Technical University of Denmark

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

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

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

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

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

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

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

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

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

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

E. Don't know.

Solution 1.

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

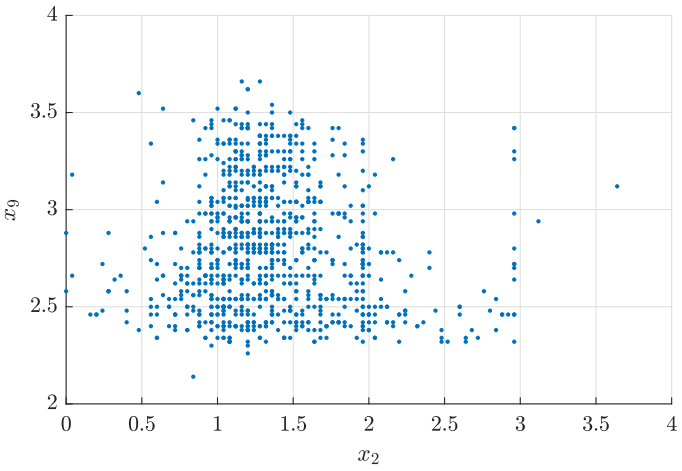

Figure 1: Scatter plot of observations x_2 and x_9 of the travel review dataset described in Table 1.

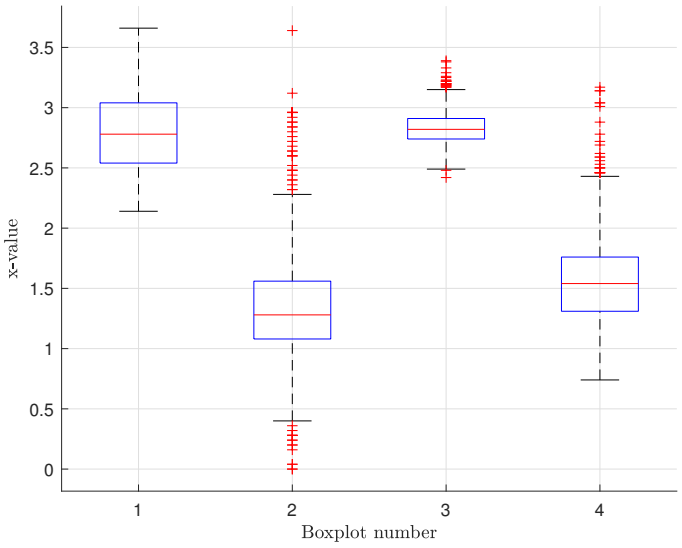

Figure 2: Four boxplots in which two of the boxplots correspond to the two variables plotted in Figure 1.

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

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

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

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

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

Which one of the following statements is true?

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

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

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

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

E. Don't know.

Solution 2. The correct answer is C. To see this, recall the variance explained by a given component k of the PCA is given by

$$\frac{\sigma_k^2}{\sum_{j=1}^{M} \sigma_j^2}$$

where M is the number of attributes in the dataset being analyzed. The values of σ_k can be read off as entry $\sigma_k = S_{kk}$ where $\boldsymbol{S}$ is the diagonal matrix of the SVD computed above. We therefore find the variance explained by components $x_2,\ x_3,\ x_4,\ x_5$ is:

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

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

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

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

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

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

E. Don't know.

Solution 3. The correct answer is A. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_1$ (i.e. column one of $\boldsymbol{V}$) is

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

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

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

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

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

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

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

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

A. 0.41

B. 1.0

C. 0.51

D. 0.83

E. Don't know.

Solution 4.

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

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

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

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

and

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

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

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

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

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

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

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

E. Don't know.

Solution 5.

