Technical University of Denmark

Written examination: August 18th 2020, 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 answers.txt.

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
A	A	A	B	A	D	D	C	A	A
11	12	13	14	15	16	17	18	19	20
D	B	A	C	C	D	B	A	B	B
21	22	23	24	25	26	27			
C	B	A	D	B	B	D			

PLEASE HAND IN YOUR ANSWERS DIGITALLY USING THE ELECTRONIC FILE answers.txt.

No.	Attribute description	Abbrev.
x_1	Tonnage (1000s of tons)	Tonnage
x_2	Passengers (100s)	Passengers
x_3	Length (100s of feet)	Length
x_4	Cabins (100s)	Cabins
x_5	Passenger Density	Density
y	Ship age (as of 2013, binarized)	Ship Age

Table 1: Description of the features of the Cruise ship dataset used in this exam. It consists of cruise ship information from www.truecruise.com and contains measurements of ship size and passenger properties. The data has been processed such that missing values have been removed and 1000s and 100s means respectively measured in thousands and hundreds. We consider the goal as predicting the ship age (binarized) as a function of the other properties. For classification, the attribute y is discrete taking values $y = 0$ (corresponding to an old ship model from 1999 or before), and $y = 1$ (corresponding to a new ship model build after 2013). There are $N = 158$ observations in total.

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

In Figure 1 is shown a scatter plot of the two attributes x_3 and x_4 from the Cruise ship dataset and in Figure 2 boxplots of the attributes x_1, x_3, x_4, x_5 (not in that order) are given. Which one of the following statements is true?

A. Attribute x_3 corresponds to boxplot 2 and x_4 corresponds to boxplot 1

B. Attribute x_3 corresponds to boxplot 1 and x_4 corresponds to boxplot 4

C. Attribute x_3 corresponds to boxplot 1 and x_4 corresponds to boxplot 3

D. Attribute x_3 corresponds to boxplot 2 and x_4 corresponds to boxplot 4

E. Don't know.

Solution 1.

The correct answer is A. 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

[1] Dataset obtained from `http://users.stat.ufl.edu/~winner/data/cruise_ship.dat`

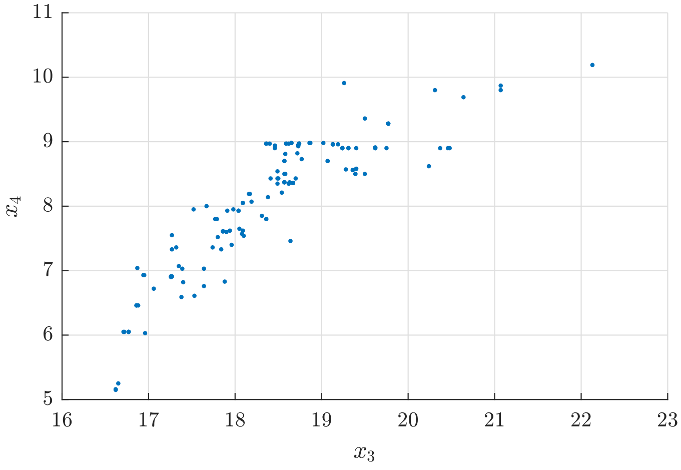

Figure 1: Scatter plot of observations x_3 and x_4 of the Cruise ship dataset described in Table 1.

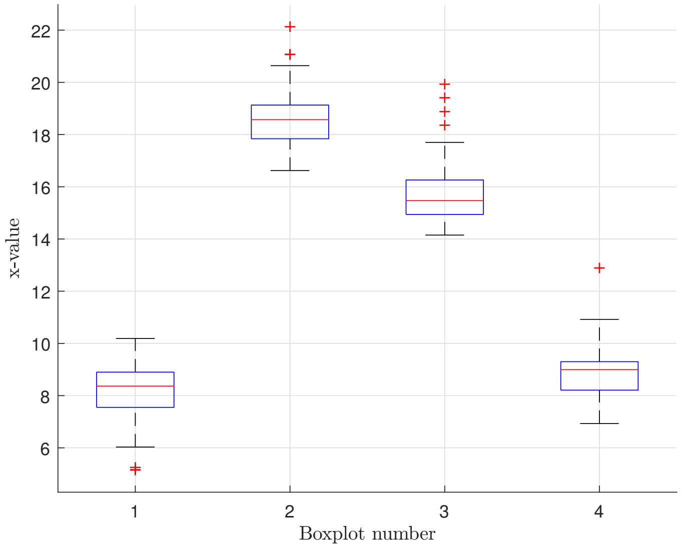

Figure 2: Four boxplots in which two of the boxplots correspond to the two variables plotted in Figure 1.

the two axis and (visually estimate) the point such that half the mass of the data is above and below. For x_3 this is 18.6 and for x_4 this is 8.4. Furhtermore, by inspecting highest and lowest value of x_3 and x_4 we can rule out all but option A.

Question 2. A Principal Component Analysis (PCA) is carried out on the Cruise ship dataset in Table 1 based on the attributes $x_1,\ x_2,\ x_3,\ x_4,\ x_5$.

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

$$\boldsymbol{V} = \begin{bmatrix} 0.37 & 0.4 & 0.84 & -0.07 & 0.0 \\ -0.05 & 0.91 & -0.41 & 0.04 & 0.01 \\ -0.54 & 0.05 & 0.25 & 0.4 & 0.7 \\ -0.53 & 0.07 & 0.13 & -0.84 & 0.03 \\ -0.54 & 0.07 & 0.24 & 0.36 & -0.72 \end{bmatrix}, \tag{1}$$

$$\boldsymbol{S} = \begin{bmatrix} 22.44 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 13.06 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 9.24 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 4.69 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.92 \end{bmatrix}.$$

Which one of the following statements is true?

A. The variance explained by the last two principal components is greater than 3.0%

B. The variance explained by the first four principal components is greater than 99.7 %

C. The variance explained by the first two principal components is less than 80.0 %

D. The variance explained by the last four principal components is greater than 40.0 %

E. Don't know.

Solution 2. The correct answer is A. 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 $x_4,\ x_5$ is:

$$\text{Var.Expl.} = \frac{\sigma_4^2 + \sigma_5^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2} = 3.27\%.$$

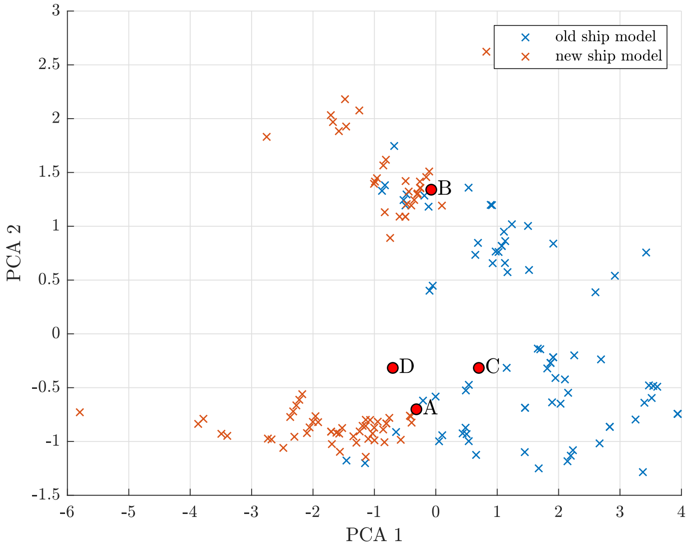

Figure 3: Projection of the Cruise ship dataset described in Table 1 projected onto the first two principal components.

