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/`

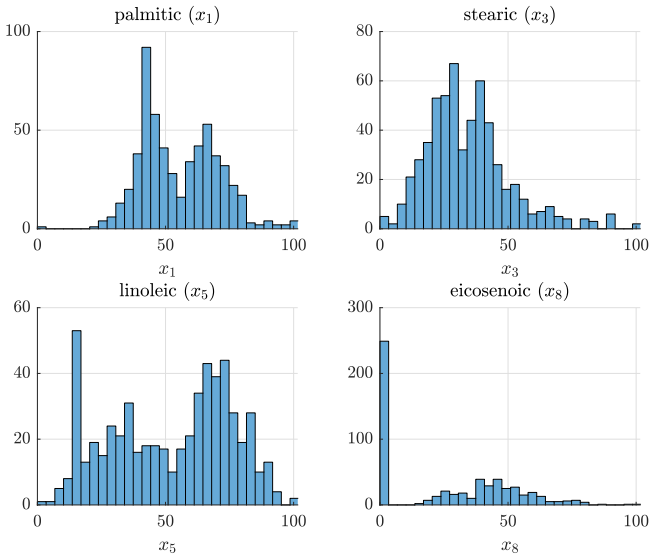

Figure 1: Plot of the observations of attributes x_1, x_3, x_5 and x_8 from the Olive Oil dataset of Table 1 as histogram plots.

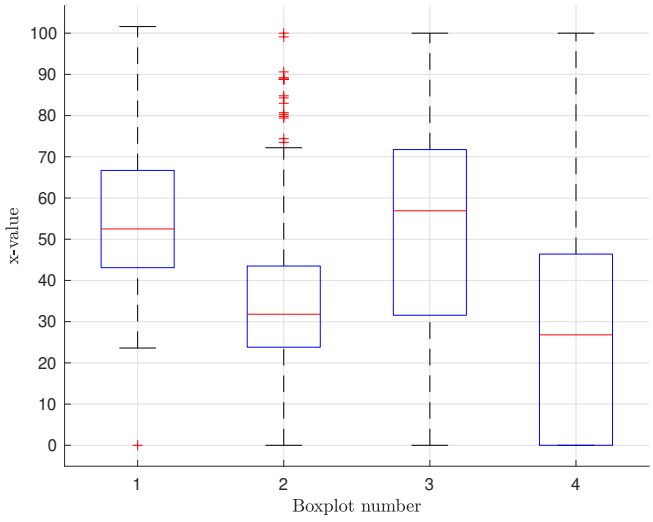

Figure 2: Boxplots corresponding to the variables plotted in Figure 1 but not necessarily in that order.

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.

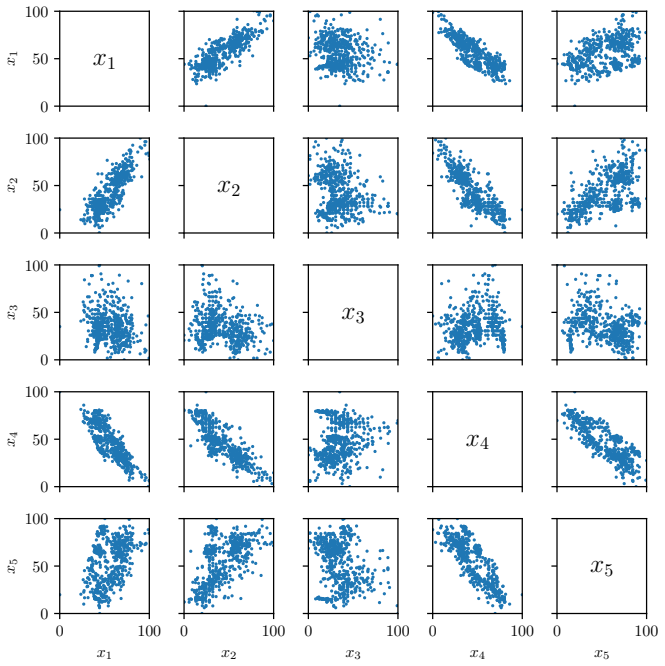

Figure 3: Scatter plot matrix for the attributes x_1, x_2, x_3, x_4, x_5 of the Olive Oil dataset of Table 1.

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.

