Technical University of Denmark

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

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

Print name and study number clearly. The exam is multiple choice. All questions have four possible answers marked by the letters A, B, C, and D as well as the answer "Don't know" marked by the letter E. Correct answer gives 3 points, wrong answer gives -1 point, and "Don't know" (E) gives 0 points.

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

Answers:

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

Name: ____

Student number: ____________________

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

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

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

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

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

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

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

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

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

E. Don't know.

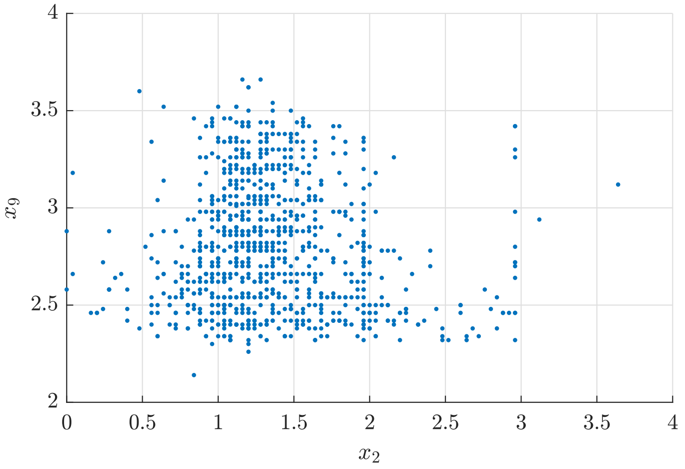

Figure 1: Scatter plot of observations x_2 and x_9 of the travel review dataset described in Table 1.

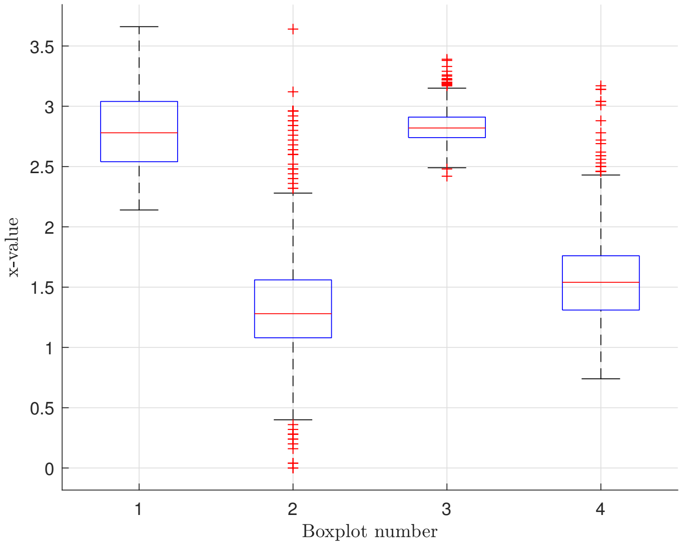

Figure 2: Four boxplots in which two of the boxplots correspond to the two variables plotted in Figure 1.

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

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

The data is standardized by (i) substracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized data matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

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

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

Which one of the following statements is true?

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

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

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

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

E. Don't know.

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

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

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

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

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

E. Don't know.

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

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

Question 4. To examine if observation o_7 may be an outlier, we will calculate the average relative density using the cityblock distance and the observations given in Table 2 only. We recall that the KNN density and average relative density (ard) for the observation $\boldsymbol{x}_i$ are given by:

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

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

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors. What is the average relative density for observation o_7 for $K = 2$ nearest neighbors?