Question 3. Consider again the Cruise ship dataset. In Figure 3 the dataset has been projected onto the two first principal components using the PCA projection described in Equation (1). Suppose a given observation from the Cruise ship dataset is

$$\boldsymbol{x} = \begin{bmatrix} 15.7 \\ 28405.5 \\ 18.5 \\ 8.8 \\ 8.8 \end{bmatrix}$$

and in addition to this the column-wise mean/standard deviation of the Cruise ship dataset, as used when processing the dataset for the PCA projection in Equation (1), is:

$$\begin{aligned} \text{mean: } & \begin{bmatrix} 15.7 & 28406.3 & 18.5 & 8.1 & 8.8 \end{bmatrix}, \\ \text{standard deviation: } & \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \end{aligned}$$

Which one of the four points indicated by the red circles in Figure 3 corresponds to observation $\boldsymbol{x}$?

A. Red circle A

B. Red circle **B**

C. Red circle **C**

D. Red circle **D**

E. Don't know.

Solution 3. To solve the problem, it is enough to compute the PCA projection onto the dataset. After subtracting the mean and dividing by the standard deviation we obtain

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} 0.0 & -0.8 & 0.0 & 0.7 & 0.0 \end{bmatrix}$$

It is then a simple matter of projecting this vector on the two PCA components. The PCA vectors are read of as being:

$$\boldsymbol{v}_1 = \begin{bmatrix} 0.37 \\ -0.05 \\ -0.54 \\ -0.53 \\ -0.54 \end{bmatrix}, \boldsymbol{v}_2 = \begin{bmatrix} 0.4 \\ 0.91 \\ 0.05 \\ 0.07 \\ 0.07 \end{bmatrix}$$

And the coordinates are given as $b_k = \tilde{\boldsymbol{x}}^\top \boldsymbol{v}_k$, i.e. $\boldsymbol{b} = \begin{bmatrix} -0.331 & -0.679 \end{bmatrix}$. From the figure it is clearly seen option A is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	3.4	4.0	3.7	1.3	9.8	4.5	4.8	6.4	7.4
o_2	3.4	0.0	3.9	1.9	3.6	7.3	2.3	1.4	3.8	4.9
o_3	4.0	3.9	0.0	4.8	5.1	9.6	5.2	4.5	6.5	7.6
o_4	3.7	1.9	4.8	0.0	3.4	6.5	0.9	2.6	2.9	3.9
o_5	1.3	3.6	5.1	3.4	0.0	9.5	4.3	4.9	6.1	7.0
o_6	9.8	7.3	9.6	6.5	9.5	0.0	5.7	6.3	3.7	2.8
o_7	4.5	2.3	5.2	0.9	4.3	5.7	0.0	2.4	2.0	3.1
o_8	4.8	1.4	4.5	2.6	4.9	6.3	2.4	0.0	3.1	4.1
o_9	6.4	3.8	6.5	2.9	6.1	3.7	2.0	3.1	0.0	1.3
o_{10}	7.4	4.9	7.6	3.9	7.0	2.8	3.1	4.1	1.3	0.0

Table 2: The pairwise Euclidian distances, $d(o_i, o_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 10 observations from the Cruise ship dataset (recall that $M = 5$). 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 $\{o_1, o_2, o_3, o_4, o_5\}$ belongs to class $y = 0$ (corresponding to an old ship model), and the red observations $\{o_6, o_7, o_8, o_9, o_{10}\}$ belongs to class $y = 1$ (corresponding to a new ship model).

Question 4. To examine if observation o_6 may be an outlier, we will calculate the K-nearest neighborhood density (i.e., inverse of the average distance to neighbors) using the Euclidean distance based on the observations given in Table 2 only. For an observation $\boldsymbol{x}_i$, recall the density is computed using the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation itself, $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$, and is denoted by $\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ What is the density for observation o_6 for $K = 1$ nearest neighbors?

A. 0.31

B. 0.36

C. 0.46

D. 0.77

E. Don't know.

Solution 4.

The density is given as:

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

Thus, to solve the problem, we only need to plug in the values. We find that the $k = 1$ neighborhood of o_6 and density is:

$$N_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = \{o_{10}\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = 0.357$$

Therefore option B is correct.

Question 5. Consider the distances in Table 2 based on 10 observations from the Cruise ship dataset. The class labels $y = 0$ and $y = 1$ (see table caption for details) will be predicted using a k-nearest neighbour classifier based on the distances given in Table 2 (ties are broken in the usual manner by considering the nearest observation from the tied classes). We will apply *hold-out cross-validation*, in which the 10 observations is split into a training and test-set, the KNN classifier is trained on the training set, and used to predict each class label in the test set. As usual, we report the error rate (averaged over the test set). Suppose $K = 3$ and the training and test sets are given as the observations:

$$\mathcal{D}^{\text{train}} = \{o_3, o_4, o_5, o_8, o_9, o_{10}\}$$
$$\mathcal{D}^{\text{test}} = \{o_1, o_2, o_6, o_7\}.$$

What is the error rate as computed on the test set?

A. error rate $= 0$

B. error rate $= \frac{1}{4}$

C. error rate $= \frac{1}{2}$

D. error rate $= \frac{3}{4}$

E. Don't know.

Solution 5.

The correct answer is A. To compute the error rate for a particular observation i in the test set

$$\{o_1, o_2, o_6, o_7\}$$

we train a model on the observations in the training set and use it to predict observation i. Doing this is simply a matter of finding the observations in the training set closest to i according to Table 2 and predict i as belonging to the majority class. Concretely, we find:

$N(o_1, k) = \{o_5, o_4, o_3\}$ all having $y = 0$,
$N(o_2, k) = \{o_8, o_4, o_5\}$, having majority of $y = 0$ observations,
$N(o_6, k) = \{o_{10}, o_9, o_8\}$all having $y = 1$,
$N(o_7, k) = \{o_4, o_9, o_8\}$, having majority of $y = 1$ observations.

In other words, all observations are classified correctly. Therefore, the error rate is $\frac{0}{4}$.

Question 6. Suppose a GMM model is applied to the Cruise ship dataset considering the 10 observations used in Table 2. The GMM is constructed as having $K = 3$ components, and the mean vector $\boldsymbol{\mu}_k$ of component k is assumed to be equal to one of the observations:

$$o_4, \; o_5, \; o_6,$$

such that $\boldsymbol{\mu}_1 = \boldsymbol{x}_4$, $\boldsymbol{\mu}_2 = \boldsymbol{x}_5$, and $\boldsymbol{\mu}_3 = \boldsymbol{x}_6$. We recall that the GMM has the overall form $p(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Assume the covariances $\boldsymbol{\Sigma}_k$ are selected to be diagonal matrices of the form $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$. When the GMM has to be evaluated we must compute K expressions of the form

$$\mathcal{N}(\boldsymbol{o}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^5 |\boldsymbol{\Sigma}_k|}} e^{\frac{-d(\boldsymbol{o}_i, \boldsymbol{\mu}_k)^2}{2\sigma_k^2}}$$

Where $|\cdot|$ is the determinant.

If $\sigma_1 = 2.6$, what is the first determinant?

A. $|\boldsymbol{\Sigma}_1| = 13.0$

B. $|\boldsymbol{\Sigma}_1| = 33.8$

C. $|\boldsymbol{\Sigma}_1| = 118.8$

D. $|\boldsymbol{\Sigma}_1| = 14116.7$

E. Don't know.

Solution 6.