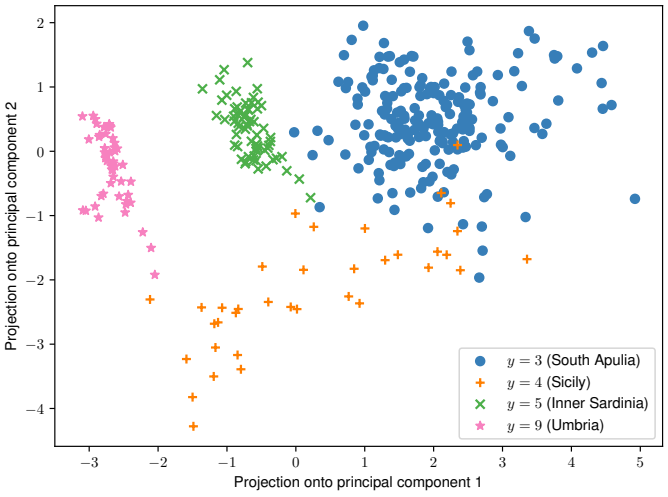

Figure 4: Scatter plot of the projection of observations belonging four classes from the Olive Oil dataset in Table 1 onto the first two principal components.

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.

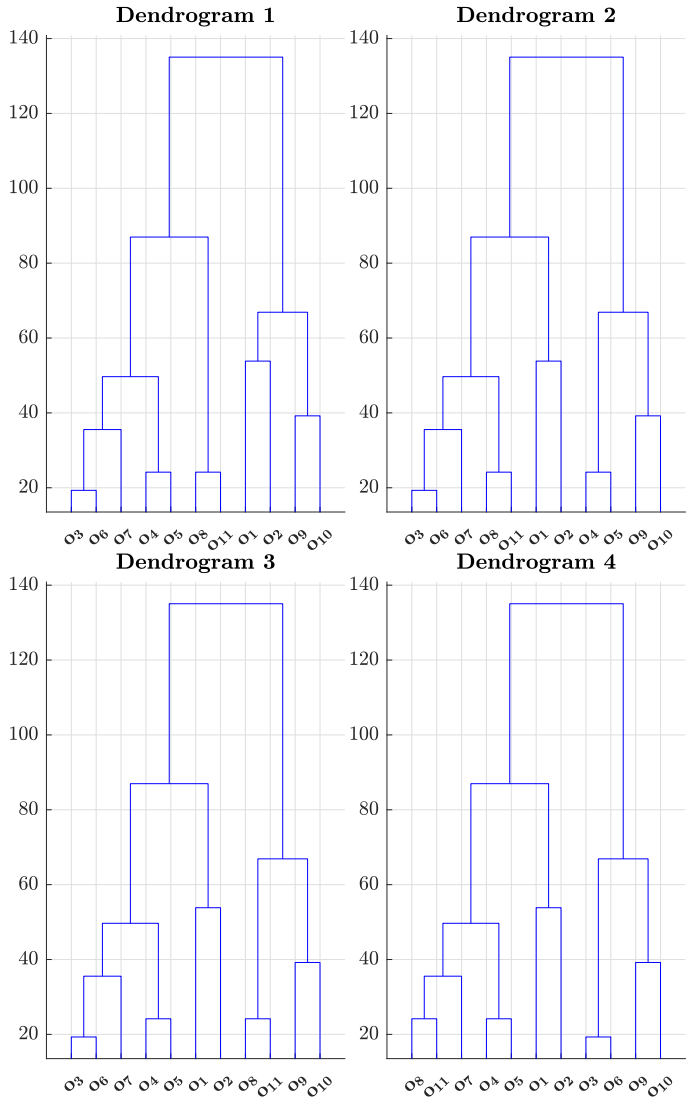

Figure 5: Proposed hierarchical clustering of the 11 observations in Table 2.

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

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

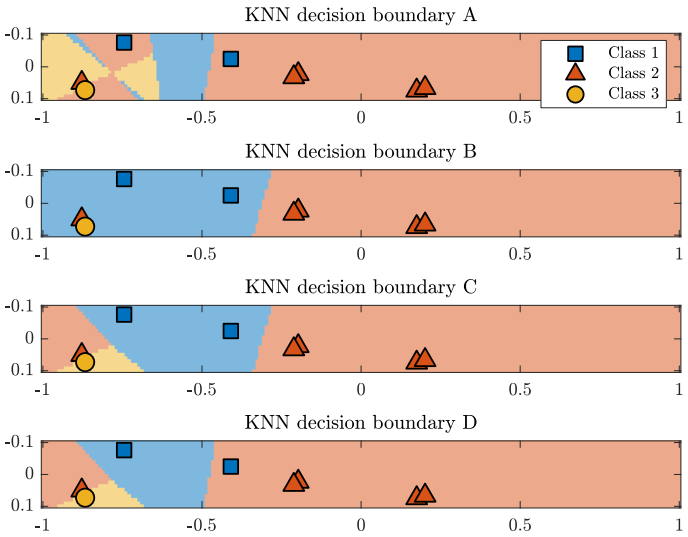

Figure 6: Decision boundaries for four KNN classifiers.

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.