A. 0.41

B. 1.0

C. 0.51

D. 0.83

E. Don't know.

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

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

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

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

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

E. Don't know.

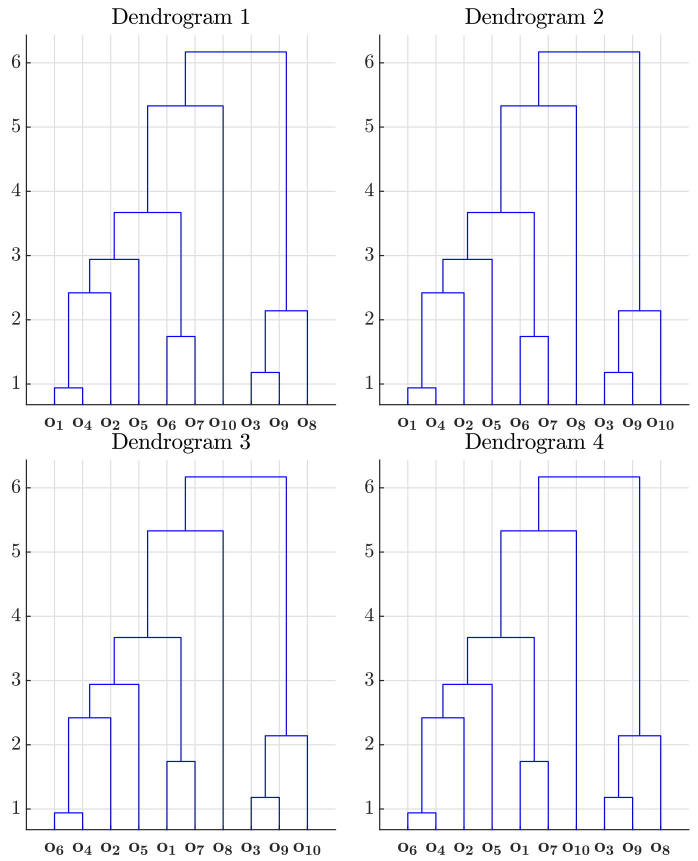
Dendrogram 1
1 2 3 4 5 6
o_1 o_4 o_2 o_5 o_6 o_7 o_{10} o_3 o_9 o_8
Dendrogram 2
1 2 3 4 5 6
o_1 o_4 o_2 o_5 o_6 o_7 o_8 o_3 o_9 o_{10}
Dendrogram 3
1 2 3 4 5 6
o_6 o_4 o_2 o_5 o_1 o_7 o_8 o_3 o_9 o_{10}
Dendrogram 4
1 2 3 4 5 6
o_6 o_4 o_2 o_5 o_1 o_7 o_{10} o_3 o_9 o_8

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

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

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

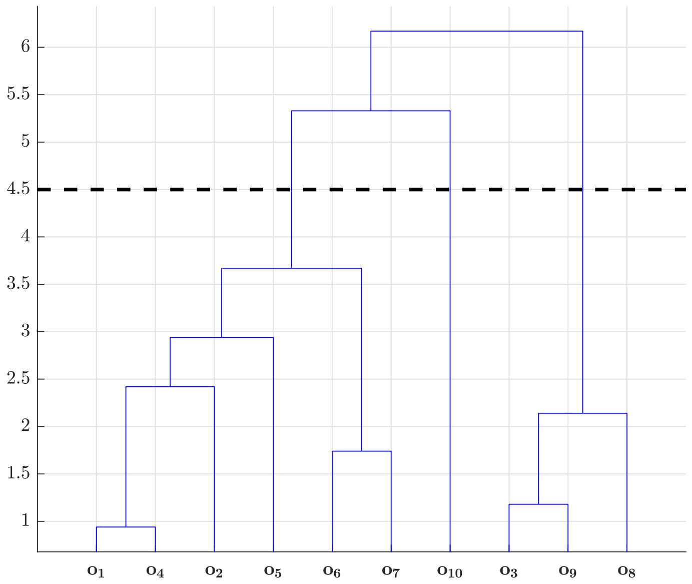
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
o_1 o_4 o_2 o_5 o_6 o_7 o_{10} o_3 o_9 o_8