The dimensionality of the multivariate normal distributions can be found in Table 2 to be $M = 5$. The determinant is then simply the product of the diagonal terms:

$$|\boldsymbol{\Sigma}_1| = (\sigma_1^2)^5$$

Plugging in the values we see option D is correct.

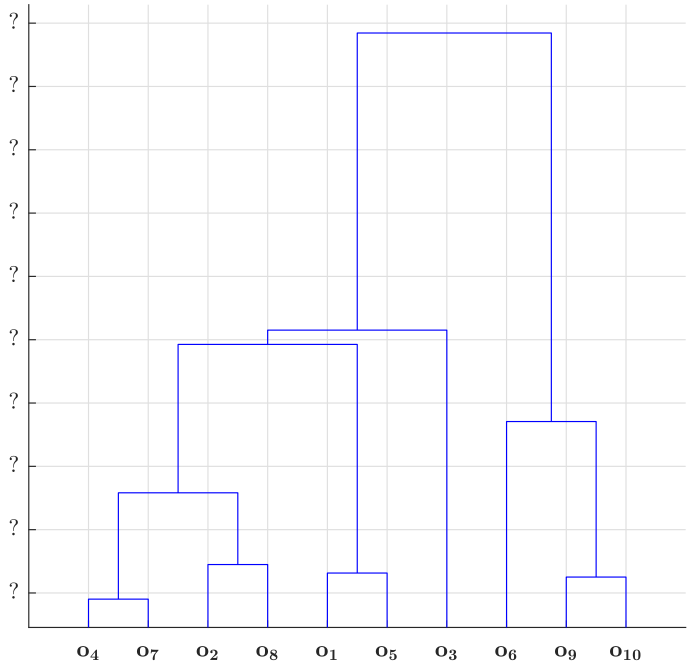

Figure 4: Hierarchical clustering of the 10 observations in Table 2.

Question 7. A hierarchical clustering is applied to the 10 observations in Table 2 using *maximum* linkage, and the result is shown in Figure 4. Recall that in a dendrogram plot, the y-levels as which branches merge (i.e., the horizontal lines) is determined by the distance of the two clusters being merged. At what level does the top-most branches merge? (i.e., what is the distance of the last two clusters being merged?)

A. Distance is 2.0

B. Distance is 3.7

C. Distance is 5.2

D. Distance is 9.8

E. Don't know.

Solution 7. The correct solution is C. The distance, when using the *maximum* linkage, is determined by the maximum pairwise distance of observations in the two clusters. From the figure the two clusters are

$$\{o_6, o_9, o_{10}\}, \{o_3, o_1, o_5, o_4, o_7, o_2, o_8\}$$

and the distances are easily read from Table 2 where the maximum distance is found to be $d(o_6, o_1) = 9.8$.

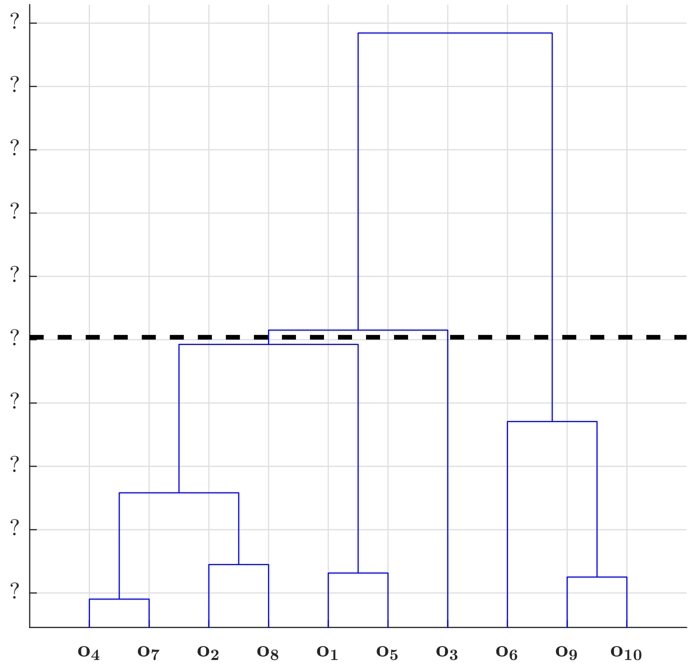

Figure 5: The dendrogram from Figure 4 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

Question 8. Consider the dendrogram from Figure 4. Suppose we apply a cutoff (indicated by the black line in Figure 5) 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

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

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

A. $\mathrm{J}[Z, Q] \approx 0.324$

B. $\mathrm{J}[Z, Q] \approx 0.343$

C. $\mathrm{J}[Z, Q] \approx 0.357$

D. $\mathrm{J}[Z, Q] \approx 0.370$

E. Don't know.

Solution 8. 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 5 given as:

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

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

$$\boldsymbol{n} = \begin{bmatrix} 1 & 4 & 0 \\ 0 & 2 & 3 \end{bmatrix}$$

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

$$S = 10, D = 17$$

From this the answer can be derived by plugging the values into the formula given in the text and answer C is correct.

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

Table 3: Binarized version of the Cruise ship 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 $\{o_1,\ o_2,\ o_3,\ o_4,\ o_5\}$ belongs to class $y = 0$ (corresponding to an old ship model), and the red observations $\{o_6,\ o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class $y = 1$ (corresponding to a new ship model).

Question 9. We again consider the Cruise ship dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce five 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 $N \times M = 10 \times 5$ binary matrix in Table 3. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_1,\ f_4,\ f_5$. If for an observations we observe

$$f_1 = 1,\ f_4 = 0,\ f_5 = 0$$

what is then the probability it is a new ship model ($y = 1$) according to the Naïve-Bayes classifier?

A. $p_{\mathbf{NB}}(y = 1|f_1 = 1,\ f_4 = 0,\ f_5 = 0) = 0$

B. $p_{\text{NB}}(y = 1|f_1 = 1,\ f_4 = 0,\ f_5 = 0) = \frac{1}{15}$

C. $p_{\text{NB}}(y = 1|f_1 = 1,\ f_4 = 0,\ f_5 = 0) = \frac{2}{5}$

D. $p_{\text{NB}}(y = 1|f_1 = 1,\ f_4 = 0,\ f_5 = 0) = \frac{8}{15}$

E. Don't know.

Solution 9. To solve this problem, we simply use the general form of the naïve-Bayes approximation and

[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.

plug in the relevant numbers. We get:

$$p_{\text{NB}}(y=1|f_1=1,\ f_4=0,\ f_5=0) = \frac{p(f_1=1|y=1)p(f_4=0|y=1)p(f_5=0|y=1)p(y=1)}{\sum_{j=0}^{1} p(f_1=1|y=j)p(f_4=0|y=j)p(f_5=0|y=j)p(y=j)}$$
$$= \frac{\frac{0}{5}\frac{4}{5}\frac{4}{5}\frac{1}{2}}{\frac{1}{5}\frac{4}{5}\frac{2}{5}\frac{1}{2}+\frac{0}{5}\frac{4}{5}\frac{4}{5}\frac{1}{2}}$$
$$= 0.$$

Therefore, answer A is correct.

Question 10. Consider the binarized version of the Cruise ship dataset shown in Table 3.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 5$ items $f_1, f_2, \ldots, f_5$. Which of the following options represents all (non-empty) itemsets with support greater than 0.35 (and only itemsets with support greater than 0.35)?

