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.

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 Itali. 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 match which boxplots?

A. Boxplot 1 is x_5, boxplot 2 is x_1, boxplot 3 is x_3 and boxplot 4 is x_8

B. Boxplot 1 is x_1, boxplot 2 is x_8, boxplot 3 is x_5 and boxplot 4 is x_3

C. Boxplot 1 is x_1, boxplot 2 is x_3, boxplot 3 is x_5 and boxplot 4 is x_8

D. Boxplot 1 is x_1, boxplot 2 is x_5, boxplot 3 is x_8 and boxplot 4 is x_3

E. Don't know.

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

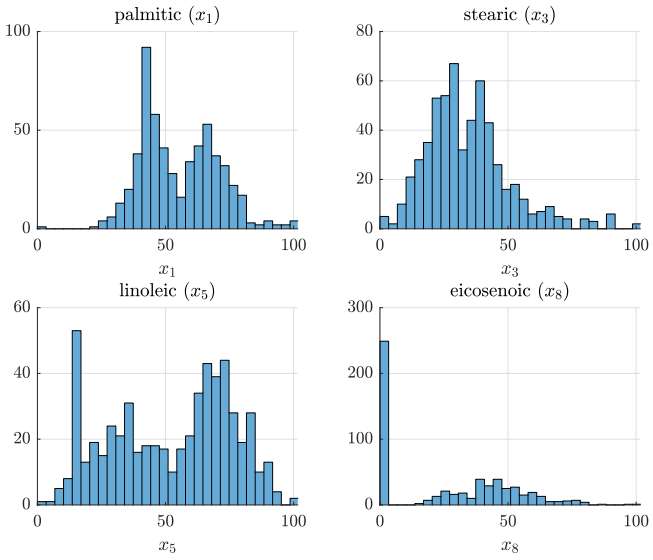

Figure 1: Plot of the observations of attributes x_1, x_3, x_5 and x_8 from the Olive Oil dataset of Table 1 as histogram plots.

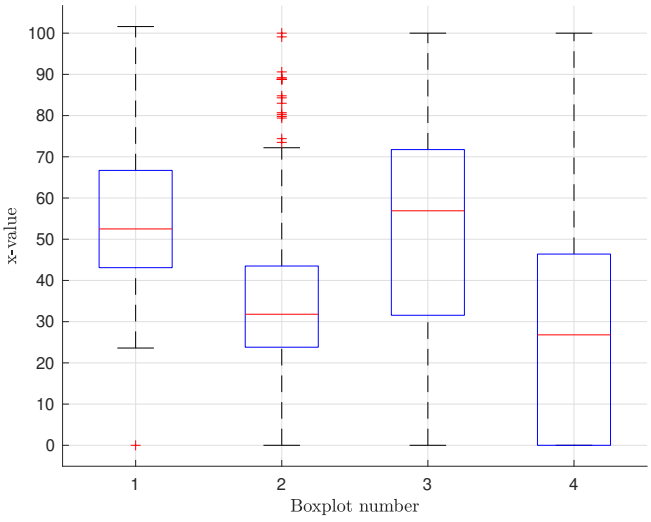

Figure 2: Boxplots corresponding to the variables plotted in Figure 1 but not necessarily in that order.

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.

