Technical University of Denmark

Written examination: Test exam (based on the Spring 2021 exam)

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.

[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 $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$

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

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

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.

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.

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.

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.

	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.

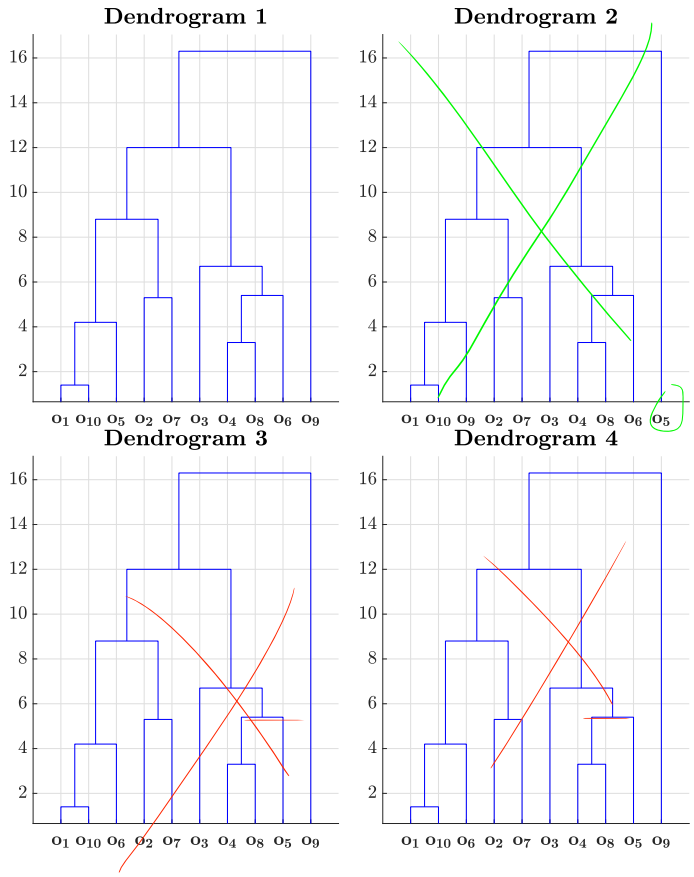

Figure 1: Proposed hierarchical clustering of the 10 observations in Table 2.

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.

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.

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.

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 Naive-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.

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.

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:

$$I : \{\text{VISIBILITY, HUMIDITY}\}.$$

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. $\text{supp}(I) = 0.415$

B. $\text{supp}(I) = 0.441$

C. $\text{supp}(I) = 0.217$

D. $\text{supp}(I) = 0.467$

E. Don't know.

	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.

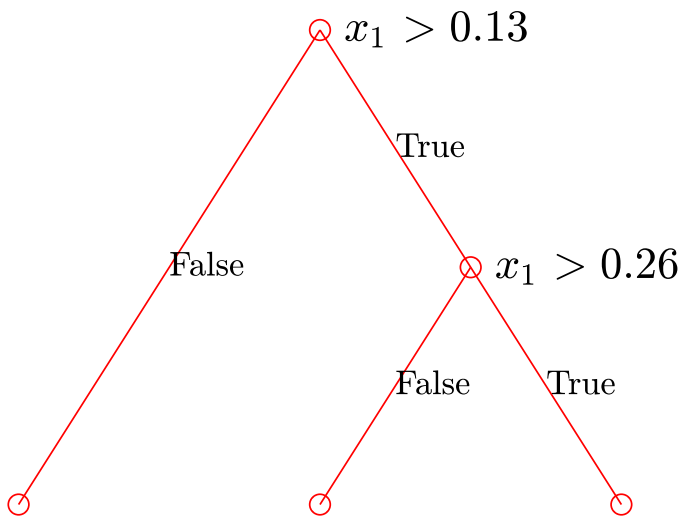

Figure 2: Structure of a regression tree. The nodes show the decision rules which determine how the observations are propagated towards the leafs of the tree.

