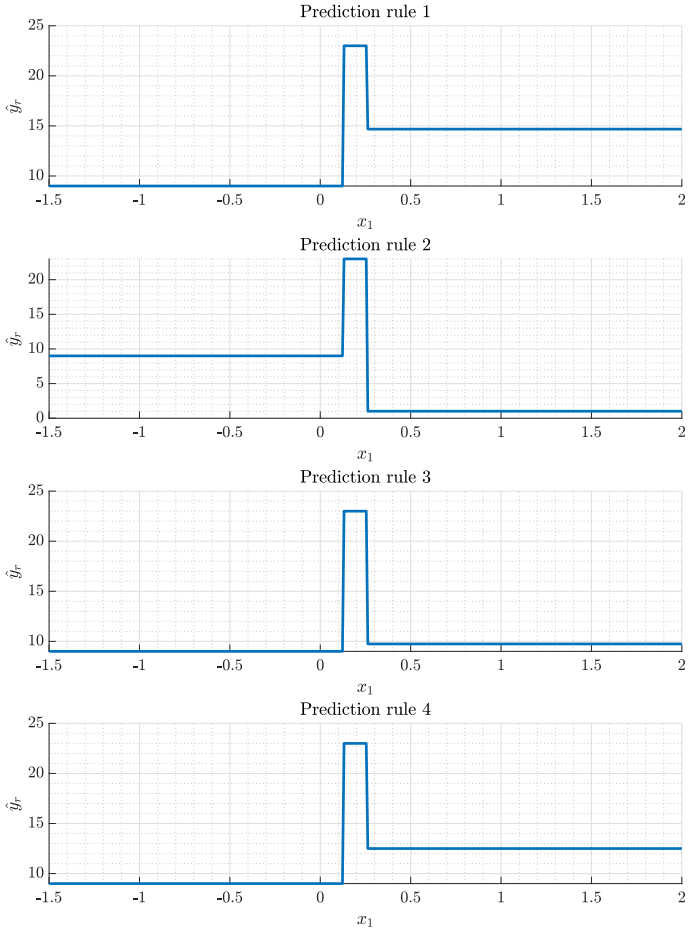

Figure 3: Possible model predictions of $\hat{y}_r$ as a function of x_1 for the decision tree illustrated in Figure 2.

observations in the training set which is assigned to the same leaf node v as the input, i.e.

$$y(v) = \frac{1}{N(v)} \sum_{i \in v} y_i$$

(see the section on regression trees in lecture notes). Therefore, we first need to find out which leaf node the observation is assigned to. To do this, start at the root and compare $x_1 = 0.4$ to the rule in the split

$$x_1 > 0.13$$

and we continue down the right branch. Continuing in this manner, we see $x_1 = 0.4$ is classified to leaf number three from the left. Then, proceeding in the same manner with the x_1 observations in Table 3, we see the observations o_5, o_6, o_7, and o_8 are also assigned to leaf three (counted from the left). According to the above the prediction is then simply the average of their y-value

$$\hat{y} = \frac{1}{4}(6 + 17 + 14 + 13)$$

or $\hat{y} = 12.5$. we compare this information in Figure 3 and note this allows us to rule out all the wrong options. Therefore, D is correct.

Question 14. In this problem, we will again consider the 8 observations from the Bicycle rental dataset shown in Table 3. Recall that Figure 2 shows the structure of the small regression tree fitted to this dataset using Hunt's algorithm along with the thereby obtained binary splitting rules. What was the purity gain Δ of the **second** split Hunt's algorithm accepted?

A. $\Delta = 101.2$

B. $\Delta = 30.64$

C. $\Delta = 17.64$

D. $\Delta = 13.0$

E. Don't know.

Solution 14.

The second split Hunt's algorithm accepted must be the non-root split, i.e. corresponding to the rule

$$x > 0.26.$$

This node, which we will call the base node of the split, partitions the observations:

$$v_0 = \{4, 5, 6, 7, 8\}$$

into the two sets

$$v_1 = \{4\}, \quad v_2 = \{5, 6, 7, 8\}$$

along the two legs of the split. The impurity of these two sets, and the impurity of all y-values at the root, is computed using the impurity measure appropriate for regression trees

$$I(v) = \frac{1}{N(v)} \sum_{i \in v_1} (y_i - y(v))^2$$

where $y(v)$ is the average of the y-values in v_i. Specifically

$$y(v_1) = 23.0, \quad y(v_2) = 12.5$$

At the base node v_0 of the split we consider we perform a similar calculation for the 5 observations and find they have a mean y-value of:

$$y(v_0) = 12.5$$

Therefore:

$$I(v_0) = 30.64, \quad I(v_1) = 0.0, \quad I(v_2) = 16.25$$

these are finally combined to the impurity gain as

$$\Delta = I(v_0) - \sum_{k=1}^{2} \frac{N(v_k)}{N} I(v_k)$$

where for instance $N(v_1) = 1$ are the number of observations in branch 1. We find by insertion that $\Delta = 17.64$ and hence C is correct.

Question 15. Consider again the Bicycle rental dataset of Table 1. Suppose we wish to predict the class label y using a multivariate regression model, and to improve performance we wish to apply Adaboost. Recall the first steps of adaboost consists of: (i) Initialize weights, (ii) select a training set (iii) fit a model to the training set. In the first round of boosting, the fitted model has an error rate ϵ when evaluated on the full dataset, and it made a correct prediction of the class membership of observation $i = 5$ and an incorrect prediction of the class membership of observation $i = 1$.

After the first round of boosting, which of the following expressions will compute the ratio of weights of observation 1, w_1 and observation 5, w_5?

A. $\frac{w_1}{w_5} = \exp\left(\frac{1-\epsilon}{\epsilon}\right)$

B. $\frac{w_1}{w_5} = \frac{1-\epsilon}{\epsilon}$

C. $\frac{w_1}{w_5} = \frac{\exp\left(\frac{1-\epsilon}{\epsilon}\right)}{\exp\left(-\frac{1-\epsilon}{\epsilon}\right)}$

D. $\frac{w_1}{w_5} = \sqrt{\frac{1-\epsilon}{\epsilon}}$

E. Don't know.

Solution 15. Recall the weights in the Adaboost algorithm, prior to normalization, are computed as $w_i(1)e^{\pm\alpha_i}$ and they are initialized as $w_i(1) = \frac{1}{N}$. Falsely classified observations are boosted up, and correctly are boosted down. Hence:

$$\frac{w_1}{w_5} = \frac{\frac{1}{N}e^{\alpha}}{\frac{1}{N}e^{-\alpha}} = e^{2\alpha}$$

Using that $\alpha = \frac{1}{2}\log\frac{1-\epsilon}{\epsilon}$ we see that option B is correct.

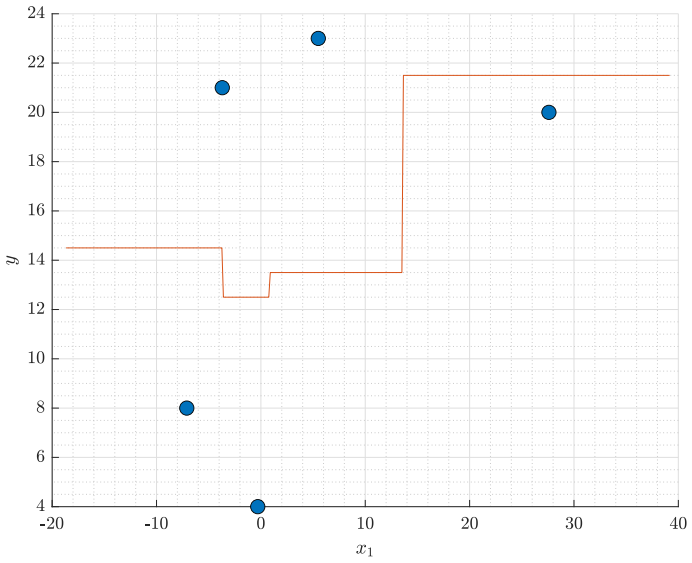

Figure 4: KNN regression model in which the red line is fitted to a small 1-dimensional dataset.

Question 16. Suppose a K-nearest neighbors regression model is fitted to a small 1-dimensional dataset with $N = 5$ observations. The predicted response is shown in Figure 4. How many neighbors (i.e. K) was used?

A. $K = 2$

B. $K = 4$

C. $K = 1$

D. $K = 3$

E. Don't know.

Solution 16.

The problem could be solved by using the definition of the KNN regression model and test various points, but it is much quicker solved using an intuitive argument. The KNN regression model consist of a series of steps, and the important information is where the discontinuities occur. If $K = 1$, the y-value has to pass through the training observations. On the other hand, if we consider the y-value at the right-most end of the x-axis, we note it is quite large consistent with it being computed using the two left-most observations, but not consisting with including the third observation from the right (or additional observations). Hence, we conclude that $K = 2$.

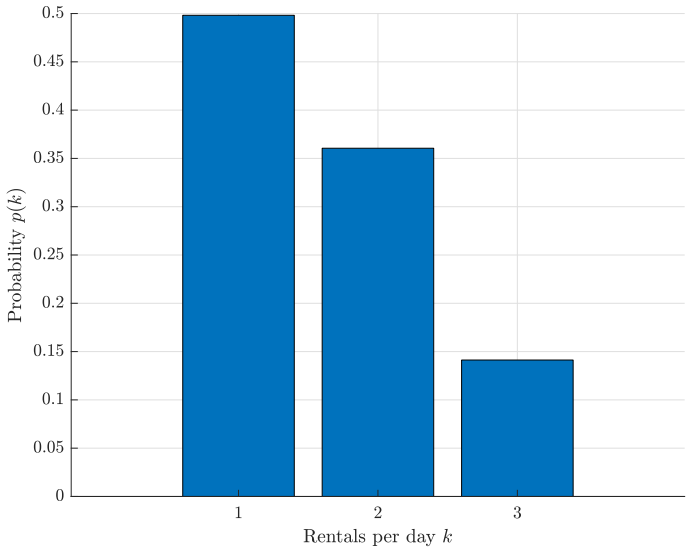

Figure 5: Probability $p(k)$ a citybike is rented exactly k times a day. The probability of $k \geq 4$ is negligible and can be ignored.

Question 17. The number of times a citybike is rented per day is an important factor in determining how often they should be replaced. Suppose the typical bike rentals per day is estimated from data, and the chance $p(k)$ a bike will be rented k times is shown in the discrete probability distribution shown in Figure 5. It is known that the mean of this distribution is 1.6, but what is the variance?

A. Variance is 3.4

B. Variance is 1.6

C. Variance is 0.2

D. Variance is 0.5

E. Don't know.

Solution 17.

The standard deviation of a discrete distribution is given by

$$\mathrm{Var}[K] = \sum_{k=1}^{3} p(k)(k-\mu)^2$$

and we are fortunate enough to be told that $\mu = 1.6$. The probabilities can be read from Figure 5 which gives us:

$$\mathrm{Var}[K] = 0.5(1-\mu)^2 + 0.36(2-\mu)^2 + 0.14(3-\mu)^2.$$

Upon insertion, we see the right answer is D.

Question 18. Which one of the following statements are true?

A. Regularization is not applicable to a logistic regression model.

B. When we apply Adaboost, the less errors a classifier makes in a given round of boosting, the more the weights will be increased for the wrongly classified observations.

C. When using McNemars test to determine if two classification models have different performance, one should apply two-level cross-validation (either hold out/K-fold or leave-one-out).

D. Let $\boldsymbol{x}_i$ be the i'th observation of a (non-standardized) dataset $\boldsymbol{X}$. Suppose we carry out a PCA analysis on $\boldsymbol{X}$ and we let $\boldsymbol{b}_i$ be the principal component coefficient vector (i.e., projection) corresponding to $\boldsymbol{x}_i$ when projected onto *all* the principal components. It is then true that $\|\boldsymbol{x}_i\| = \|\boldsymbol{b}_i\|$ (in the Euclidean norm).

E. Don't know.

Solution 18. The correct answer is B: Inspecting the adaboost algorithm, we see that as the number of error decrease, so does ϵ. When ϵ decreases, then $\alpha = \frac{1}{2}\log\frac{1-\epsilon}{\epsilon}$ will increase. Finally, note the wrongly classified observations are boosted proportional to e^{α_i}, so B is correct.

For the other options, note that regularization can easily be applied to logistic regression (same as linear regression). McNemars test only requires a set of prediction which can be obtained using normal 1-level cross-validation, and for the last problem, $\boldsymbol{x}_i$ does not have the mean-vector subtracted (whereas it is subtracted to compute $\boldsymbol{b}_i$) and so we should have no expectation the norms should be the same.

Question 19. Consider a regression problem where the goal is to predict a ratio variable y_i using the 1-dimensional input x_i. Suppose we wish to do this using a neural network with a single hidden layer (the hidden layer has a sigmoid activation function), no activation function (i.e. the identity activation function) for the output layer, and that we use the ordinary quadratic cost function suitable for regression. What is an appropriate cost function on a training set of size N (assuming all terms of the form $w_{\cdot}^{(\cdot)}$ are

weights)?

A. $\sum_{i=1}^{N}\left(\frac{w_0^{(2)}}{1+e^{-y_i}}-\frac{w_1^{(2)}}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}}}-\frac{w_2^{(2)}}{1+e^{-w_{2,0}^{(1)}-x_i w_{2,1}^{(1)}}}\right)^2$

B. $\sum_{i=1}^{N}\left(w_0^{(2)}-\frac{w_1^{(2)}}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}-y_i w_{1,2}^{(1)}}}-\frac{w_2^{(2)}}{1+e^{-w_{2,0}^{(1)}-x_i w_{2,1}^{(1)}-y_i w_{2,3}^{(1)}}}\right)^2$

C. $\sum_{i=1}^{N}\left(y_i-w_0^{(2)}-\frac{w_1^{(2)}}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}}}-\frac{w_2^{(2)}}{1+e^{-w_{2,0}^{(1)}-x_i w_{2,1}^{(1)}}}\right)^2$

D. $\sum_{i=1}^{N}\left(y_i-w_0^{(2)}-\frac{w_1^{(2)}}{w_1^{(2)}-e^{w_{1,0}^{(1)}+x_i w_{1,1}^{(1)}}}-\frac{w_2^{(2)}}{w_2^{(2)}-e^{w_{2,0}^{(1)}+x_i w_{2,1}^{(1)}}}\right)^2$

E. Don't know.

Solution 19.
For a general input x_i, the quadratic cost function will be of the form:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2$$

where $f(x)$ is the output of the neural network. If the hidden units activation is denoted z_1 and z_2 the output has the linear form

$$f(x) = w_0^{(2)} + w_1^{(2)} z_1 + w_2^{(2)} z_2$$

Finally, the output activation of the hidden units, for instance z_1, can be computed using the sigmoid activation function:

$$\sigma(w_{1,0}^{(1)} + x_i w_{1,1}^{(1)}) = \frac{1}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}}}.$$

Comparing this information with the four expression we can rule out all options except C.

	x_1	x_5	y
Mean	12.9	4.1	11.5
Standard deviation	11.9	13.1	6.9

Table 4: Column-wise mean and standard deviation computed on the Bicycle rental dataset.

Question 20. Consider once again the bicycle rental dataset described in Table 1, but this time we will limit ourselves to just the features x_1 (HOUR) and x_5 (VISIBILITY) from the full dataset $\boldsymbol{X}$. The goal is still to predict the bike rental $y = y_r$, and to achieve this we will apply ridge-regression. Recall that ridge regression determines the constant offset w_0 and the two coefficients w_1 and w_2 of x_1 and x_5 respectively, by minimizing a cost function of the form:

$$\sum_{i=1}^{N}\left(y_i - w_0 - w_1\frac{X_{i,1}-\mu_1}{\sigma_1} - w_2\frac{X_{i,5}-\mu_5}{\sigma_5}\right)^2 + \lambda(w_1^2+w_2^2).$$

In this expression, μ_k and σ_k are the mean and standard deviations of column k, and their values can be found in Table 4, along with the corresponding values for y. Assuming the regularization strength is $\lambda = 10.0$, which one of the following expressions will predict the value y for an input observation with $x_1 = 0$ and $x_5 = 1$?

A. $y = w_0 + 1.08w_1 + 0.39w_2$

B. $y = 0.14w_0 - 1.08w_1 - 0.24w_2$

C. $y = w_0 - 0.24w_2$

D. $y = 11.5 - 1.08w_1 - 0.24w_2$

E. Don't know.

Solution 20. To solve this problem, we should first apply the same transformation to the text-point $\boldsymbol{x}$ as is applied when training the model and then remember to add w_0 which is equal to the mean of $\boldsymbol{y}$. The prediction rule is

$$y = w_0 + w_1\frac{x_1-\mu_1}{\sigma_1} + w_2\frac{x_5-\mu_5}{\sigma_5}$$

The mean/standard deviations are given as the column-wise mean/standard deviations in Table 4. Inserting these values, as well as the values for x_1 and x_5, we see that answer D is correct.

Observation nr. i	$\boldsymbol{w}_1^\top \tilde{\boldsymbol{x}}_i$	$\boldsymbol{w}_2^\top \tilde{\boldsymbol{x}}_i$
1	0.03	-1.89
2	1.17	-0.89
3	1.15	-0.87
4	1.32	-0.71
5	-0.05	-1.9
6	0.64	-1.28
7	0.65	-1.27
8	1.25	-0.69

Table 5: Output of the linear transformation (prior to softmax normalization) of a multinomial regression model applied to the Bicycle rental dataset. The full dataset contains $N = 8760$ observations, but the table only contains the output for the first $i = 1, \ldots, 8$ observations.

Question 21. Consider the Bicycle rental dataset described in Table 1. Recall the dataset is comprised of $C = 3$ classes, and suppose we fit a multinomial regression model to predict the class label y_i given the $M = 8$-dimensional feature vector $\boldsymbol{x}_i$. This results in two weight-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ such that the class-label is predicted using the softmax activation as described in Section 15.3.3 in the lecture notes. Prior to softmax normalization, the output on the first 8 observations are shown in Table 5. According to the multinomial regression model, what is the probability observation $i = 1$ is assigned to the low demand class ($y = 1$)?

A. 0.01

B. 0.07

C. 0.26

D. 0.47

E. Don't know.

Solution 21. Solving this problem is simply a matter of using the definition of the multinomial regression model. The probability can be computed as

$$p(y = 1|\boldsymbol{x}_1) = \frac{e^{\boldsymbol{w}_1^\top \boldsymbol{x}_1}}{e^{\boldsymbol{w}_1^\top \boldsymbol{x}_1} + e^{\boldsymbol{w}_2^\top \boldsymbol{x}_1} + 1}$$

Inserting the number from Table 5 we see that answer D is correct.

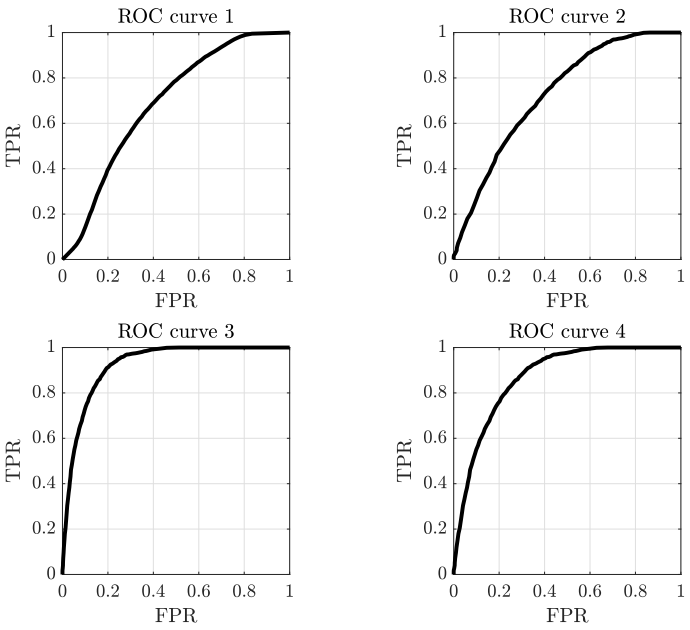

Figure 6: Candidate ROC curves for the classifier.

Question 22. We wish to predict whether an observation from the Bicycle rental dataset (see Table 1) belongs to the low demand class (or not). To accomplish this, we fit a logistic regression model to the dataset, and for each observation $\boldsymbol{x}_i, y_i$ obtain a class-probability prediction $\hat{y}_i \in [0, 1]$. We threshold the class-probability at different values θ thereby obtaining, for each value of θ, the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These are plotted as functions of θ in Figure 7. Which of the receiver-operator characteristic (ROC) plots shown in Figure 6 corresponds to these graphs?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

Solution 22. From the TP curve (left-most value) we get that the total number of positive-class observations are $P = 1652$ and from the TN curve (right-most value) we get $N = 2728$. The simplest approach is to compute a point on the ROC curve. Most values will do, however we choose the point corresponding to $\theta = 0.5$, at which the number of false positives is $FP = 570$ and true positives is $TP = 1285$. We

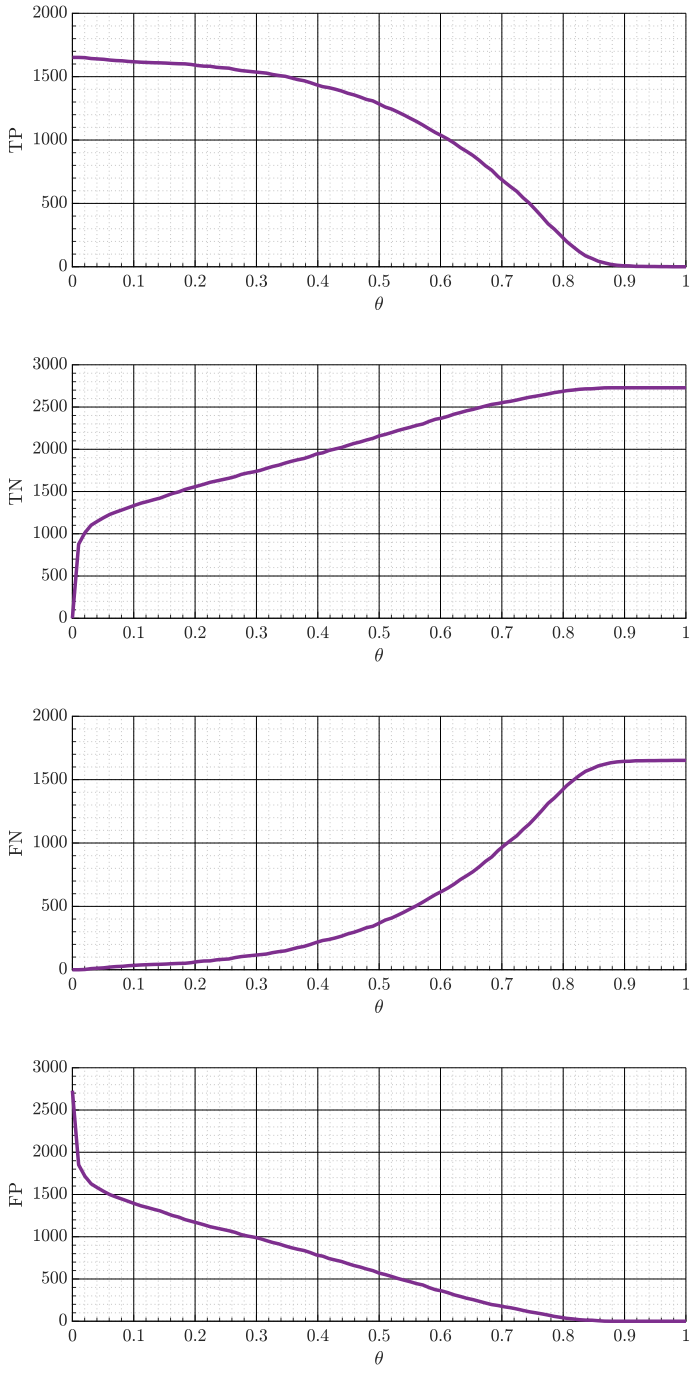

Figure 7: TP, TN, FN, and FP as functions of threshold value θ.

therefore see that the following point must lie on the ROC curve:

$$(\text{fpr}, \text{tpr}) = (\frac{FP}{N}, \frac{TP}{P}) = (0.21, 0.78)$$

this rules out all options except D.

Question 23. Which one of the following statements is true?

A. Suppose hold-out cross-validated backward selection is applied to select which features to include in a linear regression model. Each time a new model is selected by backward selection, the training error for that model will be smaller than (or equal to) the training error in the previous step.

B. Consider how Bagging and Boosting makes predictions in a binary classification task. Recall that both bagging and boosting train multiple classifiers on the same dataset. The only difference between the methods is how the training sets used to train the classifiers is sampled from the full dataset. Both sample the datasets with replacement, but bagging sample them uniformly, whereas Adaboost sample them according to weights which are iteratively updated.

C. In terms of a bias-variance trade-off, a logistic regression model with a well-tuned regularization parameter has a negligible bias but a fairly high variance.

D. When comparing two classifiers, leave-one-out cross-validation is a suitable cross-validation method to use in conjunction with McNemars test.

E. Don't know.

Solution 23. The correct answer is D: McNemars test can be used for any cross-validation procedure as long as the methods are tested on the same sets of observations. This is guaranteed for leave-one-out cross-validation. A is wrong because backward selection will remove features, and is thereby guaranteed to increase the training error as the models become less and less expressive. AdaBoost uses a weighted combination of classifiers and therefore B is wrong. Finally, regardless of regularization parameter, logistic regression is a fairly inflexible (and therefore biased) model type.

Question 24. Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In Figure 8 is given 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Each observation is colored and marked in terms of which cluster it came from in the Gaussian Mixture model.

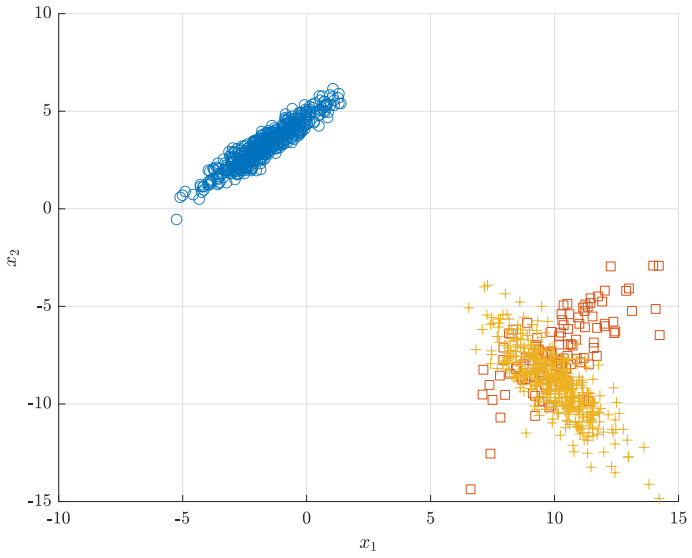

Figure 8: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

Which one of the following GMM densities was used to

generate the data?

A.

$$p(\boldsymbol{x}) = \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right) + \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right)$$

B.

$$p(\boldsymbol{x}) = \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right)$$

C.

$$p(\boldsymbol{x}) = \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right)$$

D.

$$p(\boldsymbol{x}) = \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right)$$

E. Don't know.

Solution 24.

D The three components in the candidate GMM densities can be matched to the colored observations by their mean values. Then, by considering the basic properties of the covariance matrices, we can easily rule out all options except A. Alternatively, in Figure 9 is shown the densities for densities corresponding to option B (upper left), C (upper right) and D (bottom center).

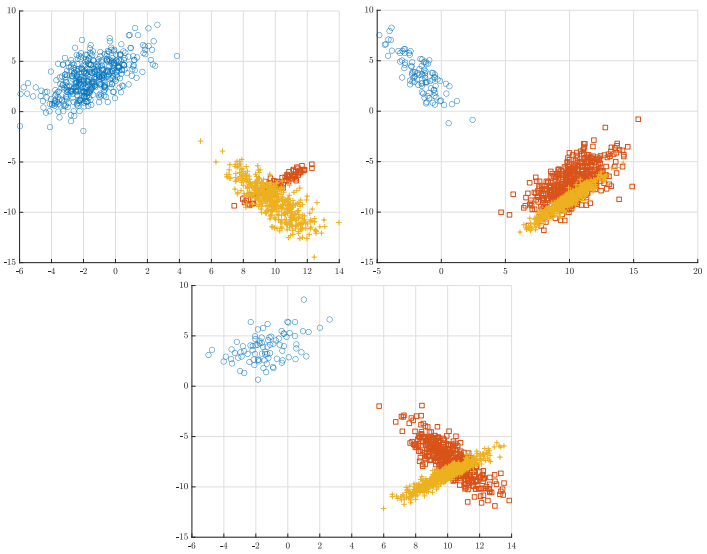

Figure 9: GMM mixtures corresponding to alternative options.

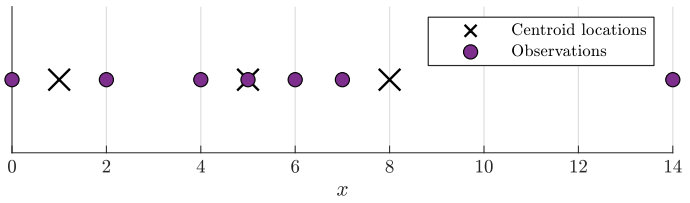

Figure 10: A small 1-dimensional dataset and initial values of centroids.

Question 25. Consider a small dataset comprised of $N = 7$ one-dimensional observations shown as the filled circles in Figure 10.

Suppose a k-means algorithm is applied to the dataset with $K = 3$ and using Euclidean distances. We will assume the location of the centroids are initialized to the values indicated by the crosses in Figure 10. After initialization, the k-means algorithm is evaluated for one step, comprised of assigning observations to centroids and updating the location of the centroids. After the first step, what will be the new location of the centroids?