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

E. Don't know.

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

A. The accuracy is: 0.42

B. The accuracy is: 0.685

C. The accuracy is: 0.338

D. The accuracy is: 0.097

E. Don't know.

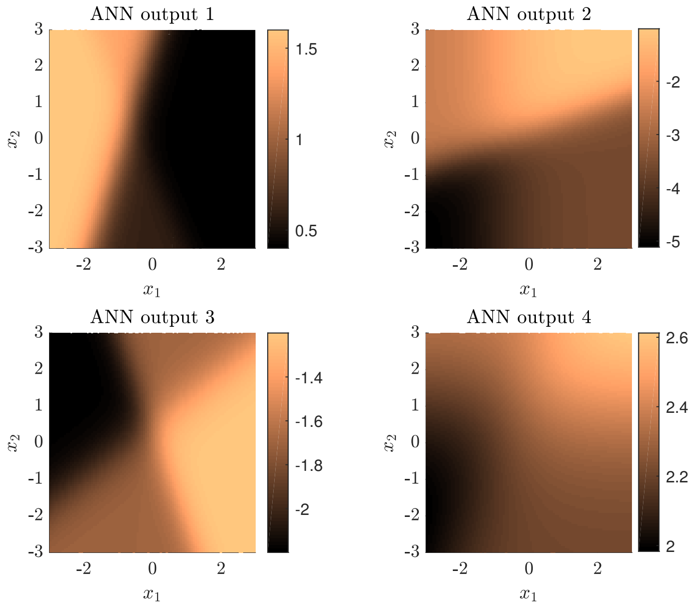

Figure 5: Suggested outputs of an ANN trained on the two attributes x_1 and x_2 from the travel review dataset to predict y.

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

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

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

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

$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -0.3 \\ 0.5 \end{bmatrix}, \quad w_0^{(2)} = 2.2.$$

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

A. ANN output 1

B. ANN output 2

C. ANN output 3

D. ANN output 4

E. Don't know.

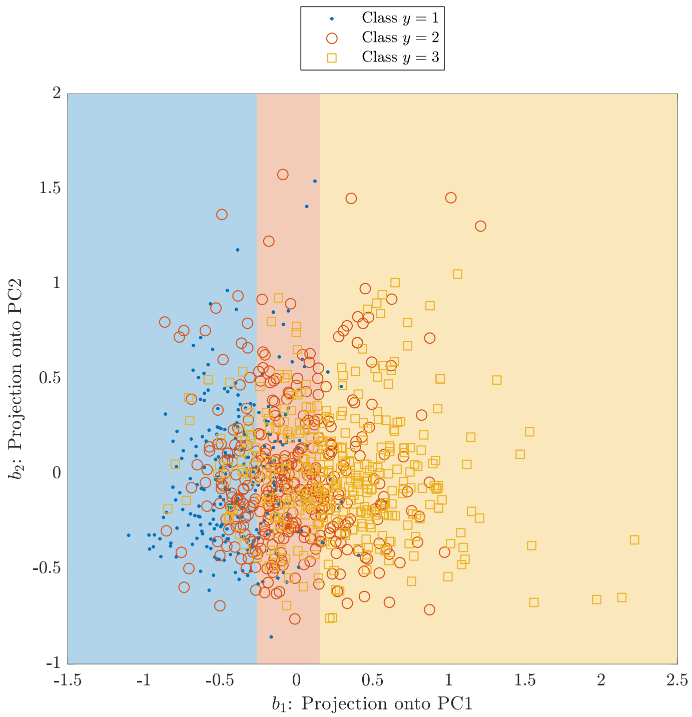

Figure 6: Output of a logistic regression classifier trained on observations from the travel review dataset.

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

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

and then use the softmax transformation in the form:

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

Question 13. We again consider the travel review dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce 9 new, binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[2], and we thereby arrive at the $N \times M = 10 \times 9$ binary matrix in Table 4. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_2,\ f_4,\ f_5$. If for an observations we observe

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

E. Don't know.

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

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

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

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

C. The confidence is 1

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

E. Don't know.

[2] Note that in association mining, we would normally also include features f_i such that $f_i = 1$ if the corresponding feature is less than the median; for brevity we will not consider features of this kind in this problem