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.

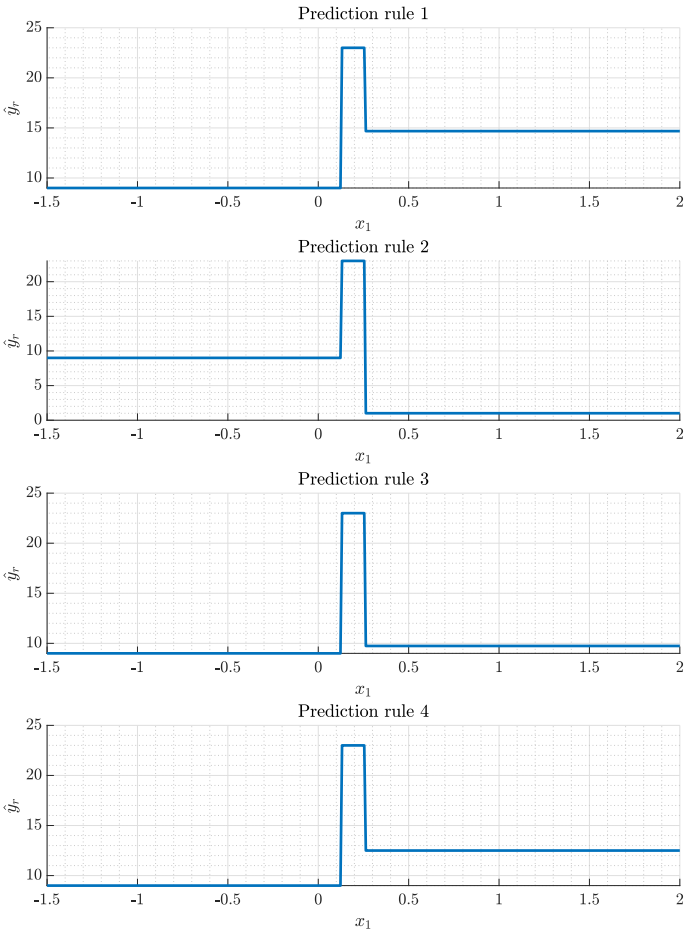

Figure 3: Possible model predictions of $\hat{y}_r$ as a function of x_1 for the decision tree illustrated in Figure 2.

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.

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.

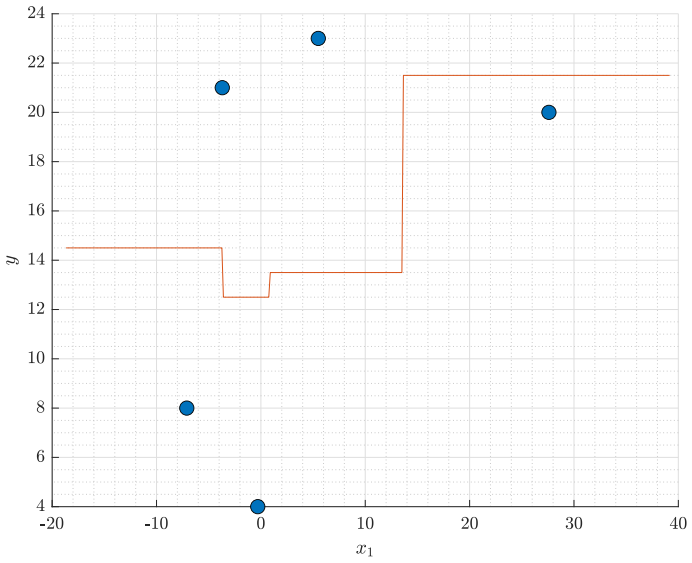

Figure 4: KNN regression model in which the red line is fitted to a small 1-dimensional dataset.

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.