A. $\mu_1 = 1$, $\mu_2 = \frac{11}{2}$, and $\mu_3 = 14$.

B. $\mu_1 = 4$, $\mu_2 = 6$, and $\mu_3 = 7$.

C. $\mu_1 = 1$, $\mu_2 = 5$, and $\mu_3 = \frac{21}{2}$.

D. $\mu_1 = 2$, $\mu_2 = \frac{11}{2}$, and $\mu_3 = \frac{21}{2}$.

E. Don't know.

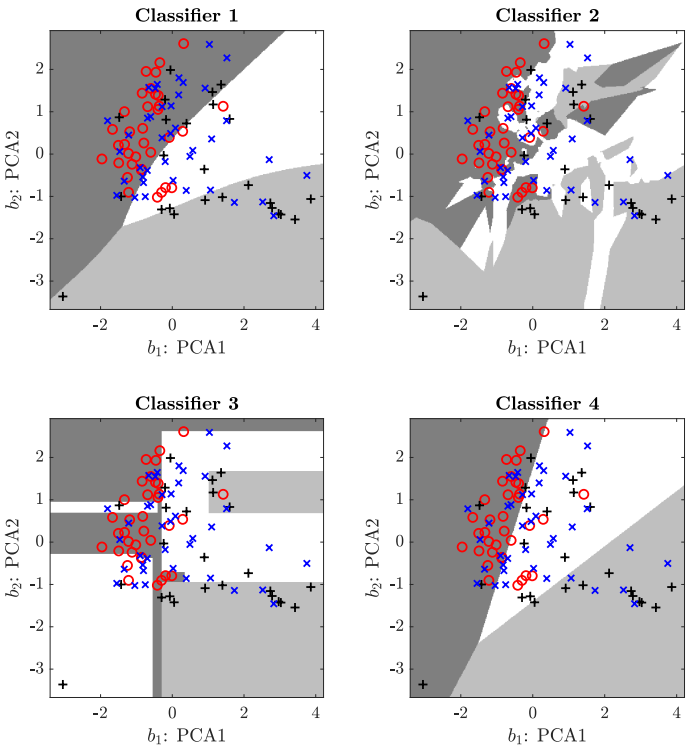

Figure 11: Decision boundaries for four different classifiers trained on the Bicycle rental dataset when projected onto the first two principal components.

Solution 25. The location of the observations and centroids is first read from Figure 10. When this is done, each observation is assigned to the nearest centroid. The observations are thereby partitioned into the three clusters $\{0, 2\}$, $\{4, 5, 6\}$, $\{7, 14\}$.

The new location of the centroids are simply the mean of the observation in each of these three sets. Doing this, we see C is the correct answer.

Question 26. We will consider a subset of the Bicycle rental dataset (described in Table 1) after it has been projected onto the first two principal components b_1 and b_2 given in Equation (1), thereby giving rise to a smaller two-dimensional dataset.

We will consider the following four classifiers:

MREG: Multinomial regression

ANN: Artificial neural network with 5 hidden units

CT: Classification tree with regular axis-aligned splits ($b_i < c$)

KNN: K-nearest neighbours with $K = 3$

Suppose the classifiers are trained on the two-dimensional dataset and the decision boundary for each of the four classifiers is given in Figure 11. Which one of the following statements is correct?

A. Classifier 1 corresponds to ANN,
Classifier 2 corresponds to KNN,
Classifier 3 corresponds to CT,
Classifier 4 corresponds to MREG.

B. Classifier 1 corresponds to **CT**,
Classifier 2 corresponds to **MREG**,
Classifier 3 corresponds to **KNN**,
Classifier 4 corresponds to **ANN**.

C. Classifier 1 corresponds to **MREG**,
Classifier 2 corresponds to **CT**,
Classifier 3 corresponds to **KNN**,
Classifier 4 corresponds to **ANN**.

D. Classifier 1 corresponds to **KNN**,
Classifier 2 corresponds to **ANN**,
Classifier 3 corresponds to **CT**,
Classifier 4 corresponds to **MREG**.

E. Don't know.

Solution 26. To solve this problem, we have to use our intuition about what the typical decision boundaries for the different methods look like:

- A KNN method will have decision boundaries dictated by the nearest neighbors. That is, points (x, y) where the nearest K neighbors are in one class must be in the same class and therefore the boundaries will be fairly complex and respect the data distribution well.
- A decision tree has axis aligned splits, therefore the boundaries must be vertical or horizontal
- A multivariate regression model must have linear boundaries
- An artificial neural network with few hidden units can have some non-linearity, but otherwise have boundaries of limited complexity and consisting of relatively simple shapes

It is easy to see this rules out all but option A.

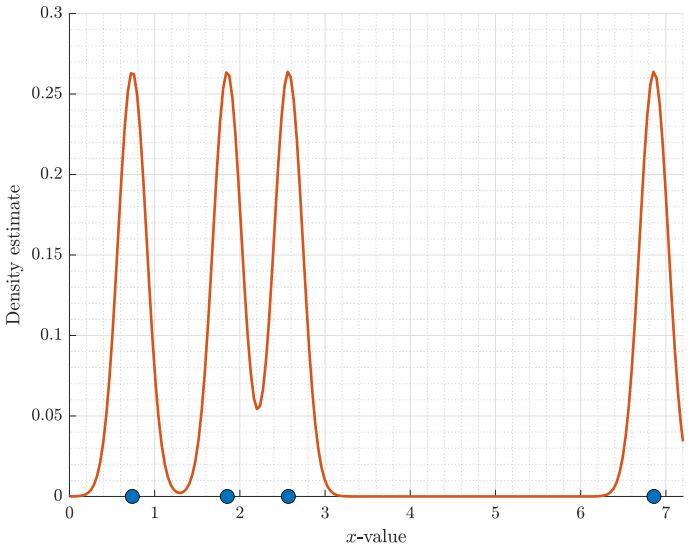

Figure 12: Plot of the density function of a kernel density estimator applied to a 1-dimensional dataset using a Gaussian kernel with kernel width $\lambda = 0.168$. Only a subset of the dataset, indicated by the circles, is shown.

Question 27. A small 1-dimensional dataset of N observations, along with the kernel density estimate, is shown in Figure 12. The kernel is the usual Gaussian kernel with kernel width $\lambda = 0.168$ (i.e., the individual Gaussian components in the KDE have variance $\sigma^2 = \lambda^2$). Note the x-axis has been truncated so not all observations are shown. How many observations were in the dataset?

A. $N = 9$

B. $N = 6$

C. $N = 21$

D. $N = 17$

E. Don't know.

Solution 27. In Figure 12 we see the density $p(x)$ of the KDE. Recall the general formula for a KDE:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}(x|x_i, \sigma^2).$$

Since the kernel width $\lambda = \sigma$ is fairly small relative to the component distance, the density at an x-value corresponding to the peak of one of the components, i.e. x_i, will only be determined by the density of that component (and none of the other ones). In other words we get:

$$\begin{aligned} p(x_i) &\approx \frac{1}{N}\mathcal{N}(x_i|x_i, \sigma^2) \\ &= \frac{1}{N}\frac{1}{\sqrt{2\pi}\sigma}. \end{aligned}$$

We can read off $p(x_i) \approx 0.26$ from the figure and solve to find $N = \frac{1}{\sqrt{2\pi}\sigma p(x)}$. By doing so we see that A is correct.

Technical University of Denmark

Written examination: December 18th 2018, 9 AM - 1 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.

Please hand in your answers using the electronic file. Only use this page in the case where digital handin is unavailable. In case you have to hand in the answers using the form on this sheet, please follow these instructions:

Print name and study number clearly. 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.

The individual questions are answered by filling in the answer fields with one of the letters A, B, C, D, or E.

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

USE ONLY THIS PAGE FOR HAND IN IF YOU ARE UNABLE TO HAND IN DIGITALLY.

No.	Attribute description	Abbrev.
x_1	intercolumnar distance	interdist
x_2	upper margin	upperm
x_3	lower margin	lowerm
x_4	exploitation	exploit
x_5	row number	row nr.
x_6	modular ratio	modular
x_7	interlinear spacing	interlin
x_8	weight	weight
x_9	peak number	peak nr.
x_{10}	modular ratio/ interlinear spacing	mr/is
y	Who copied the text?	Copyist

Table 1: Description of the features of the Avila Bible dataset used in this exam. The dataset has been extracted from images of the 'Avila Bible', an XII century giant Latin copy of the Bible. The prediction task consists in associating each pattern to one of three copyist (copyist refers to the monk who copied the text in the bible), indicated by the y-value. Note that only a subset of the dataset is used. The dataset used here consist of $N = 525$ observations and the attribute y is discrete taking values $y = 1,\ 2,\ 3$ corresponding to the three different copyists.

Question 1.

The main dataset used in this exam is the Avila Bible dataset[1] shown in Table 1.

In Figure 1 and Figure 2 are shown respectively percentile plots and boxplots of the Avila Bible dataset based on the attributes $x_2,\ x_3,\ x_9,\ x_{10}$ found in Table 1. Which percentile plots match which boxplots?

A. Boxplot 1 is mr/is, Boxplot 2 is lowerm, Boxplot 3 is upperm and Boxplot 4 is peak nr.

B. Boxplot 1 is upperm, Boxplot 2 is lowerm, Boxplot 3 is peak nr. and Boxplot 4 is mr/is

C. Boxplot 1 is upperm, Boxplot 2 is peak nr., Boxplot 3 is mr/is and Boxplot 4 is lowerm

D. Boxplot 1 is mr/is, Boxplot 2 is lowerm, Boxplot 3 is peak nr. and Boxplot 4 is upperm

E. Don't know.

[1]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Avila`

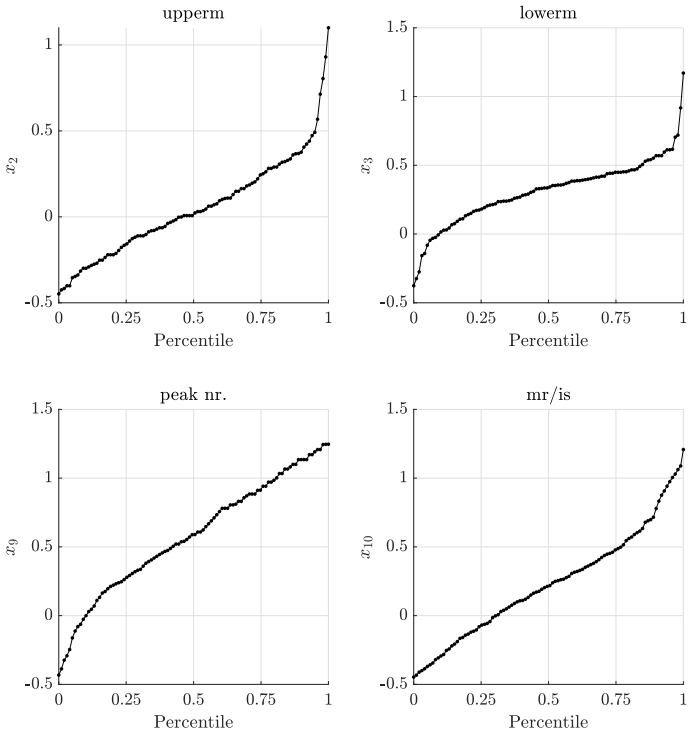

Figure 1: Plot of observations $x_2,\ x_3,\ x_9,\ x_{10}$ of the Avila Bible dataset of Table 1 as percentile plots.

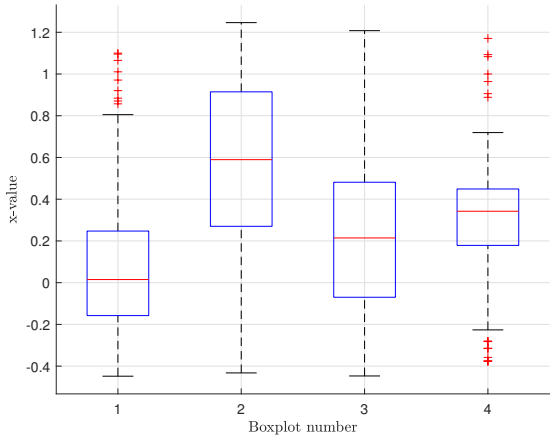

Figure 2: Boxplots corresponding to the variables plotted in Figure 1 but not necessarily in that order.

Solution 1. The correct answer is C. To see this, recall that by the definition of a boxplot the horizontal red line indicates the 50th percentile. We can read these of from the percentile plots by observing the values corresponding to 0.5. These are:

$$x_2 = 0.0, \quad x_3 = 0.3, \quad x_9 = 0.6, \quad x_{10} = 0.2.$$

In a similar manner, we know the upper-part of the box must correspond to the 75th percentile. These can also be read of from the percentile plots (the value corresponding to 0.75) and are:

$$x_2 = 0.2, \quad x_3 = 0.4, \quad x_9 = 0.9, \quad x_{10} = 0.5.$$

Taken together these rule out all but option C.

Question 2.

A Principal Component Analysis (PCA) is carried out on the Avila Bible dataset in Table 1 based on the attributes $x_1,\ x_3,\ x_5,\ x_6,\ x_7$.

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

$$\boldsymbol{V} = \begin{bmatrix} 0.04 & -0.12 & -0.14 & 0.35 & 0.92 \\ 0.06 & 0.13 & 0.05 & -0.92 & 0.37 \\ -0.03 & -0.98 & 0.08 & -0.16 & -0.05 \\ -0.99 & 0.03 & 0.06 & -0.02 & 0.07 \\ -0.07 & -0.05 & -0.98 & -0.11 & -0.11 \end{bmatrix} \tag{1}$$

$$\boldsymbol{S} = \begin{bmatrix} 14.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 8.19 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 7.83 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 6.91 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 6.01 \end{bmatrix}$$

Which one of the following statements is true?

A. The variance explained by the first principal component is greater than 0.45

B. The variance explained by the first four principal components is less than 0.85

C. The variance explained by the last four principal components is greater than 0.56

D. The variance explained by the first three principal components is less than 0.75

E. Don't know.

Solution 2. The correct answer is A. To see this, recall 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 components x_1 is:

$$\text{Var.Expl.} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_3^2 + \sigma_5^2 + \sigma_6^2 + \sigma_7^2} = 0.4942.$$

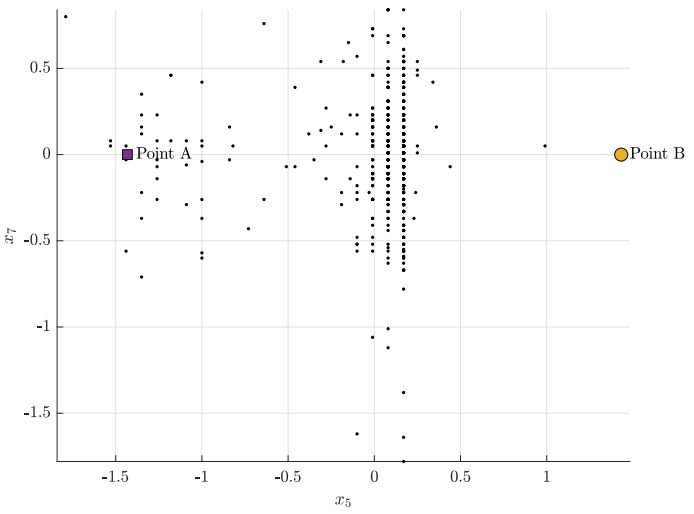

Figure 3: Black dots show attributes x_5 and x_7 of the Avila Bible dataset from Table 1. The two points corresponding to the colored markers indicate two specific observations A, B.

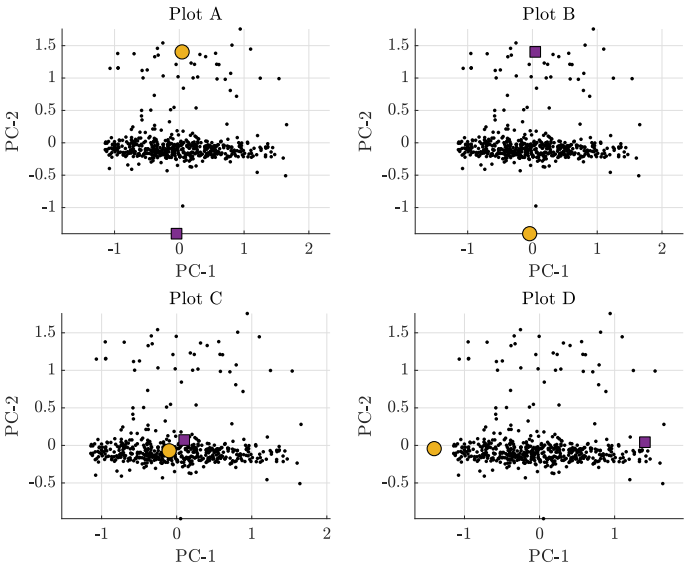

Figure 4: Candidate plots of the observations and path shown in Figure 3 projected onto the first two principal components considered in Equation (1). The colored markers still refer to points A and B, now in the coordinate system corresponding to the PCA projection.

Question 3.

Consider again the PCA analysis fo the Avila Bible dataset. In Figure 3 the features x_5 and x_7 from Table 1 are plotted as black dots. We have indicated two special observations as colored markers (Point A and Point B).

We can imagine that the dataset, along with the two special observations, is projected onto the first two principal component directions given in $\boldsymbol{V}$ as computed earlier (see Equation (1)). Which one of the four plots in Figure 4 shows the correct PCA projection?

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 3. Since we don't know the exact values of most of the x_i-coordinates, it is easier to work with the difference between observation A and B in Figure 3 and translate them into the difference in the PCA projections. Notice from Figure 3 we can immediately compute:

$$\Delta \boldsymbol{x} = \boldsymbol{x}_{\text{end}} - \boldsymbol{x}_{\text{start}} = \begin{bmatrix} 0.0 \\ 0.0 \\ 2.86 \\ 0.0 \\ 0.0 \end{bmatrix}$$

(this corresponds to the vector going from Point A to Point B). Then, all we need is to compute the PCA projection of this vector as:

$$\Delta \boldsymbol{b} = \left((\Delta \boldsymbol{x})^\top \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \right)^\top = \begin{bmatrix} -0.09 \\ -2.8 \end{bmatrix}$$

Which should be the vector beginning at Point A and terminating at B in the PCA projected plots. This rules out all plots except option B.

Question 4. To examine if observation o_4 may be an outlier, we will calculate the average relative density based on euclidean distance and the observations given in Table 2 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_4 for $K = 2$ nearest neighbors?

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	2.91	0.63	1.88	1.02	1.82	1.92	1.58	1.08	1.43
o_2	2.91	0.0	3.23	3.9	2.88	3.27	3.48	4.02	3.08	3.47
o_3	0.63	3.23	0.0	2.03	1.06	2.15	2.11	1.15	1.09	1.65
o_4	1.88	3.9	2.03	0.0	2.52	1.04	2.25	2.42	2.18	2.17
o_5	1.02	2.88	1.06	2.52	0.0	2.44	2.38	1.53	1.71	1.94
o_6	1.82	3.27	2.15	1.04	2.44	0.0	1.93	2.72	1.98	1.8
o_7	1.92	3.48	2.11	2.25	2.38	1.93	0.0	2.53	2.09	1.66
o_8	1.58	4.02	1.15	2.42	1.53	2.72	2.53	0.0	1.68	2.06
o_9	1.08	3.08	1.09	2.18	1.71	1.98	2.09	1.68	0.0	1.48
o_{10}	1.43	3.47	1.65	2.17	1.94	1.8	1.66	2.06	1.48	0.0

Table 2: The pairwise Euclidian distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 10 observations from the Avila Bible dataset (recall $M = 10$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1 (the data has been standardized). The colors indicate classes such that the black observations $\{o_1,\ o_2,\ o_3\}$ belongs to class C_1 (corresponding to copyist one), the red observations $\{o_4,\ o_5,\ o_6,\ o_7,\ o_8\}$ belongs to class C_2 (corresponding to copyist two), and the blue observations $\{o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to copyist three).

A. 1.0

B. 0.71

C. 0.68

D. 0.36

E. Don't know.

Solution 4.

To solve the problem, first observe the $k = 2$ neighborhood of o_4 and density is:

$$N_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = \{o_6, o_1\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = 0.685$$

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

$$N_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = \{o_4, o_{10}\}, \quad N_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = \{o_3, o_5\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = 0.704, \text{density}_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = 1.212.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem.

Question 5.

Suppose a GMM model is applied to the Avila Bible dataset in the processed version shown in Table 2. The GMM is constructed as having $K = 3$ components, and each component k of the GMM is fitted by letting it's mean vectors $\boldsymbol{\mu}_k$ be equal to the location of the observations:

$$o_7, \; o_8, \; o_9$$

(i.e. each observation corresponds to exactly one mean vector) and setting the covariance matrix equal to $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix:

$$\mathcal{N}(\boldsymbol{o}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} e^{\frac{-d(\boldsymbol{o}_i, \boldsymbol{\mu}_k)^2}{2\sigma^2}}$$

where $|\cdot|$ is the determinant. The components of the GMM are weighted evenly.

If $\sigma = 0.5$, and denoting the density of the GMM as $p(\boldsymbol{x})$, what is the density as evaluated at observation o_3?

A. $p(o_3) = 0.048402$

B. $p(o_3) = 0.076$

C. $p(o_3) = 0.005718$

D. $p(o_3) = 0.114084$

E. Don't know.

Solution 5.

Since the mixture components are weighted equally, the density of the test observation becomes:

$$p(\boldsymbol{o}_i) = \sum_{k=1}^{3} \frac{1}{3} \mathcal{N}(\boldsymbol{o}_i | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}).$$

To correctly evaluate this density, we also need to know the dimensionality of the multivariate normal distributions. This can be found in Table 2 to be $M = 10$. The density of a single mixture component is therefore:

$$\mathcal{N}(\boldsymbol{o}_i | \boldsymbol{o}_j, \sigma^2 \boldsymbol{I}) = \frac{1}{(2\pi\sigma^2)^{\frac{10}{2}}} e^{-\frac{d(\boldsymbol{o}_i, \boldsymbol{o}_j)^2}{2\sigma^2}}.$$

where the distances can be found in Table 2. Plugging in the values we see option C is correct.

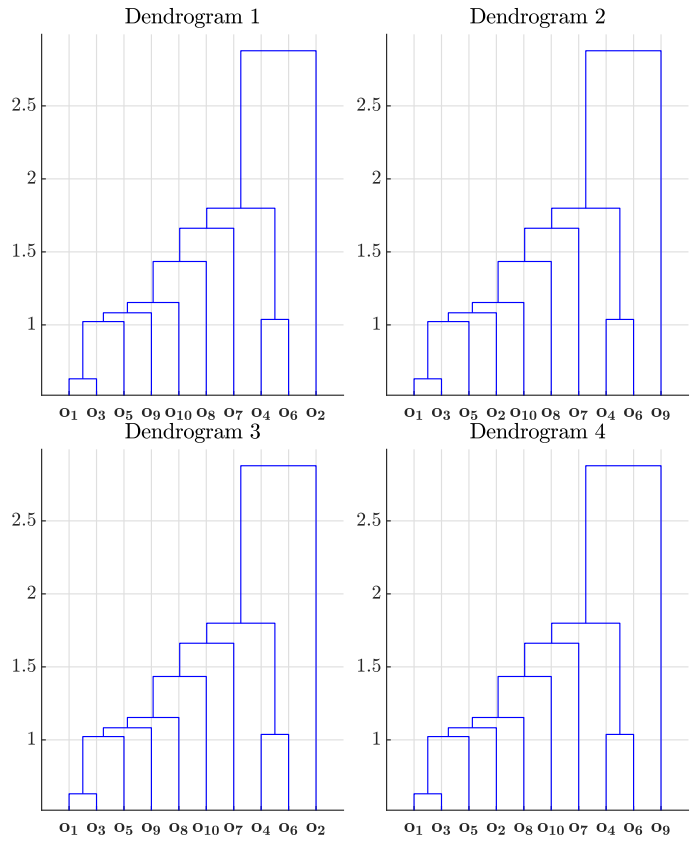

Figure 5: Proposed hierarchical clustering of the 10 observations in Table 2.

Question 6. A hierarchical clustering is applied to the 10 observations in Table 2 using *minimum* linkage. Which of the dendrograms shown in Figure 5 corresponds to the clustering?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

Solution 6. The correct solution is C. We can rule out the other solutions by observing the first merge operation at which they diverge from the correct solution.

- In dendrogram 1, merge operation number 5 should have been between the sets $\{f_8\}$ and $\{f_9, f_5, f_1, f_3\}$ at a height of 1.15, however in dendrogram 1 merge number 5 is between the sets $\{f_{10}\}$ and $\{f_9, f_5, f_1, f_3\}$.
- In dendrogram 2, merge operation number 4 should have been between the sets $\{f_9\}$ and $\{f_5, f_1, f_3\}$ at a height of 1.08, however in dendrogram 2 merge number 4 is between the sets $\{f_2\}$ and $\{f_5, f_1, f_3\}$.
- In dendrogram 4, merge operation number 4 should have been between the sets $\{f_9\}$ and $\{f_5, f_1, f_3\}$ at a height of 1.08, however in dendrogram 4 merge number 4 is between the sets $\{f_2\}$ and $\{f_5, f_1, f_3\}$.

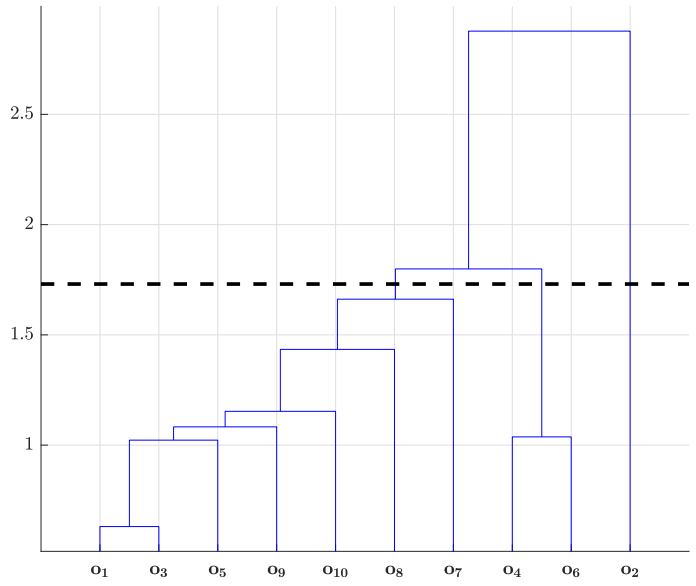

Figure 6: Dendrogram 1 from Figure 5 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

Question 7.

Consider dendrogram 1 from Figure 5. Suppose we apply a cutoff (indicated by the black line) thereby generating three clusters. We wish to compare the quality of this clustering, Q, to the ground-truth clustering, Z, indicated by the colors in Table 2. Recall the *normalized mutual information* of the two clusterings Z and Q is defined as

$$\text{NMI}[Z, Q] = \frac{\text{MI}[Z, Q]}{\sqrt{H[Z]}\sqrt{H[Q]}}$$

where MI is the *mutual information* and H is the entropy. Assuming we always use an entropy based on the natural logarithm,

$$H = -\sum_{i=1}^{n} p_i \log p_i, \quad \log(e) = 1,$$

what is the normalized mutual information of the two clusterings?

A. $\text{NMI}[Z, Q] \approx 0.313$

B. $\text{NMI}[Z, Q] \approx 0.302$

C. $\text{NMI}[Z, Q] \approx 0.32$

D. $\text{NMI}[Z, Q] \approx 0.274$

E. Don't know.

Solution 7. To compute the MI, we will use the relation

$$\text{MI}[p, q] = H[P] + H[Q] - H[P, Q]$$

Where P is the clustering corresponding to the colors in Table 2 and Q the clustering obtained by cutting the dendrogram in Figure 6:

$$\{4, 6\},\ \{1, 3, 5, 7, 8, 9, 10\},\ \{2\}$$

From this information we can define the matrix of probabilities $p(i, j)$ such that

$$p(i, j) = \frac{\text{Observations in cluster } i \text{ in } P \text{ and } j \text{ in } Q}{N}$$
$$= \frac{1}{N}\begin{bmatrix} 0 & 2 & 1 \\ 2 & 3 & 0 \\ 0 & 2 & 0 \end{bmatrix}$$

From these we can define the probabilities corresponding to the clustering of P and Q as: $p_P(i) = \sum_j p(i, j)$, $p_Q(j) = \sum_i p(i, j)$. The mutual information is then

$$\begin{aligned} \text{MI} &= H[P] + H[Q] - H[P, Q] \\ &= 0.802 + 1.03 - 1.557 \\ &= 0.274. \end{aligned}$$

We can then simply use the equation for the NMI given in the problem to see answer B is correct.

Question 8. Consider the distances in Table 2 based on 10 observations from the Avila Bible dataset. The class labels C_1, C_2, C_3 (see table caption for details) will be predicted using a k-nearest neighbour classifier based on the distances given in Table 2. Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 1-nearest neighbour classifier (i.e. $k = 1$). What is the error rate computed for all $N = 10$ observations?

A. error rate $= \frac{4}{10}$

B. error rate $= \frac{9}{10}$

C. error rate $= \frac{2}{10}$

D. error rate $= \frac{6}{10}$

E. Don't know.

Solution 8.

The correct answer is D. To see this, recall that leave-one-out cross-validation means we train a total of $N = 10$ models, each model being tested on a single observation and trained on the remaining such that each observation is used for testing exactly once.