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

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

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

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

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

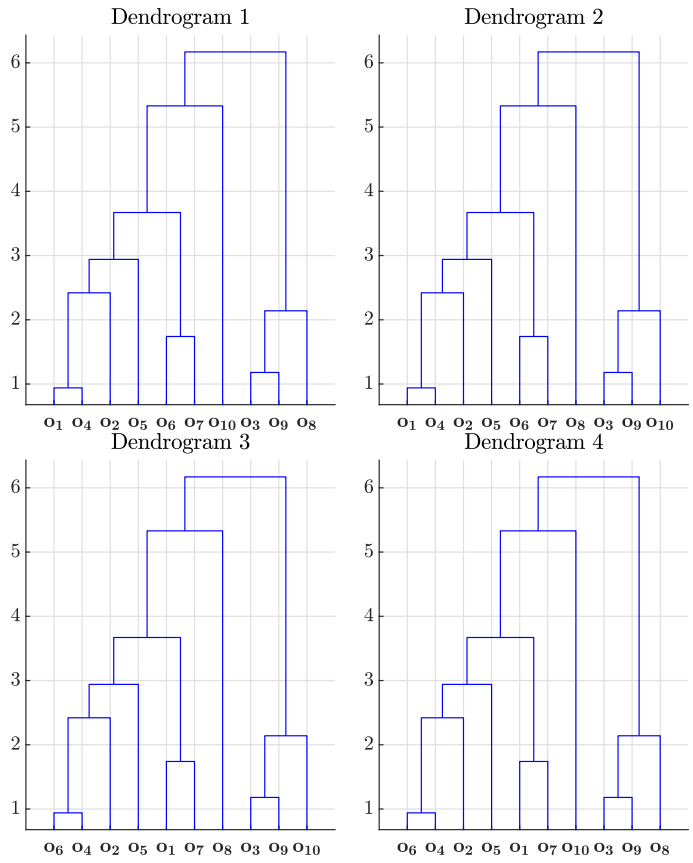

Figure 3: Proposed hierarchical clustering of the 10 observations in Table 2.

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

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

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

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

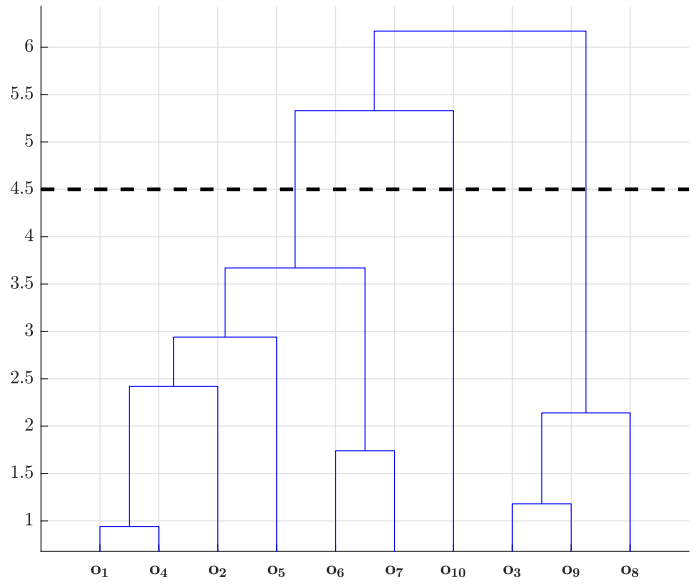

Figure 4: Dendrogram 1 from Figure 3 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

$$S = 4, D = 17$$

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

and the impurity I_0 from

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

In particular we obtain:

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

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

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

A. The accuracy is: 0.42

B. The accuracy is: 0.685

C. The accuracy is: 0.338

D. The accuracy is: 0.097

E. Don't know.

Solution 9.

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

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

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

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

Therefore, the accuracy is:

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

and answer A is correct.