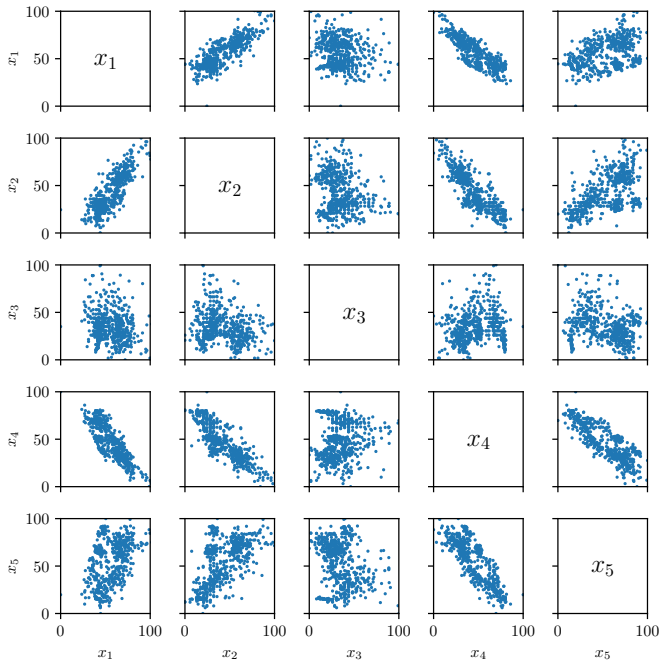

Figure 3: Scatter plot matrix for the attributes x_1, x_2, x_3, x_4, x_5 of the Olive Oil dataset of Table 1.

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.

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.

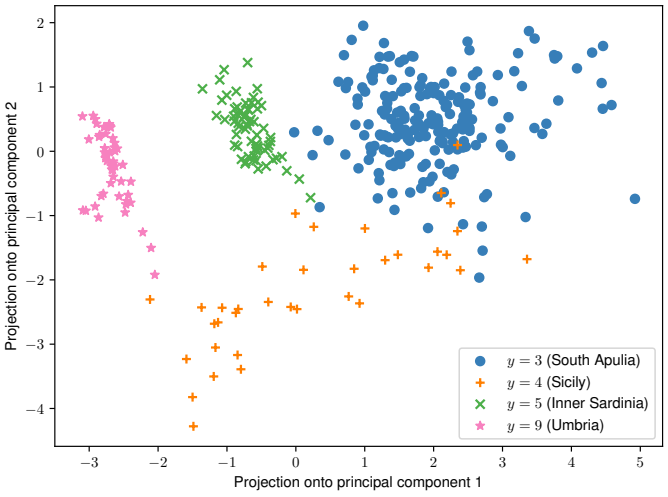

Figure 4: Scatter plot of the projection of observations belonging four classes from the Olive Oil dataset in Table 1 onto the first two principal components.

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.

	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.