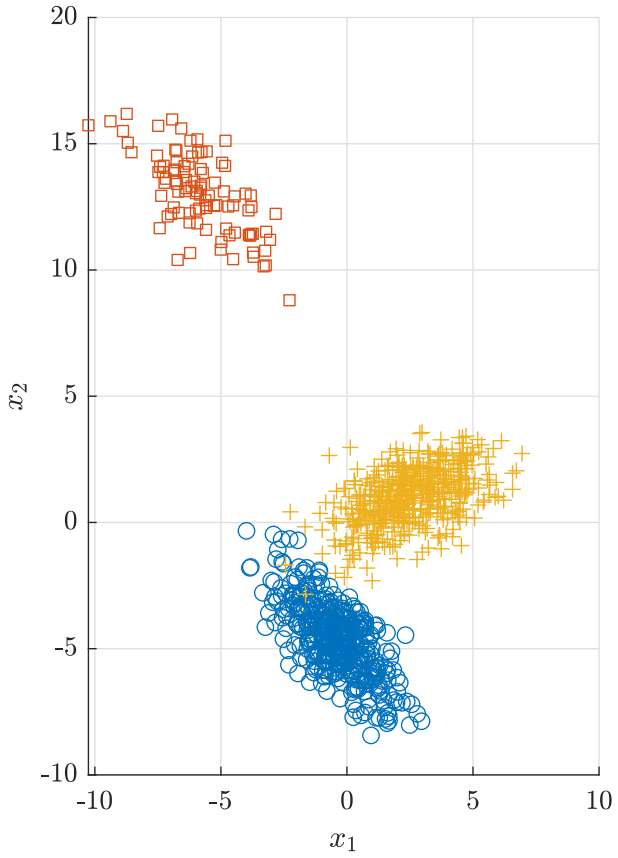

Figure 7: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

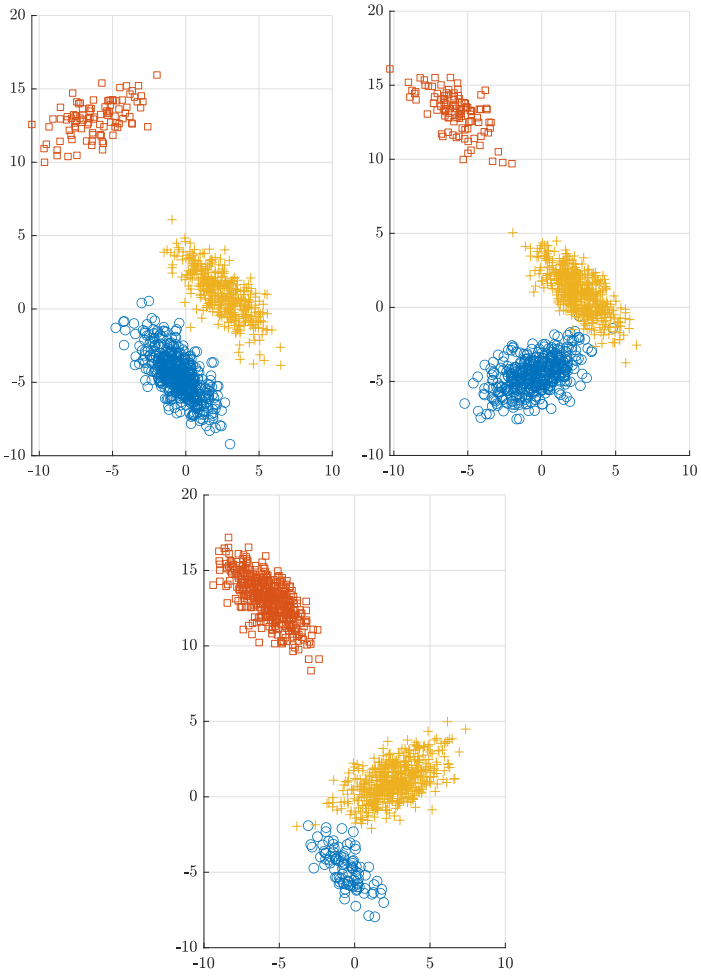

Figure 8: GMM mixtures corresponding to alternative options.

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.