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

E. Don't know.

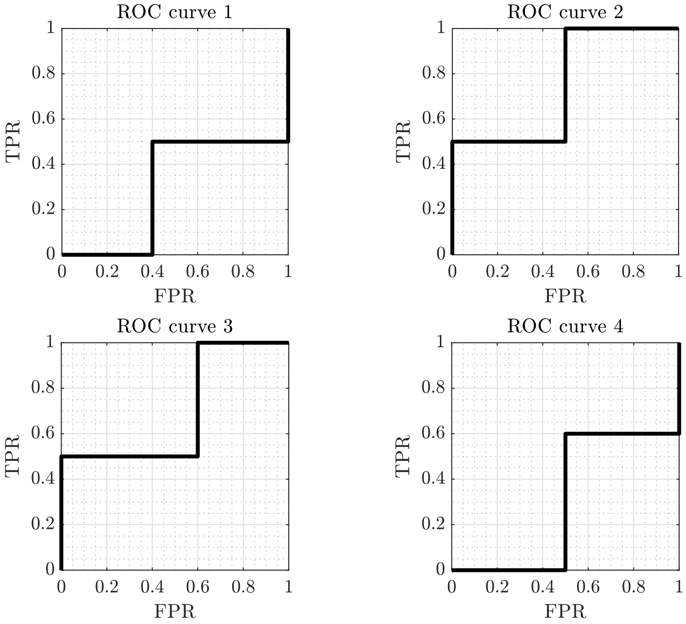

Figure 7: Proposed ROC curves for the neural network classifier with predictions/true class labels given in Table 5

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

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

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

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

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

Question 19. Consider again the travel review dataset in Table 1. We would like to predict a resort's rating using a linear regression, and since we would like the model to be as interpretable as possible we will use variable selection to obtain a parsimonious model. We limit ourselves to the five features $x_1,\ x_6,\ x_7,\ x_8,\ x_9$ and in Table 6 we have pre-computed the estimated training and test error for different variable combinations of the dataset. Which of the following statements is correct?

A. Forward selection will select attributes x_6

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

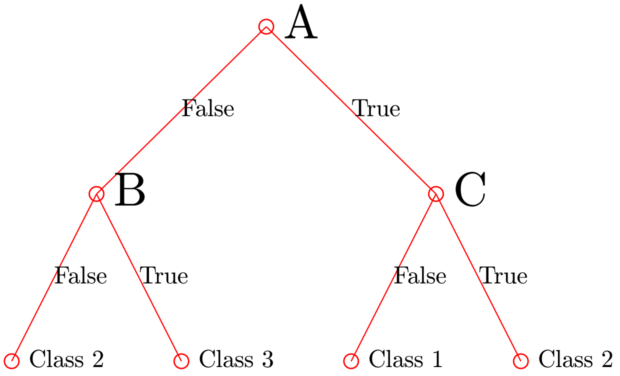

Figure 8: Example classification tree.

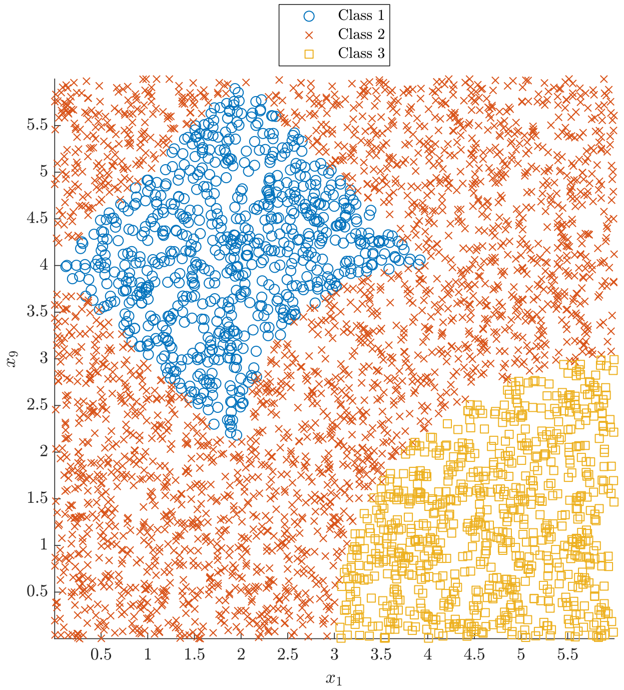

Figure 9: classification boundary.

Question 21. We consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 8 resulting in a partition into classes indicated by the colors/markers in Figure 9. What is the correct rule assignment to the nodes in the decision tree?