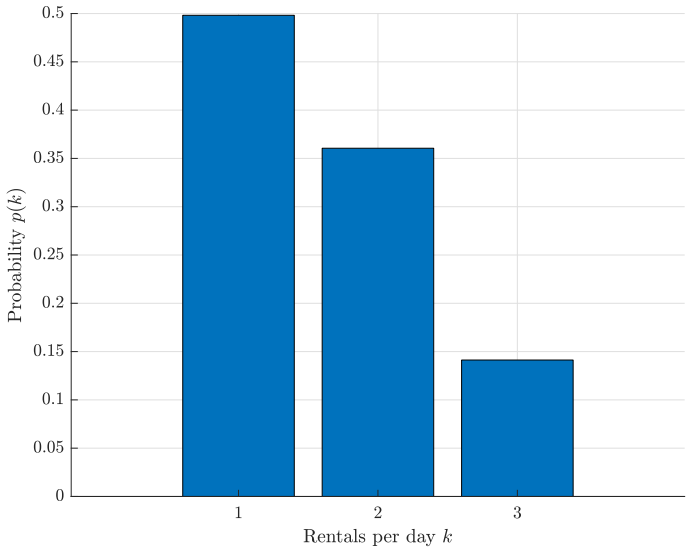

Figure 5: Probability $p(k)$ a citybike is rented exactly k times a day. The probability of $k \geq 4$ is negligible and can be ignored.

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.

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.

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.

	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.

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.

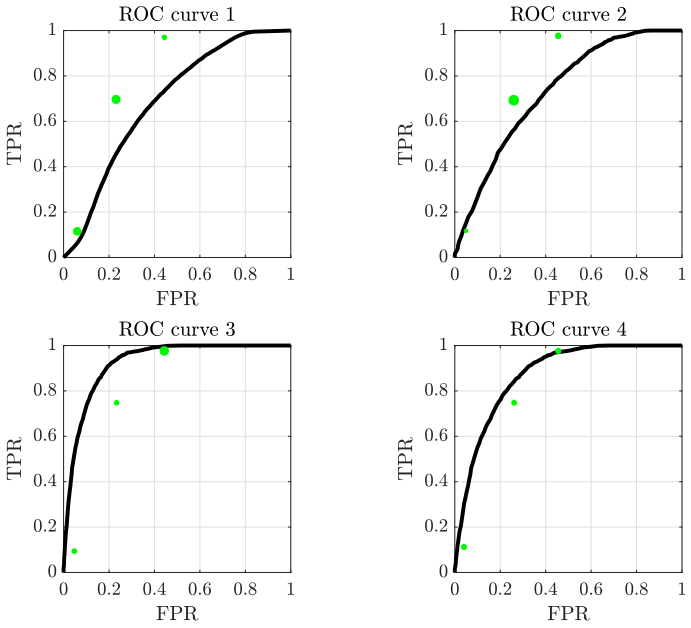

Figure 6: Candidate ROC curves for the classifier.

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.