The model considered is KNN classifier with $k = 1$. To figure out the error for a particular observation i (i.e. the test set for this fold), we train a model on the other observations and predict on observation i. To do that, simply find the observation different than i closest to i according to Table 2 and predict i as belonging to it's class. Concretely, we find: $N(o_i, k) = \{o_3\}$, $N(o_i, k) = \{o_5\}$, $N(o_i, k) = \{o_1\}$, $N(o_i, k) = \{o_6\}$, $N(o_i, k) = \{o_1\}$, $N(o_i, k) = \{o_4\}$, $N(o_i, k) = \{o_{10}\}$, $N(o_i, k) = \{o_3\}$, $N(o_i, k) = \{o_1\}$, and $N(o_i, k) = \{o_1\}$.

The error is then found by observing how often the class label of the observation in the neighborhood agrees with the true class label. We find this happens for observations

$$\{o_1, o_3, o_4, o_6\}$$

and the remaining observations are therefore erroneously classified, in other words, the classification error is $\frac{6}{10}$.

Question 9.

Suppose we wish to build a classification tree based on Hunt's algorithm where the goal is to predict Copyist which can belong to three classes, $y = 1$, $y = 2$, $y = 3$. The first split we consider is a two-way split based on the value of x_9 into the intervals indicated in Table 3. For each interval, we count how many observations belong to each of the three classes and the result is indicated in Table 3. Suppose we use the *classification error* impurity measure, what is then the purity gain Δ?

x_9-interval	$y = 1$	$y = 2$	$y = 3$
$x_9 \leq 0.13$	108	112	56
$0.13 < x_9$	58	75	116

Table 3: Proposed split of the Avila Bible dataset based on the attribute x_9. We consider a 2-way split where for each interval we count how many observations belonging to that interval has the given class label.

A. $\Delta \approx 0.485$

B. $\Delta \approx 0.078$

C. $\Delta \approx 0.566$

D. $\Delta \approx 1.128$

E. Don't know.

Solution 9. Recall the information gain Δ is given as:

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

These quantities are easiest computed by forming the matrix R_{ki}, defined as the number of observations in split k belonging to class i:

$$R = \begin{bmatrix} 108 & 112 & 56 \\ 58 & 75 & 116 \end{bmatrix}.$$

We obtain $N(r) = \sum_{ki} R_{ki} = 525$ as the total number of observations and the number of observations in each branch is simply:

$$N(v_k) = \sum_i R_{ki}.$$

Next, the impurities $I(v_k)$ is computed from the probabilities

$$p_i = \frac{R_{ki}}{N(v_k)}$$

and the impurity I_0 from

$$p_i = \frac{\sum_k R_{ki}}{N(r)}.$$

In particular we obtain:

$$I_0 = 0.644, I(v_1) = 0.594,\ I(v_2) = 0.534.$$

Combining these we see that $\Delta = 0.078$ and therefore option B is correct.

Question 10. Consider the split in Table 3. Suppose we build a classification tree with *only* this split and evaluate it on the same data it was trained on. What is the accuracy?

A. Accuracy is: 0.64

B. Accuracy is: 0.29

C. Accuracy is: 0.35

D. Accuracy is: 0.43

E. Don't know.

Solution 10. We will first form the matrix R_{ki}, defined as the number of observations in split k belonging to class i:

$$R = \begin{bmatrix} 108 & 112 & 56 \\ 58 & 75 & 116 \end{bmatrix}.$$

From this we obtain $N = \sum_{ki} R_{ki} = 525$ as the total number of observations. For each split, the number of observations in the largest classes, n_k, is:

$$n_1 = \max_i R_{ik} = 112,\ n_2 = \max_i R_{ik} = 116.$$

Therefore, the accuracy is:

$$\text{Accuracy:} \quad \frac{112 + 116}{525}$$

and answer D is correct.

Question 11. Suppose s_1 and s_2 are two text documents containing the text:

$$s_1 = \begin{Bmatrix} \text{the bag of words representation} \\ \text{should not give you a hard time} \end{Bmatrix}$$
$$s_2 = \begin{Bmatrix} \text{remember the representation should} \\ \text{be a vector} \end{Bmatrix}$$

The documents are encoded using a bag-of-words encoding assuming a total vocabulary size of $M = 10000$. No stopwords lists or stemming is applied to the dataset. What is the cosine similarity between documents s_1 and s_2?

A. cosine similarity of s_1 and s_2 is 0.047619

B. cosine similarity of s_1 and s_2 is 0.000044

C. cosine similarity of s_1 and s_2 is 0.000400

D. cosine similarity of s_1 and s_2 is 0.436436

E. Don't know.

Solution 11. The correct answer is D. Since we are computing the cosine similarity, the length of the vocabulary is irrelevant. We then observe that document s_1 contains $n_1 = 12$ unique words and document s_2 contains $n_2 = 7$ unique words, and the two documents have $f_{11} = 4$ words in common. The cosine similarity is therefore:

$$\cos(s_1, s_2) = \frac{f_{11}}{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|} = \frac{f_{11}}{\sqrt{n_1}\sqrt{n_2}} \approx 0.44.$$

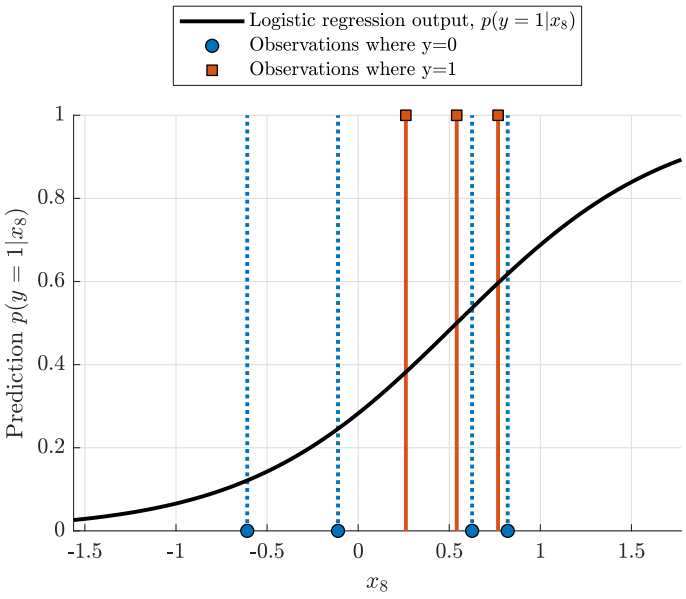

Figure 7: Output of a logistic regression classifier trained on 7 observations from the dataset.

Question 12. Consider again the Avila Bible dataset. We are particularly interested in predicting whether a bible copy was written by copyist 1, and we therefore wish to train a logistic regression classifier to distinguish between copyist one vs. copyist two and three.

To simplify the setup further, we select just 7 observations and train a logistic regression classifier using only the feature x_8 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term). To be consistent with the lecture notes, we label the output as $y = 0$ (corresponding to copyist one) and $y = 1$ (corresponding to copyist two and three).

In Figure 7 is shown the predicted output probability an observation belongs to the positive class, $p(y = 1|x_8)$. What are the weights?

A. $\begin{bmatrix} -0.93 \\ 1.72 \end{bmatrix}$

B. $\begin{bmatrix} -2.82 \\ 0.0 \end{bmatrix}$

C. $\begin{bmatrix} 1.36 \\ 0.4 \end{bmatrix}$

D. $\begin{bmatrix} -0.65 \\ 0.0 \end{bmatrix}$

E. Don't know.

Solution 12. The solution is easily found by simply computing the predicted $\hat{y} = p(y = 1|x_8)$-value for an appropriate choice of x_8. Notice that

$$p(y = 1|x_8) = \sigma(\tilde{\boldsymbol{x}}_8^T \boldsymbol{w})$$

If we select $x_8 = 1$ and select the weights as in option A we find $p(y = 1|x_8) = 0.69$, in good agreement with the figure. On the other hand, for the weights in option C we obtain $\hat{y} = 0.85$, for D that $\hat{y} = 0.34$ and finally for B that $\hat{y} = 0.06$. We can therefore conclude that A is correct.

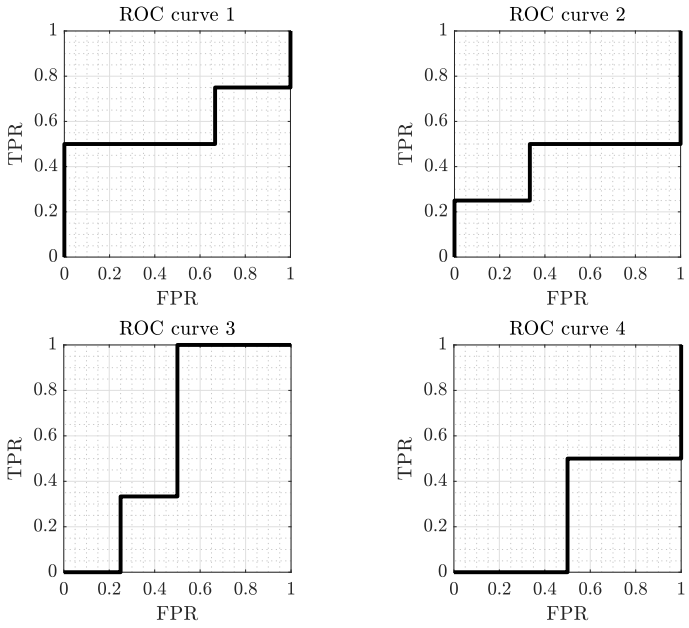

Figure 8: Proposed ROC curves for the logistic regression classifier in Figure 7.

Question 13.

To evaluate the classifier Figure 7, we will use the *area under curve* (AUC) of the *reciever operator characteristic* (ROC) curve as computed on the 7 observations in Figure 7. In Figure 8 is given four proposed ROC curves, which one of the curves corresponds to the classifier?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

Solution 13. To compute the AUC, we need to compute 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

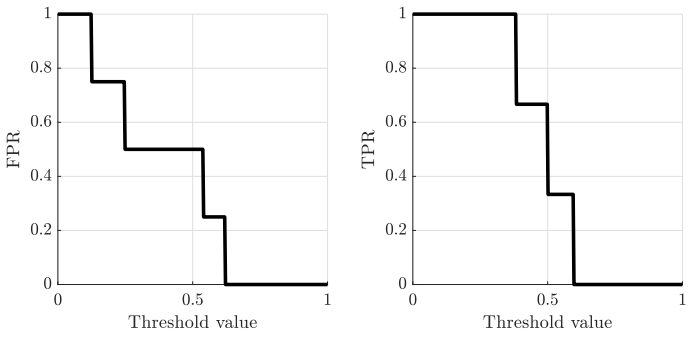

Figure 9: TPR, FPR curves for the logistic regression classifier in Figure 7.

but in fact belongs to class 0, divided by the total number of observations in the *negative* class.

This procedure is then repeated for different threshold values to obtain the curves shown in Figure 9. The ROC curve is then obtained by plotting these two curves against each other. I.e. for each threshold value, the point

$$(x, y) = (\text{FPR}, \text{TPR})$$

is on the AUC curve. This rules out all options except C.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
o_1	1	1	0	0	0	1	0	0	0	1
o_2	1	0	0	0	0	0	0	0	0	0
o_3	1	1	0	0	0	1	0	0	0	1
o_4	0	1	1	1	0	0	0	1	1	0
o_5	1	1	0	0	0	1	0	0	0	1
o_6	0	1	1	1	0	0	1	1	1	0
o_7	1	1	1	0	0	1	1	1	1	0
o_8	0	1	1	1	0	1	1	0	0	1
o_9	0	0	0	0	1	1	1	0	1	1
o_{10}	1	0	0	0	0	1	1	1	1	0

Table 4: Binarized version of the Avila Bible dataset. Each of the features f_i are obtained by taking a feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$).The colors indicate classes such that the black observations $\{o_1,\ o_2,\ o_3\}$ belongs to class C_1 (corresponding to copyist one), the red observations $\{o_4,\ o_5,\ o_6,\ o_7,\ o_8\}$ belongs to class C_2 (corresponding to copyist two), and the blue observations $\{o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to copyist three).

Question 14. We again consider the Avila Bible dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce 10 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 $N \times M = 10 \times 10$ binary matrix in Table 4. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_1,\ f_2,\ f_6$. If for an observations we observe

$$f_1 = 1,\ f_2 = 1,\ f_6 = 0$$

what is then the probability that $y = 1$ according to the Naïve-Bayes classifier?

A. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{50}{77}$

B. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{25}{43}$

C. $p_{\mathbf{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{5}{11}$

D. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{10}{19}$

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

Solution 14. To solve this problem, we simply use the general form of the naïve-Bayes approximation and plug in the relevant numbers. We get:

$$p_{\text{NB}}(y=1|f_1=1,\ f_2=1,\ f_6=0) =$$
$$\frac{p(f_1=1|y=1)p(f_2=1|y=1)p(f_6=0|y=1)p(y=1)}{\sum_{j=1}^{3} p(f_1=1|y=j)p(f_2=1|y=j)p(f_6=0|y=j)p(y=j)}$$
$$= \frac{\frac{1}{1}\frac{2}{3}\frac{1}{3}\frac{3}{10}}{\frac{1}{1}\frac{2}{3}\frac{1}{3}\frac{3}{10} + \frac{2}{5}\frac{1}{1}\frac{2}{5}\frac{1}{2} + \frac{1}{2}\frac{0}{1}\frac{0}{1}\frac{1}{5}}$$
$$= \frac{5}{11}.$$

Therefore, answer C is correct.

Question 15.

Consider the binarized version of the Avila Bible dataset shown in Table 4.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 10$ items $f_1, f_2, \ldots, f_{10}$. Which of the following options represents all (non-empty) itemsets with support greater than 0.55 (and only itemsets with support greater than 0.55)?

A. $\{f_1\}$, $\{f_2\}$, $\{f_6\}$, $\{f_7\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_6, f_{10}\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_6\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_6\}$, $\{f_7\}$, $\{f_8\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_2\}$, $\{f_2, f_3\}$, $\{f_2, f_4\}$, $\{f_3, f_4\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_2, f_7\}$, $\{f_3, f_7\}$, $\{f_6, f_7\}$, $\{f_2, f_8\}$, $\{f_3, f_8\}$, $\{f_7, f_8\}$, $\{f_2, f_9\}$, $\{f_3, f_9\}$, $\{f_6, f_9\}$, $\{f_7, f_9\}$, $\{f_8, f_9\}$, $\{f_1, f_{10}\}$, $\{f_2, f_{10}\}$, $\{f_6, f_{10}\}$, $\{f_2, f_3, f_4\}$, $\{f_1, f_2, f_6\}$, $\{f_2, f_3, f_7\}$, $\{f_2, f_3, f_8\}$, $\{f_2, f_3, f_9\}$, $\{f_6, f_7, f_9\}$, $\{f_2, f_8, f_9\}$, $\{f_3, f_8, f_9\}$, $\{f_7, f_8, f_9\}$, $\{f_1, f_2, f_{10}\}$, $\{f_1, f_6, f_{10}\}$, $\{f_2, f_6, f_{10}\}$, $\{f_2, f_3, f_8, f_9\}$, $\{f_1, f_2, f_6, f_{10}\}$

D. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_6\}$, $\{f_7\}$, $\{f_8\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_2\}$, $\{f_2, f_3\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_6, f_7\}$, $\{f_7, f_9\}$, $\{f_8, f_9\}$, $\{f_2, f_{10}\}$, $\{f_6, f_{10}\}$, $\{f_1, f_2, f_6\}$, $\{f_2, f_6, f_{10}\}$

E. Don't know.

Solution 15. 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 0.55, an itemset needs to be contained in 6 rows. It is easy to see this rules out all options except *B*.

Question 16. We again consider the binary matrix from Table 4 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 10$ items $f_1, \ldots, f_{10}$.

What is the *confidence* of the rule $\{f_1, f_3, f_8, f_9\} \to \{f_2, f_6, f_7\}$

A. Confidence is $\frac{1}{10}$

B. Confidence is 1

C. Confidence is $\frac{1}{2}$

D. Confidence is $\frac{3}{20}$

E. Don't know.

Solution 16. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_1, f_3, f_8, f_9\} \cup \{f_2, f_6, f_7\})}{\text{support}(\{f_1, f_3, f_8, f_9\})} = \frac{\frac{1}{10}}{\frac{1}{10}} = 1.$$

Therefore, answer B is correct.

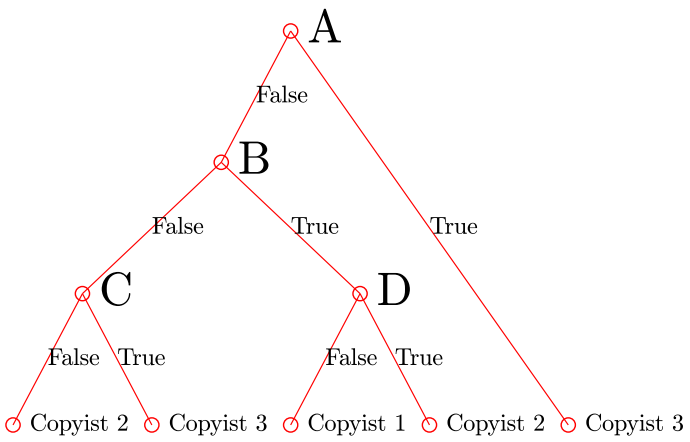

Figure 10: Example classification tree.

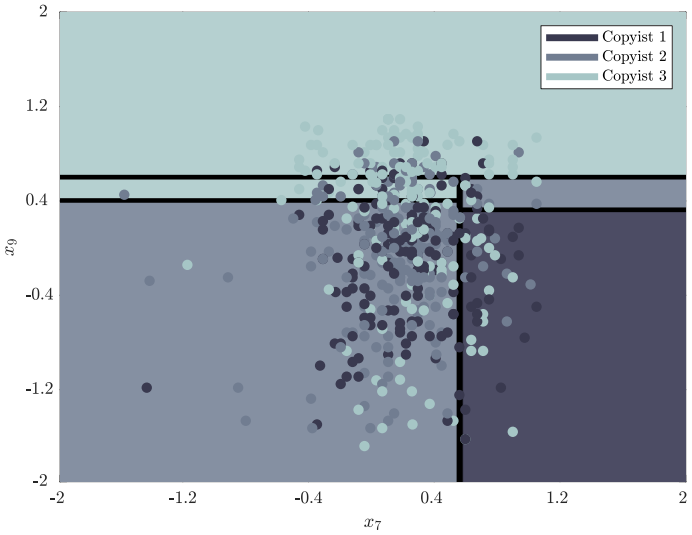

Figure 11: classification boundary.

Question 17.

Consider again the Avila Bible dataset. Suppose we train a decision tree to classify which of the 3 classes, Copyist 1, Copyist 2, Copyist 3, an observation belongs to. Since the attributes of the dataset are continuous, we will consider binary splits of the form $x_i \geq z$ for different values of i and z, and for simplicity we limit ourselves to the attributes x_7 and x_9. Suppose the trained decision tree has the form shown in Figure 10, and that according to the tree the predicted label assignment for the $N = 525$ observations are as given in Figure 11, what is then the correct rule assignment

to the nodes in the decision tree?

A. ***A***: $x_7 \geq 0.5$, ***B***: $x_9 \geq 0.54$, ***C***: $x_9 \geq 0.35$, ***D***: $x_9 \geq 0.26$

B. ***A***: $x_7 \geq 0.5$, ***B***: $x_9 \geq 0.26$, ***C***: $x_9 \geq 0.54$, ***D***: $x_9 \geq 0.35$

C. *A*: $x_9 \geq 0.54$, *B*: $x_7 \geq 0.5$, *C*: $x_9 \geq 0.35$, *D*: $x_9 \geq 0.26$

D. ***A***: $x_9 \geq 0.26$, ***B***: $x_7 \geq 0.5$, ***C***: $x_9 \geq 0.35$, ***D***: $x_9 \geq 0.54$

E. Don't know.

Solution 17.

This problem is solved by using the definition of a decision tree and observing what classification rule each of the assignment of features to node names in the decision tree will result in. I.e. beginning at the top of the tree, check if the condition assigned to the node is met and proceed along the true or false leg of the tree.

The resulting decision boundaries for each of the options are shown in Figure 12 and it follows answer C is correct.

Question 18. We will again consider the binarized version of the Avila Bible dataset already encountered in Table 4, however we will now only consider the first $M = 6$ features f_1, f_2, f_3, f_4, f_5, f_6.

We wish to apply the Apriori algorithm (the specific variant encountered in chapter 19 of the lecture notes) to find all itemsets with support greater than $\varepsilon = 0.15$. Suppose at iteration $k = 3$ we know that:

$$L_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Recall the key step in the Apriori algorithm is to construct L_3 by first considering a large number of candidate itemsets C'_3, and then rule out some of them using the downwards-closure principle thereby saving many (potentially costly) evaluations of support. Suppose L_2 is given as above, which of the following itemsets does the Apriori algorithm *not* have to evaluate the support of?

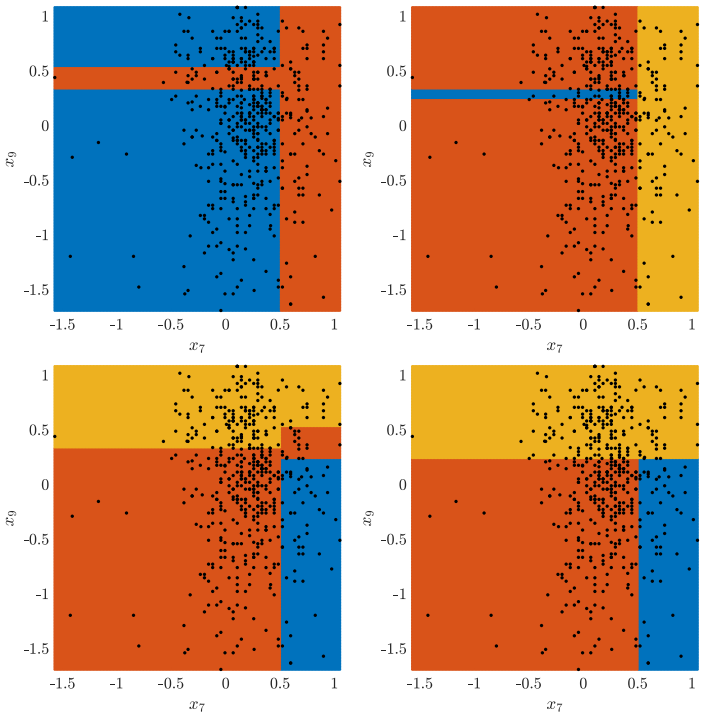

Figure 12: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

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

B. $\{f_1, f_2, f_6\}$

C. $\{f_2, f_3, f_6\}$

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

E. Don't know.

Solution 18. Recall the Apriori algorithm obtain L_3 from L_2 in three steps. First, the Apriori algorithm construct C'_3 by, for each itemset I in L_2, loop over all items not already in I and consider all such combinations where I is enlarged by a single item as a candidate

itemset in C_3'. Specifically we get:

$$C_3' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The downwards closure principle is then applied by removing and itemset I in C_3' if I contains a subset of 2 items not found in L_2. We thereby get:

$$C_3 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Finally, L_3 is constructed from C_3 by removing those itemsets with a support lower than ε. Thus, the itemsets we don't have to compute support from are those itemsets found in C_3' but not in C_3, or as an even simpler criteria, those which have a subset of size 2 not found in L_2. This rules out all options except D.

Question 19.

Consider again the Avila Bible dataset in Table 1. We would like to predict the copyist using a linear regression, 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 5 features $x_1,\ x_5,\ x_6,\ x_8,\ x_9$ and in Table 5 we have pre-computed the estimated training and test error for different variable combinations of the dataset. Which of the following statements is correct?

A. Backward selection will select attributes x_1

B. Backward selection will select attributes $x_1,\ x_5,\ x_6,\ x_8$

C. Forward selection will select attributes $x_1,\ x_8$

D. Forward selection will select attributes $x_1,\ x_5,\ x_6,\ x_8$

E. Don't know.

Solution 19.

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

Forward selection: The method is initialized with the set $\{\}$ having an error of 4.163.

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 $\{x_1\}$, $\{x_5\}$, $\{x_6\}$, $\{x_8\}$, $\{x_9\}$. Since the lowest error of the available sets is 3.252, which is lower than 4.163, we update the current selected variables to $\{x_1\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_5\}$, $\{x_1, x_6\}$, $\{x_5, x_6\}$, $\{x_1, x_8\}$, $\{x_5, x_8\}$, $\{x_6, x_8\}$, $\{x_1, x_9\}$, $\{x_5, x_9\}$, $\{x_6, x_9\}$, $\{x_8, x_9\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Backward selection: The method is initialized with the set $\{x_1, x_5, x_6, x_8, x_9\}$ having an error of 5.766.

Feature(s)	Training RMSE	Test RMSE
none	3.429	4.163
x_1	3.043	3.252
x_5	3.303	4.52
x_6	3.424	4.274
x_8	3.399	4.429
x_9	2.866	5.016
$x_1,\ x_5$	3.001	3.44
$x_1,\ x_6$	3.031	3.423
$x_5,\ x_6$	3.297	4.641
$x_1,\ x_8$	3.017	3.42
$x_5,\ x_8$	3.299	4.485
$x_6,\ x_8$	3.396	4.519
$x_1,\ x_9$	2.644	4.267
$x_5,\ x_9$	2.645	5.495
$x_6,\ x_9$	2.787	5.956
$x_8,\ x_9$	2.71	5.536
$x_1,\ x_5,\ x_6$	2.988	3.607
$x_1,\ x_5,\ x_8$	3.0	3.453
$x_1,\ x_6,\ x_8$	3.007	3.574
$x_5,\ x_6,\ x_8$	3.292	4.61
$x_1,\ x_5,\ x_9$	2.523	4.704
$x_1,\ x_6,\ x_9$	2.562	5.184
$x_5,\ x_6,\ x_9$	2.544	6.552
$x_1,\ x_8,\ x_9$	2.517	4.686
$x_5,\ x_8,\ x_9$	2.628	5.532
$x_6,\ x_8,\ x_9$	2.629	6.569
$x_1,\ x_5,\ x_6,\ x_8$	2.988	3.614
$x_1,\ x_5,\ x_6,\ x_9$	2.425	5.725
$x_1,\ x_5,\ x_8,\ x_9$	2.491	4.734
$x_1,\ x_6,\ x_8,\ x_9$	2.433	5.687
$x_5,\ x_6,\ x_8,\ x_9$	2.53	6.597
$x_1,\ x_5,\ x_6,\ x_8,\ x_9$	2.398	5.766

Table 5: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y in the avila dataset using different combinations of the features $x_1,\ x_5,\ x_6,\ x_8,\ x_9$.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_5, x_6, x_8, x_9\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_5, x_6, x_8\}$, $\{x_1, x_5, x_6, x_9\}$, $\{x_1, x_5, x_8, x_9\}$, $\{x_1, x_6, x_8, x_9\}$, $\{x_5, x_6, x_8, x_9\}$. Since the lowest error of the available sets is 3.614, which is lower than 5.766, we update the current selected variables to $\{x_1, x_5, x_6, x_8\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_5, x_6, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_5, x_6\}$, $\{x_1, x_5, x_8\}$, $\{x_1, x_6, x_8\}$, $\{x_5, x_6, x_8\}$, $\{x_1, x_5, x_9\}$, $\{x_1, x_6, x_9\}$, $\{x_5, x_6, x_9\}$, $\{x_1, x_8, x_9\}$, $\{x_5, x_8, x_9\}$, $\{x_6, x_8, x_9\}$. Since the lowest error of the available sets is 3.453, which is lower than 3.614, we update the current selected variables to $\{x_1, x_5, x_8\}$

Step $i = 3$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_5, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_5\}$, $\{x_1, x_6\}$, $\{x_5, x_6\}$, $\{x_1, x_8\}$, $\{x_5, x_8\}$, $\{x_6, x_8\}$, $\{x_1, x_9\}$, $\{x_5, x_9\}$, $\{x_6, x_9\}$, $\{x_8, x_9\}$. Since the lowest error of the available sets is 3.42, which is lower than 3.453, we update the current selected variables to $\{x_1, x_8\}$

Step $i = 4$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1\}$, $\{x_5\}$, $\{x_6\}$, $\{x_8\}$, $\{x_9\}$. Since the lowest error of the available sets is 3.252, which is lower than 3.42, we update the current selected variables to $\{x_1\}$

Step $i = 5$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1\}$ 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 20.

Consider the Avila Bible dataset from Table 1. We wish to predict the copyist based on the attributes *upperm* and *mr/is*.

$p(\tilde{x}_2, \tilde{x}_{10}\|y)$	$y=1$	$y=2$	$y=3$
$\tilde{x}_2=0,\ \tilde{x}_{10}=0$	0.19	0.3	0.19
$\tilde{x}_2=0,\ \tilde{x}_{10}=1$	0.22	0.3	0.26
$\tilde{x}_2=1,\ \tilde{x}_{10}=0$	0.25	0.2	0.35
$\tilde{x}_2=1,\ \tilde{x}_{10}=1$	0.34	0.2	0.2

Table 6: Probability of observing particular values of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on y.

Therefore, suppose the attributes have been binarized such that $\tilde{x}_2 = 0$ corresponds $x_2 \leq -0.056$ (and otherwise $\tilde{x}_2 = 1$) and $\tilde{x}_{10} = 0$ corresponds $x_{10} \leq -0.002$ (and otherwise $\tilde{x}_{10} = 1$). Suppose the probability for each of the configurations of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on the copyist y are as given in Table 6. and the prior probability of the copyists is

$$p(y=1) = 0.316,\ p(y=2) = 0.356,\ p(y=3) = 0.328.$$

Using this, what is then the probability an observation was authored by copyist 1 given that $\tilde{x}_2 = 1$ and $\tilde{x}_{10} = 0$?

A. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.25$

B. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.313$

C. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.262$

D. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.298$

E. Don't know.

Solution 20. The problem is solved by a simple application of Bayes' theorem:

$$p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = \frac{p(\tilde{x}_2=1,\tilde{x}_{10}=0|y=1)p(y=1)}{\sum_{k=1}^{3} p(\tilde{x}_2=1,\tilde{x}_{10}=0|y=k)p(y=k)}$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_2=1,\tilde{x}_{10}=0|y)$ in Table 6. Inserting the values we see option D is correct.

Variable	$t=1$	$t=2$	$t=3$	$t=4$
y_1	1	2	2	2
y_2	1	2	2	1
y_3	2	2	2	1
y_4	1	1	1	2
y_5	1	1	1	1
y_6	2	2	2	1
y_7	1	2	2	1
y_8	2	1	1	2
y_9	2	2	2	2
y_{10}	1	1	2	2
y_{11}	2	2	1	2
y_{12}	2	1	1	2
y_1^{test}	2	1	1	2
y_2^{test}	2	2	1	2
ϵ_t	0.583	0.657	0.591	0.398
α_t	-0.168	-0.325	-0.185	0.207

Table 7: Tabulation of each of the predicted outputs of the AdaBoost classifiers, as well as the intermediate values α_t and ϵ_t, when the AdaBoost algorithm when evaluated for $T = 4$ steps. Note the table includes the prediction of the two test points in Figure 13.

Question 21.

Consider again the Avila Bible dataset of Table 1. Suppose we limit ourselves to $N = 12$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_6 and x_8. We wish to apply a KNN classification model ($K = 2$) to this dataset and apply AdaBoost to improve the performance. During the first $T = 4$ rounds of boosting, we obtain the decision boundaries shown in Figure 13. The figure also contains two test observations (marked by a cross and a square).

The prediction of the intermediate AdaBoost classifiers, as well as the values of α_t and ϵ_t, are given in Table 7. Given this information, how will the AdaBoost classifier, as obtained by combining the $T = 4$ weak classifiers, classify the two test observations?

A. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}1 & 1\end{bmatrix}$

B. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}2 & 1\end{bmatrix}$

C. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}1 & 2\end{bmatrix}$

D. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}2 & 2\end{bmatrix}$

E. Don't know.

Solution 21.

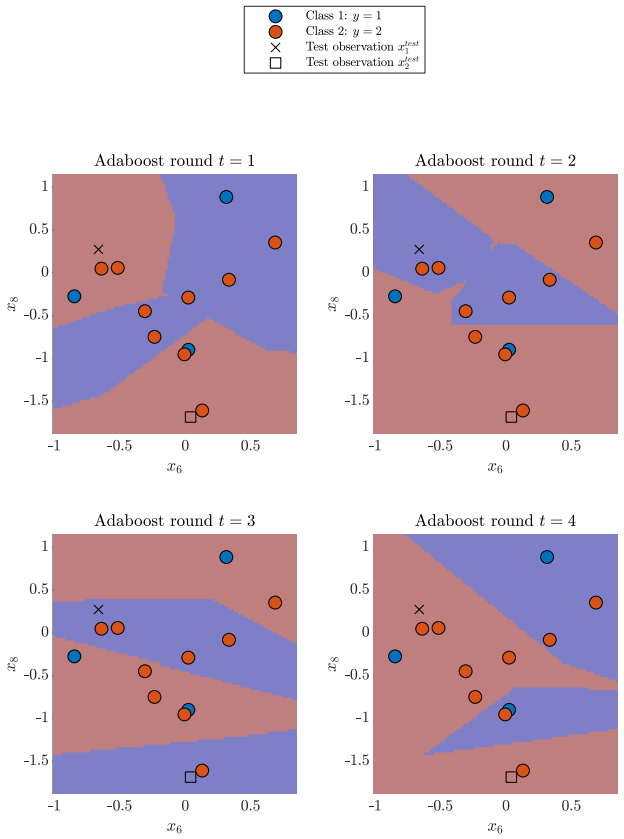

Figure 13: Decision boundaries for a KNN classifier for the first $T = 4$ rounds of boosting. Notice that in addition to the training data, the plot also indicate the location of two test points.

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

$$f^*(\boldsymbol{x}) = \arg\max_{y=1,2} \sum_{t=1}^{T} \alpha_t \delta_{f_t(\boldsymbol{x}),y}.$$

In other words, the classification rule is obtained by summing the α_t where $f_t(\boldsymbol{x}) = 1$ (as F_1) and those where $f_t(\boldsymbol{x}) = 2$ (as F_2) and then selecting the y corresponding to the largest value. We get for the two test points:

$$\begin{aligned}
F_1(\boldsymbol{x}_1^{\text{test}}) &= \alpha_2 + \alpha_3 = -0.51 \\
F_2(\boldsymbol{x}_1^{\text{test}}) &= \alpha_1 + \alpha_4 = 0.039 \\
F_1(\boldsymbol{x}_2^{\text{test}}) &= \alpha_3 = -0.185 \\
F_2(\boldsymbol{x}_2^{\text{test}}) &= \alpha_1 + \alpha_2 + \alpha_4 = -0.286.
\end{aligned}$$

Therefore, we get

$$\begin{bmatrix} \tilde{y}_1^{\text{test}} \\ \tilde{y}_2^{\text{test}} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and option B is correct.

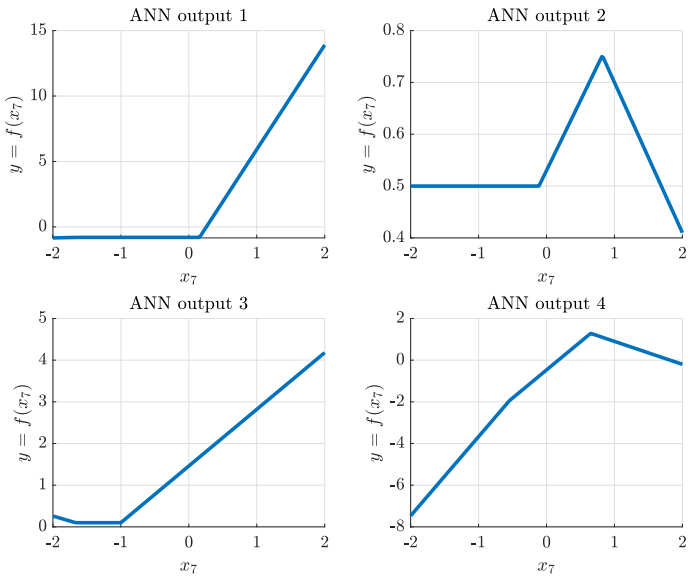

Figure 14: Suggested activation curves for an ANN applied to the feature x_7 from Avila Bible dataset.

Question 22.

We will consider an artificial neural network (ANN) applied to the Avila Bible dataset described in Table 1 and trained to predict based on just the feature x_7; that is, the neural network is a function that maps from a single real number to a single real number: $f(x_7) = y$

Suppose the neural network takes the form:

$$f(x, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x]\boldsymbol{w}_j^{(1)}).$$

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer and the weights are given as:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -1.8 \\ -1.1 \end{bmatrix}$$
$$\boldsymbol{w}_2^{(1)} = \begin{bmatrix} -0.6 \\ 3.8 \end{bmatrix}$$
$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -0.1 \\ 2.1 \end{bmatrix},$$
$$w_0^{(2)} = -0.8.$$

Which of the curves in Figure 14 will then correspond to the function f?

A. ANN output 4

B. ANN output 1

C. ANN output 3

D. ANN output 2

E. Don't know.

Solution 22.

It suffices to compute the activation of the neural network at $x_7 = 2$. The activation of each of the two hidden neurons is:

$$n_1 = h^{(1)}(\begin{bmatrix} 1 & 2 \end{bmatrix} \boldsymbol{w}_1^{(1)}) = 0$$
$$n_2 = h^{(1)}(\begin{bmatrix} 1 & 2 \end{bmatrix} \boldsymbol{w}_2^{(1)}) = 7.$$

The final output is then computed by a simple linear transformation:

$$f(x, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x]\boldsymbol{w}_j^{(1)})$$
$$= w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} n_j = 13.9.$$

This rules out all options except B.

Question 23. Suppose a neural network is trained to translate documents. As part of training the network, we wish to select between four different ways to encode the documents (i.e., $S = 4$ models) and estimate the generalization error of the optimal choice. In the outer loop we opt for $K_1 = 3$-fold cross-validation, and in the inner $K_2 = 4$-fold cross-validation. The time taken to *train* a single model is 20 minutes, and this can be assumed constant for each fold. If the time taken to test a model is negligible, what is the total time required for the 2-level cross-validation procedure?

A. 1020 minutes

B. 2040 minutes

C. 300 minutes

D. 960 minutes

E. Don't know.

Solution 23. Going over the 2-level cross-validation algorithm we see the total number of models to be *trained* is:

$$K_1(K_2S + 1) = 51$$

Multiplying by the time taken to train a single model we obtain a total training time of 1020 minutes and therefore answer A is correct.

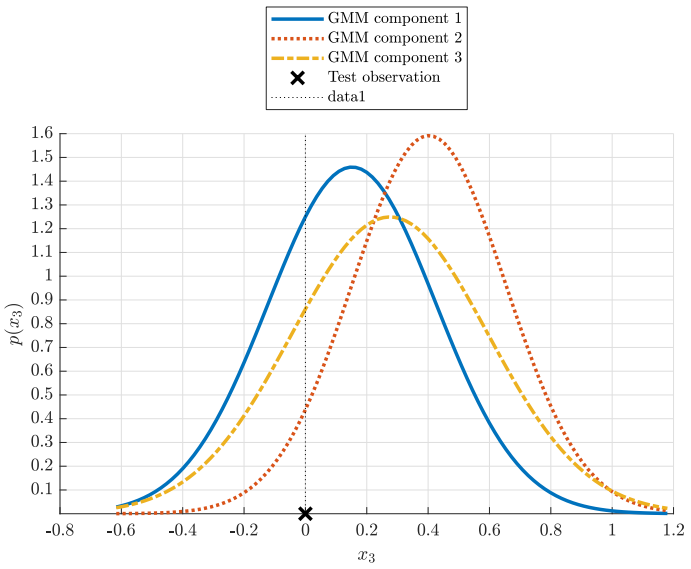

Figure 15: Mixture components in a GMM mixture model with $K = 3$.

Question 24.

We wish to apply the EM algorithm to fit a 1D GMM mixture model to the single feature x_3 from the Avila Bible dataset. At the first step of the EM algorithm, the $K = 3$ mixture components has densities as indicated by each of the curves in Figure 15 (i.e. each curve is a normalized, Gaussian density $\mathcal{N}(x; \mu_k, \sigma_k)$). In the figure, we have indicated the x_3-value of a single observation i from the dataset as a black cross.

Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step. We assume the weights of the components are

$$\boldsymbol{\pi} = \begin{bmatrix} 0.15 & 0.53 & 0.32 \end{bmatrix}$$

and the mean/variances of the components are those indicated in the figure.

According to the EM algorithm, what is the (approximate) probability the black cross is assigned to mixture component 3 (γ_{ik})?

A. 0.4

B. 0.86

C. 0.28

D. 0.58

E. Don't know.

Solution 24.

Recall γ_{ik} is the posterior probability that observation i is assigned to mixture component 3 which can easily be obtained using Bayes' theorem. We see that:

$$\gamma_{i,3} = \frac{p(x_i|z_{i,3}=1)\pi_3}{\sum_{k=1}^{3} p(x_i|z_{ik}=1)\pi_k}.$$

To use Bayes' theorem, we need to read off the probabilities from Figure 15. These are (approximately):

$$p(x_i|z_{i1}=1) = 1.25$$
$$p(x_i|z_{i2}=1) = 0.45$$
$$p(x_i|z_{i3}=1) = 0.85$$

Combining these with the class-assignment probabilities we obtain:

$$\gamma_{i,3} = 0.39.$$

Note this answer is not *exactly* the answer given in the question because we lost precision when we read off the probabilities from the figure. However, the answer is close enough to the answer 0.4 (and far enough away from the other) we can conclude the solution is A.

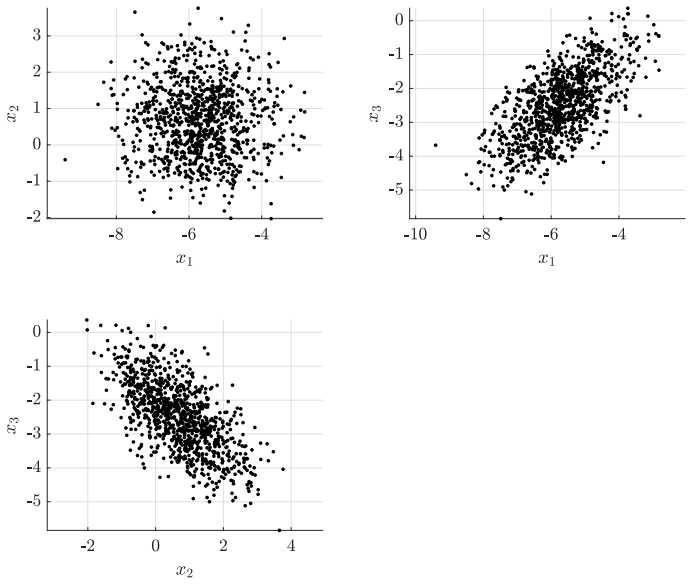

Figure 16: Scatter plot of each pairs of attributes of a vectors $\boldsymbol{x}$ drawn from a multivariate normal distribution of 3 dimensions.

Question 25. Consider a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ and mean μ and suppose we generate 1000 random samples from it:

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Plots of each pair of coordinates of the draws $\boldsymbol{x}$ is shown in Figure 16. What is the most plausible covariance matrix?

A. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.65 & -0.65 \\ 0.65 & 1.0 & 0.0 \\ -0.65 & 0.0 & 1.0 \end{bmatrix}$

B. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & 0.65 \\ 0.0 & 1.0 & -0.65 \\ 0.65 & -0.65 & 1.0 \end{bmatrix}$

C. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & -0.65 & 0.0 \\ -0.65 & 1.0 & 0.65 \\ 0.0 & 0.65 & 1.0 \end{bmatrix}$

D. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & -0.65 \\ 0.0 & 1.0 & 0.65 \\ -0.65 & 0.65 & 1.0 \end{bmatrix}$

E. Don't know.

Solution 25. To solve this problem, recall that the correlation between coordinates x_i, x_j of an observation drawn from a multivariate normal distribution is

positive if $\Sigma_{ij} > 0$, negative if $\Sigma_{ij} < 0$ and zero if $\Sigma_{ij} \approx 0$. Furthermore, recall positive correlation in a scatter plot means the points (x_i, x_j) tend to lie on a line sloping upwards, negative correlation means it is sloping downwards and zero means the data is axis-aligned.

We can therefore use the scatter plots of variables x_i, x_j to read of the sign off Σ_{ij} (or whether it is zero). This rules out all but option B.

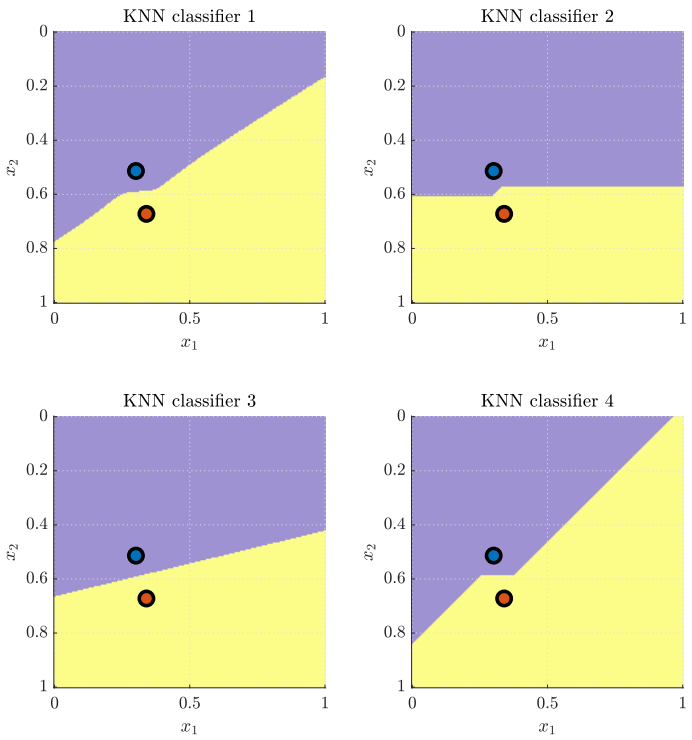

Figure 17: Decision boundaries for a KNN classifier, $K = 1$, computed for the two observations marked by circles (the colors indicate class labels), but using four different p-distances $d_p(\cdot, \cdot)$ to compute k-neighbors.

Question 26.

We consider a K-nearest neighbor (KNN) classifier with $K = 1$. Recall in a KNN classifier, we find the nearest neighbors by computing the distances using a distance measure $d(\boldsymbol{x}, \boldsymbol{y})$. For this problem, we will consider KNN classifiers based on different distance measures based on p-norms

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}}, p \geq 1$$

and what decision surfaces they induce.

In Figure 17 are shown four different decision boundaries obtained by training the KNN ($K = 1$) classifiers using the training observations (marked by the two circles in the figure):

$$\boldsymbol{x}_1 = \begin{bmatrix} 0.301 \\ 0.514 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 0.34 \\ 0.672 \end{bmatrix}$$

and with corresponding class labels $y_1 = 0$ and $y_2 = 1$, but with distance measures based on $p = 1,\ 2,\ 4,\ \infty$ (not necessarily plotted in that order).

Which norms were used in the four KNN classifiers?

A. KNN classifier 1 corresponds to $p = \infty$, KNN classifier 2 corresponds to $p = 2$, KNN classifier 3 corresponds to $p = 4$, KNN classifier 4 corresponds to $p = 1$

B. KNN classifier 1 corresponds to $p = 4$, KNN classifier 2 corresponds to $p = 2$, KNN classifier 3 corresponds to $p = 1$, KNN classifier 4 corresponds to $p = \infty$

C. KNN classifier 1 corresponds to $p = 4$, KNN classifier 2 corresponds to $p = 1$, KNN classifier 3 corresponds to $p = 2$, KNN classifier 4 corresponds to $p = \infty$

D. KNN classifier 1 corresponds to $p = \infty$, KNN classifier 2 corresponds to $p = 1$, KNN classifier 3 corresponds to $p = 2$, KNN classifier 4 corresponds to $p = 4$

E. Don't know.

Solution 26.

To solve this problem, one could simply consider points on the decision boundary and verify under which norm they had the same distance to the two test-observations. As this may feel a bit ad-hoc, we will here present a more general solution:

For simplicity, notice that (i) translating (moving) the coordinate system does not affect the distance calculation (ii) rotating the coordinate system by $\frac{\pi}{2}$ only corresponds to interchanging the role of x and y (iii) reflecting the coordinate system around an axis will not change the distance computation.

We can therefore consider the case where the location of the coordinates of the two points are $(-x, -y)$ and (x, y) where $y > x \geq 0$. Consider a point (d, h) *on* the decision boundary. The criteria for being so is:

$$(|d + x|^p + |h + y|^p)^{\frac{1}{p}} = (|d - x|^p + |h - y|^p)^{\frac{1}{p}}$$

Suppose now that $p = 2$. This is the standard Euclidean distance, and it is clear the decision boundary is a straight line passing through $(0, 0)$ and perpendicular to the vector (x, y).

For the other choices of p, suppose we limit ourselves to the case where $d < x$. For $p = 1$ we obtain:

$$|-d + x|^p + |-h + y|^p = |d + x|^p + |h + y|^p$$

Since the quantities within the absolute value operators are positive this becomes:

$$-d + x + -h + y = d + x + h + y$$

and therefore $d = -h$. That is, the decision boundary (for small d) must be a straight line at an 45-degree angle to the coordinate system.

For the case $p = \infty$, assume that $d + x < y$ (which is always possible since $y > x$). We then get:

$$\max\{|-d + x|, |-h + y|\} = \max\{|d + x|, |h + y|\}$$

One can either approach this expression with some algebra, but notice if $h = 0$ we get:

$$\max\{|-d + x|, |-0 + y|\} = \max\{|d + x|, |0 + y|\}$$

Therefore, if d is so small that $d + x < y$ this is trivially satisfied. In other words, when $d + x < y$ the horizontal line $h = 0$ is a solution.

Finally, the case $p = 4$ can be obtained by the process of illumination, or by noting the decicion bounday must look somewhat like a crossover between the $p = 2$ and $p = \infty$ case. We can therefore rule out all possibilities except C.

Question 27. Consider a small dataset comprised of $N = 9$ observations

$$x = \begin{bmatrix} 0.1 & 0.3 & 0.5 & 1.0 & 2.2 & 3.0 & 4.1 & 4.4 & 4.7 \end{bmatrix}.$$

Suppose a k-means algorithm is applied to the dataset with $K = 4$ and using Euclidian distances. At a given stage of the algorithm the data is partitioned into the blocks:

$$\{0.1, 0.3\},\ \{0.5, 1\},\ \{2.2, 3, 4.1\},\ \{4.4, 4.7\}$$

What clustering will the k-means algorithm eventually converge to?

A. $\{0.1, 0.3, 0.5, 1\}$, $\{2.2\}$, $\{\}$, $\{3, 4.1, 4.4, 4.7\}$

B. $\{0.1, 0.3\}$, $\{0.5, 1\}$, $\{2.2, 3\}$, $\{4.1, 4.4, 4.7\}$

C. $\{0.1, 0.3\}$, $\{0.5\}$, $\{1, 2.2\}$, $\{3, 4.1, 4.4, 4.7\}$

D. $\{0.1, 0.3\}$, $\{0.5, 1, 2.2, 3\}$, $\{4.1, 4.4\}$, $\{4.7\}$

E. Don't know.

Solution 27. Recall the K-means algorithm iterates between assigning the observations to their nearest centroids, and then updating the centroids to be equal to the average of the observations assigned to them. Therefore, the subsequent steps in the K-means algorithm are:

Step $t = 1$: The centroids are computed to be:

$$\mu_1 = 0.2,\ \mu_2 = 0.75,\ \mu_3 = 3.1,\ \mu_4 = 4.55.$$

And the updated assignment of observations to nearest centroids results in the clustering:

$$\{0.1, 0.3\},\ \{0.5, 1\},\ \{2.2, 3\},\ \{4.1, 4.4, 4.7\}.$$

Step $t = 2$: The centroids are computed to be:

$$\mu_1 = 0.2,\ \mu_2 = 0.75,\ \mu_3 = 2.6,\ \mu_4 = 4.4.$$

And the updated assignment of observations to nearest centroids results in the clustering:

$$\{0.1, 0.3\},\ \{0.5, 1\},\ \{2.2, 3\},\ \{4.1, 4.4, 4.7\}.$$

At this point, the centroids are no longer changing and the algorithm terminates. Hence, B is correct.

Technical University of Denmark

Written examination: December 17th 2019, 9 AM - 1 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 5 PM and file 2 no later than 5:15 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 ultimately potentially count as attempt of cheating the exam.

Answers:

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

No.	Attribute description	Abbrev.
x_1	Month (1-12)	MONTH
x_2	$PM_{2.5}$ concentration ($\mu g/m^3$)	$PM_{2.5}$
x_3	PM_{10} concentration ($\mu g/m^3$)	PM_{10}
x_4	NO_2 concentration ($\mu g/m^3$)	NO_2
x_5	SO concentration ($\mu g/m^3$)	CO
x_6	O_3 concentration ($\mu g/m^3$)	O_3
x_7	Temperature (degree Celsius)	TEMP
x_8	Pressure (hPa)	PRES
x_9	Dew point temperature (degree Celsius)	DEWP
x_{10}	Precipitation/rainfall (mm)	RAIN
x_{11}	Wind speed (m/s)	WSPM
y	SO_2 concentration ($\mu g/m^3$)	pollution level

Table 1: Description of the features of the Beijing air pollution dataset used in this exam. It consists of measurements from 12 air-quality sites provided by the China Meteorological Administration. The measurements were taken hourly (March 1st, 2013 to February 28th, 2017), but we will only consider data from 2014, subsampled to every 8 hours, and with missing values removed. We consider the goal as predicting the SO_2 level both as regression and classification task. For regression tasks, y_r will refer to the continuous value in $\mu g/m^3$. For classification, the attribute y is discrete taking values $y = 1$ (corresponding to a light pollution level), $y = 2$ (corresponding to a medium pollution level), and $y = 3$ (corresponding to a high pollution level). There are $N = 981$ observations in total.

Question 1. The main dataset used in this exam is the Beijing air pollution dataset[1] described in Table 1. Table 2 contains summary statistics of four attributes from the Beijing air pollution dataset. Which boxplots

[1]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data`

	Mean	Std	$x_{p=25\%}$	$x_{p=50\%}$	$x_{p=75\%}$
$PM_{2.5}$	85.58	78.09	26	66	121.25
PM_{10}	113.2	85.18	48.75	97	156.25
NO_2	55.89	31.8	30	51	76.25
O_3	54.4	61.72	8	33	75

Table 2: Summary statistics of four attributes from the Beijing air pollution 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.

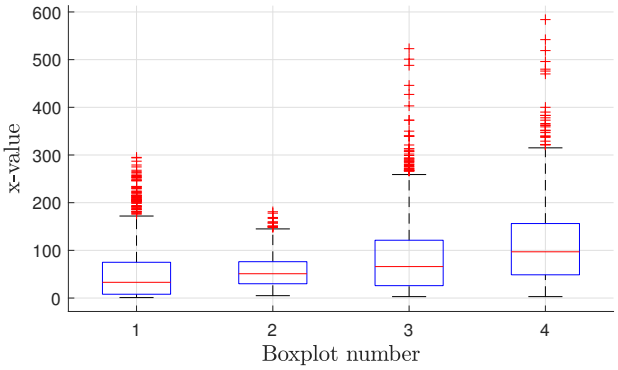

Figure 1: Boxplots corresponding to the variables with summary statistics indicated in Table 2 but not necessarily in that order.

in Figure 1 match which attributes?

A. Attribute $PM_{2.5}$ corresponds to boxplot 3 PM_{10} corresponds to boxplot 4 NO_2 corresponds to boxplot 1 and O_3 corresponds to boxplot 2

B. Attribute $PM_{2.5}$ corresponds to boxplot 4 PM_{10} corresponds to boxplot 3 NO_2 corresponds to boxplot 2 and O_3 corresponds to boxplot 1

C. Attribute $PM_{2.5}$ corresponds to boxplot 3 PM_{10} corresponds to boxplot 4 NO_2 corresponds to boxplot 2 and O_3 corresponds to boxplot 1

D. Attribute $PM_{2.5}$ corresponds to boxplot 1 PM_{10} corresponds to boxplot 3 NO_2 corresponds to boxplot 2 and O_3 corresponds to boxplot 4

E. Don't know.

Solution 1. 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 in turns can be matched to the boxplots in

Figure 1 by observing the median is the horizontal red line and the 25'th and 75'th percentiles corresponds to the top and bottom of the boxes. This easily rules out all options except C.

Question 2. A Principal Component Analysis (PCA) is carried out on the Beijing air pollution dataset in Table 1 based on the attributes x_1, x_3, x_5, x_8, x_{10}, x_{11}.

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 $\boldsymbol{USV}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.1 & -0.45 & -0.55 & 0.67 & -0.2 & 0.01 \\ -0.63 & -0.02 & -0.01 & -0.05 & -0.44 & -0.64 \\ -0.67 & 0.07 & 0.03 & 0.13 & -0.12 & 0.72 \\ -0.09 & 0.69 & 0.03 & 0.6 & 0.32 & -0.2 \\ 0.06 & -0.35 & 0.83 & 0.41 & -0.09 & -0.03 \\ 0.37 & 0.44 & 0.05 & 0.04 & -0.8 & 0.17 \end{bmatrix} \quad (1)$$

$$\boldsymbol{S} = \begin{bmatrix} 43.67 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 33.47 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 31.15 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 30.36 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 27.77 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 13.86 \end{bmatrix}.$$

Which one of the following statements is true?

A. The variance explained by the first five principal components is less than 0.9

B. The variance explained by the first three principal components is less than 0.715

C. The variance explained by the first principal component is less than 0.3

D. The variance explained by the last two principal components is less than 0.15

E. Don't know.

Solution 2. The correct answer is B. To see this, recall 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 components x_1, x_2, x_3 is:

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2 + \sigma_6^2} = 0.6796.$$

Question 3. Consider again the PCA analysis for the Beijing air pollution dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (1). Which one of the following statements is true?

A. An observation with a low value of **PM**$_{10}$, a high value of **PRES**, and a low value of **WSPM** will typically have a negative value of the projection onto principal component number 5.

B. An observation with a high value of **PM**$_{10}$, a high value of **CO**, and a low value of **WSPM** will typically have a positive value of the projection onto principal component number 1.

C. An observation with a low value of **MONTH**, a low value of **PRES**, and a low value of **RAIN** will typically have a positive value of the projection onto principal component number 4.

D. An observation with a high value of MONTH, and a low value of RAIN will typically have a negative value of the projection onto principal component number 3.

E. Don't know.

Solution 3. The correct answer is D. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_3$ (i.e. column three of $\boldsymbol{V}$) is

$$b_3 = \boldsymbol{x}^\top \boldsymbol{v}_3 = \begin{bmatrix} x_1 & x_3 & x_5 & x_8 & x_{10} & x_{11} \end{bmatrix} \begin{bmatrix} -0.55 \\ -0.01 \\ 0.03 \\ 0.03 \\ 0.83 \\ 0.05 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and negative, this occurs if x_1, x_{10} has large magnitude and the sign convention given in option D.