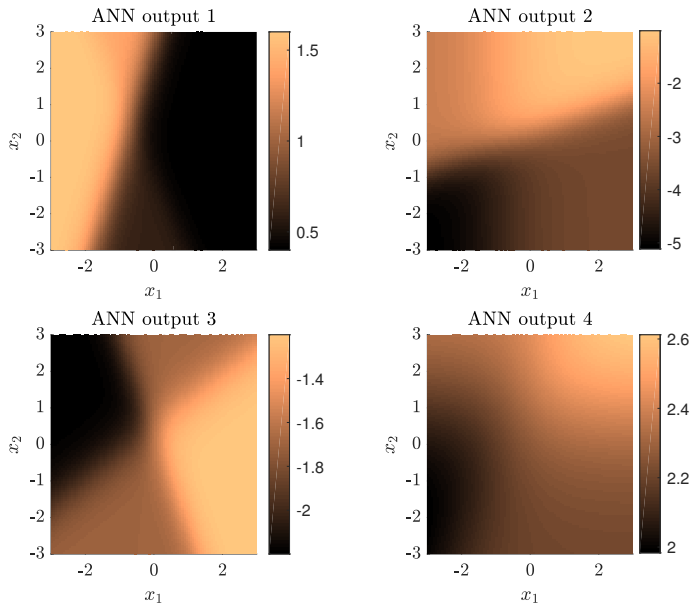

Figure 5: Suggested outputs of an ANN trained on the two attributes x_1 and x_2 from the travel review dataset to predict y.

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

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

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

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

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

A. ANN output 1

B. ANN output 2

C. ANN output 3

D. ANN output 4

E. Don't know.

Solution 10.

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

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

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

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

This rules out all options except D.

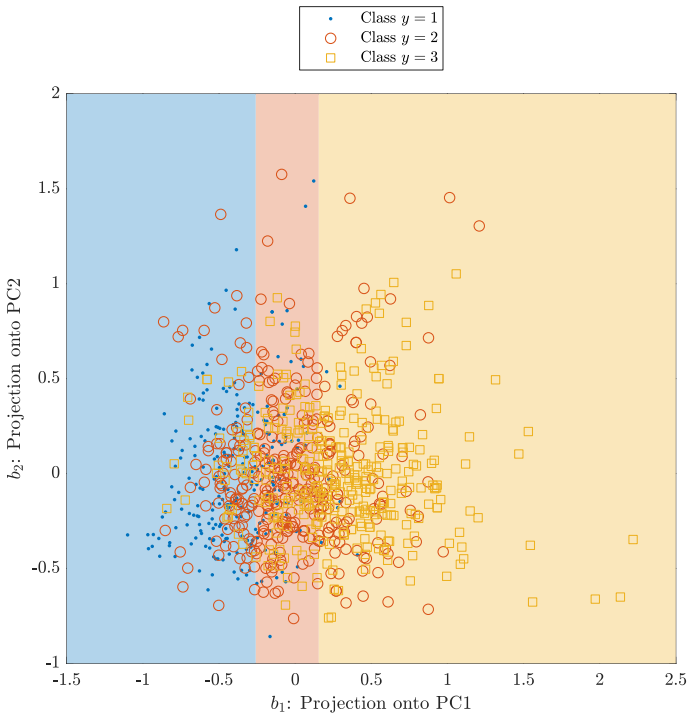

Figure 6: Output of a logistic regression classifier trained on observations from the travel review dataset.

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

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

and then use the softmax transformation in the form:

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

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

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

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

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

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

E. Don't know.

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

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

The projections onto the four options are, in order,

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

Therefore, answer A is correct.

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

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

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

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

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

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

E. Don't know.

Solution 14. Recall the support of an itemset is the number of rows containing all items in the itemset divided by the total number of rows. Therefore, to have a support of 0.15, an itemset needs to be contained in 2 rows. It is easy to see this rules out all options except *A*.

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

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

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

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

C. The confidence is 1

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

E. Don't know.

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

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

Therefore, answer B is correct.

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

Therefore, answer B is correct.

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

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

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

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

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

E. Don't know.

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

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

and therefore option B is correct.