A. $\{f_5\}$

B. $\{f_2\}$, $\{f_5\}$

C. $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_5\}$

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

E. Don't know.

Solution 10. 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 0.35, an itemset needs to be contained in 4 rows. It is easy to see this rules out all options except *A*.

Question 11. We again consider the binary matrix from Table 3 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 5$ items $f_1, \ldots, f_5$.

What is the *confidence* of the rule $\{f_1\} \to \{f_2\}$?

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

B. The confidence is $\frac{1}{5}$

C. The confidence is $\frac{1}{3}$

D. The confidence is 1

E. Don't know.

Solution 11. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_1\} \cup \{f_2\})}{\text{support}(\{f_1\})} = \frac{\frac{1}{10}}{\frac{1}{10}} = 1.$$

Therefore, answer D is correct.

Fold	M_1, M_2	$M_1, \overline{M}_2$	$\overline{M}_1, M_2$	$\overline{M}_1, \overline{M}_2$
1	47	2	0	3
2	45	3	1	4
3	47	1	1	4

Table 4: Outcome of cross-validation. Rows are combination of outcomes of the two models.

Question 12.

We will consider the Cruise ship dataset and two models for predicting the class label y. Specifically, let M_1 be a $K = 1$ nearest neighbor classification model and M_2 a $K = 5$ nearest neighbor classification model. To compare them statistically, we perform $K = 3$ fold cross-validation, and for each fold we record the number of observations where both models are correct (as M_1, M_2), M_1 is correct and M_2 wrong (as $M_1, \overline{M_2}$), and so on. The outcome can be found in Table 4.

Suppose we compute a Jeffrey confidence interval

$$\theta_L = \text{cdf}_B^{-1}\left(\frac{\alpha}{2}|a,b\right),$$
$$\theta_U = \text{cdf}_B^{-1}\left(1-\frac{\alpha}{2}|a,b\right)$$

for the accuracy θ of model M_2. What is the estimated mean performance $\hat{\theta} = \mathbb{E}[\theta]$ of M_2 according to the test?

A. $\mathbb{E}[\theta] = 0.877$

B. $\mathbb{E}[\theta] = 0.890$

C. $\mathbb{E}[\theta] = 0.918$

D. $\mathbb{E}[\theta] = 0.924$

E. Don't know.

Solution 12.

Since the cross-validation folds are non-overlapping, we can easily find the number of times model M_2 makes a correct prediction as the sum of columns 1 and 3 in Table 4 or $n^+ = 141$. Similarly, the sum of all entries in column 2 and 4 are the number of wrong guesses or $n^- = 17$.

The mean performance difference can now be simply computed as $\frac{a}{a+b}$ where $a = n^+ + \frac{1}{2}$ and $a = n^- + \frac{1}{2}$. Therefore, B is correct.

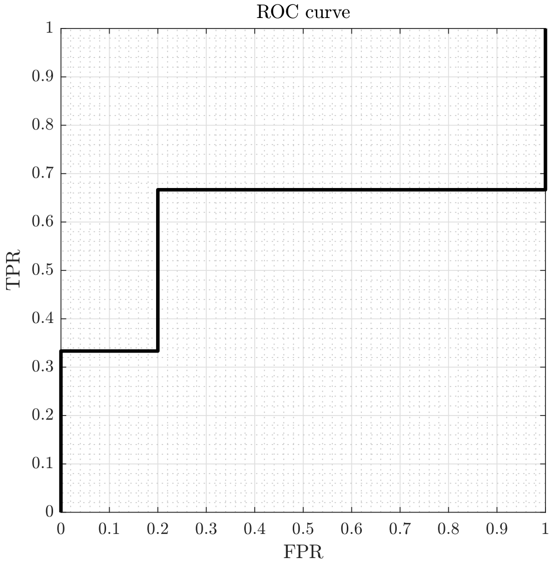

Figure 6: ROC curve for a neural network classifier, where the predictions and true class labels are one of the options in Figure 7.

Question 13. A neural network classifier is trained to distinguish between two classes $y \in \{0, 1\}$ and produce class-probability $\hat{y}$. The *receiver operator characteristic* (ROC) curve of the network when evaluated on a test set with $N = 8$ observations is shown in Figure 6. Suppose we plot the predictions on the $N = 8$ test observations by their $\hat{y}$ value along the x-axis and indicate the class labels by either a black circle (class $y = 0$) or red cross ($y = 1$), which one of the subplots in Figure 7 then corresponds to the ROC curve in Figure 6?

A. Prediction A

B. Prediction B

C. Prediction C

D. Prediction D

E. Don't know.

Solution 13. The correct answer is A. To see this, recall that the ROC curve is computed from the false positive rate (FPR) and true positive rate (TPR) for particular choices of threshold value $\hat{y}$. To compute e.g. the TPR, one assumes every observation predicted to belong to class 1 with a probability higher than $\hat{y}$ is actually assigned to class one. We then divide the total number of observations belonging to class one *and which are predicted to belong to class 1* with the number of observations in the *positive* class.

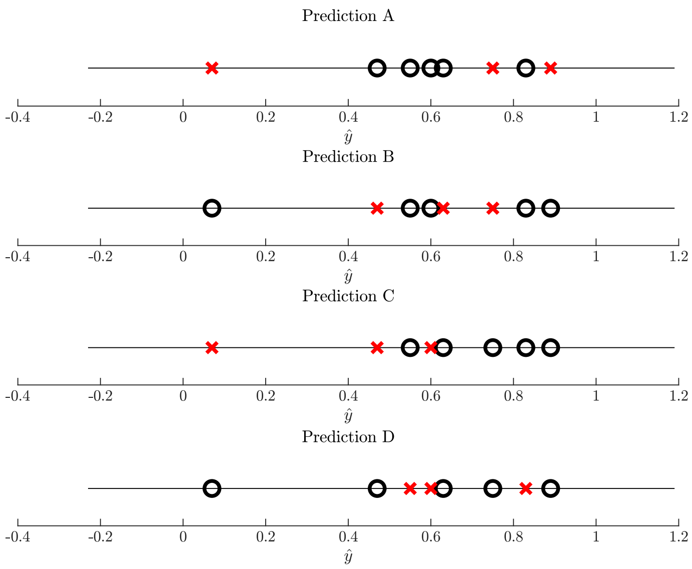

Figure 7: Four candidate predictions for the ROC curve in Figure 6. The observations are plotted horizontally, such that the position on the x-axis indicate the predicted value $\hat{y}_i$, and the marker/color indicate the class membership, such that the black circles indicate the observation belongs to class $y_i = 0$ and red crosses to $y_i = 1$.

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 ROC curve. This rules out all options except A. For completeness, we have included the ROC curves for all options in Figure 9.

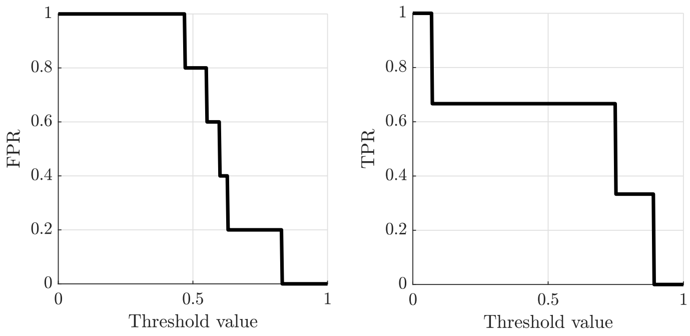

Figure 8: TPR, FPR curves for the classifier.