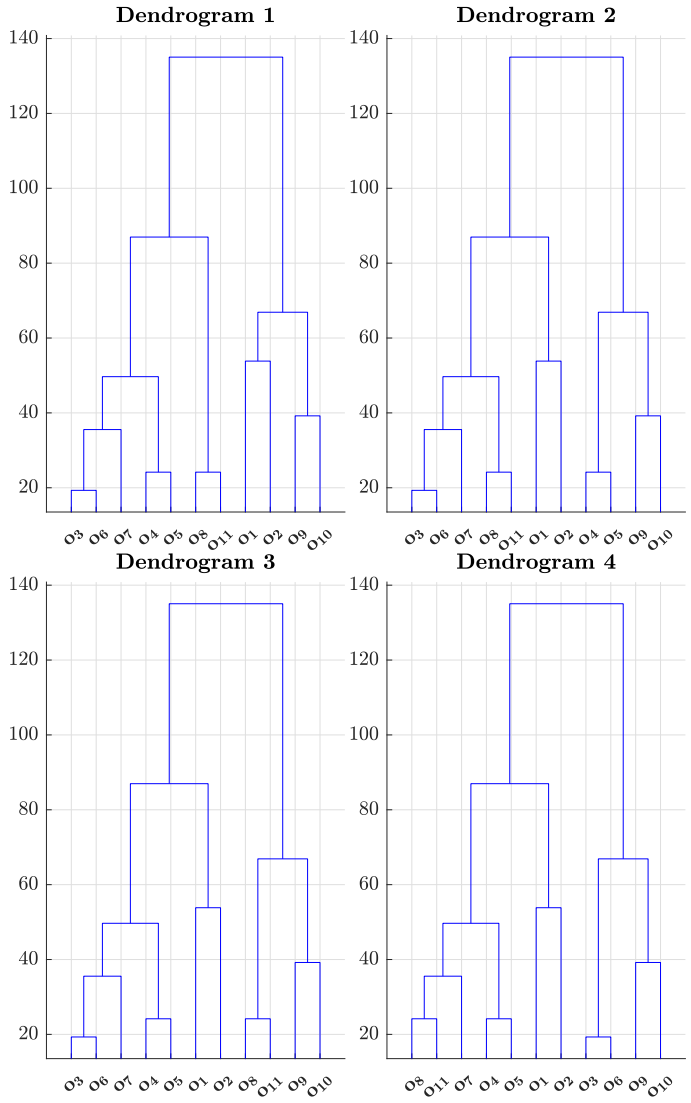

Figure 5: Proposed hierarchical clustering of the 11 observations in Table 2.

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.

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.

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.

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

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

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

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

E. Don't know.

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.

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 $\{f_6, f_8\} \rightarrow \{f_3, f_5\}$?

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.

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.

	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.

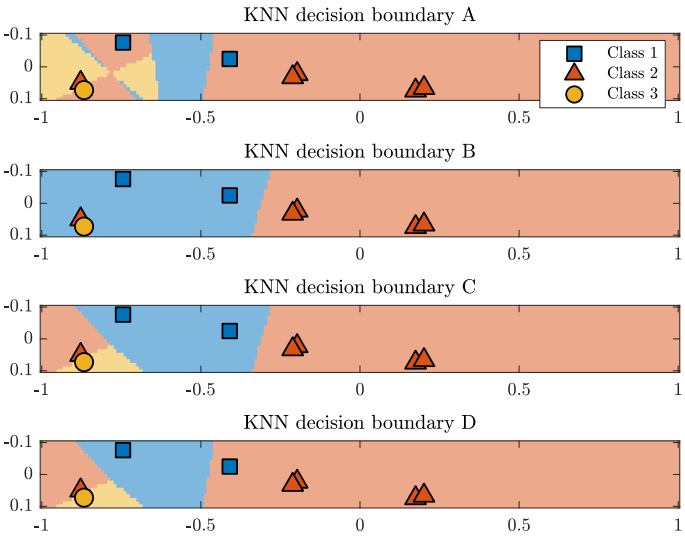

Figure 6: Decision boundaries for four KNN classifiers.

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 red squares (class 1), black 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.

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.

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 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N$ and corresponding targets $y_1, y_2, \ldots y_N$, 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.

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.

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.

	$i = 1$	$i = 2$	$i = 3$	$i = 4$
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 $\phi(x) = \begin{bmatrix} \cos(\frac{\pi}{2}x) & \sin(\frac{\pi}{2}x) \end{bmatrix}^\mathsf{T}$. Find the weights $\boldsymbol{w}^* = \begin{bmatrix} w_1^* & w_2^* \end{bmatrix}^\mathsf{T}$ 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.

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.

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.

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.

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.

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.