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

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

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

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

E. Don't know.

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

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

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

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

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

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

A. 208 models

B. 100 models

C. 200 models

D. 104 models

E. Don't know.

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

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

and suppose that the identified values of the mixture weights are

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

and the parameters of the multivariate normal densities:

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

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

A. 0.49

B. 0.31

C. 0.08

D. 0.68

E. Don't know.

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

Table 9: For each of the $N = 7$ observations (first column), the table indicate the true class labels y^{true} (second column) and the predicted outputs of the AdaBoost classifier (third column) which is also shown in Figure 10.

Question 24. Consider again the travel review dataset of Table 1. Suppose we limit ourselves to $N = 7$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_4 and x_6. We use a KNN classification model ($K = 1$) to this dataset and apply AdaBoost to improve the performance. After the first $T = 1$ round of boosting, we obtain the decision boundaries shown in Figure 10 (the predictions of the $T = 1$ weaker classifiers and the true class labels is also tabulated in Table 9).

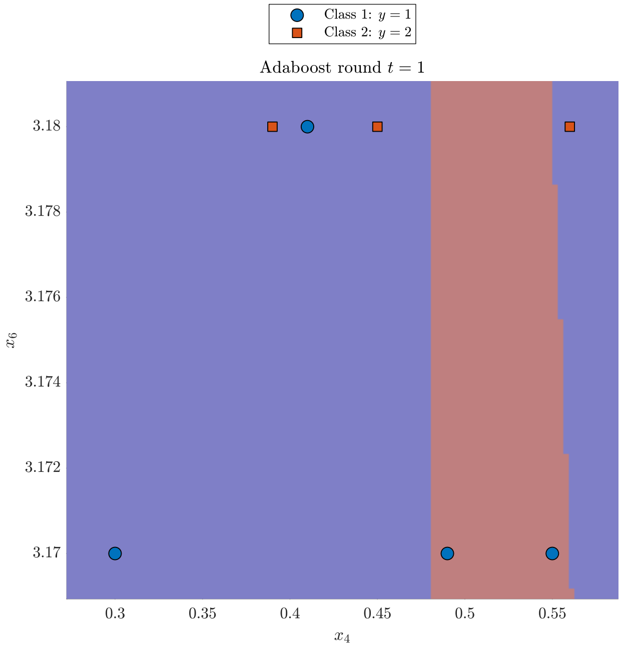

Figure 10: Decision boundaries for a KNN classifier for the first $T = 1$ rounds of boosting.

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

A. $\begin{bmatrix} 0.25 & 0.1 & 0.1 & 0.1 & 0.25 & 0.1 & 0.1 \end{bmatrix}$

B. $\begin{bmatrix} 0.388 & 0.045 & 0.045 & 0.045 & 0.388 & 0.045 & 0.045 \end{bmatrix}$

C. $\begin{bmatrix} 0.126 & 0.15 & 0.15 & 0.15 & 0.126 & 0.15 & 0.15 \end{bmatrix}$

D. $\begin{bmatrix} 0.066 & 0.173 & 0.173 & 0.173 & 0.066 & 0.173 & 0.173 \end{bmatrix}$

E. Don't know.

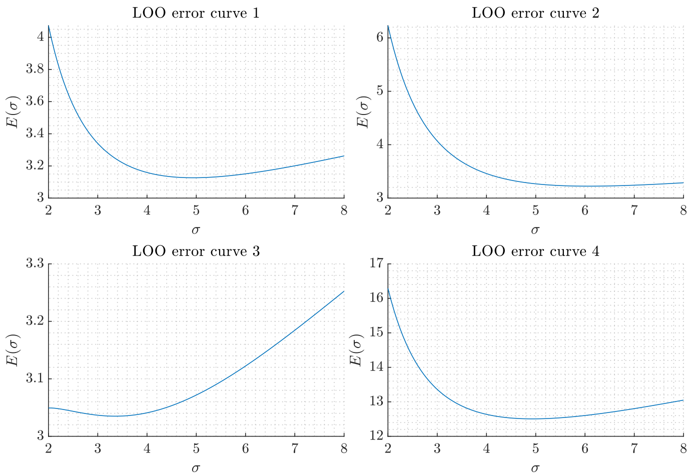

Figure 11: Estimated negative log-likelihood as obtained using LOO cross-validation on a small, $N = 4$ one-dimensional dataset as a function of kernel width σ.