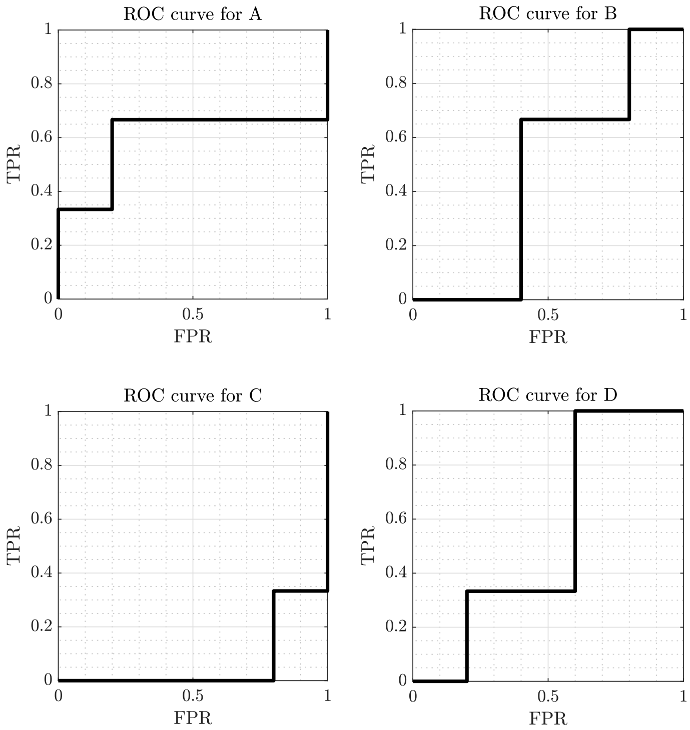

Figure 9: ROC curves for all options.

Question 14. Which one of the following statements regarding the Gaussian Mixture Model (GMM) is correct?

A. When fitting the GMM with the EM-algorithm we are guaranteed to find the optimal clustering.

B. The number of clusters to use for the GMM can be determined by observing how similar the covariance matrices of each cluster are.

C. Increasing the number of clusters K will in general increase the log-likelihood of the training data $\log P(\boldsymbol{X}^{\text{train}}|\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}, \boldsymbol{\pi})$.

D. The covariance matrix of each cluster in the GMM fitted using the EM-algorithm is in general asymmetric, i.e. $\boldsymbol{\Sigma}_k \neq \boldsymbol{\Sigma}_k^\top$.

E. Don't know

Solution 14. The EM-algorithm is prone to issues of local minima and it is therefore recommended to fit the GMM using multiple restarts. We are therefore not guaranteed to find the optimal clustering solution when running the EM-algorithm. The number of clusters can be determined using information criteria such as AIC/BIC or by the use of cross-validation but not by observing the shape of the covariance matrices of the clusters. In particular, we can imagine a clustering problem where multiple clusters have similar covariance matrices but different locations justifying their use. Increasing the number of clusters will in general enable the GMM to better fit to the training data and thereby increase the log-likelihood $\log P(\boldsymbol{X}^{\text{train}}|\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}, \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}, \boldsymbol{\pi})$. The covariance matrices are always symmetric. In particular, we observe from the update of each cluster's covariance matrix according to the EM-algorithm that it is given by $\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{i=1}^N \gamma_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top$ such that $\boldsymbol{\Sigma}_k^\top = \left[\frac{1}{N_k}\sum_{i=1}^N \gamma_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top\right]^\top = \frac{1}{N_k}\sum_{i=1}^N \gamma_{ik}\left[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top\right]^\top = \frac{1}{N_k}\sum_{i=1}^N \gamma_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top$. As a result, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k^\top$.

$p(\hat{x}_1, \hat{x}_2, y)$	$y = 0$	$y = 1$
$\hat{x}_1 = 0,\ \hat{x}_2 = 0$	0.09	0.3
$\hat{x}_1 = 0,\ \hat{x}_2 = 1$	0.06	0.07
$\hat{x}_1 = 1,\ \hat{x}_2 = 0$	0.15	0
$\hat{x}_1 = 1,\ \hat{x}_2 = 1$	0.2	0.13

Table 5: Probability of observing particular values of $\hat{x}_1$, $\hat{x}_2$, and y.

Question 15. Consider the Cruise ship dataset from Table 1. Suppose the attributes have been binarized such that $\hat{x}_1 = 0$ corresponds to $x_1 \leq 15.47$ (and otherwise $\hat{x}_1 = 1$) and $\hat{x}_2 = 0$ corresponds to $x_2 \leq 28405.53$ (and otherwise $\hat{x}_2 = 1$). Suppose the probability for each of the configurations of $\hat{x}_1$, $\hat{x}_2$ and ship age y are as given in Table 5.

What is then the probability y corresponds to a new ship model given an observation has $\hat{x}_1 = 0$?

A. $p(y = 1|\hat{x}_1 = 0) = 37.00\%$

B. $p(y = 1|\hat{x}_1 = 0) = 52.00\%$

C. $p(y = 1|\hat{x}_1 = 0) = 71.15\%$

D. $p(y = 1|\hat{x}_1 = 0) = 74.00\%$

E. Don't know.

Solution 15. Recall the marginalization rule which gives:

$$\sum_{x_2} p(x_1, x_2, y) = p(x_1, y)$$

We can therefore first marginalize out the $\hat{x}_2$-value. Next we marginalize out x_1 to get the prior probability $p(y)$

$$\sum_{y} p(x_1, y) = p(x_1)$$

From this we can obtain the conditional probability using the product rule $p(y|x) = p(x,y)/p(x)$. In our case we obtain the values:

$$p(\hat{x}_1 = 0, y = 1) = 0.37$$

and

$$p(\hat{x}_1 = 0) = p(\hat{x}_1 = 0, y = 0) + p(\hat{x}_1 = 0, y = 1) = 0.15 + 0.37 = 0$$

and so we obtain $p(y = 1|\hat{x}_1 = 0) = \frac{0.37}{0.52} = 71.2\%$.

Question 16. Suppose s_1 and s_2 are two text documents containing the text:

$$s_1 = \{\text{the cruise will take your breath away}\}$$
$$s_2 = \{\text{wash your hands before the buffet }\}$$

The documents are encoded using a bag-of-words encoding assuming a total vocabulary size of $M = 10000$. No stopwords lists or stemming is applied to the dataset. What is the SMC similarity between documents s_1 and s_2?

A. SMC similarity of s_1 and s_2 is 0.0002

B. SMC similarity of s_1 and s_2 is 0.1538

C. SMC similarity of s_1 and s_2 is 0.1818

D. SMC similarity of s_1 and s_2 is 0.9991

E. Don't know.

Solution 16.

The correct answer is D. Since we are computing the SMC, we also have to count 0-0 matches. We then observe that document s_1 contains $f_{10} = 5$ words that are not in documents s_2 and similarly, document s_2 contains $f_{01} = 4$ words that are not in document s_1, and the two documents have $f_{11} = 2$ words in common. Given this we compute $f_{00} = M - f_{01} - f_{10} - f_{11} = 9989$. All in all this means the SMC is:

$$\text{SMC}(s_1, s_2) = \frac{f_{11} + f_{00}}{M} \approx 0.9991.$$

Question 17. Which one of the following statements in regards to machine learning approaches discussed in the course is correct?

A. Regression as opposed to classification requires the input data $\boldsymbol{X}$ to be continuous.

B. For the AdaBoost algorithm if the weighted error rate $\epsilon_t = 0.5$ at round t then the classifier trained at round t will have no influence on the predictions of the ensemble classifier.