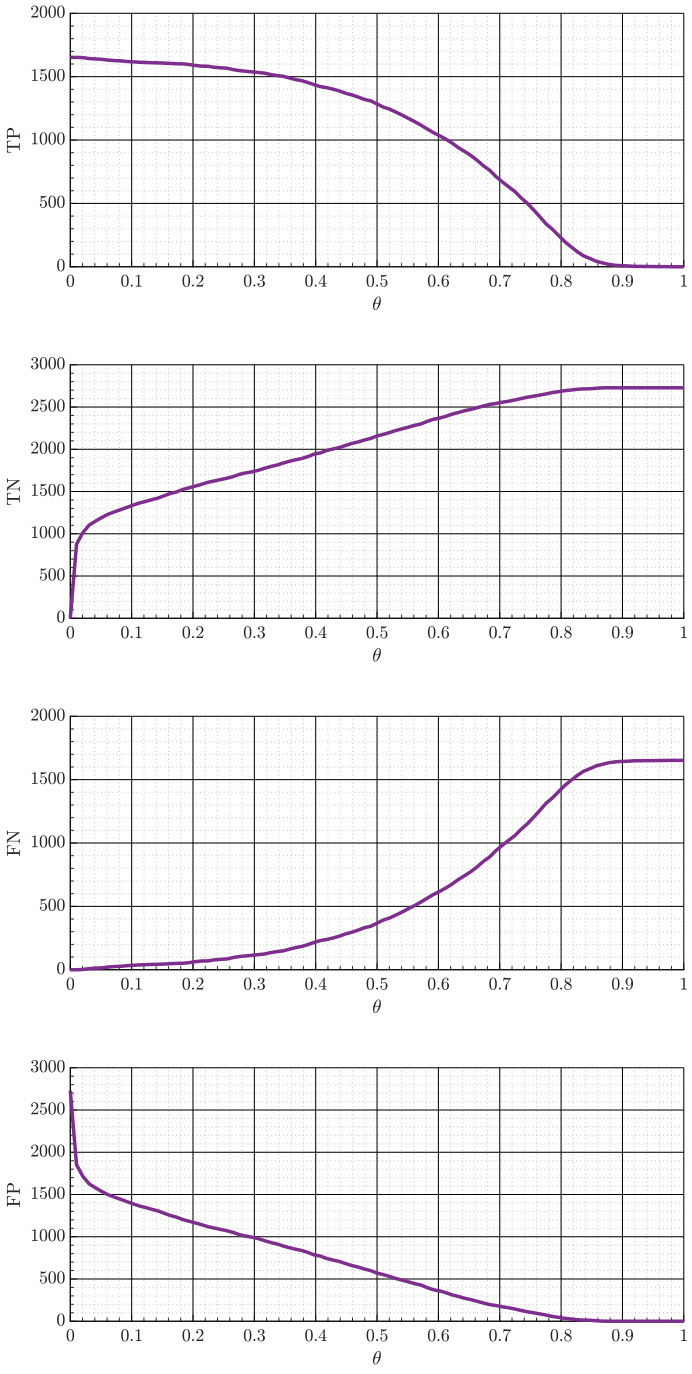

Figure 7: TP, TN, FN, and FP as functions of threshold value θ.

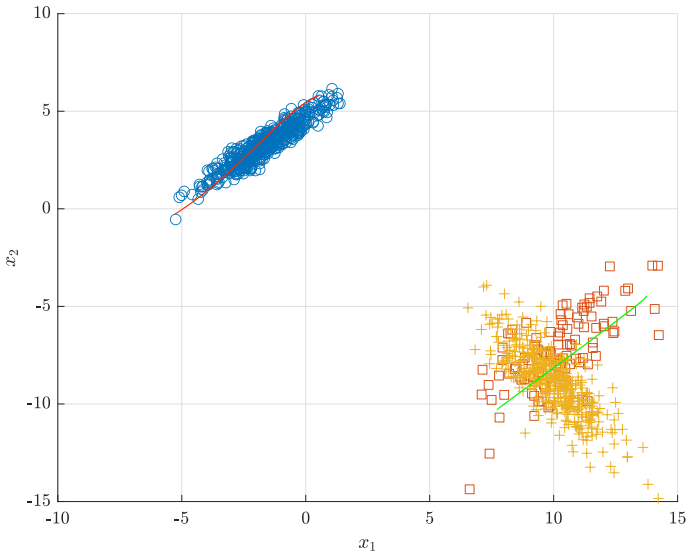

Figure 8: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

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.

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.
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.

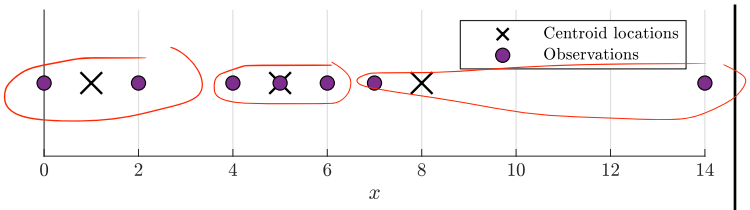

Figure 9: A small 1-dimensional dataset and initial values of centroids.