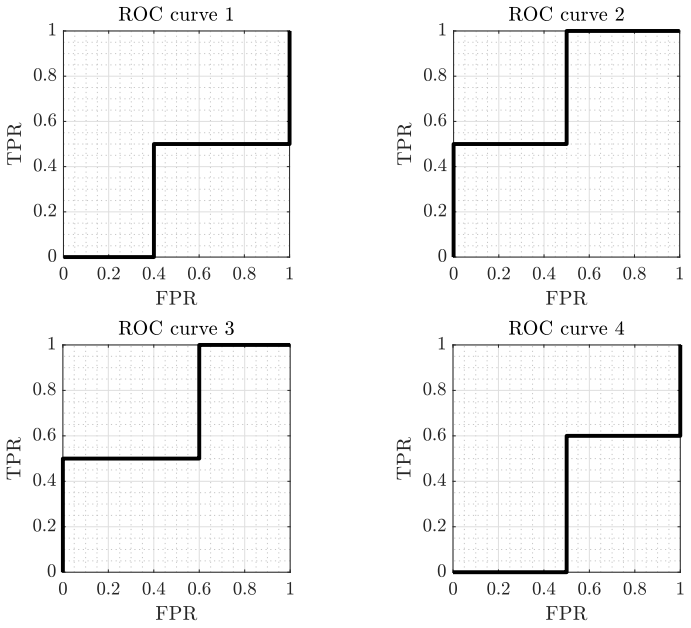

Figure 7: Proposed ROC curves for the neural network classifier with predictions/true class labels given in Table 5

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

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

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

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

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

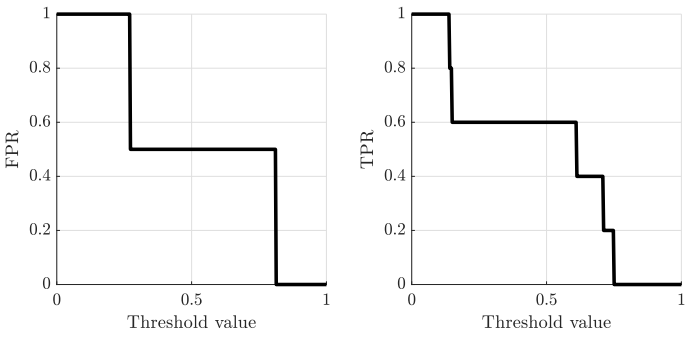

Figure 8: TPR, FPR curves for the classifier.

Solution 18. To compute the AUC, we need to compute the false positive rate (FPR) and true positive rate (TPR) for particular choices of threshold value $\hat{y}$. To compute e.g. the TPR, one assumes every observation predicted to belong to class 1 with a probability higher than $\hat{y}$ is actually assigned to class one. We then divide the total number of observations belonging to class one *and which are predicted to belong to class 1* with the number of observations in the *positive* class.

Similarly for the FPR, where we now count the number of observations that are assigned to class one *but in fact belongs to class 0*, divided by the total number of observations in the *negative* class.

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

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

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

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

A. Forward selection will select attributes x_6

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

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

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

E. Don't know.

Solution 19.

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

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

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1\}$, $\{x_6\}$, $\{x_7\}$, $\{x_8\}$, $\{x_9\}$. Since the lowest error of the available sets is 4.57, which is lower than 5.528, we update the current selected variables to $\{x_6\}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

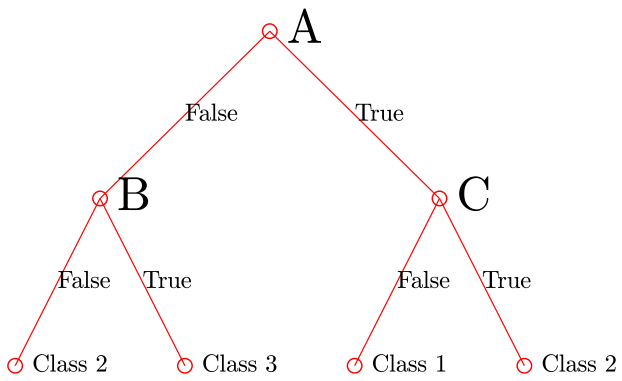

Figure 9: Example classification tree.