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

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

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

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

Which of the curves in Figure 11 shows the LOO estimate of the generalization error $E(\sigma)$?

A. LOO curve 1

B. LOO curve 2

C. LOO curve 3

D. LOO curve 4

E. Don't know.

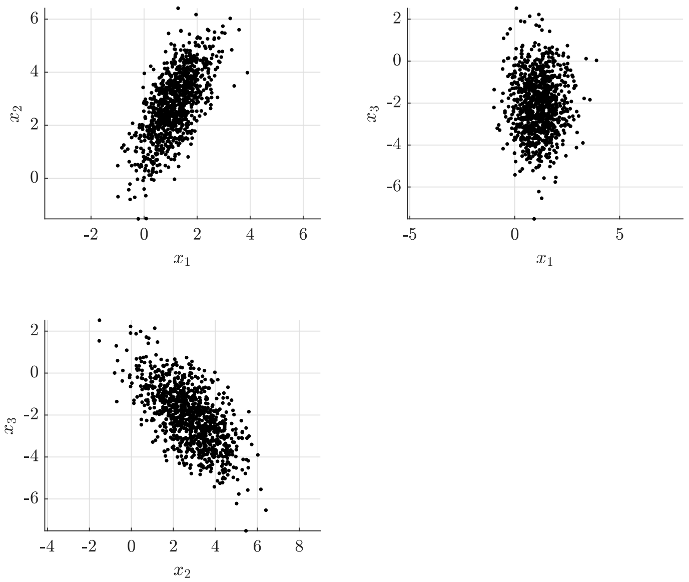

Figure 12: Scatter plots of all pairs of attributes of a vector $\boldsymbol{x}$ when $\boldsymbol{x}$ is a random vector distributed as a multivariate normal distribution of 3 dimensions.

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

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

Plots of each pair of coordinates of the draws $\boldsymbol{x}$ is shown in Figure 12. One of the following covariance matrices was used to generate the data:

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

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

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

A. The correlation between x_1 and x_2 is 0.647

B. The correlation between x_1 and x_2 is -0.611

C. The correlation between x_1 and x_2 is 0.747

D. The correlation between x_1 and x_2 is 0.56

E. Don't know.

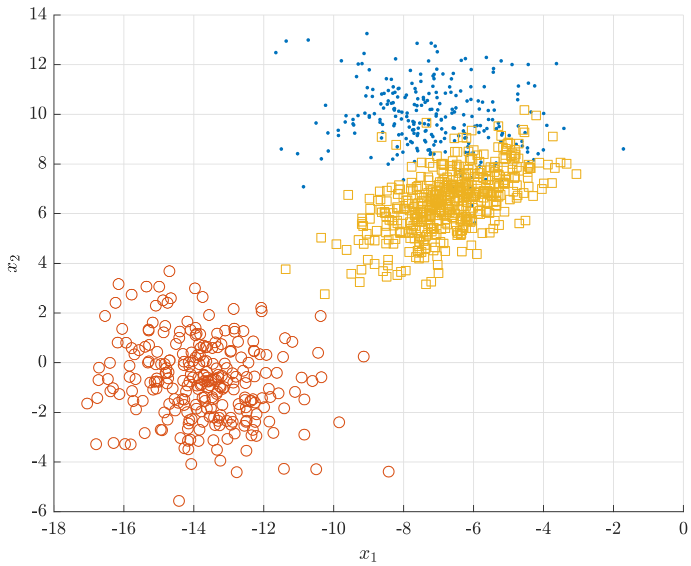

Figure 13: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

Question 27. 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 13 is given 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Each observation is colored and marked in terms of which cluster it came from in the Gaussian Mixture.
Which one of the following GMM densities was used to generate the data?

A.

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

B.

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

C.

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

D.

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

E. Don’t know.