C. Regularized least squares regression (i.e., ridge regression) is in particular expected to be useful when we have a very large number of (independent) observations relative to the number of features, i.e. when $N \gg M$ as opposed to when $N < M$.

D. k-means clustering using Euclidean distances is a contiguity-based clustering approach.

E. Don't know

Solution 17. Regression and classification problems differ not in regards to the input data $\boldsymbol{X}$ but in regards to the outputs y being nominal (classification) or continuous (regression). For the Adaboost algorithm if $\epsilon_t = 0.5$ we have that $\alpha_t = 0.5 \log \frac{1-\epsilon_t}{\epsilon_t} = 0$. As such, when we execute the ensemble voting we will ignore the classifier. Regularized least squares regression is beneficial when the variance can be reduced without introducing too much bias. The variance of a model is high when we have few (independent) observations as compared to features. However, when we have many (independent) observations the parameters of the regression model is very well defined and thus we cannot expect to gain much by regularizing the model in comparison to what is introduced in terms of bias. The k-means clustering procedure is a center-based approach in which the assignment of observations to clusters is based on the distance to the center of the clusters (where cluster centers correspond to the mean of the assigned points to the cluster when considering the Euclidean distance measure).

	$x_1 \leq 14.94$	$x_1 \leq 15.47$	$x_1 \leq 16.26$
$y = 0$	8	24	45
$y = 1$	34	59	74

Table 6: Proposed split of the Cruise ship dataset based on the attribute x_1. 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 18. Suppose we wish to build a classification tree based on Hunt's algorithm where the goal is to predict Ship Age which can belong to two classes, $y = 0$, $y = 1$. The number of observations in each of the classes are:

$$n_{y=0} = 79,\ n_{y=1} = 79.$$

We consider binary splits based on the value of x_1 of the form $x_1 < z$ for three different values of z. In Table 6 we have indicated the number of observations in each of the two 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 $x_1 \leq 16.26$ is $\Delta \approx 0.1835$

B. The impurity gain of the split $x_1 \leq 16.26$ is $\Delta \approx 0.2254$

C. The impurity gain of the split $x_1 \leq 16.26$ is $\Delta \approx 0.2426$

D. The impurity gain of the split $x_1 \leq 16.26$ is $\Delta \approx 0.2468$

E. Don't know.

Solution 18. 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 $x_1 \leq 16.26$ as:

$$R = \begin{bmatrix} 45 & 34 \\ 74 & 5 \end{bmatrix}.$$

We obtain $N(r) = \sum_{ki} R_{ki} = 158$ 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:

$$I_0 = 0.5, I(v_1) = 0.378,\ I(v_2) = 0.128.$$

and

$$N(v_1) = 119,\ N(v_2) = 39.$$

Combining these we see that $\Delta = 0.184$ and therefore option A is correct.

Question 19. We will consider an artificial neural network (ANN) trained on the Cruise ship dataset described in Table 1 to predict the class label y based on attributes $x_1, \dots, x_5$. The neural network has a single hidden layer containing $n_h = 3$ units, and uses the sigmoid activation function to predict the class label y. For the hidden layer we will use the tanh non-linear activation function. How many parameters has to be trained to fit the neural network?

A. Network contains 21 parameters

B. Network contains 22 parameters

C. Network contains 24 parameters

D. Network contains 26 parameters

E. Don't know.

Solution 19.

Each hidden unit has as many input units weights are there are features $M = 5$ plus one (the bias), therefore they contribute with

$$(M+1)n_h$$

weights. The sigmoid is computed deterministically as a single output neuron (connected to all hidden neurons and including a bias term):

$$n_h + 1$$

Adding these two numbers together gives option B.

Question 20. Consider a small dataset comprised of $N = 10$ observations

$$x = \begin{bmatrix} 0.1 & 0.7 & 1.0 & 1.4 & 3.1 & 3.8 & 4.1 & 4.3 & 4.9 & 5.0 \end{bmatrix}.$$

Suppose a k-means algorithm is applied to the dataset with $K = 3$ and using Euclidian distances. At a given stage of the algorithm the data is partitioned into the blocks:

$$\{0.1, 0.7, 1.0\}, \{1.4, 3.1, 3.8, 4.1\}, \{4.3, 4.9, 5.0\}$$

What clustering will the k-means algorithm eventually converge to?

A. $\{0.1, 0.7, 1.0\}$, $\{1.4, 3.1, 3.8\}$, $\{4.1, 4.3, 4.9, 5.0\}$

B. $\{0.1, 0.7, 1.0, 1.4\}$**,** $\{3.1, 3.8\}$**,** $\{4.1, 4.3, 4.9, 5.0\}$

C. $\{0.1.0, 0.7\}$, $\{1.0, 1.4, 3.1, 3.8, 4.1, 4.3\}$, $\{4.9, 5.0\}$

D. $\{0.1, 0.7, 1.0, 1.4\}$, $\{3.1, 3.8, 4.1\}$, $\{4.3, 4.9, 5.0\}$

E. Don't know.

Solution 20. 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. Therefore, the subsequent steps in the K-means algorithm are:

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

$$\mu_1 = 0.6,\ \mu_2 = 3.1,\ \mu_3 = 4.73333.$$

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

$$\{0.1, 0.7, 1, 1.4\}, \{3.1, 3.8\}, \{4.1, 4.3, 4.9, 5\}.$$

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

$$\mu_1 = 0.8,\ \mu_2 = 3.45,\ \mu_3 = 4.575.$$

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

$$\{0.1, 0.7, 1, 1.4\}, \{3.1, 3.8\}, \{4.1, 4.3, 4.9, 5\}.$$

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

Question 21. Consider again the Cruise ship dataset in Table 1. We would like to predict a ship's age using a linear regression considering the output variable before it was binarized. 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_2, x_3, x_4, and x_5 and in Table 7 we have pre-computed the estimated training and test error for the different variable combinations. Which of the following statements is correct?

A. Forward selection will select attributes x_2, x_3, x_4

B. Forward selection will select attributes x_2, x_3, x_4, x_5

C. Backward selection will select attributes x_2, x_3, x_4, x_5

D. Backward selection will select attributes x_2, x_3, x_4

E. Don't know.

Solution 21.

The correct answer is C. 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 1.251.

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 $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$. Since the lowest error of the available sets is 0.609, which is lower than 1.251, we update the current selected variables to $\{x_5\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_5\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1, x_4\}$, $\{x_2, x_4\}$, $\{x_3, x_4\}$, $\{x_1, x_5\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$, $\{x_4, x_5\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Feature(s)	Training RMSE	Test RMSE
none	0.882	1.251
x_1	0.627	2.35
x_2	0.839	1.273
x_3	0.82	0.921
x_4	0.786	0.712
x_5	0.463	0.609
x_1, x_2	0.627	2.376
x_1, x_3	0.432	2.122
x_2, x_3	0.785	0.956
x_1, x_4	0.448	1.619
x_2, x_4	0.659	0.522
x_3, x_4	0.784	0.71
x_1, x_5	0.179	1.602
x_2, x_5	0.441	0.629
x_3, x_5	0.251	0.736
x_4, x_5	0.46	0.675
x_1, x_2, x_3	0.406	2.303
x_1, x_2, x_4	0.432	1.434
x_1, x_3, x_4	0.406	1.819
x_2, x_3, x_4	0.622	0.485
x_1, x_2, x_5	0.166	1.698
x_1, x_3, x_5	0.13	1.378
x_2, x_3, x_5	0.226	0.77
x_1, x_4, x_5	0.175	1.542
x_2, x_4, x_5	0.44	0.603
x_3, x_4, x_5	0.203	0.502
x_1, x_2, x_3, x_4	0.403	2.035
x_1, x_2, x_3, x_5	0.128	1.421
x_1, x_2, x_4, x_5	0.166	1.691
x_1, x_3, x_4, x_5	0.091	1.058
x_2, x_3, x_4, x_5	0.049	0.317
x_1, x_2, x_3, x_4, x_5	0.047	0.424

Table 7: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y_r in the Cruise ship dataset using different combinations of the features x_1, x_2, x_3, x_4, and x_5.