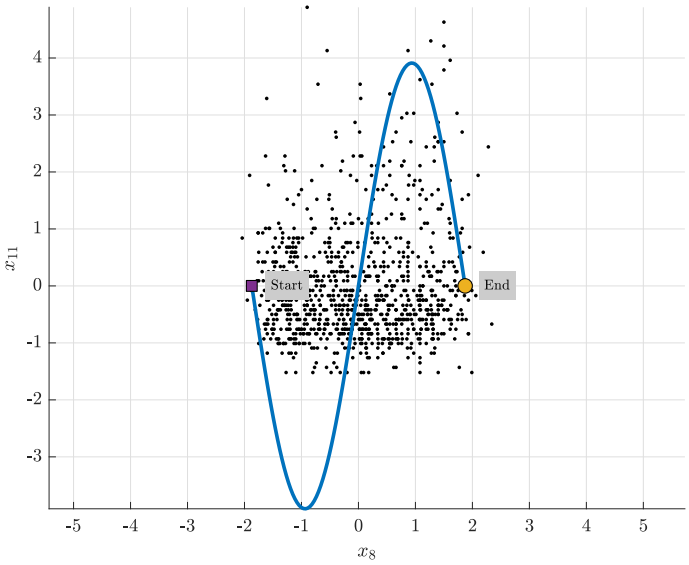

Figure 2: Black dots show attributes x_8 and x_{11} of the Beijing air pollution dataset from Table 1. The other attributes are kept fixed while x_8 and x_{11} are varied and thereby trace out the path indicated by the blue line, starting at the purple square and ending at the yellow circle.

Question 4. Consider again the Beijing air pollution dataset. In Figure 3 the features x_8 and x_{11} from Table 1 are plotted as black dots. Recall the data is temporally ordered, and suppose over a period of time the measurements undergoes an evolution indicated by the path here shown as a blue line which begins at the purple square, and ends at the yellow circle and where the other features can be considered fixed.

We can imagine the dataset, along with the path, is projected onto the first two principal components given in Equation (1). Which one of the four plots in Figure 2 shows the path?

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 4. Since we don't know the exact values of most of the x_i-coordinates, it is easier to work with the difference between the start and end-points in Figure 2 and translate them into the difference of the start

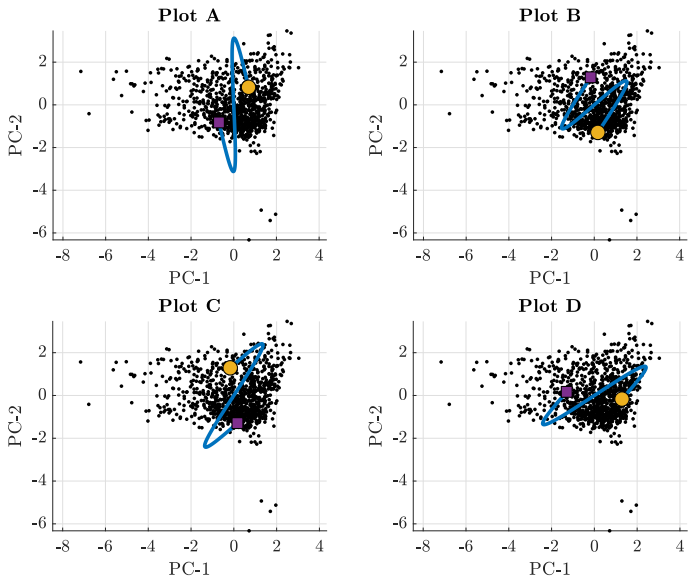

Figure 3: Candidate plots of the observations and path shown in Figure 2 projected onto the first two principal components considered in Equation (1). The start point is indicated by the purple square and the end point by the yellow circle.

and end points in the PCA projections. Notice from Figure 2 we can immediately compute:

$$\Delta \boldsymbol{x} = x_{\text{end}} - x_{\text{start}} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 3.74 \\ 0.0 \\ 0.0 \end{bmatrix}$$

(this corresponds to the vector going from the start to end points). Then, all we need is to compute the PCA projection of this vector as:

$$\Delta \boldsymbol{b} = (\Delta \boldsymbol{x})^\top \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} -0.34 \\ 2.58 \end{bmatrix}$$

Which should be the vector beginning at the start-point and terminating at the end-point in the PCA projected plots. This rules out all plots except option C.

Question 5. Consider the Beijing air pollution dataset (but for this problem in the non-standardized version). The empirical covariance matrix of the first 5 attributes $x_1, \ldots, x_5$ is:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 12 & -29 & -21 & -12 & -317 \\ -29 & 6104 & 6026 & 1557 & 67964 \\ -21 & 6026 & 7263 & 1701 & 70892 \\ -12 & 1557 & 1701 & 1012 & 25415 \\ -317 & 67964 & 70892 & 25415 & 1212707 \end{bmatrix}.$$

What is the empirical correlation of MONTH and $\text{PM}_{2.5}$?

A. -5.38516

B. -0.0199

C. -0.10715

D. -0.0004

E. Don't know.

Solution 5. Recall the correlation is defined as

$$\text{cor}[x, y] = \frac{\text{cov}[x, y]}{\text{std}[x]\,\text{std}[y]}$$

Next, by definition the diagonal elements of the covariance matrix are estimates of the variance and the off-diagonal elements are estimates of the covariance, i.e. for $i \neq j$:

$$\hat{\Sigma}_{ii} = \text{Var}[x_i], \quad \hat{\Sigma}_{ij} = \text{cov}[x_i, x_j]$$

Therefore we get:

$$\text{cor}[x_i, y_j] = \frac{\hat{\Sigma}_{ij}}{\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}}}$$

Then finally observe from Table 1 we get that MONTH corresponds to x_1 and $\text{PM}_{2.5}$ corresponds to x_2 so $i = 1$ and $j = 2$ and therefore by simple insertion option C is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	4.2	8.3	3.9	3.8	4.6	6.3	4.8	7.1	4.9
o_2	4.2	0.0	7.4	2.6	3.0	3.2	5.3	3.1	6.6	4.6
o_3	8.3	7.4	0.0	6.3	7.1	5.5	2.8	5.4	2.4	5.3
o_4	3.9	2.6	6.3	0.0	1.5	1.6	4.1	1.8	5.3	2.4
o_5	3.8	3.0	7.1	1.5	0.0	2.4	4.9	2.8	5.8	3.2
o_6	4.6	3.2	5.5	1.6	2.4	0.0	3.7	1.7	4.8	2.3
o_7	6.3	5.3	2.8	4.1	4.9	3.7	0.0	3.8	1.9	3.6
o_8	4.8	3.1	5.4	1.8	2.8	1.7	3.8	0.0	4.9	2.1
o_9	7.1	6.6	2.4	5.3	5.8	4.8	1.9	4.9	0.0	4.4
o_{10}	4.9	4.6	5.3	2.4	3.2	2.3	3.6	2.1	4.4	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 10 observations from the Beijing air pollution dataset (recall that $M = 11$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1. The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belongs to class C_1 (corresponding to a light pollution level), the red observations $\{o_3,\ o_4,\ o_5,\ o_6\}$ belongs to class C_2 (corresponding to a medium pollution level), and the blue observations $\{o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to a high pollution level).

Question 6. To examine if observation o_5 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_5 for $K = 2$ nearest neighbors?

A. 0.41

B. 0.82

C. 1.0

D. 0.51

E. Don't know.

Solution 6.

To solve the problem, first observe the $k = 2$ neighborhood of o_5 and density is:

$$N_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = \{o_4, o_6\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = 0.513$$

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

$$N_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = \{o_5, o_6\}, \quad N_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = \{o_4, o_8\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = 0.645, \text{density}_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = 0.606.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem.

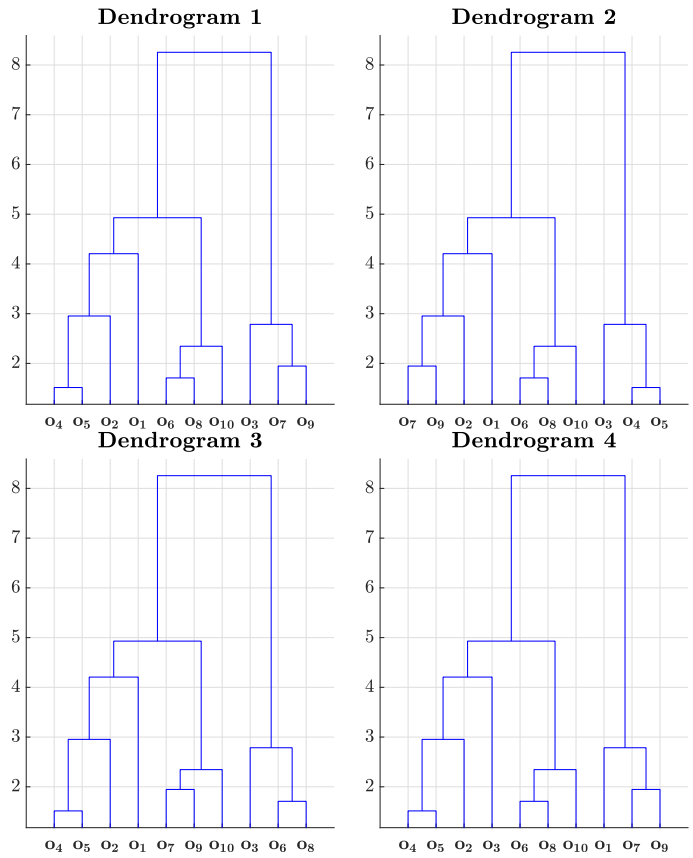

Figure 4: Proposed hierarchical clustering of the 10 observations in Table 3.

Question 7. A hierarchical clustering is applied to the 10 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 1 should have been between the sets $\{f_4\}$ and $\{f_5\}$ at a height of 1.52, however in dendrogram 2 merge number 1 is between the sets $\{f_7\}$ and $\{f_9\}$.
- In dendrogram 3, merge operation number 2 should have been between the sets $\{f_6\}$ and $\{f_8\}$ at a height of 1.71, however in dendrogram 3 merge number 2 is between the sets $\{f_7\}$ and $\{f_9\}$.
- In dendrogram 4, merge operation number 5 should have been between the sets $\{f_3\}$ and $\{f_7, f_9\}$ at a height of 2.79, however in dendrogram 4 merge number 5 is between the sets $\{f_1\}$ and $\{f_7, f_9\}$.

Question 8. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two binary vectors of dimension $N = 1500$ such that $\boldsymbol{x}_1$ has one non-zero element and $\boldsymbol{x}_2$ has 1498 non-zero elements. What are the possible range of values of the Jaccard similarities of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$?

A. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00242]$

B. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00067]$

C. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00074]$

D. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00206]$

E. Don't know.

Solution 8. To solve this problem, recall the Jaccard similarity is defined as

$$\frac{n_{11}}{N - n_{00}}$$

note it is possible for n_{11}, the number of coordinates where both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are non-zero, to be zero (in the case $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are sorted in ascending and descending order respectively) and hence the Jaccard similarity can be zero. On the other extreme, the maximal Jaccard similarity is obtained when n_{11} is as large as possible while n_{00} is as small as possible. This evidently occurs when the two arrays are sorted in the same order. In this case the overlap is

$$n_{11} = \min\{1, 1498\} = 1$$

and similarly

$$n_{00} = \min\{N - 1, N - 1498\} = 2.$$

Corresponding to a Jaccard similarity of 0.00067. We therefore see that B is correct.

Question 9. Consider again the Beijing air pollution dataset in Table 1. We would like to predict a pollution level using a linear regression, 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_2, x_4, x_6, x_9, and x_{11} and in Table 4 we have pre-computed the estimated

Feature(s)	Training RMSE	Test RMSE
none	2.235	2.851
x_2	2.096	2.232
x_4	1.902	1.793
x_6	2.214	2.351
x_9	2.183	3.227
x_{11}	2.235	2.83
$x_2,\ x_4$	1.9	1.797
$x_2,\ x_6$	2.081	2.597
$x_4,\ x_6$	1.777	2.785
$x_2,\ x_9$	1.606	3.09
$x_4,\ x_9$	1.724	2.243
$x_6,\ x_9$	2.087	2.307
$x_2,\ x_{11}$	2.046	2.754
$x_4,\ x_{11}$	1.87	2.143
$x_6,\ x_{11}$	2.214	2.37
$x_9,\ x_{11}$	2.177	3.058
$x_2,\ x_4,\ x_6$	1.773	2.838
$x_2,\ x_4,\ x_9$	1.574	2.81
$x_2,\ x_6,\ x_9$	1.605	3.187
$x_4,\ x_6,\ x_9$	1.691	2.698
$x_2,\ x_4,\ x_{11}$	1.868	2.188
$x_2,\ x_6,\ x_{11}$	2.003	3.738
$x_4,\ x_6,\ x_{11}$	1.723	3.472
$x_2,\ x_9,\ x_{11}$	1.483	4.246
$x_4,\ x_9,\ x_{11}$	1.714	2.418
$x_6,\ x_9,\ x_{11}$	2.081	2.159
$x_2,\ x_4,\ x_6,\ x_9$	1.549	3.174
$x_2,\ x_4,\ x_6,\ x_{11}$	1.676	4.227
$x_2,\ x_4,\ x_9,\ x_{11}$	1.469	3.944
$x_2,\ x_6,\ x_9,\ x_{11}$	1.459	5.017
$x_4,\ x_6,\ x_9,\ x_{11}$	1.667	3.146
$x_2,\ x_4,\ x_6,\ x_9,\ x_{11}$	1.406	5.006

Table 4: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y_r in the Beijing air pollution dataset using different combinations of the features $x_2,\ x_4,\ x_6,\ x_9$, and x_{11}.

training and test error for the different variable combinations. Which of the following statements is correct?

A. Backward selection will select attributes $x_6,\ x_9,\ x_{11}$

B. Backward selection will select attributes $x_4,\ x_6,\ x_9,\ x_{11}$

C. Forward selection will select attributes $x_6,\ x_9,\ x_{11}$

D. Forward selection will select attributes $x_4,\ x_6,\ x_9,\ x_{11}$

E. Don't know.

Solution 9.

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

Forward selection: The method is initialized with the set $\{\}$ having an error of 2.851.

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 $\{x_2\}$, $\{x_4\}$, $\{x_6\}$, $\{x_9\}$, $\{x_{11}\}$. Since the lowest error of the available sets is 1.793, which is lower than 2.851, we update the current selected variables to $\{x_4\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_4\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_2, x_4\}$, $\{x_2, x_6\}$, $\{x_4, x_6\}$, $\{x_2, x_9\}$, $\{x_4, x_9\}$, $\{x_6, x_9\}$, $\{x_2, x_{11}\}$, $\{x_4, x_{11}\}$, $\{x_6, x_{11}\}$, $\{x_9, x_{11}\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Backward selection: The method is initialized with the set $\{x_2, x_4, x_6, x_9, x_{11}\}$ having an error of 5.006.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_2, x_4, x_6, x_9, x_1 1\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_2, x_4, x_6, x_9\}$, $\{x_2, x_4, x_6, x_{11}\}$, $\{x_2, x_4, x_9, x_{11}\}$, $\{x_2, x_6, x_9, x_{11}\}$, $\{x_4, x_6, x_9, x_{11}\}$. Since the lowest error of the available sets is 3.146, which is lower than 5.006, we update the current selected variables to $\{x_4, x_6, x_9, x_{11}\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_4, x_6, x_9, x_1 1\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_2, x_4, x_6\}$, $\{x_2, x_4, x_9\}$, $\{x_2, x_6, x_9\}$, $\{x_4, x_6, x_9\}$, $\{x_2, x_4, x_{11}\}$, $\{x_2, x_6, x_{11}\}$, $\{x_4, x_6, x_{11}\}$, $\{x_2, x_9, x_{11}\}$, $\{x_4, x_9, x_{11}\}$, $\{x_6, x_9, x_{11}\}$. Since the lowest error of the available sets is 2.159, which is lower than 3.146, we update the current selected variables to $\{x_6, x_9, x_{11}\}$

Step $i = 3$ The available variable sets to choose between is obtained by taking the current variable set $\{x_6, x_9, x_1 1\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_2, x_4\}$, $\{x_2, x_6\}$, $\{x_4, x_6\}$, $\{x_2, x_9\}$, $\{x_4, x_9\}$, $\{x_6, x_9\}$, $\{x_2, x_{11}\}$, $\{x_4, x_{11}\}$, $\{x_6, x_{11}\}$, $\{x_9, x_{11}\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}
o_1	0	0	0	0	0	1	1	1	0	1	1
o_2	1	0	0	1	0	1	1	0	1	1	1
o_3	1	1	1	1	1	0	0	0	0	1	0
o_4	0	1	0	1	0	0	0	1	0	1	0
o_5	0	0	0	0	0	1	0	1	1	1	0
o_6	0	1	1	1	1	0	0	0	1	1	0
o_7	1	1	1	1	1	0	0	1	0	1	0
o_8	0	1	1	1	1	0	0	0	1	1	0
o_9	1	1	1	1	1	0	0	1	0	1	0
o_{10}	0	1	1	1	1	0	0	1	0	1	0

Table 5: Binarized version of the Beijing air pollution dataset. Each of the features f_i are obtained by taking a feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). The colors indicate classes such that the black observations $\{o_1,\ o_2\}$ belongs to class C_1 (corresponding to a light pollution level), the red observations $\{o_3,\ o_4,\ o_5,\ o_6\}$ belongs to class C_2 (corresponding to a medium pollution level), and the blue observations $\{o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to a high pollution level).

Question 10. We again consider the Beijing air pollution dataset from Table 1 and the $N = 10$ observations we already encountered in Table 3. The data is processed to produce 11 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 $N \times M = 10 \times 11$ 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

$$p(f_2 = 0,\ f_{11} = 0|y = 2).$$

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

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

[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.

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

A. $p(f_2 = 0,\ f_{11} = 0|y = 2) = \frac{1}{3}$

B. $p(f_2 = 0,\ f_{11} = 0|y = 2) = \frac{3}{5}$

C. $p(f_2 = 0,\ f_{11} = 0|y = 2) = 0$

D. $p(f_2 = 0,\ f_{11} = 0|y = 2) = 1$

E. Don't know.

Solution 10. Of the observations in class $y = 2$ only 1 have simultaneously $f_2 = 0,\ f_{11} = 0$. As this class contains *four* observations, we see the answer is

$$\frac{1+\alpha}{4+2\alpha} = \frac{2}{6}$$

Therefore, answer A is correct.

Question 11. Consider the binarized version of the Beijing air pollution dataset shown in Table 5.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 11$ items $f_1, f_2, \ldots, f_{11}$. Which of the following options represents all (non-empty) itemsets with support greater than 0.65 (and only itemsets with support greater than 0.65)?

A. $\{f_4\}$, $\{f_{10}\}$, $\{f_4, f_{10}\}$

B. $\{f_2\}$, $\{f_4\}$, $\{f_{10}\}$, $\{f_2, f_4\}$, $\{f_2, f_{10}\}$, $\{f_4, f_{10}\}$, $\{f_2, f_4, f_{10}\}$

C. $\{f_2\}$, $\{f_4\}$, $\{f_{10}\}$, $\{f_2, f_4\}$, $\{f_2, f_{10}\}$, $\{f_4, f_{10}\}$

D. $\{f_{10}\}$

E. Don't know.

Solution 11. 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 0.65, an itemset needs to be contained in 7 rows. It is easy to see this rules out all options except B.

Question 12. We again consider the binary matrix from Table 5 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 11$ items $f_1, \ldots, f_{11}$.

What is the *confidence* of the rule $\{f_1, f_3, f_4, f_5, f_8\} \rightarrow \{f_2, f_{10}\}$?

A. The confidence is 1

B. The confidence is $\frac{1}{5}$

C. The confidence is $\frac{2}{7}$

D. The confidence is $\frac{9}{20}$

E. Don't know.

Solution 12. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_1, f_3, f_4, f_5, f_8\} \cup \{f_2, f_{10}\})}{\text{support}(\{f_1, f_3, f_4, f_5, f_8\})} = \frac{\frac{1}{5}}{\frac{1}{5}} = 1.$$

Therefore, answer A is correct.

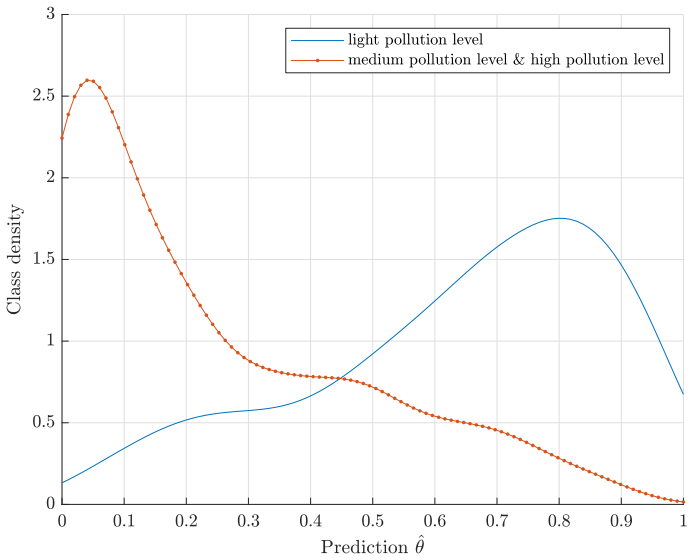

Figure 5: Class density (as function of the predictions of a logistic regression classifier $\hat{\theta}$) of the two-class problem of predicting *light pollution level* vs. *medium pollution level & high pollution level*.

Question 13. A logistic regression classifier is applied to the Beijing air pollution dataset described in Table 1 to solve the binary classification problem of *light pollution level* (positive class) vs. *medium pollution level & high pollution level* (negative class). The output of the classifier is the class-assignment probability $\hat{\theta}$, and for each threshold value θ_0 we assign observations with $\hat{\theta} > \theta_0$ to the positive class *light pollution level* (and otherwise to the negative class *medium pollution level & high pollution level*).

Suppose the class-density for each class is as indicated in Figure 5, which of the reciever operator characteristic (ROC) curves in Figure 5 corresponds to the classifier?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

Solution 13. Recall we compute the ROC curve from the false positive rate (FPR) and true positive rate (TPR) for particular choices of threshold value $\hat{\theta}_0$

$$(x, y) = (\text{FPR}, \text{TPR}).$$

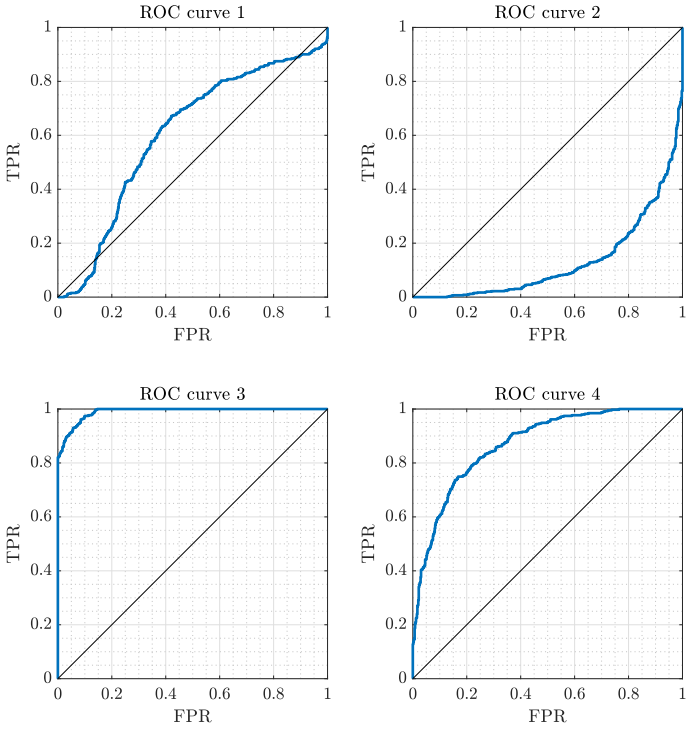

Figure 6: Proposed ROC curves for the two-class classifier described in Figure 5.

Furthermore, compute e.g. the TPR, one assumes every observation predicted to belong to *medium pollution level & high pollution level* (positive class) if $\hat{\theta} > \theta_0$ and otherwise to *light pollution level* (negative class)

We then divide the total number of observations belonging to positive class *and which are predicted to belong to the positive class* 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 the positive class *but in fact belongs to the negative class*, divided by the total number of observations in the *negative* class. For concreteness, we have inserted the true values of the TPR and FPR as function of θ_0 in Figure 7.

To reason about the options, first, observe that the location of the two humps in Figure 5 implies an AUC well over 0.5: Consider for instance a value of $_0$ between them in which the TPR is much larger than the FPR.

Next, consider the case where θ_0 is very small and increases, and recall this corresponds to the point $(1, 1)$ on the ROC curve. Since for low values of θ_0 the negative class has a large density (and the positive class has a low but non-zero density), this implies the FPR has to decrease much less rapidly than the TPR, because the number negative predictions falsely

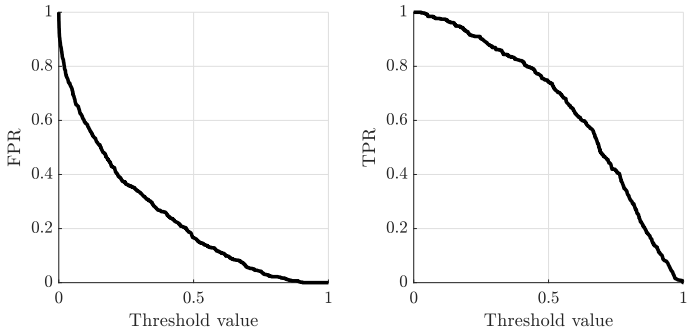

Figure 7: TPR, FPR curves for the classifier.

classified as positive decreases much more rapidly than the number of positive class members predicted to be negative. In other words, the ROC curve must initially stay above the line of identity.

At the same time, this rules out the case that e.g. the FPR decreases while the TPR remains at 1, as the two classes overlap. In other words, if the FPR ¡ 1 then so will the TPR. These observations together allows us to rule out all options except D.

	$i=1$	$i=2$	$i=3$	$i=4$
x_i	2	5	6	7
y_i	6	7	7	9

Table 6: Simple 1d regression dataset

Question 14. Consider the small 1d dataset shown in Table 6 comprised of $N = 4$ observations and where the goal is to predict y_i given x_i. Suppose we apply ridge regression to the problem in the form described in the lecture notes, Section 14.1.

If $\lambda = 2$, what is the ridge regression cost function assuming the weight-vector is

$$\boldsymbol{w} = \begin{bmatrix} 0.6 \end{bmatrix}$$

i.e. $E_\lambda(\boldsymbol{w}, w_0)$?

A. $E_\lambda(\boldsymbol{w}, w_0) = 1.205$

B. $E_\lambda(\boldsymbol{w}, w_0) = 1.97$

C. $E_\lambda(\boldsymbol{w}, w_0) = 1.033$

D. $E_\lambda(\boldsymbol{w}, w_0) = 2.662$

E. Don't know.

Solution 14. The cost function is defined as

$$E_\lambda(\boldsymbol{w}, w_0) = \sum_{i=1}^{4} (y_i - \hat{y}_i)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w}.$$

Where $\hat{y}_i$ are the predictions. According to Section 14.1 these are computed from the *standardized* feature matrix as:

$$\hat{y} = \frac{x - \mu}{\sigma} w + \mathbb{E}[y].$$

Where μ and σ is the mean and standard deviations of x as computed on the training set in Table 6, i.e. $\mu = 5.0$ and $\sigma = 2.16$. Since $\mathbb{E}[y] = \frac{1}{N}\sum_{i=1}^{N} y_i = 7.25$ we find that the predicted values of y are:

$$\hat{\boldsymbol{y}} = \begin{bmatrix} 6.417 & 7.25 & 7.528 & 7.805 \end{bmatrix}.$$

Inserting these in the cost function we get:

$$E_\lambda(\boldsymbol{w}, w_0) = 0.42^2 + 0.25^2 + 0.53^2 + 1.19^2 + \lambda 0.36 = 2.662$$

hence D is correct.

	1	2	3	4	5	6
x_7	-1.76	-0	0.06	0.08	0.65	1.3
y_r	12	6	8	10	4	2

Table 7: Values of x_7 and the corresponding value of y_r.

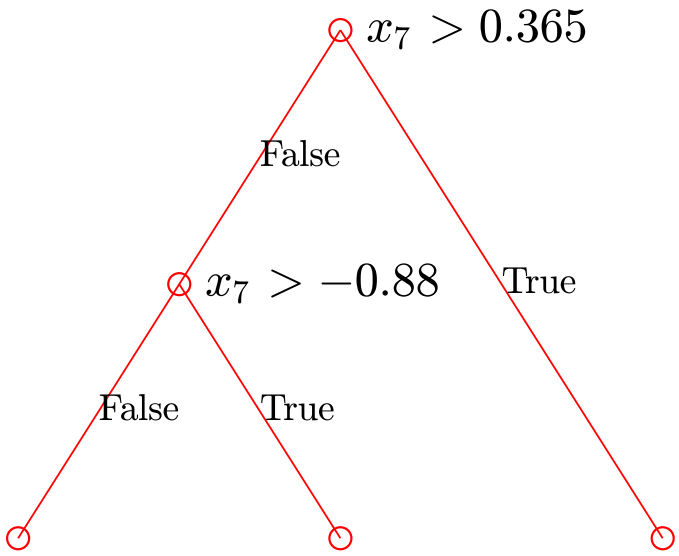

Figure 8: Structure of decision tree. The goal is to determine the splitting rules.

Question 15. We will consider the first 6 observations of the Beijing air pollution dataset shown in Table 3. Table 7 shows their corresponding value of x_7 and y_r. We fit a small regression tree to this dataset, the structure (and binary splitting rules) is depicted in Figure 8. What is the predicted value $\hat{y}_r$ as evaluated at $x_7 = 0.5$?

A. $\hat{y}_r = 2.85$

B. $\hat{y}_r = 3.0$

C. $\hat{y}_r = 2.52$

D. $\hat{y}_r = 0.98$

E. Don't know.

Solution 15.