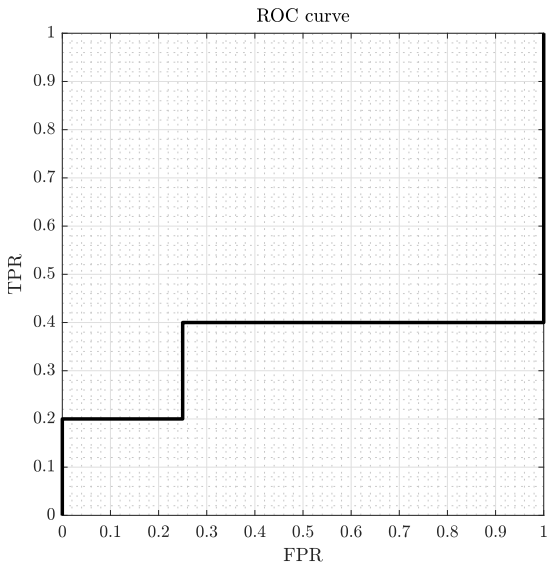

Figure 9: ROC curve for a neural network classifier, where the predictions and true class labels are one of the options in Figure 10.

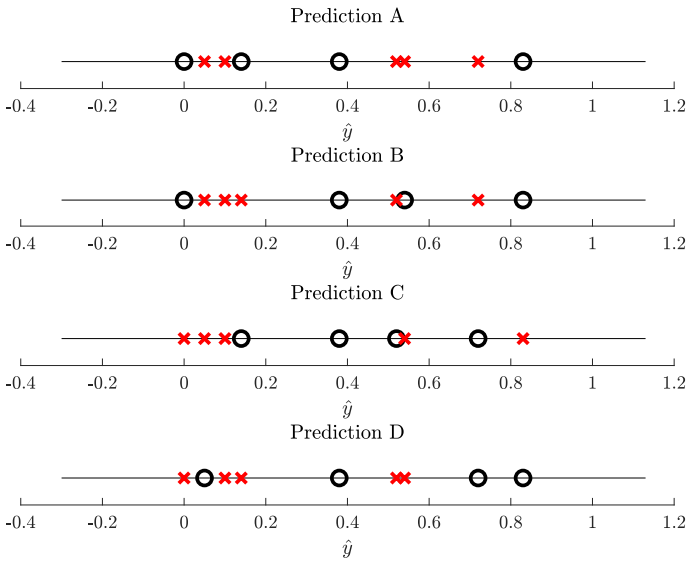

Figure 10: Four candidate predictions for the ROC curve in Figure 9. The observations are plotted horizontally, such that the position on the x-axis indicate the predicted value $\hat{y}_i$, and the marker/color indicate the class membership, such that the black circles indicate the observation belongs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

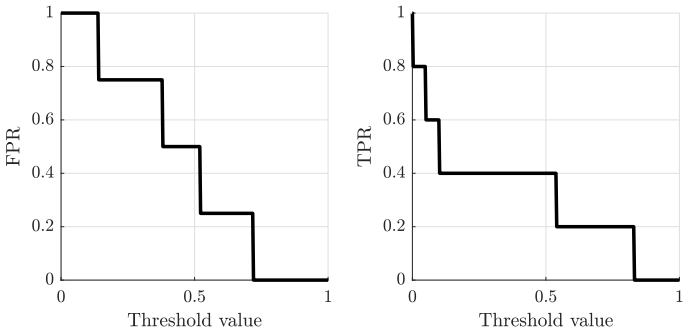

Figure 11: TPR, FPR curves for the classifier.

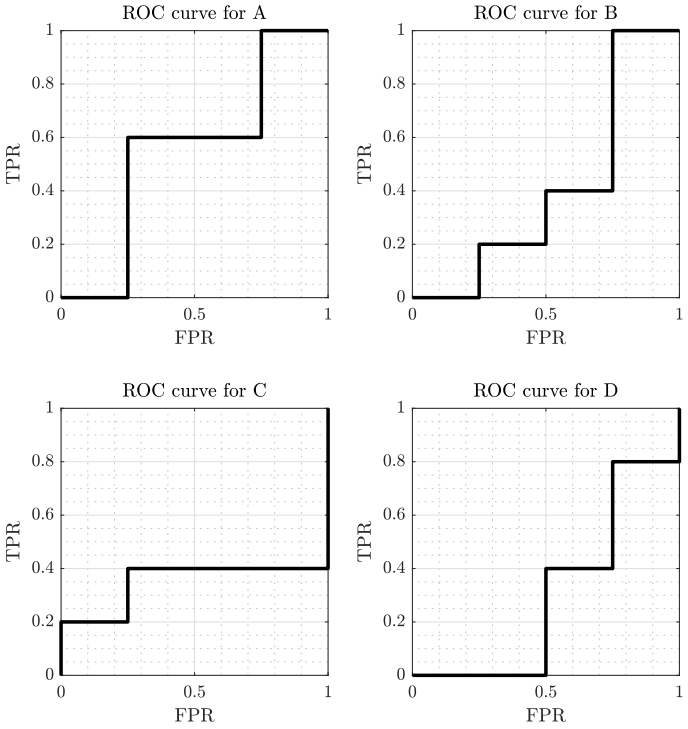

Figure 12: ROC curves for all options.