Backward selection: The method is initialized with the set $\{x_1, x_2, x_3, x_4, x_5\}$ having an error of 0.424.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_2, x_3, x_4, x_5\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_2, x_3, x_4\}$, $\{x_1, x_2, x_3, x_5\}$, $\{x_1, x_2, x_4, x_5\}$, $\{x_1, x_3, x_4, x_5\}$, $\{x_2, x_3, x_4, x_5\}$. Since the lowest error of the available sets is 0.317, which is lower than 0.424, we update the current selected variables to $\{x_2, x_3, x_4, x_5\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_2, x_3, x_4, x_5\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, $\{x_1, x_3, x_4\}$, $\{x_2, x_3, x_4\}$, $\{x_1, x_2, x_5\}$, $\{x_1, x_3, x_5\}$, $\{x_2, x_3, x_5\}$, $\{x_1, x_4, x_5\}$, $\{x_2, x_4, x_5\}$, $\{x_3, x_4, x_5\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

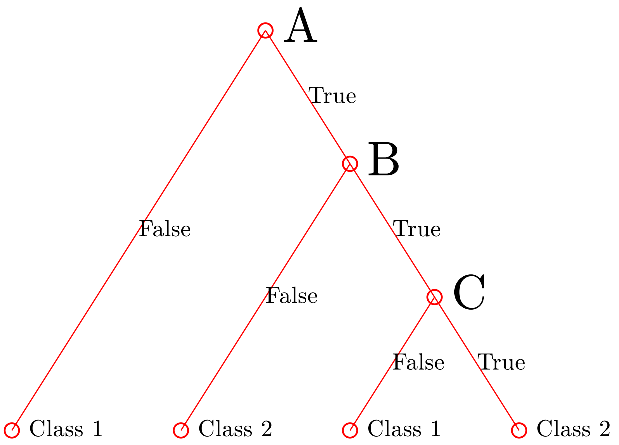

Figure 10: Example classification tree.

Question 22. We consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 10 resulting in a partition into classes indicated by the colors/markers in Figure 11. What is the correct rule assignment to the nodes in the decision tree?

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

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

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

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

E. Don't know.

Solution 22.

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.

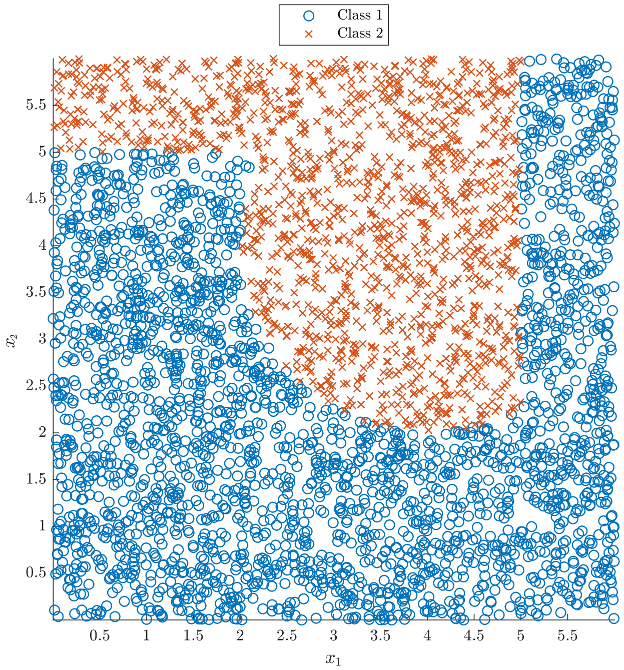

Figure 11: classification boundary.

The resulting decision boundaries for each of the options are shown in Figure 12 and it follows answer B is correct.

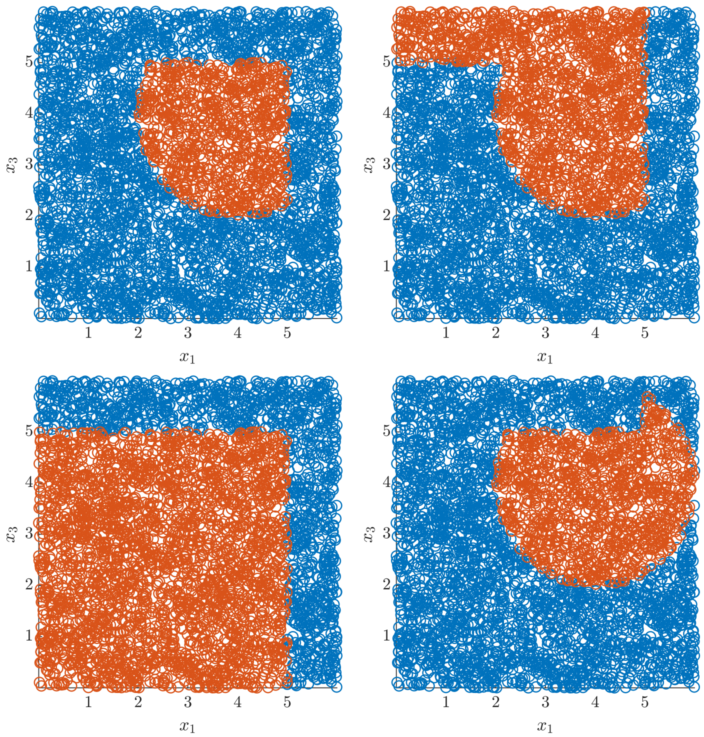

Figure 12: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

	ANN		Ridge reg.	
	n_h^*	E_1^{test}	λ^*	E_2^{test}
Outer fold 1	1	0.581	0.01	0.226
Outer fold 2	1	0.531	0.01	0.438
Outer fold 3	1	0.5	0.01	0.25
Outer fold 4	1	0.406	0.01	0.094
Outer fold 5	1	0.484	0.01	0.387

Table 8: Result of applying two-level cross-validation to a neural network model and a ridge 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.

a	$\nu = 1$	$\nu = 2$	$\nu = 3$	$\nu = 4$	$\nu = 5$
0.025	-12.706	-4.303	-3.182	-2.776	-2.571
0.05	-6.314	-2.92	-2.353	-2.132	-2.015
0.075	-4.165	-2.282	-1.924	-1.778	-1.699
0.1	-3.078	-1.886	-1.638	-1.533	-1.476
0.9	3.078	1.886	1.638	1.533	1.476
0.925	4.165	2.282	1.924	1.778	1.699
0.95	6.314	2.92	2.353	2.132	2.015
0.975	12.706	4.303	3.182	2.776	2.571

Table 9: Pre-computed values of the inverse student-t cumulative density function (CDF). Let $p_{\text{stud-t}}(x|\nu, \mu, \sigma)$ be the student-t distribution and $\text{cdf}^{-1}(a|\nu, \mu, \sigma) = \mu + \sigma \text{cdf}^{-1}(a|\nu, \mu = 0, \sigma = 1)$ the inverse CDF, then the table shows values of $\text{cdf}^{-1}(a|\nu, \mu = 0, \sigma = 1)$ for different values of ν (horizontally) and a (vertically).