The predicted value for a given input is computed as the average y-value of those observations in the training set which is assigned to the same leaf node v as the input, i.e.

$$y(v) = \frac{1}{N(v)} \sum_{i \in v} y_i$$

(see the section on regression trees in lecture notes). Therefore, we first need to find out which leaf node the observation is assigned to. To do this, start at the root and compare $x_7 = 0.5$ to the rule in the split

$$x_7 > 0.365$$

and we continue down the right branch. Continuing in this manner, we see $x_7 = 0.5$ is classified to leaf number one from the left. Then, proceeding in the same manner with the x_7 observations in Table 7, we see the observations o_5, and o_6 are also assigned to leaf one (counted from the left). According to the above the prediction is then simply the average of their y-value

$$\hat{y} = \frac{1}{2}(4 + 2)$$

or $\hat{y} = 3.0$, hence B is correct.

Question 16. In this problem, we will again consider the 6 observations from the Beijing air pollution dataset shown in Table 7. Recall Figure 8 shows the structure of the small regression tree fitted to this dataset using Hunt's algorithm along with the thereby obtained binary splitting rules. What was the purity gain Δ of the first split Hunt's algorithm accepted?

A. $\Delta = 11.67$

B. $\Delta = 3.67$

C. $\Delta = 8.0$

D. $\Delta = 56.0$

E. Don't know.

Solution 16.

The first split Hunt's algorithm accepted must be the split at the root, i.e.

$$x > 0.365$$

Partitioning the observations in Table 7 according to this split results in the two sets

$$v_1 = \{1, 2, 3, 4\}, \quad v_2 = \{5, 6\}$$

at the two legs. The impurity of these two sets, and the impurity of all y-values, is computed using the impurity measure appropriate for regression trees

$$I(v) = \frac{1}{N(v)} \sum_{i \in v_1} (y_i - y(v))^2$$

where $y(v)$ is the average of the y-values in v_i. Specifically

$$y(v_1) = 9.0, \quad y(v_2) = 3.0$$

And therefore, with a similar calculation for the set at the root node v_0 corresponding to all 6 observations,

$$y(v_0) = 7$$

Therefore:

$$I(v_0) = 11.67, \quad I(v_1) = 5.0, \quad I(v_2) = 1.0$$

these are finally combined to the impurity gain as

$$\Delta = I(v_0) - \sum_{k=1}^{2} \frac{N(v_k)}{N} I(v_k)$$

where for instance $N(v_1) = 4$ are the number of observations in branch 1. We find by insertion that $\Delta = 8.0$ and hence C is correct.

Question 17. Consider once more the Beijing air pollution dataset treated as a regression problem where the goal is to predict y_r. We wish to do this using KNN regression using $K = 3$ neighbors. We will simplify the problem by only considering the first $N = 6$ observations whose pairwise distances are given in Table 3, and their corresponding y_r-value can be found in Table 7.

Suppose we evaluate the leave-one-out estimate of the generalization error defined as

$$E = \frac{1}{N}\sum_{i=1}^{N} L(y_{r,i}, \hat{y}_{r,i})$$

where $y_{r,i}$ is the y_r-value of observation i, $\hat{y}_{r,i}$ is the predicted value and L is the standard squared (Euclidian) loss.

It is too time-consuming to compute the full LOO estimate of the generalization error, but what is the contribution from observation $i = 1$?

A. $L(y_{r,1}, \hat{y}_{r,1}) = 6.667$

B. $L(y_{r,1}, \hat{y}_{r,1}) = 28.444$

C. $L(y_{r,1}, \hat{y}_{r,1}) = 6.0$

D. $L(y_{r,1}, \hat{y}_{r,1}) = 7.111$

E. Don't know.

Solution 17.

First, notice that by a simple lookup in Table 7 that $y_{r,1} = 12$. To compute the predicted value, note that the K-nearest neighbors to observation $i = 1$ (but not including 1 itself) are observations

$$\{o_2, o_4, o_5\}$$

according to Table 3. The predicted value is then the mean of the corresponding y-values according to table Table 7 or

$$\hat{y}_{r,1} \approx 6.667$$

We can then simply compute the squared loss as:

$$L(y_{r,1}, \hat{y}_{r,1}) = (y_{r,1} - \hat{y}_{r,1})^2 = 28.444$$

and therefore option B is correct.

Fold	M_1/M_2	$M_1/\overline{M}_2$	$\overline{M}_1/M_2$	$\overline{M}_1/\overline{M}_2$
1	134	40	24	47
2	141	31	26	48
3	131	23	25	66
4	132	30	25	58

Table 8: Outcome of cross-validation. Rows are combination of outcomes of the two models.

Question 18. We will consider the Beijing air pollution dataset, and compare two models for predicting the class label y. Specifically, let M_1 be a $K = 1$ nearest neighbor classification model and M_2 a $K = 5$ nearest neighbor classification model. To compare them statistically, we perform $K = 4$ fold cross-validation, and for each fold we record the number of observations where both models are correct (as M_1/M_2), M_1 is correct and M_2 wrong (as $M_1/\overline{M_2}$), and so on. The outcome can be found in Table 8.

These results are sufficient to perform the McNemar test to compare the performance difference, i.e. the difference in accuracy, of model M_1 and M_2. According to the McNemar test, what is the estimated difference in accuracy

$$\hat{\theta} = \text{acc}(M_1) - \text{acc}(M_2)$$

of the two models?

A. $\hat{\theta} = 0.75$

B. $\hat{\theta} = 0.07$

C. $\hat{\theta} = 0.11$

D. $\hat{\theta} = 0.02$

E. Don't know.

Solution 18.

While the accuracies can easily be computed explicitly from the information in Table 8, a simpler solution (which is more true to the lecture notes) is to observe the differences in accuracies is

$$\hat{\theta} = \frac{n_{12} - n_{21}}{N}$$

Where n_{12} are the number of times model M_1 is correct and M_2 is false (i.e. the sum of column 2) and similarly n_{21} is the sum of column 3 and finally $N = 981$ is the number of observations. We find

$$\hat{\theta} = \frac{124 - 100}{N} = 0.02$$

and therefore D is correct.

Question 19. We will again consider the result of the two KNN models in Table 8 as evaluated over the $K = 4$ folds. What is the Jeffreys $\alpha = 0.05$ confidence interval $[\theta_L, \theta_U]$ of the model M_2?

A.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 538.5, b = 443.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 538.5, b = 443.5)$$

B.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 638.5, b = 343.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 638.5, b = 343.5)$$

C.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 538.5, b = 219.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 538.5, b = 219.5)$$

D.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 662.5, b = 319.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 662.5, b = 319.5)$$

E. Don't know.

Solution 19.

Since the cross-validation folds are non-overlapping, we can easily find the number of times model M_2 makes a correct prediction as the sum of columns 1 and 3 in Table 8 or $n^+ = 638$. Similarly, the sum of all entries in column 2 and 4 are the number of wrong guesses or $n^- = 343$.

The lower limit of the Jeffreys interval is now defined as

$$\theta_L = \text{cdf}_B^{-1}\left(\frac{\alpha}{2} \,\middle|\, a = n^+ + \frac{1}{2}, b = n^- + \frac{1}{2}\right)$$

and the upper limit can be found from the same expression by replacing $\frac{\alpha}{2}$ with $1 - \frac{\alpha}{2}$. Therefore, B is correct.

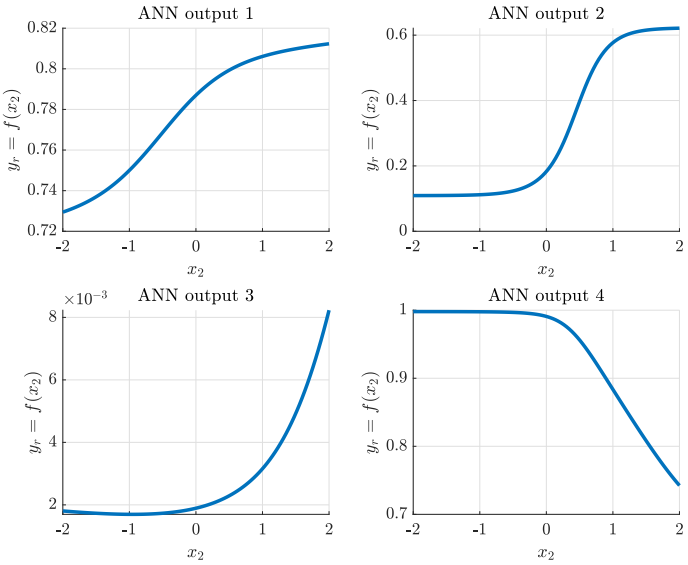

Figure 9: Suggested outputs of an ANN trained on the attribute x_2 from the Beijing air pollution dataset to predict y_r.

Question 20. ***Notice:*** *The version of question 20 in the main exam set contains a minor misprint in the axis on Figure 8 and the text to Figure 8. The misprint has been corrected in this version. Use this version when answering the question.*

We will consider an artificial neural network (ANN) trained on the Beijing air pollution dataset described in Table 1 to predict y_r from the attribute x_2. Suppose the neural network takes the form:

$$f(x, \boldsymbol{w}) = h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x]\boldsymbol{w}_j^{(1)})\right).$$

where the activation functions are selected as $h^{(1)}(x) = \sigma(x)$ (the logistic sigmoid activation function) and $h^{(2)}(x) = \sigma(x)$ (the logistic sigmoid activation function) and the weights are given as:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -0.5 \\ -0.1 \end{bmatrix}, \quad \boldsymbol{w}_2^{(1)} = \begin{bmatrix} 0.9 \\ 2.0 \end{bmatrix},$$
$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -1.0 \\ 0.4 \end{bmatrix}, \quad w_0^{(2)} = 1.4.$$

Which one of the curves in Figure 9 will then corre-

spond to the function f?

A. ANN output 1

B. ANN output 2

C. ANN output 3

D. ANN output 4

E. Don't know.

Solution 20.

It suffices to compute the activation of the neural network at $x_2 = -2$. The activation of each of the two hidden neurons is:

$$n_1 = h^{(1)}(\begin{bmatrix}1 & -2\end{bmatrix} \boldsymbol{w}_1^{(1)}) = 0.426$$
$$n_2 = h^{(1)}(\begin{bmatrix}1 & -2\end{bmatrix} \boldsymbol{w}_2^{(1)}) = 0.043.$$

The final output is then computed as:

$$f(x, \boldsymbol{w}) = h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}(\begin{bmatrix}1 & x_1 & x_2\end{bmatrix} \boldsymbol{w}_j^{(1)})\right)$$
$$= h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} n_j\right) = 0.729.$$

This rules out all options except A.

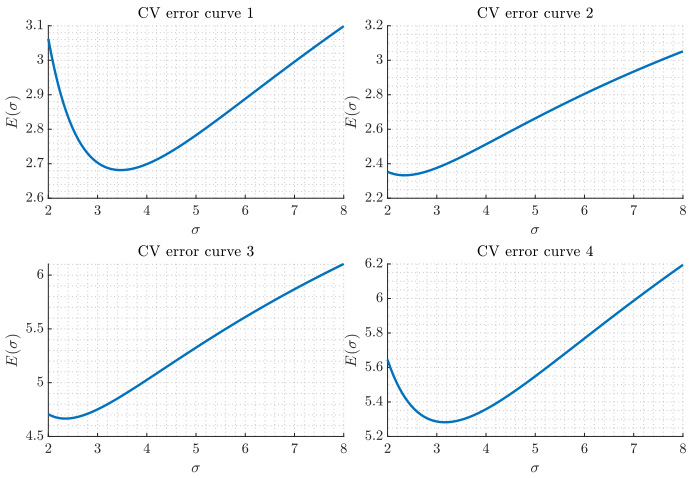

Figure 10: Estimated negative log-likelihood as obtained using hold-out cross validation on a small, $N = 3$ one-dimensional dataset as a function of kernel width σ.

Question 21. Consider the following $N = 3$ observations of the attribute CO from the Beijing air pollution dataset described in Table 1.

$$x_5 : \quad \begin{bmatrix}4.5 & -0.5 & 1.2\end{bmatrix}.$$

Suppose we apply a Kernel Density Estimator (KDE) to the dataset with kernel width σ (i.e., σ is the standard deviation of the Gaussian kernels), and we wish to find σ by using hold-out cross validation (CV) using the average (per observation) negative log-likelihood

$$E(\sigma) = \frac{-1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \log p_\sigma(x_i).$$

We construct the hold out split by considering the first 2 observations a training set and the last observations as a test set.

Which of the cross validation curves in Figure 10 shows the cross-validation estimate of the generalization error $E(\sigma)$?

A. CV error curve 1

B. CV error curve 2

C. CV error curve 3

D. CV error curve 4

E. Don't know.

Solution 21. To solve the problem, we will compute the hold-out cross-validation estimate of the generalization error at $\sigma = 2$. To do so, recall the density at each observation i, when the KDE is fitted on the other $N-1$ observations, is:

$$p_\sigma(x_i) = \frac{1}{N-1}\sum_{j\neq i}\mathcal{N}(x_i|x_j, \sigma = 2)$$

Therefore, training on the two first observations and testing on the last simply corresponds to evaluating this expression for $i = 3$ i.e.:

$$p_\sigma(x_3) = 0.095$$

The CV hold-out error is the average of the test set, but since the test set only contains a single observation it is equal to minus the log of the above expression. In other words

$$\begin{aligned} E(\sigma = 2) &= \frac{1}{N^{\text{test}}}\sum_{i=1}^{N^{\text{test}}} -\log p_\sigma(x_i) \\ &= -\log p_\sigma(x_3) = 2.353. \end{aligned}$$

Therefore, the correct answer is B.

Variable	y^{true}	$t = 1$
y_1	1	1
y_2	1	2
y_3	1	1
y_4	2	1
y_5	2	1
y_6	2	2
y_7	2	2

Table 9: For each of the $N = 7$ observations (first column), the table indicate the true class labels y^{true} (second column) and the predicted outputs of the AdaBoost classifier (third column) for the first round of boosting $t = 1$.

Question 22. Consider again the Beijing air pollution dataset of Table 1. Suppose we limit ourselves to $N = 7$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_6 and x_9. We use a KNN classification model ($K = 1$) to this dataset and apply AdaBoost to improve the performance. After the first round of boosting, we obtain predictions and the true class labels as tabulated in Table 9. Given this information, how will the AdaBoost update the weights $\boldsymbol{w}$?

A. $[0.173 \quad 0.103 \quad 0.173 \quad 0.103 \quad 0.103 \quad 0.173 \quad 0.173]$

B. $[0.146 \quad 0.138 \quad 0.146 \quad 0.138 \quad 0.138 \quad 0.146 \quad 0.146]$

C. $[0.125 \quad 0.167 \quad 0.125 \quad 0.167 \quad 0.167 \quad 0.125 \quad 0.125]$

D. $[0.102 \quad 0.198 \quad 0.102 \quad 0.198 \quad 0.198 \quad 0.102 \quad 0.102]$

E. Don't know.

Solution 22.

We first observe the AdaBoost classifier at $t = 1$ mis-classify observations:

$$\{y_2, y_4, y_5\}$$

Since the weights are just $w_i = \frac{1}{N}$, we therefore get:

$$\epsilon_{t=1} = \sum_i w_i(t)(1 - \delta_{f_t(x_i),y_i}) = 0.429$$

From this, we compute α_t as

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} = 0.144$$

Scaling the observations corresponding to the misclassified weights as $w_i e^{\alpha_t}$ and those corresponding to the correctly classified weights as $w_i e^{-\alpha_t}$ and normalizing the new weights to sum to one then give answer C.

Question 23. Consider the Beijing air pollution dataset from Table 1 consisting of $N = 981$ observations, and suppose the attribute O_3 concentration ($\mu g/m^3$) has been binarized into low and high values. We still consider the goal to predict the pollution level. Given the following information

- Of the 391 observations with light pollution level, 64 had a high value of O_3 concentration ($\mu g/m^3$)
- Of the 241 observations with medium pollution level, 66 had a high value of O_3 concentration ($\mu g/m^3$)
- Of the 349 observations with high pollution level, 206 had a high value of O_3 concentration ($\mu g/m^3$)

and supposing a particular observation has a low value of O_3 concentration ($\mu g/m^3$), what is the probability of observing medium pollution level?

A. 0.271

B. 0.192

C. 0.044

D. 0.141

E. Don't know.

Solution 23. The problem is solved by applying Bayes rule. Introducing the binary variable x such that $x = 1$ if an observation has a high value of O_3 concentration ($\mu g/m^3$) (and otherwise $x = 0$) the question asked is equivalent to computing $p(y = 2|x = 0)$. Applying Bayes' theorem we get:

$$p(y = 2|x = 0) = \frac{p(x = 0|y = 2)p(y = 2)}{\sum_{k=1}^{3} p(x = 0|y = k)p(y = k)}$$

Recall that $p(x = 0|y) = p(x = 1|y)$, we can obtain the required probabilities from each of the three bullet points above. We obtain:

- $p(y = 1) = \frac{391}{N}$ and $p(x = 1|y = 1) = \frac{64}{391}$.
- $p(y = 2) = \frac{241}{N}$ and $p(x = 1|y = 2) = \frac{66}{241}$.
- $p(y = 3) = \frac{349}{N}$ and $p(x = 1|y = 3) = \frac{206}{349}$.

Plugging these into Bayes theorem, and using that $p(x = 0|y) = 1 - p(x = 1|y)$ because x is binary, we see $p(y = 2|x = 0) = 0.271$ and hence that option A is correct.

Question 24. Consider again the Beijing air pollution dataset. We consider a multinomial regression model applied to the dataset projected onto the first two principal directions, i.e. applied to observations of the form $\boldsymbol{x} = \begin{bmatrix} b_1 & b_2 \end{bmatrix}^\top$ where b_1 and b_2 are the coordinates of the PCA projections.

In the notation of the lecture notes, suppose the weight-vectors in the multinomial regression model are

$$w_1 = \begin{bmatrix} 0.04 \\ 1.32 \\ -1.48 \end{bmatrix}, \quad w_2 = \begin{bmatrix} -0.03 \\ 0.7 \\ -0.85 \end{bmatrix}.$$

What is the class-assignment probability vector $\tilde{\boldsymbol{y}}$ for the input observation with coordinates $b_1 = -5.52$, $b_2 = -4.69$?

A. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.77 & 0.23 & 0.0 \end{bmatrix}^\top$

B. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.26 & 0.39 & 0.35 \end{bmatrix}^\top$

C. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.16 & 0.24 & 0.6 \end{bmatrix}^\top$

D. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.22 & 0.07 & 0.72 \end{bmatrix}^\top$

E. Don't know.

Solution 24. Let $\boldsymbol{b}$ be the input vector. Then:

$$\tilde{\boldsymbol{b}} = \begin{bmatrix} 1.0 \\ -5.52 \\ -4.69 \end{bmatrix}.$$

Recall the class-assignment probability vector is computed as

$$P(y = k|\boldsymbol{x}) = \begin{cases} \frac{e^{\hat{y}_k}}{1+\sum_{k'=1}^{2} e^{\hat{y}_{k'}}}. & \text{if } k \leq 2 \\ \frac{1}{1+\sum_{k'=1}^{2} e^{\hat{y}_{k'}}} & \text{if } k = 3. \end{cases}$$

in the case of multinomial regression we have

$$\hat{y}_1 = \tilde{\boldsymbol{b}}^T \boldsymbol{w}_1 \approx -0.305 \quad \hat{y}_2 = \tilde{\boldsymbol{b}}^T \boldsymbol{w}_2 \approx 0.093$$

Simply inserting these number we get that the first coordinate of the class-assignment probability vector is:

$$p(y = 1|\boldsymbol{x}) = \frac{e^{\hat{y}_1}}{1 + e^{\hat{y}_1} + e^{\hat{y}_2}} = 0.26$$

(and similar for the other values of y). From this B is evidently correct.

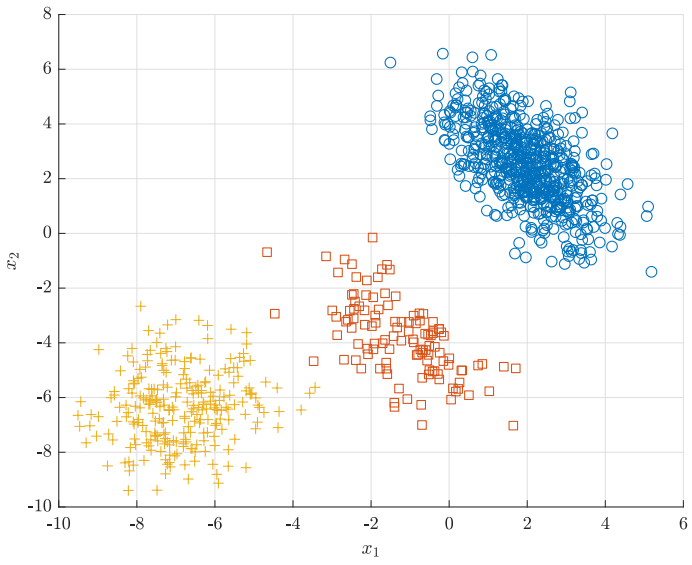

Figure 11: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

Question 25. Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In Figure 11 is given 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Each observation is colored and marked in terms of which cluster it came from in the Gaussian Mixture model.
Which one of the following GMM densities was used to

generate the data?

A.

$$p(\boldsymbol{x}) = \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right) + \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right)$$

B.

$$p(\boldsymbol{x}) = \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right) + \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right) + \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right)$$

C.

$$p(\boldsymbol{x}) = \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right) + \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right)$$

D.

$$p(\boldsymbol{x}) = \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right) + \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right)$$

E. Don't know.

Solution 25.

B The three components in the candidate GMM densities can be matched to the colored observations by their mean values. Then, by considering the basic properties of the covariance matrices, we can easily rule out all options except A. Alternatively, in Figure 12 is shown the densities for densities corresponding to option B (upper left), C (upper right) and D (bottom center).

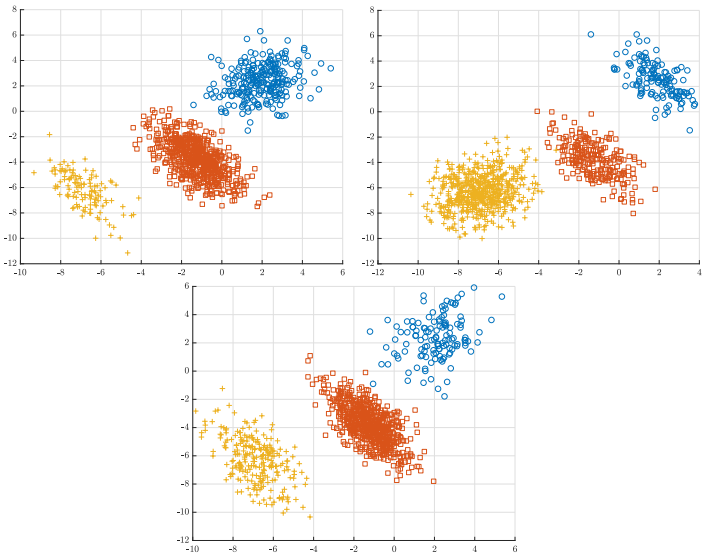

Figure 12: GMM mixtures corresponding to alternative options.

Question 26. Consider the following four classifiers:

MREG: Multinomial regression

ANN: Artificial neural network with 5 hidden units

CT: Classification tree with regular axis-aligned splits ($b_i < c$)

KNN: K-nearest neighbours with $K = 3$

Suppose the classifiers are trained on a subset of the Beijing air pollution dataset described in Table 1 after it has been projected onto the first two principal components b_1 and b_2 from Equation (1). The decision boundary for each of the four classifiers is given in Figure 13. Which one of the following statements is correct?

A. Classifier 1 corresponds to **ANN**,
Classifier 2 corresponds to **CT**,
Classifier 3 corresponds to **MREG**,
Classifier 4 corresponds to **KNN**.

B. Classifier 1 corresponds to KNN,
Classifier 2 corresponds to CT,
Classifier 3 corresponds to MREG,
Classifier 4 corresponds to ANN.

C. Classifier 1 corresponds to **CT**,
Classifier 2 corresponds to **MREG**,

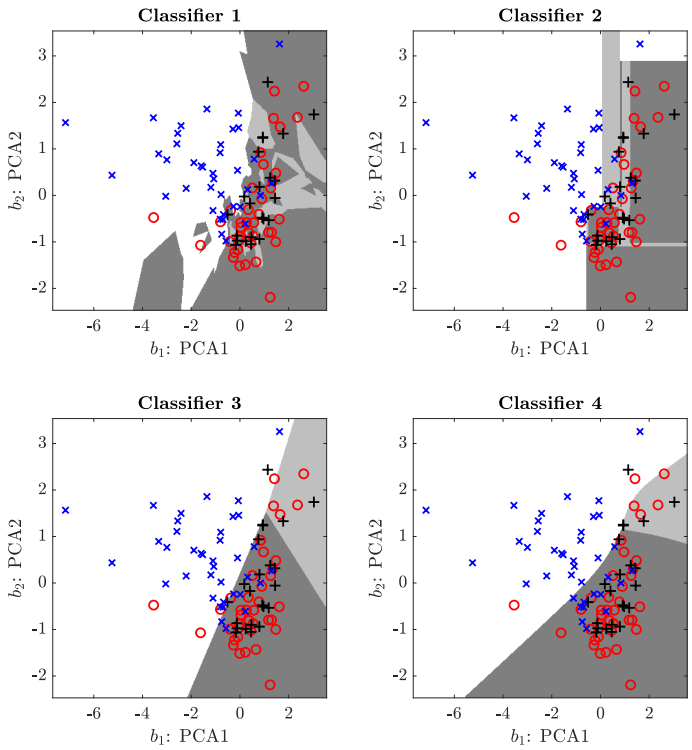

Figure 13: Decision boundaries for four different classifiers trained on the Beijing air pollution dataset when projected onto the first two principal components.

Classifier 3 corresponds to **ANN**,

Classifier 4 corresponds to **KNN**.

D. Classifier 1 corresponds to **ANN**,

Classifier 2 corresponds to **KNN**,

Classifier 3 corresponds to **MREG**,

Classifier 4 corresponds to **CT**.

E. Don't know.

Solution 26. To solve this problem, we have to use our intuition about what the typical decision boundaries for the different methods look like:

- A KNN method will have decision boundaries dictated by the nearest neighbors. That is, points (x, y) where the nearest K neighbors are in one class must be in the same class and therefore the boundaries will be fairly complex and respect the data distribution well.
- A decision tree has axis aligned splits, therefore the boundaries must be vertical or horizontal
- A multivariate regression model must have linear boundaries
- An artificial neural network with few hidden units can have some non-linearity, but otherwise have boundaries of limited complexity and consisting of relatively simple shapes

It is easy to see this rules out all but option B.

Question 27. Consider a small dataset comprised of $N = 10$ observations

$$x = \begin{bmatrix} 0.4 & 0.5 & 1.1 & 2.2 & 2.6 & 3.0 & 3.6 & 3.7 & 4.9 & 5.0 \end{bmatrix}.$$

Suppose a k-means algorithm is applied to the dataset with $K = 3$ and using Euclidian distances. The algorithm is initialized with K cluster centers located at

$$\mu_1 = 2.4, \ \mu_2 = 3.3, \ \mu_3 = 3.5$$

What clustering will the k-means algorithm eventually converge to?

A. $\{0.4, 0.5, 1.1, 2.2\}$, $\{2.6, 3, 3.6\}$, $\{3.7, 4.9, 5\}$

B. $\{0.4, 0.5, 1.1\}$, $\{2.2, 2.6, 3, 3.6\}$, $\{3.7, 4.9, 5\}$

C. $\{0.4, 0.5, 1.1, 2.2, 2.6\}$, $\{3, 3.6, 3.7, 4.9\}$, $\{5\}$

D. $\{0.4, 0.5, 1.1, 2.2, 2.6\}$, $\{3, 3.6, 3.7\}$, $\{4.9, 5\}$

E. Don't know.

Solution 27. Recall the K-means algorithm iterates between assigning the observations to their nearest centroids, and then updating the centroids to be equal to the average of the observations assigned to them. Given the initial centroids, the K-means algorithm assign observations to the nearest centroid resulting in the partition:

$$\{0.4, 0.5, 1.1, 2.2, 2.6\}, \{3\}, \{3.6, 3.7, 4.9, 5\}.$$

Therefore, the subsequent steps in the K-means algorithm are:

Step $t = 1$: The centroids are computed to be:

$$\mu_1 = 1.36, \ \mu_2 = 3, \ \mu_3 = 4.3.$$

And the updated assignment of observations to nearest centroids results in the clustering:

$$\{0.4, 0.5, 1.1\}, \{2.2, 2.6, 3, 3.6\}, \{3.7, 4.9, 5\}.$$

Step $t = 2$: The centroids are computed to be:

$$\mu_1 = 0.666667,\ \mu_2 = 2.85,\ \mu_3 = 4.53333.$$

And the updated assignment of observations to nearest centroids results in the clustering:

$$\{0.4, 0.5, 1.1\},\ \{2.2, 2.6, 3, 3.6\},\ \{3.7, 4.9, 5\}.$$

At this point, the centroids are no longer changing and the algorithm terminates. Hence, B is correct.

Technical University of Denmark

Written examination: December 16th 2020, 9:00–13:30.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: $4\frac{1}{2}$ 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`-file (with `bmp`, `png`, `tiff` and `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)[1].

Failing to timely upload both documents will count as not having handed in the exam.

Guessing on an answer is not allowed for the online exam, as each answer has to include an accompanying argumentation in writing for the answer.

Questions, where the answers in the `answers.txt` file (file 1) differ from the explanation (file 2) or where explanations are insufficient or unreadable, will be treated as "Don't know". Systematic discrepancy between the answers in the two hand-in files will potentially count as an attempt of cheating the exam.