Question 25. Consider a small dataset comprised of $N = 7$ one-dimensional observations shown as the filled circles in Figure 9.

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 9. 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.

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 10. 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**.

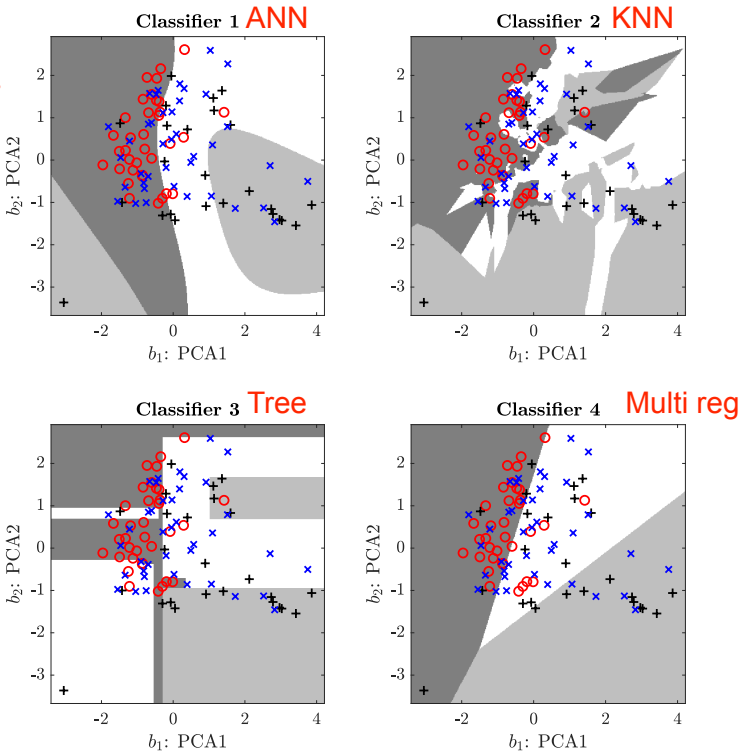

Figure 10: Decision boundaries for four different classifiers trained on the Bicycle rental dataset when projected onto the first two principal components.

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.

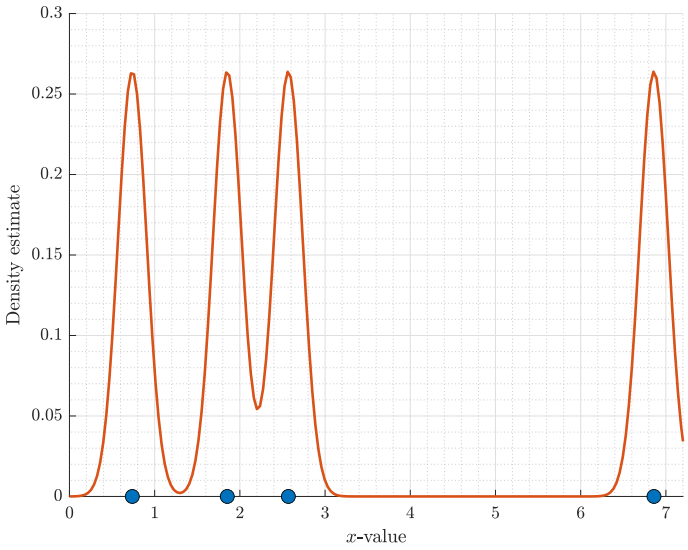

Figure 11: Plot of the density function of a kernel density estimator applied to a 1-dimensional dataset using a Gaussian kernel with kernel width $\lambda = 0.168$. Only a subset of the dataset, indicated by the circles, is shown.

Question 27. A small 1-dimensional dataset of N observations, along with the kernel density estimate, is shown in Figure 11. 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.