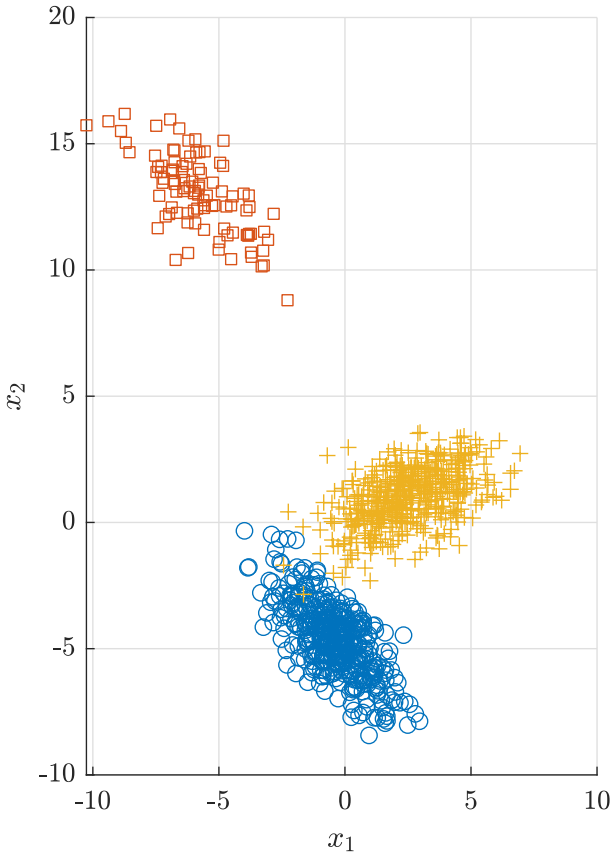

Figure 7: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

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 9. 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 9 then corresponds to the ROC curve in Figure 8?

A. Prediction A

B. Prediction B

C. Prediction C

D. Prediction D

E. Don't know.

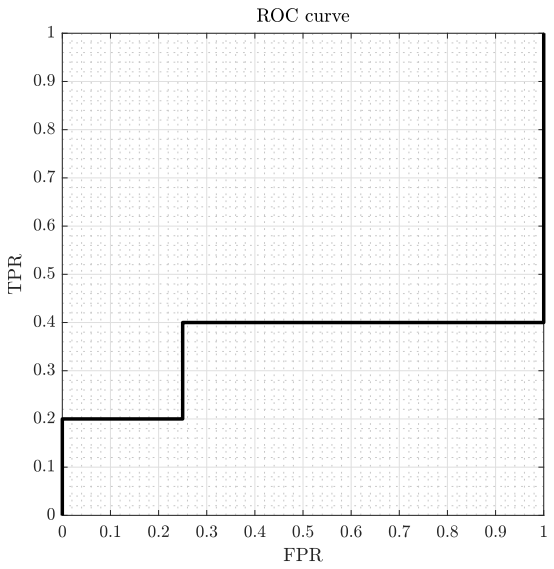

Figure 8: ROC curve for a neural network classifier, where the predictions and true class labels are one of the options in Figure 9.

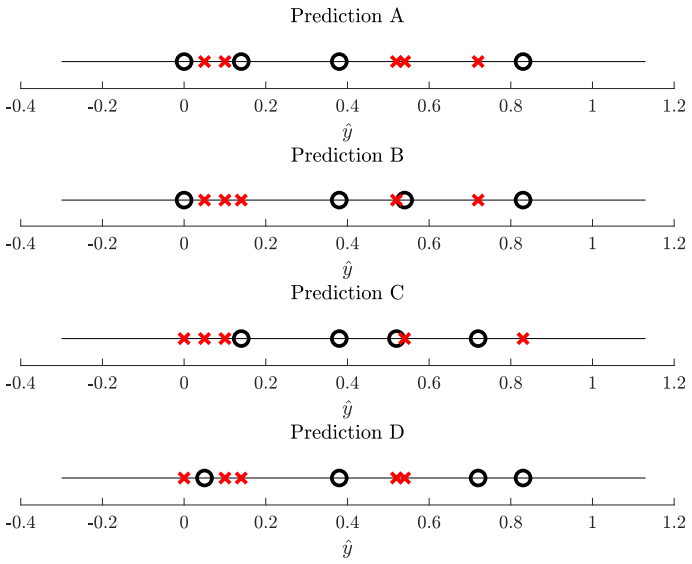

Figure 9: Four candidate predictions for the ROC curve in Figure 8. The observations are plotted horizontally, such that the position on the x-axis indicate the predicted value $\hat{y}_i$, and the marker/color indicate the class membership, such that the black circles indicate the observation belongs to class $y_i = 0$ and red crosses to $y_i = 1$.