Only in the exceptional case, where the exam submission system stops working, you should send your two files in a single email to `exam-02450@compute.dtu.dk`. Files sent to the email address will only be accepted in the rare event that the exam submission system fails.

In the event that there is an error in the exam, then you should contact Morten Mørup (`mmor@dtu.dk`) using your study/DTU email.

Answers:

1	2	3	4	5	6	7	8	9	10
C	B	C	A	C	D	A	C	C	B
11	12	13	14	15	16	17	18	19	20
C	D	B	C	C	A	D	B	A	B

[1]The *original* file format must be either zip or PDF.

21	22	23	24	25	26	27
D	A	B	C	C	D	C

No.	Attribute description	Abbrev.
x_1	Bill[2] length (millimeters)	Bill length
x_2	Bill depth (millimeters)	Bill depth
x_3	Flipper length (millimeters)	Flipper length
x_4	Body mass (grams)	Mass
x_5	Penguin sex (1=male, 2=female)	Sex
x_6	Study year (2007, 2008, or 2009)	Year
y		Species

Table 1: Description of the features of the Palmer Penguins dataset used in this exam. The dataset is collected in the Palmer Archipelago (Antarctica) and contains $M = 6$ attributes for three different species of penguins. The dataset has been pre-processed such that data objects with missing values have been removed. The binary attribute *Sex* is encoded as an integer, where a male penguin takes the value $x_5 = 1$ and a female penguin takes the value $x_5 = 2$. For classification, the objective is to predict the species, and the output variable y is taking values $y = 1$ (corresponding to an Adelie), $y = 2$ (corresponding to a Gentoo), and $y = 3$ (corresponding to a Chinstrap). After removing missing values, there are $N = 333$ observations in total.

Question 1. The main dataset used in this exam is the Palmer Penguins dataset[3] described in Table 1. We will consider the type of an attribute *as the highest level* it obtains in the type-hierarchy (nominal, ordinal, interval and ratio). Which one of the following statements is true about the types of the attributes $x_1, \ldots, x_6$ and the output y in the Palmer Penguins dataset?

A. All attributes except x_5 (*Sex*) are ratio.

B. x_5 (*Sex*) and x_6 (*Year*) are both ordinal.

C. x_5 (*Sex*) and y (*Species*) are both nominal.

D. x_4 (*Mass*) and x_6 (*Year*) are both interval.

E. Don't know.

Solution 1. The problem is solved by simply thinking about what the attributes represent and comparing them to the definition in the different types. Recall that

- Nominal is a type that only allow comparison (equal or different)
- Ordinal allows ordering (but not differences)
- Interval allows differences but no (physically well-defined) zero
- Ratio is a type with a zero with a well-defined meaning

With these definitions, we see that

x_1 (***Bill length***) is ratio

x_2 (***Bill depth***) is ratio

x_3 (***Flipper length***) is ratio

x_4 (***Mass***) is ratio

x_5 (***Sex***) is nominal

x_6 (***Year***) is interval

y (***Species***) is nominal

and therefore option C is correct.

[2] Bill is synonymous for beak.

[3] Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. `https://allisonhorst.github.io/palmerpenguins/`

	mean	$x_{p=25\%}$	$x_{p=50\%}$	$x_{p=75\%}$
Bill length	43.99	39.45	44.50	48.63
Bill depth	17.16	15.60	17.30	18.70
Flipper length	200.97	190.00	197.00	213.00
Mass	4207.06	3550.00	4050.00	4781.25

Table 2: Summary statistics of four attributes from the Palmer Penguins 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.

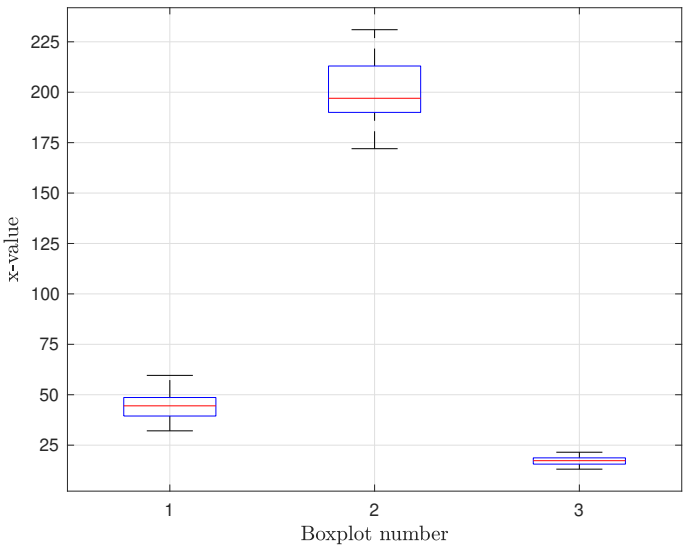

Figure 1: Boxplots corresponding to three of the variables with summary statistics indicated in Table 2 but not necessarily in that order.

Question 2. Table 2 contains the summary statistics of the four first attributes (x_1, x_2, x_3, x_4) from the Palmer Penguins dataset. Which boxplots in Figure 1 match which attributes?

A. Attribute *Bill length* corresponds to boxplot 1, *Bill depth* corresponds to boxplot 2, and *Flipper length* corresponds to boxplot 3.

B. Attribute *Bill length* corresponds to boxplot 1, *Flipper length* corresponds to boxplot 2, and *Bill depth* corresponds to boxplot 3.

C. Attribute *Flipper length* corresponds to boxplot 1, *Mass* corresponds to boxplot 2, and *Bill length* corresponds to boxplot 3.

D. Attribute *Bill depth* corresponds to boxplot 1, *Flipper length* corresponds to boxplot 2, and *Bill length* corresponds to boxplot 3.

E. Don't know.

Solution 2. We can read off the medians (red line) of the boxplots. We observe that

- Boxplot 1 has median between 25 and 50, and the only attribute with a median in this interval is *Bill length*.
- Boxplot 2 has median between 175 and 200, and the only attribute with a median in this interval is *Flipper length*.
- Boxplot 3 has median below 25, and the only attribute with a median below this value is *Flipper depth*.

Therefore B is the only correct answer.

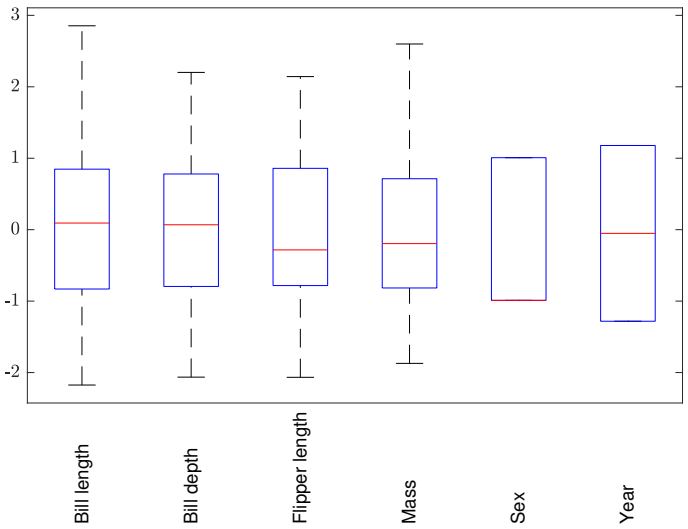

Figure 2: Boxplots of the six attributes (x_1, x_2, x_3, x_4, x_5, x_6) from the Palmer Penguins dataset. The attributes are standardized.

Question 3. Figure 2 shows boxplots of the six attributes (x_1, x_2, x_3, x_4, x_5, x_6) from the Palmer Penguins dataset. The attributes are standardized (i.e., the mean has been subtracted and the attributes divided by their standard deviations). Which one of the following statements about the original dataset can be concluded from the boxplots (i.e., based only on information regarding the data provided by Figure 2)?

A. The variance of *Bill length* is larger than the variance of *Flipper length*.

B. For *Flipper length* the mean and median values coincide.

C. There are more male penguins than female penguins.

D. *Bill length* and *Bill depth* have positive correlation.

E. Don't know.

Solution 3.

- A is false, since the boxplots are based on standardized data (with unit variance) and therefore it cannot inform about the variance of the original data (in fact bill length has smaller variance than flipper length).
- B is false, since you cannot read of the mean from a boxplot.
- C is true, since the boxplot shows that the median is close to -1. As males are encoded as 1 and female as 2, the negative median after standardization indicates that more had the lower value (in fact there are 168 males and 165 females).
- D is false, since boxplots do not inform about correlation, but only summarizes the individual feature (in fact it turns out that the two features have negative correlation).

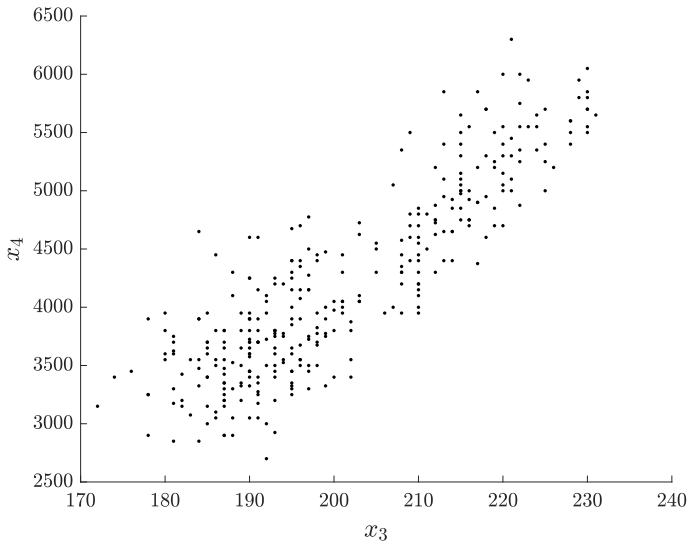

Figure 3: Black dots show attributes x_3 and x_4 of the Palmer Penguins dataset.

Question 4. Figure 3 shows a scatterplot for the two attributes x_3 (*Flipper length*) and x_4 (*Mass*) of the Palmer Penguins dataset. The two attributes have a positive correlation coefficient $\rho = 0.87$ and covariance $\text{cov}(x_3, x_4) = 9852$. Which of the following are the best estimates for the variance of x_3 and variance of x_4?

A. $\sigma^2_{x_3} = 196$ **and** $\sigma^2_{x_4} = 648025$

B. $\sigma^2_{x_3} = 38$ and $\sigma^2_{x_4} = 298$

C. $\sigma^2_{x_3} = 648025$ and $\sigma^2_{x_4} = 196$

D. $\sigma^2_{x_3} = 298$ and $\sigma^2_{x_4} = 38$

E. Don't know.

Solution 4. From Figure 3 we can observe that the spread of x_3 is smaller than the spread of x_4, i.e., $\sigma_{x_3} < \sigma_{x_4}$. This means that we can rule out answer C and D.

The correlation coefficient between x_3 and x_4 is defined as

$$\rho = \frac{\text{cov}(x_3, x_4)}{\sigma_{x_3}\sigma_{x_4}}.$$

If we use the number from A, we find that

$$\rho = \frac{9852}{\sqrt{196}\sqrt{648025}} \approx 0.87 \ ,$$

which means that this is the correct answer.

Question 5. A Principal Component Analysis (PCA) is carried out on the Palmer Penguins dataset in Table 1 based on the attributes x_1, x_2, x_3, x_4. 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 $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.45 & -0.60 & -0.64 & 0.15 \\ -0.40 & -0.80 & 0.43 & -0.16 \\ 0.58 & -0.01 & 0.24 & -0.78 \\ 0.55 & -0.08 & 0.59 & 0.58 \end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix} 30.19 & 0.0 & 0.0 & 0.0 \\ 0.0 & 16.08 & 0.0 & 0.0 \\ 0.0 & 0.0 & 11.07 & 0.0 \\ 0.0 & 0.0 & 0.0 & 5.98 \end{bmatrix}.$$

Which one of the following statements is *correct*?

A. The first principal component accounts for less than 50 percent of the variance.

B. The first two principal components account for more than 90 percent of the variance.

C. The first three principal components account for more than 95 percent of the variance.

D. The last principal component accounts for more than 3 percent of the variance.

E. Don't know.

Solution 5. The correct answer is D. To see this, recall 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 components x_1, x_2, x_3 is:

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2} = 0.9731.$$

Question 6. Consider again the principal component analysis described in Question 5. Recall that the d'th column of $\boldsymbol{V}$ defines the d'th principal component direction. Which one of the following statements is *correct*?

A. Any single observation with a high negative projection onto the second principal component will in general have high negative values for all four features x_1, x_2, x_3, x_4.

B. The first column vector of $\boldsymbol{V}$ is longer than the second column vector of $\boldsymbol{V}$ measured by the Euclidean norm (2-norm).

C. The first principal component primarily separates penguins with relatively short *Flipper length* and high *Body mass* from penguins with relatively long *Flipper length* and low *Body mass*.

D. The fourth principal component primarily separates penguins with relatively short *Flipper length* and high *Body mass* from penguins with relatively long *Flipper length* and low *Body mass*.

E. Don't know.

Solution 6.

- A is false: since all the coordinates in the second PC are negative, an observation with high negative values for all features will have a high positive projection onto the second principal component.
- B is false: since all column vectors have unit length.
- C is false: The first PC has nearly the same positive value for both *Flipper length* (x_3) and *Body mass* (x_4). Therefore, flipping the values of x_3 and x_4 would nearly give the same projection.
- D is true: The forth PC has a large negative coefficient for flipper length (x_3) and a large positive coefficient for body mass (x_4). Therefore, the forth PC primarily separates penguins with relatively low x_3 (short flipper length) and high x_4 (high body mass) from penguins with relatively high x_3 (long flipper length) and low x_4 (low body mass).

Question 7. Based on the principal component analysis described in Question 5, Figure 4 shows 2D scatter plots for the Palmer Penguin dataset projected onto different combinations of the principal components. The class labels y (penguin species *Adelie, Gentoo, Chinstrap*) are indicated in the plots.

Consider a new observation $\boldsymbol{x}$ for which the four features x_1, x_2, x_3, x_4 are standardized (by subtracting the mean from each column and dividing by the standard deviations), resulting in the point $\tilde{\boldsymbol{x}}$ with coordinates:

$$\tilde{x}_1 = -1,\ \tilde{x}_2 = -1,\ \tilde{x}_3 = -1,\ \tilde{x}_4 = 1.$$

Consider the point projected onto the principal components. Furthermore, consider a k-nearest neighbor classification with $k = 1$ using the Euclidean distance measure and using the projected dataset shown in the individual 2D plots as the training data. In which one of the following 2D projections (shown in Figure 4) will the 1-nearest neighbor classifier classify the point $\tilde{\boldsymbol{x}}$ as a *Chinstrap*?

A. Data projected onto PC1 and PC4.

B. Data projected onto PC2 and PC4.

C. Data projected onto PC2 and PC3.

D. Data projected onto PC3 and PC4.

E. Don't know.

Solution 7. The point $\tilde{\boldsymbol{x}}^T = [-1, -1, -1, 1]$ is projected to the vector with coordinates

$$\tilde{\boldsymbol{x}}^T \boldsymbol{V} = [-0.08, 1.33, 0.56, 1.37]^T$$

which can be seen to only being closest to a '+' in one figure (PC1 vs. PC4), which is illustrated in Figure 5.

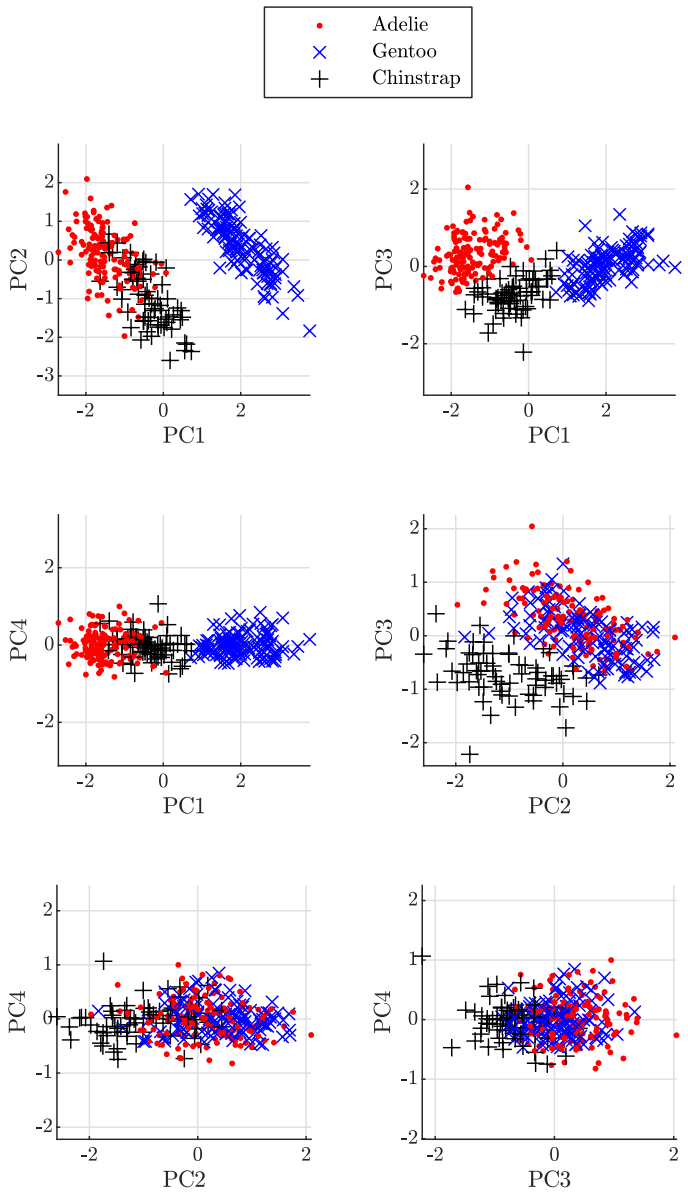

Figure 4: Projection of the Palmer Penguins dataset onto different combinations of two principals components obtained from the principal component analysis described in question 5. The three species of penguins are indicated with an *Adelie* as a red dot, a *Gentoo* as a blue 'x' and a *Chinstrap* as a black '+'.

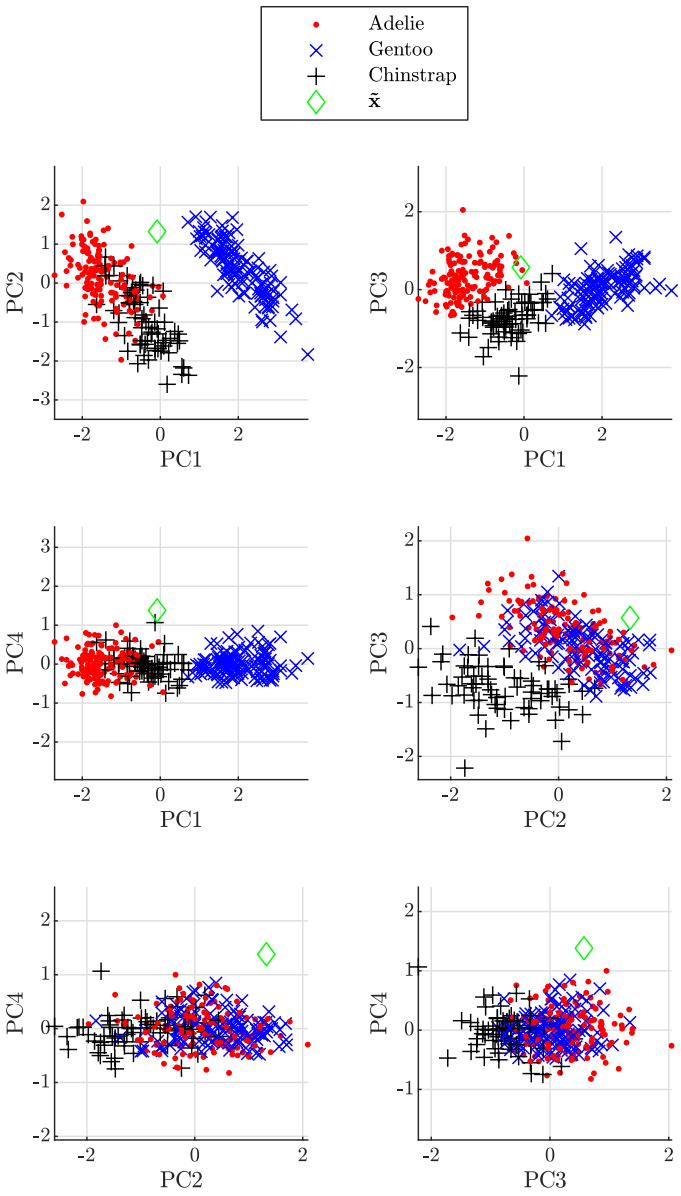

Figure 5: The point $\tilde{\boldsymbol{x}}$ (green diamond) project onto the different combinations of two principals components.

Figure 6: Confusion matrix obtained by comparing the actual classes in the Palmer Penguins dataset with three cluster obtained by the k-means algorithm for $k = 3$.

Question 8. The k-means algorithm with $k = 3$ was applied to the Palmer Penguins dataset. Figure 6 shows the confusion matrix, where the cluster assignment obtained by the k-means algorithm is compared to the ground truth class label.

We wish to compare the quality of the k-means clustering, Z, to the ground truth clustering, Q. What is the Rand index (Rand similarity) between Z and Q?

A. $R(Z, Q) = 0.72$

B. $R(Z, Q) = 0.79$

C. $R(Z, Q) = 0.87$

D. $R(Z, Q) = 0.96$

E. Don't know.

Solution 8. First note that the confusion matrix is equivalent to the counting matrix $\boldsymbol{n}$ that can be used to calculate the Rand similarity.

We can calculate the number of times Z and Q agrees two observations are S or are not in D the same cluster by

$$\begin{aligned} S &= \sum_{k=1}^{3}\sum_{m=1}^{3} \frac{n_{km}(n_{km}-1)}{2} \\ &= \frac{114\cdot 113 + 32\cdot 31 + 119\cdot 118 + 8\cdot 7 + 60\cdot 59}{2} \\ &= 15756 \end{aligned}$$

and

$$\begin{aligned} D &= \frac{N(N-1)}{2} - \sum_{k=1}^{K}\frac{n_k^Z(n_k^Z-1)}{2} - \sum_{m=1}^{M}\frac{n_m^Q(n_m^Q-1)}{2} + S \\ &= \frac{333\cdot 332}{2} - \frac{146\cdot 145 + 119\cdot 118 + 68\cdot 67}{2} \\ &\quad - \frac{122\cdot 121 + 119\cdot 118 + 92\cdot 91}{2} + 15756 \\ &= 32562. \end{aligned}$$

We then find the Rand similarity to be

$$R(Z, Q) = \frac{S + D}{\frac{1}{2}N(N-1)} = \frac{15756 + 32562}{\frac{1}{2}\cdot 333\cdot 332} = 0.87.$$

Question 9. Which one of the following machine-learning models is not trained by fitting any parameters?

A. Linear regression applied to data with one attribute.

B. Neural network with no hidden layers.

C. K nearest neighbor classification using the $K = 1$ nearest neighbour classification rule.

D. Multinomial regression applied to data with three output classes.

E. Don't know.

Solution 9. KNN is instance based with no learned parameters when K is fixed.

Question 10. Three of the following actions will typically reduce the amount of over-fitting, and one of them will typically increase it. Which option will typically *increase* the amount of over-fitting?

A. Reduce the number of attributes.

B. Reduce the amount of training data.

C. Select a less complex model.

D. Add model regularisation.

E. Don't know.

Solution 10. Reducing the amount of training data is the only option that will typically increase the amount of over-fitting. The other choices will typically decrease over-fitting.

$p(\hat{x}_1, \hat{x}_2 \mid y)$	$y=1$	$y=2$	$y=3$
$\hat{x}_1=0,\ \hat{x}_2=0$	0.23	0.15	0.07
$\hat{x}_1=0,\ \hat{x}_2=1$	0.75	0	0.01
$\hat{x}_1=1,\ \hat{x}_2=0$	0	0.85	0.16
$\hat{x}_1=1,\ \hat{x}_2=1$	0.02	0	0.76

Table 3: Probability of observing particular values of $\hat{x}_1$ and $\hat{x}_2$ conditional on y.

Question 11. Consider the Palmer Penguins dataset from Table 1. We wish to predict the species based on the attributes *Bill length* and *Bill depth* using a Bayes classifier. Suppose the attributes have been binarized such that $\hat{x}_1 = 0$ corresponds to $x_1 \leq 44.5$ (and otherwise $\hat{x}_1 = 1$) and $\hat{x}_2 = 0$ corresponds to $x_2 \leq 17.3$ (and otherwise $\hat{x}_2 = 1$). Suppose the probability for each of the configurations of $\hat{x}_1$ and $\hat{x}_2$ conditional on the species y are as given in Table 3 and the prior probability of the species are

$$p(y=1) = 0.44,\ p(y=2) = 0.36,\ p(y=3) = 0.20.$$

Using this, what is then the probability an observation is *Chinstrap* ($y=3$) given that $\hat{x}_1 = 1$ and $\hat{x}_2 = 0$?

A. $p(y=3|\hat{x}_1=1,\hat{x}_2=0) = 0.033$

B. $p(y=3|\hat{x}_1=1,\hat{x}_2=0) = 0.085$

C. $p(y=3|\hat{x}_1=1,\hat{x}_2=0) = 0.095$

D. $p(y=3|\hat{x}_1=1,\hat{x}_2=0) = 0.158$

E. Don't know.

Solution 11. The problem is solved by a simple application of Bayes' theorem:

$$p(y=3|\tilde{x}_1=1,\tilde{x}_2=0) = \frac{p(\tilde{x}_1=1,\tilde{x}_2=0|y=3)p(y=3)}{\sum_{k=1}^{3} p(\tilde{x}_1=1,\tilde{x}_2=0|y=k)p(y=k)}$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_1=1,\tilde{x}_2=0|y)$ in Table 3. Inserting the values we see option D is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0	725	800	150	1000	525	600	500	400	850
o_2	725	0	75	575	275	1250	1325	226	325	125
o_3	800	75	0	650	200	1325	1400	300	400	51
o_4	150	575	650	0	850	675	750	350	250	700
o_5	1000	275	200	850	0	1525	1600	500	600	150
o_6	525	1250	1325	675	1525	0	75	1025	925	1375
o_7	600	1325	1400	750	1600	75	0	1100	1000	1450
o_8	500	226	300	350	500	1025	1100	0	100	350
o_9	400	325	400	250	600	925	1000	100	0	450
o_{10}	850	125	51	700	150	1375	1450	350	450	0

Table 4: The pairwise Euclidean distances, $d(o_i, o_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 10 observations from the Palmer Penguins dataset (recall that $M = 4$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1. The colors indicate classes such that the black observations $\{o_1,\ o_2,\ o_3,\ o_4,\ o_5\}$ belong to class C_1 (corresponding to an Adelie), the red observations $\{o_6,\ o_7\}$ belong to class C_2 (corresponding to a Gentoo), and the blue observations $\{o_8,\ o_9,\ o_{10}\}$ belong to class C_3 (corresponding to a Chinstrap). The distances are rounded to integers.

Question 12. Table 4 shows the pairwise Euclidean distances between 10 observations $o_1, o_2, \ldots, o_{10}$ from the Palmer Penguins dataset. To examine if observation o_2 may be an outlier, we will calculate the average relative density using the Euclidean distance based on the observations given in Table 4 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_2 for $K = 2$ nearest neighbors?

A. 0.01

B. 0.37

C. 0.68

D. 0.73

E. Don't know.

Solution 12.

To solve the problem, first observe the $k = 2$ neighborhood of o_2 and density is:

$$N_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = \{o_3, o_{10}\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = 0.01$$

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

$$N_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = \{o_{10}, o_2\}, \quad N_{\boldsymbol{X}_{\setminus 10}}(\boldsymbol{x}_{10}) = \{o_3, o_2\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = 0.016, \text{density}_{\boldsymbol{X}_{\setminus 10}}(\boldsymbol{x}_{10}) = 0.011.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem.

Question 13. Again, consider the pairwise Euclidean distances between the 10 observations $o_1, o_2, \ldots, o_{10}$ from the Palmer Penguins dataset in Table 4. Consider the two clusters

$$C_2 = \{o_6, o_7\}$$
$$C_3 = \{o_8, o_9, o_{10}\}$$

What is the distance between C_2 and C_3 using average linkage as the linkage function?

A. $d(C_2, C_3) \approx 1108.3$

B. $d(C_2, C_3) \approx 1145.8$

C. $d(C_2, C_3) \approx 1183.3$

D. $d(C_2, C_3) \approx 1450.0$

E. Don't know.

Solution 13. We have that the average linkage is given by

$$\begin{aligned} d(C_2, C_3) &= \frac{\sum_{x \in C_1, y \in C_2} ||x - y||_2}{|C_1||C_2|} \\ &= \frac{1025 + 925 + 1375 + 1100 + 1000 + 1450}{2 \cdot 3} \end{aligned}$$

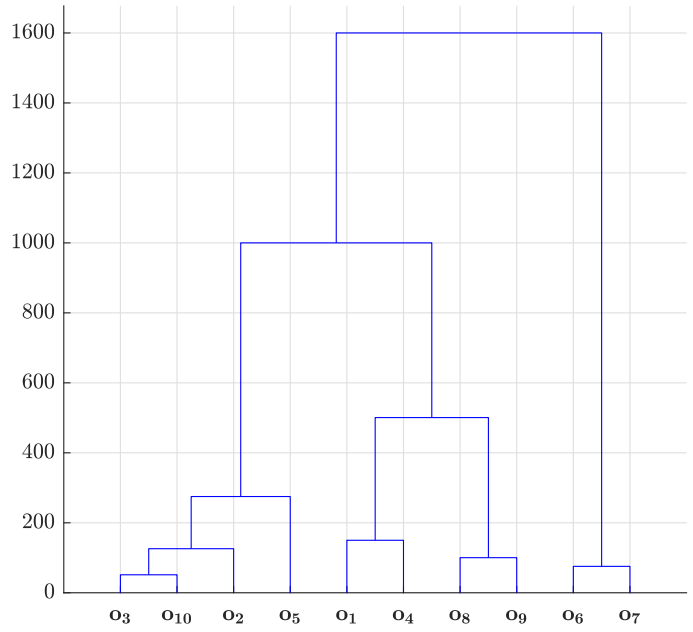

Figure 7: Hierarchical clustering of the 10 observations in Table 4.