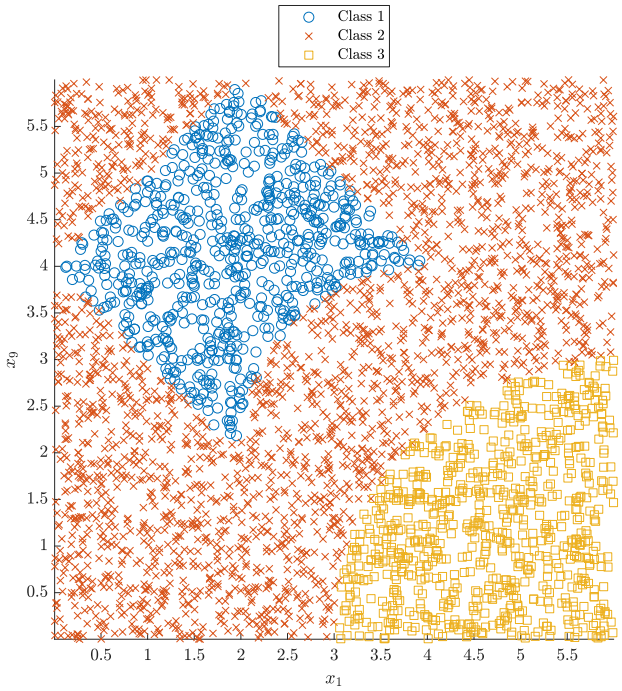

Figure 10: classification boundary.

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

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

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

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

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

E. Don't know.

Solution 21.

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

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

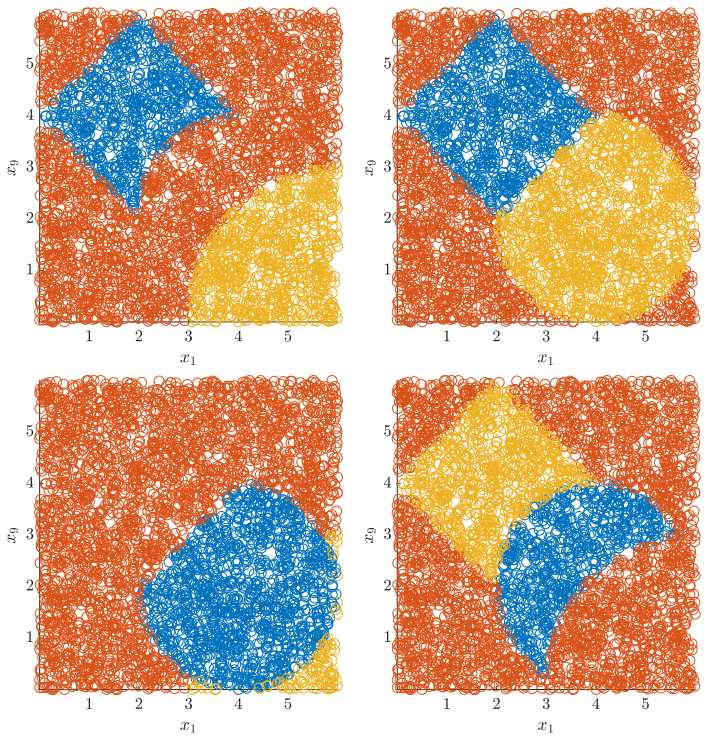

Figure 11: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

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

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

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

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

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

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

A. 208 models

B. 100 models

C. 200 models

D. 104 models

E. Don't know.

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

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

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

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

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

and suppose that the identified values of the mixture weights are

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

and the parameters of the multivariate normal densities:

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

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

A. 0.49

B. 0.31

C. 0.08

D. 0.68

E. Don't know.

Solution 23.

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

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

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

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

Combining these with the class-assignment probabilities we obtain:

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

and conclude the solution is B.

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

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

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

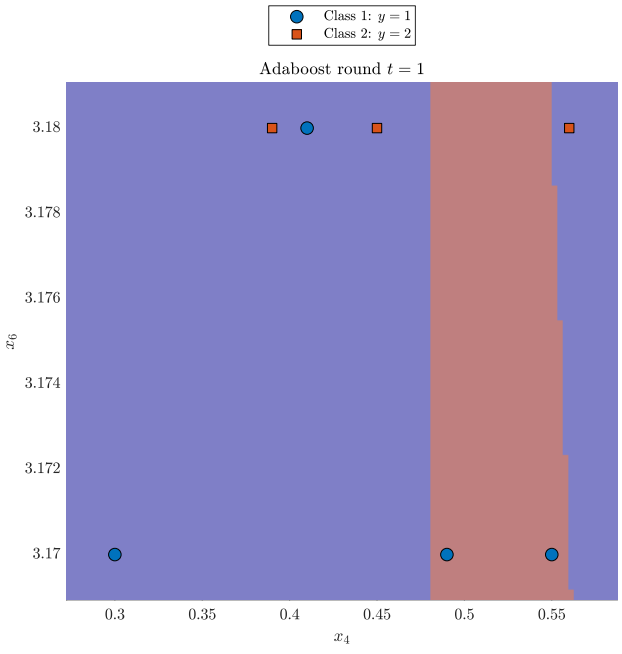

Figure 12: Decision boundaries for a KNN classifier for the first $T = 1$ rounds of boosting.

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

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

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

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

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

E. Don't know.

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

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

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

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

From this, we compute α_t as

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

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

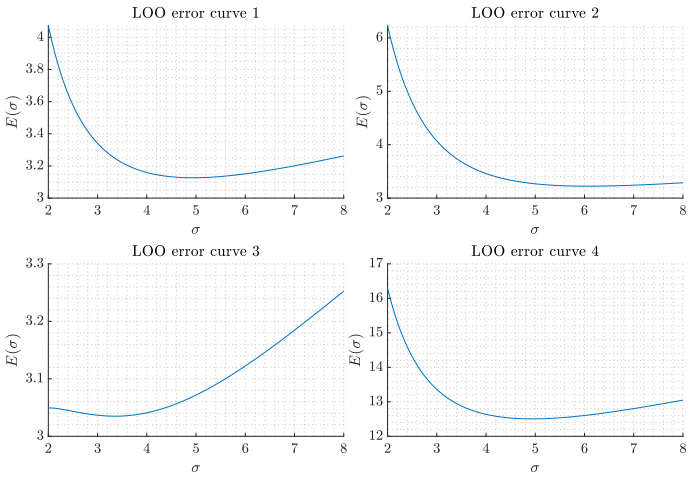

Figure 13: Estimated negative log-likelihood as obtained using LOO cross-validation on a small, $N = 4$ one-dimensional dataset as a function of kernel width σ.

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

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

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

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

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

A. LOO curve 1

B. LOO curve 2

C. LOO curve 3

D. LOO curve 4

E. Don't know.

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

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

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

These values are approximately:

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

The LOO error is then:

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

Therefore, the correct answer is A.