Question 14. A hierarchical clustering is applied to the 10 observations in Table 4 using *maximum* linkage, and the result is shown in Figure 7. Which one of the following set of clusters *cannot* be obtained from the dendrogram by applying a valid cutoff?

A. $\{o_3, o_{10}, o_2, o_5\}, \{o_1, o_4\}, \{o_8, o_9\}, \{o_6, o_7\}$

B. $\{o_3, o_{10}, o_2\}, \{o_5\}, \{o_1, o_4\}, \{o_8, o_9\}, \{o_6, o_7\}$

C. $\{o_3, o_{10}\}, \{o_2\}, \{o_5\}, \{o_1, o_4\}, \{o_8, o_9\}, \{o_6, o_7\}$

D. $\{o_3, o_{10}\}, \{o_2\}, \{o_5\}, \{o_1\}, \{o_4\}, \{o_8, o_9\}, \{o_6, o_7\}$

E. Don't know.

Solution 14.

- A can be achieved by a cutoff at 400.
- B can be achieved by a cutoff at 200.
- C cannot be achieved, as o_2 cannot be split from $\{o_3, o_{10}, o_2\}$ without also splitting $\{o_1, o_4\}$. This is hence the correct answer.
- D can be achieved by by a cutoff at 125.
- Don't know.

	f_1	f_2	f_3	f_4	x_5
o_{11}	0	0	1	0	1
o_{12}	1	0	1	0	1
o_{13}	0	1	0	0	1
o_{14}	1	0	1	1	1
o_{15}	1	0	0	0	1
o_{16}	0	0	0	1	1
o_{17}	1	0	1	0	2
o_{18}	0	1	0	0	2
o_{19}	0	0	1	0	2
o_{20}	1	0	0	0	2

Table 5: Binarized version of 10 observations from the Palmer Penguins dataset. Each of the features f_i are obtained by taking a feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$). The binary feature x_5 indicates the sex of the penguin.

Question 15. Table 5 shows $N = 10$ observations from the Palmer Penguins dataset. The data is processed to produce four new binary features such that $f_i = 1$ corresponds to a value x_i greater than the median. Here, we make the following conditional independence assumption about the binary features

$$p(f_1, f_2, f_3, f_4|x_5) = p(f_1|x_5)p(f_2, f_3|x_5)p(f_4|x_5).$$

Using this assumption and Bayes' rule, we obtain the classifier

$$p(x_5|f_1, f_2, f_3, f_4) = \frac{p(f_1|x_5)p(f_2, f_3|x_5)p(f_4|x_5)p(x_5)}{\sum_{k=1}^{2} p(f_1|x_5 = k)p(f_2, f_3|x_5 = k)p(f_4|x_5 = k)p(x_5 = k)}.$$

Consider a new observations o_{21} with the following values for the binary features:

	f_1	f_2	f_3	f_4
o_{21}	0	0	1	0

What is the probability that o_{21} is classified as *male* ($x_5 = 1$) using the classifier above?

A. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{1}{10}$

B. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{1}{5}$

C. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{1}{2}$

D. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{10}{19}$

E. Don't know.

Solution 15. From Table 5 we find that

- $p(x_5 = 1) = \frac{6}{10}$
- $p(f_1 = 0|x_5 = 1) = \frac{3}{6}$
- $p(f_2 = 0, f_3 = 1|x_5 = 1) = \frac{3}{6}$
- $p(f_4 = 0|x_5 = 1) = \frac{4}{6}$
- $p(x_5 = 2) = \frac{4}{10}$
- $p(f_1 = 0|x_5 = 2) = \frac{2}{4}$
- $p(f_2 = 0, f_3 = 1|x_5 = 2) = \frac{2}{4}$
- $p(f_4 = 0|x_5 = 2) = \frac{4}{4}$

Therefore we find that

$$p(x_5 = 1|f_1 = 0, f_2 = 1, f_3 = 1, f_4 = 0) = \frac{\frac{3}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{6}{10}}{\frac{3}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{6}{10} + \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{4} \cdot \frac{4}{10}} = \frac{1}{2}$$

Question 16. We again consider the Palmer Penguins dataset from Table 1 and the $N = 10$ observations we already encountered in Table 5. Recall that, the data is processed to produce four new binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[4], and we thereby arrive at the $N \times M = 10 \times 4$ binary matrix in Table 5. In this exercise, we do not consider attribute x_5.

Then the matrix can be considered as representing $N = 10$ transactions $o_{11}, o_{12}, \ldots, o_{20}$ and $M = 4$ items $f_1, f_2, \ldots, f_4$. Which of the following options represents all (non-empty) itemsets with support greater than 0.25 (and only itemsets with support greater than 0.25)?

A. $\{f_1\}$, $\{f_3\}$, $\{f_1, f_3\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_1, f_3\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_1, f_3\}$, $\{f_1, f_4\}$, $\{f_3, f_4\}$, $\{f_1, f_3, f_4\}$

D. $\{f_1\}$, $\{f_3\}$

E. Don't know.

Solution 16. 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 0.25, an itemset needs to be contained in 3 rows. It is easy to see this rules out all options except *A*.

[4]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 17. We again consider the binary matrix from Table 5 as a market basket problem consisting of $N = 10$ transactions $o_{11}, \ldots, o_{20}$ and $M = 4$ items $f_1, \ldots, f_4$.

What is the *confidence* of the rule $\{f_1, f_4\} \to \{f_3\}$?

A. The confidence is $\frac{1}{10}$

B. The confidence is $\frac{3}{10}$

C. The confidence is $\frac{7}{10}$

D. The confidence is 1

E. Don't know.

Solution 17. The confidence of the rule is computed as

$$\frac{\text{support}(\{f_1, f_4\} \cup \{f_3\})}{\text{support}(\{f_1, f_4\})} = \frac{\frac{1}{10}}{\frac{1}{10}} = 1.$$

Question 18. Consider classifying the Palmer Penguins dataset according to the species attribute y. The dataset contains 146 observations for $y = 1$ (*Adelie*), 119 observation for $y = 2$ (*Gentoo*) and 68 observation for $y = 3$ (*Chinstrap*).

During training of a decision tree, the classification error has been used as impurity measure

$$\text{classError}(v) = 1 - \max_c p(c|v)$$

where $p(c|v)$ denotes the fraction of observations belonging to class c at a given node v. The tree is constructed by Hunts algorithm using the purity gain

$$\Delta = \text{classError}(\textit{parent}) - \sum_{k=1}^{2} \frac{N(v_k)}{N} \text{classError}(v_k)$$

where N is the total number of observations at the parent node and $N(v_k)$ is the number of observations associated with the k^{th} child node, v_k.

After the first split you learn that the left node contains all the *Adelie* and *Gentoo* observations while the right node contains all the *Chinstrap* observations. What is the purity gain for the split?

A. The purity gain is $\frac{1}{5}$.

B. The purity gain is $\frac{68}{333}$.

C. The purity gain is $\frac{68}{265}$.

D. The purity gain is $\frac{265}{333}$.

E. Don't know.

Solution 18. With I denoting the class error, we obtain:

- $I(\textit{parent}) = 1 - \frac{146}{333}$
- $I(\textit{left}) = 1 - \frac{146}{146+119} = 1 - \frac{146}{265}$
- $I(\textit{right}) = 1 - \frac{68}{68} = 0$
- $\Delta = \left(1 - \frac{146}{333}\right) - \frac{265}{333} \cdot \left(1 - \frac{146}{265}\right) - 0 = 1 - \frac{265}{333} = \frac{68}{333}$

Variable	y^{true}	$t=1$
y_1	2	2
y_2	1	1
y_3	1	2
y_4	1	1
y_5	2	2
y_6	2	2
y_7	2	2

Table 6: For each of the $N = 7$ observations (first column), the table indicates the true class labels y^{true} (second column) and the predicted outputs of the AdaBoost classifier (third column) for the first round of boosting $t = 1$.

Question 19. Consider again the Palmer Penguins dataset of Table 1. Suppose we limit ourselves to $N = 7$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_1 and x_2. We use a KNN classification model ($K = 3$) to this dataset and apply AdaBoost to improve the performance. After the first round of boosting, we obtain predictions and the true class labels as tabulated in Table 6. Given this information, how will the AdaBoost update the weights $\boldsymbol{w}$?

A. $[0.083 \quad 0.083 \quad 0.5 \quad 0.083 \quad 0.083 \quad 0.083 \quad 0.083]$

B. $[0.165 \quad 0.165 \quad 0.008 \quad 0.165 \quad 0.165 \quad 0.165 \quad 0.165]$

C. $[0.152 \quad 0.152 \quad 0.085 \quad 0.152 \quad 0.152 \quad 0.152 \quad 0.152]$

D. $[0.01 \quad 0.01 \quad 0.937 \quad 0.01 \quad 0.01 \quad 0.01 \quad 0.01]$

E. Don't know.

Solution 19.

We first observe the AdaBoost classifier at $t = 1$ mis-classify observations:

$$\{y_3\}$$

Since the weights are just $w_i = \frac{1}{N}$, we therefore get:

$$\epsilon_{t=1} = \sum_i w_i(t)(1 - \delta_{f_t(x_i),y_i}) = 0.143$$

From this, we compute α_t as

$$\alpha_t = \frac{1}{2} \log \frac{1 - \epsilon_t}{\epsilon_t} = 0.896$$

Scaling the observations corresponding to the mis-classified weights as $w_i e^{\alpha_t}$ and those corresponding to the correctly classified weights as $w_i e^{-\alpha_t}$ and normalizing the new weights to sum to one then give answer A.

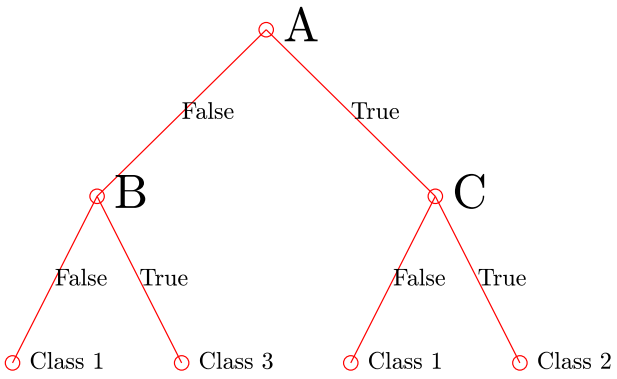

Figure 8: Example classification tree.

Question 20. We consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 8 resulting in a partition into classes indicated by the colors/markers in Figure 9. What is the correct rule assignment to the nodes in the decision tree?

A. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 3$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 4 \end{bmatrix} \right\|_1 < 3$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_1 < 3$

B. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 4 \end{bmatrix} \right\|_1 < 3$, **$\boldsymbol{B}$:** $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_1 < 3$, **$\boldsymbol{C}$:** $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 3$

C. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 4 \end{bmatrix} \right\|_1 < 3$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 3$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_1 < 3$

D. $\boldsymbol{A}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\|_1 < 3$, $\boldsymbol{B}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right\|_1 < 3$, $\boldsymbol{C}$: $\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 6 \\ 4 \end{bmatrix} \right\|_1 < 3$

E. Don't know.

Solution 20.

This problem is solved by using the definition of a decision tree and observing what classification rule each of the assignment of features to node names in the decision tree will result in. I.e., beginning at the top of the tree, check if the condition assigned to the node is met and proceed along the true or false leg of the tree.

The resulting decision boundaries for each of the options are shown in Figure 10 and it follows answer B is correct.

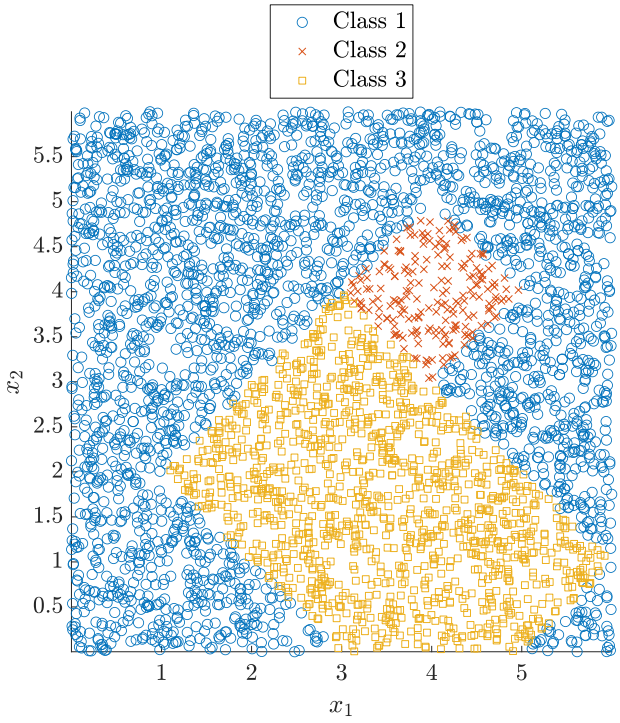

Figure 9: Classification boundaries.

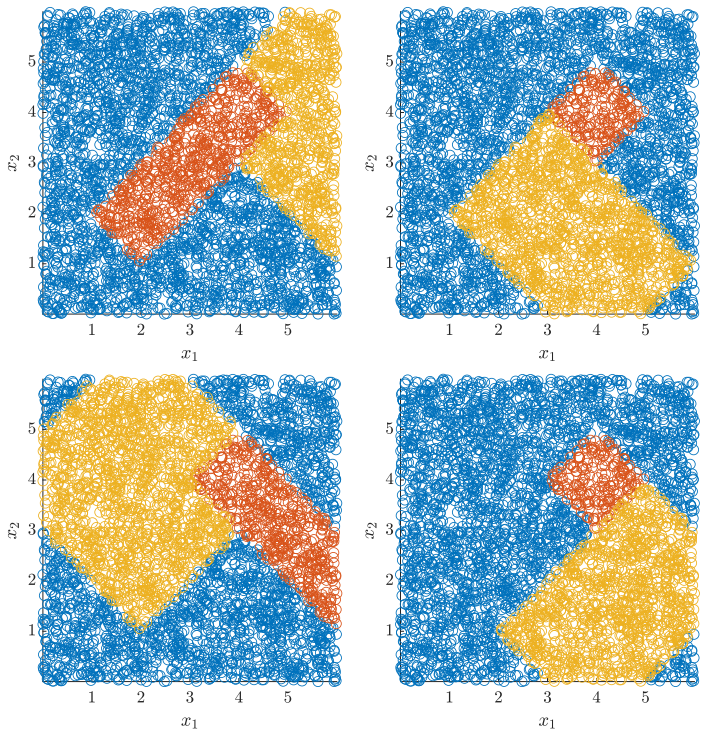

Figure 10: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

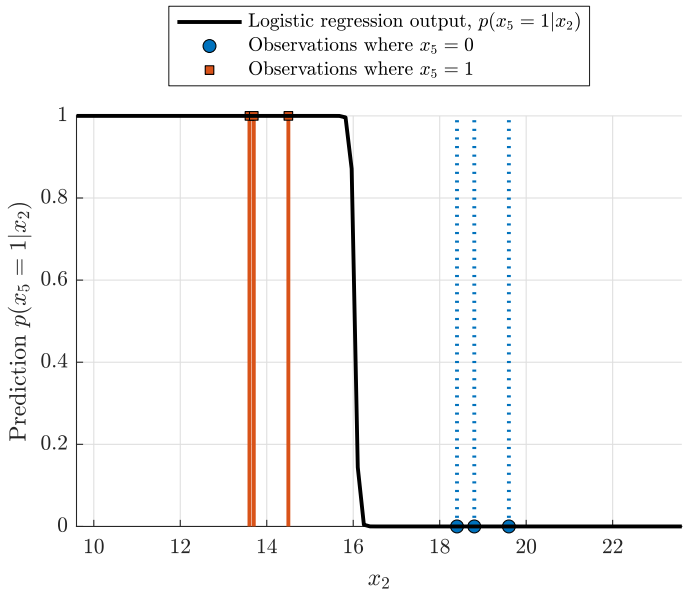

Figure 11: Output of a logistic regression classifier trained on 6 observations from the dataset.

Question 21. Consider again the Palmer Penguins dataset. To simplify the setup further, we select just 6 observations and train a logistic regression classifier using only the feature x_2 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term). As output, we want to predict the Penguin sex (x_5). To be consistent with the lecture notes, we label the output as $x_5 = 0$ (corresponding to *male*) and $x_5 = 1$ (corresponding to *female*).

In Figure 11 is shown the predicted output probability of an observation belonging to the positive class, $p(x_5 = 1|x_2)$. What are the weights?

A. $\boldsymbol{w} = \begin{bmatrix} 423.49 \\ 48.16 \end{bmatrix}$

B. $\boldsymbol{w} = \begin{bmatrix} 0.0 \\ -46.21 \end{bmatrix}$

C. $\boldsymbol{w} = \begin{bmatrix} 0.0 \\ -27.89 \end{bmatrix}$

D. $\boldsymbol{w} = \begin{bmatrix} 418.94 \\ -26.12 \end{bmatrix}$

E. Don't know.

Solution 21. The solution is easily found by simply computing the predicted $\hat{x}_5 = p(x_5 = 1|x_2)$-value for an appropriate choice of x_2. Notice that

$$p(x_5 = 1|x_2) = \sigma(\tilde{\boldsymbol{x}}_2^T \boldsymbol{w})$$

If we select $x_2 = 16$ and select the weights as in option D we find $\hat{x}_5 = p(x_5 = 1|x_2) = 0.735$, in good agreement with the figure. On the other hand, for the weights in option A we obtain $\hat{x}_5 = 1$, for C that $\hat{x}_5 = 0$ and finally for B that $\hat{x}_5 = 0$. We can therefore conclude that D is correct.

Question 22. We fit a GMM to a single feature x_2 from the Palmer Penguins dataset. Recall the density of a 1D GMM is

$$p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \sigma_k^2)$$

and suppose that the identified values of the mixture weights are

$$w_1 = 0.13,\ w_2 = 0.55,\ w_3 = 0.32$$

and the parameters of the multivariate normal densities:

$$\mu_1 = 18.347,\ \mu_2 = 14.997,\ \mu_3 = 18.421$$
$$\sigma_1 = 1.2193,\ \sigma_2 = 0.986,\ \sigma_3 = 1.1354.$$

According to the GMM, what is the probability an observation at $x_0 = 15.38$ is assigned to cluster $k = 2$?

A. 0.975

B. 0.389

C. 0.213

D. 0.042

E. Don't know.

Solution 22. Recall γ_{ik} is the posterior probability that observation i is assigned to mixture component 2 which can easily be obtained using Bayes' theorem. We see that:

$$\gamma_{i,2} = \frac{p(x_i|z_{i,2} = 1)\pi_2}{\sum_{k=1}^{3} p(x_i|z_{ik} = 1)\pi_k}.$$

To use Bayes' theorem, we need to compute the probabilities using the normal density. These are:

$$p(x_i|z_{i1} = 1) = 0.017$$
$$p(x_i|z_{i2} = 1) = 0.375$$
$$p(x_i|z_{i3} = 1) = 0.01$$

Combining these with the class-assignment probabilities we obtain:

$$\gamma_{i,2} = 0.975$$

and conclude the solution is A.

Fold	M_1/M_2	$M_1/\overline{M}_2$	$\overline{M}_1/M_2$	$\overline{M}_1/\overline{M}_2$
1	86	8	7	10
2	65	15	11	20
3	79	5	17	10

Table 7: Outcome of cross-validation. Rows are combination of outcomes of the two models.

Question 23. We will consider the Palmer Penguins dataset and two models for predicting the class label y. Specifically, let M_1 be a $K = 1$ nearest neighbor classification model and M_2 a $K = 5$ nearest neighbor classification model. To compare them statistically, we perform $K = 3$ fold cross-validation, and for each fold we record the number of observations where both models are correct (as M_1/M_2), M_1 is correct and M_2 wrong (as $M_1/\overline{M_2}$), and so on. The outcome can be found in Table 7.

We want to test if the two classifiers perform differently. The null hypothesis is that the models have the same performance. Given the values of the binomial cumulative distribution function in Table 8 and assuming that the null hypothesis is true, what is the p-value from McNemar's test (rounded to two decimals)?

A. 0.23

B. 0.45

C. 0.84

D. 0.90

E. Don't know.

Solution 23. Since the cross-validation folds are non-overlapping, we can easily find by summing columns 2 and 3 in Table 7.

- n_1 the total number of times that M_1 is correct and M_2 is incorrect, and
- n_2 the total number of times that M_1 is incorrect and M_2 is correct.

We find that $n_1 = 28$ and $n_2 = 35$. We can calculate the p-value from the CDF of the binomial distribution as

$$p = 2\text{cdf}_{\text{binom}}(m = \min\{n_1, n_2\}|\theta = 1/2, N = n_1 + n_2)$$
$$= 2\text{cdf}_{\text{binom}}(m = 28|\theta = 1/2, N = 63) \approx 2 \cdot 0.225 \approx 0.45$$

Therefore, B is correct.

	$m=14$	$m=21$	$m=28$	$m=35$
$N=30$	0.428	0.992	1.000	1.000
$N=33$	0.243	0.960	1.000	1.000
$N=36$	0.121	0.879	1.000	1.000
$N=39$	0.054	0.739	0.998	1.000
$N=42$	0.022	0.561	0.990	1.000
$N=45$	0.008	0.383	0.964	1.000
$N=48$	0.003	0.235	0.903	1.000
$N=51$	0.001	0.131	0.799	0.998
$N=54$	0.000	0.067	0.658	0.990
$N=57$	0.000	0.031	0.500	0.969
$N=60$	0.000	0.014	0.349	0.922
$N=63$	0.000	0.006	0.225	0.843
$N=66$	0.000	0.002	0.134	0.731
$N=69$	0.000	0.001	0.074	0.595
$N=72$	0.000	0.000	0.038	0.453
$N=75$	0.000	0.000	0.018	0.322
$N=78$	0.000	0.000	0.008	0.214
$N=81$	0.000	0.000	0.004	0.133
$N=84$	0.000	0.000	0.001	0.078
$N=87$	0.000	0.000	0.001	0.043

Table 8: Values of the binomial cumulative distribution function $\text{cdf}_{\text{binom}}(m|N,\theta=\frac{1}{2})$ for different values of the number of successes m and the number of trials N.

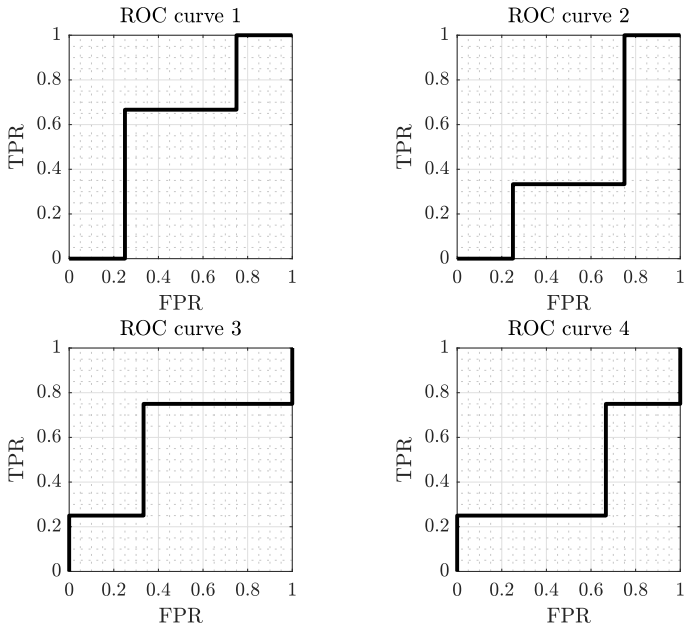

Figure 12: Proposed ROC curves for the neural network classifier with predictions/true class labels given in Table 9

y	1	0	0	1	1	0	1
$\hat{y}$	0.01	0.05	0.14	0.3	0.31	0.36	0.91

Table 9: Small binary classification dataset of $N=7$ observations along with the predicted class probability $\hat{y}$.

Question 24. A neural network classifier is trained to distinguish between two classes $y \in \{0,1\}$ in a small dataset consisting of $N=7$ observations. Suppose the true class label y and predicted probability an observation belongs to class 1, $\hat{y}$, is as given in Table 9.

To evaluate the classifier, we will use the *area under curve* (AUC) of the *reciever operator characteristic* (ROC) curve. In Figure 12 is given four proposed ROC curves, which one of the curves corresponds to the classifier?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

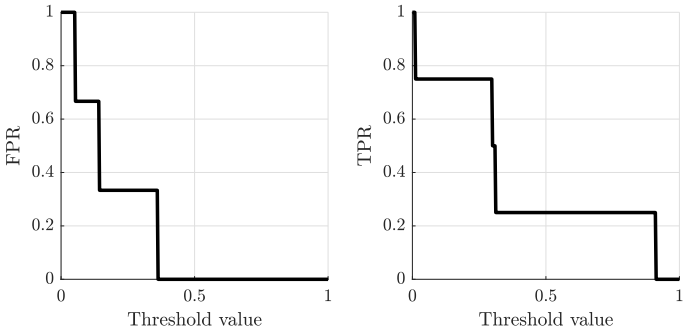

Figure 13: TPR, FPR curves for the classifier.

Solution 24. To compute the AUC, we need to compute 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.

This procedure is then repeated for different threshold values to obtain the curves shown in Figure 13. The ROC curve is then obtained by plotting these two curves against each other. I.e. for each threshold value, the point

$$(x, y) = (\text{FPR}, \text{TPR})$$

is on the AUC curve. This rules out all options except A.

Question 25. We will consider an artificial neural network (ANN) trained on the Palmer Penguins dataset described in Table 1 to predict the class label y based on the first four attributes $x_1, \ldots, x_4$. The neural network has a single hidden layer containing $n_h = 6$ units, and will use the softmax activation function (specifically, we will use the over-parameterized softmax function described in section 15.3.2 *(Neural networks for multi-class classification)* of the lecture notes) to predict the class label y since it is a multi-class problem. For the hidden layer we will use the tanh non-linear activation function. How many parameters has to be trained to fit the neural network?

A. The network contains 36 parameters

B. The network contains 42 parameters

C. The network contains 51 parameters

D. The network contains 72 parameters

E. Don't know.

Solution 25. Each hidden unit has as many input unit weights are there are features $M = 4$ plus one (the bias), therefore they contribute with

$$(M + 1)n_h$$

weights. The softmax is computed deterministically from $C = 3$ units (as many as there are classes in the dataset), and each has as many weights as there are hidden units plus one (the bias):

$$(n_h + 1)C$$

Adding these two numbers together gives option C.

Question 26. Suppose that you have a deep neural network that can binary classify whether an image contains a penguin or not. If an image contains a penguin, the network will correctly classify it as a penguin with probability 97%. If an image does not contain a penguin, the network will classify it as a penguin with probability 3%. You apply the classifier to a dataset where 1% of the images contain a penguin. What is the probability that a random image from this dataset contains a penguin given that it is classified as a penguin?

A. ≈ 0.97

B. ≈ 0.75

C. ≈ 0.50

D. ≈ 0.25

E. Don't know.

Solution 26. Let A denote the event that the classifier classifies a picture as a penguin, and let B denote the event that a picture contains a penguin. From the text we are told

- $P(A|B) = 0.97$
- $P(A|\bar{B}) = 0.03$
- $P(B) = 0.01$
- $P(\bar{B}) = 1 - P(B)$

Using Bayes rule, we find that

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \\ &= \frac{0.97 \cdot 0.01}{0.97 \cdot 0.01 + 0.03 \cdot (1 - 0.01)} \\ &\approx 0.25 \end{aligned}$$

Question 27. We use a multinomial regression model to predict the species (y) from the six attributes $x_1, \ldots, x_6$ for the Palmer Penguins dataset in Table 1. We apply least squares regularization (i.e., L_2 regularization) to the weights of the multinomial regression model and use two-level cross-validation to select the optimal value of the regularization constant $\lambda \in \{0.001, 0.01, 0.1, 1.0\}$. In the outer fold, we use 3-fold cross-validation, and in the inner fold, we use leave-one-out cross-validation. With this setup and $N = 333$ observations $o_1, \ldots, o_{333}$ in the dataset, how many times is o_1 used for training a model?

A. 999

B. 1768

C. 1770

D. 2667

E. Don't know.

Solution 27. Consider a single observation o_0 in the Palmer Penguins dataset:

- In the outer fold, we have three training sets $\{D_1^{\text{par}}, D_2^{\text{par}}, D_3^{\text{par}}\}$ each of size 222.
- o_0 is part of two of the training sets $\{D_1^{\text{par}}, D_2^{\text{par}}, D_3^{\text{par}}\}$.
- In the inner fold, we split D_i^{par} into training sets $\{D_1^{\text{train}}, \ldots, D_{222}^{\text{train}}\}$. If $o_2 \in D_i^{\text{par}}$ then o_2 is part of 221 of these training sets.
- We train 4 models on each of the training sets D_j^{train}.
- Finally, we also train a single model for the optimal choice of λ on each of D_i^{par}.

This means that o_0 is used $2 \cdot (221 \cdot 4 + 1) = 1770$ times for training a model.