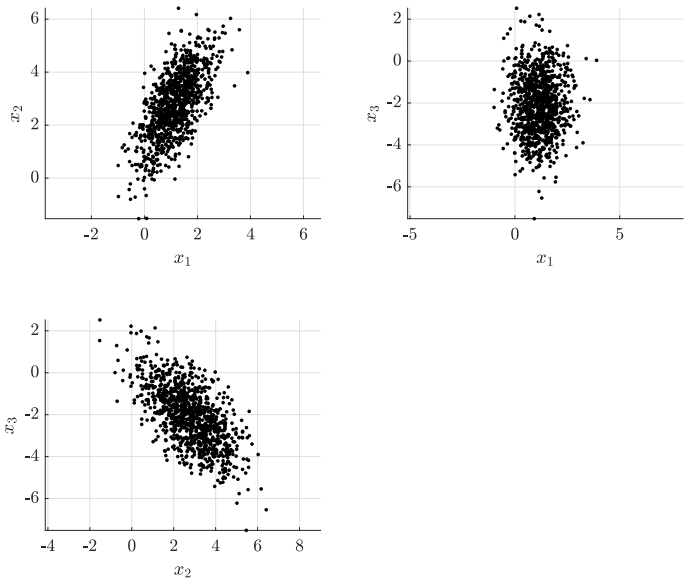

Figure 14: Scatter plots of all pairs of attributes of a vector $\boldsymbol{x}$ when $\boldsymbol{x}$ is a random vector distributed as a multivariate normal distribution of 3 dimensions.

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

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

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

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

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

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

A. The correlation between x_1 and x_2 is 0.647

B. The correlation between x_1 and x_2 is -0.611

C. The correlation between x_1 and x_2 is 0.747

D. The correlation between x_1 and x_2 is 0.56

E. Don't know.

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

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

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

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

The correlation is then given as:

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

and therefore answer A is correct.

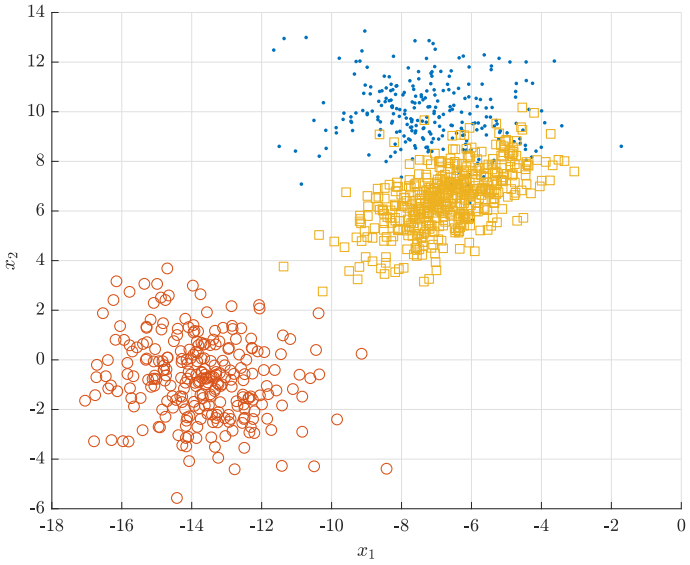

Figure 15: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

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

generate the data?

A.

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

B.

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

C.

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

D.

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

E. Don't know.

Solution 27.

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

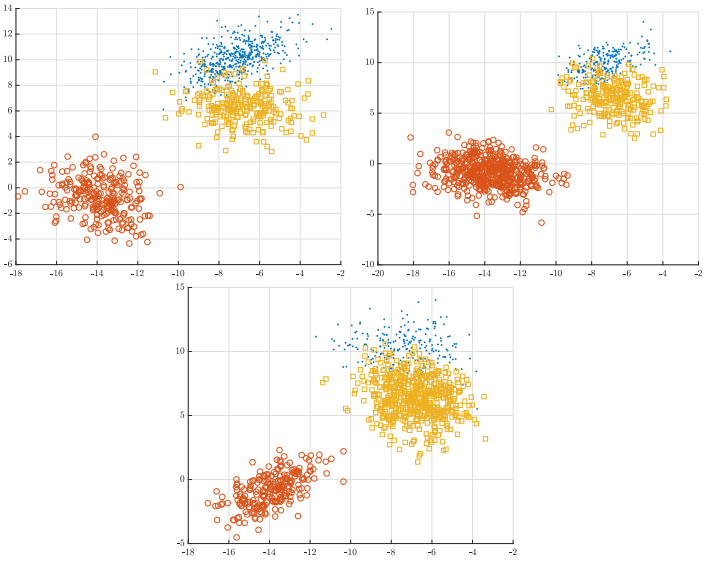

Figure 16: GMM mixtures corresponding to alternative options.