Question 23. Suppose we wish to compare a neural network model and a regularized ridge regression model on the Cruise ship dataset. For the neural network, we wish to find the optimal number of hidden neurons n_h, and for the ridge regression model the optimal value of λ. We therefore opt for a two-level cross-validation approach where for each outer fold, we determine the optimal number of hidden units (or regularization strength) using an inner cross-validation loop with $K_2 = 3$ folds. The tested values are:

$$\lambda : \{0.01, 0.06, 0.32, 1.78, 10\}$$
$$n_h : \{1, 2, 3\}.$$

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

training time: 20 and testing time: 5

and the time taken to train/test a single ridge regression model is

training time: 8 and testing time: 1,

what is approximately the time taken to compose the table?

A. 1970.0 ms

B. 4080.0 ms

C. 7140.0 ms

D. 7480.0 ms

E. Don't know.

Solution 23. Going over the 2-level cross-validation algorithm we see the total number of models to be *trained* for a two-level cross-validation procedure is:

$$K_1(K_2 S + 1)$$

where S is the number of parameters to test. For the two models we have $S = 5$ and $S = 3$ respectively. We therefore see we have to train 50 and 80 models. As the number of models to test is the same as those trained, it is easy to multiply these numbers by the time consumed and see the answer is

$$50 \times 20 + 50 \times 5 + 80 \times 8 + 80 \times 1$$

and therefore option A is correct.

Question 24. Consider once more the two-level cross-validation results shown in Table 8. Suppose we wish to compare the two regression models by computing the confidence interval of the estimated performance difference. Assuming the assumptions of the test is satisfied, what is the $\alpha = 0.05$ confidence interval of this difference? (Hint: Table 9 shows values of the inverse cumulative density function of the student-t distribution, and for the test we have following the notation of the book $\hat{z} = 0.221$, and $\tilde{\sigma} = \sqrt{\frac{1}{K-1}\mathrm{Var}[z_i]} = 0.054$, which may be of help)

A. $[\theta_L, \theta_U] = [-0.15, 0.15]$

B. $[\theta_L, \theta_U] = [0.11, 0.34]$

C. $[\theta_L, \theta_U] = [-0.04, 0.48]$

D. $[\theta_L, \theta_U] = [0.07, 0.37]$

E. Don't know.

Solution 24. To compute the confidence interval, we follow the recipe in the lecture notes. First, we compute the 5 differences in performance between the generalization error estimated in the $K = 5$ folds: $z_i = E^{\text{test}}_{1,i} - E^{\text{test}}_{2,i}$. Given these, we compute:

$$\hat{z} = \mathbb{E}[z_i] = 0.221$$

and

$$\tilde{\sigma} = \sqrt{\frac{1}{K-1}\mathrm{Var}[z_i]} = 0.054.$$

The interval can then be found as, recalling $\nu = K - 1$:

$$\theta_L = \hat{z} + \tilde{\sigma}\mathrm{cdf}^{-1}(\frac{1}{2}\alpha|\nu, \mu = 0, \sigma = 1),$$

and

$$\theta_U = \hat{z} + \tilde{\sigma}\mathrm{cdf}^{-1}(1 - \frac{1}{2}\alpha|\nu, \mu = 0, \sigma = 1)$$

where the values for the inverse student-t distribution can be found in the table as -2.776 and 2.776 respectively. Plugging in the numbers we see option D is correct.

Variable	y^{true}	$t = 1$
y_1	0	1
y_2	1	1
y_3	1	1
y_4	0	0
y_5	0	0

Table 10: For each of the $N = 5$ observations (first column), the table indicate the true class labels y^{true} (second column) and the predicted outputs of the AdaBoost classifier (third column) for the first round of boosting $t = 1$.

Question 25. Consider again the Cruise ship dataset of Table 1. Suppose we limit ourselves to $N = 5$ observations from the original dataset and only consider the features x_1 and x_2. We use a KNN classification model ($K = 2$) to this dataset and apply AdaBoost to improve the performance. After the first round of boosting, we obtain predictions and the true class labels as tabulated in Table 10. Given this information, how will the AdaBoost update the weights $\boldsymbol{w}$?

A. $[0.136 \quad 0.216 \quad 0.216 \quad 0.216 \quad 0.216]$

B. $[0.500 \quad 0.125 \quad 0.125 \quad 0.125 \quad 0.125]$

C. $[0.900 \quad 0.025 \quad 0.025 \quad 0.025 \quad 0.025]$

D. $[0.016 \quad 0.246 \quad 0.246 \quad 0.246 \quad 0.246]$

E. Don't know.

Solution 25.

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

$$\{y_1\}$$

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

$$\epsilon_{t=1} = \sum_i w_i(t)(1 - \delta_{f_t(x_i), y_i}) = 0.2$$

From this, we compute α_t as

$$\alpha_t = \frac{1}{2}\log\frac{1 - \epsilon_t}{\epsilon_t} = 0.693$$

Scaling the observations corresponding to the misclassified 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 B.

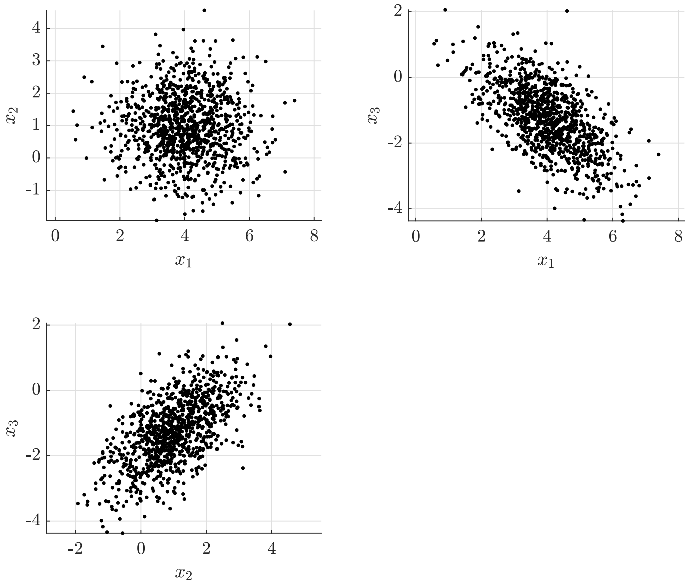

Figure 13: Scatter plot of all pairs of attributes of a vectors $\boldsymbol{x}$ drawn from a multivariate normal distribution of 3 dimensions.

Question 26. Consider a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$ 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 13. What is the most plausible covariance matrix?

A. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & 0.6 \\ 0.0 & 1.0 & -0.6 \\ 0.6 & -0.6 & 1.0 \end{bmatrix}$

B. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & -0.6 \\ 0.0 & 1.0 & 0.6 \\ -0.6 & 0.6 & 1.0 \end{bmatrix}$

C. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & -0.6 & 0.0 \\ -0.6 & 1.0 & 0.6 \\ 0.0 & 0.6 & 1.0 \end{bmatrix}$

D. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & -0.6 & 0.6 \\ -0.6 & 1.0 & 0.0 \\ 0.6 & 0.0 & 1.0 \end{bmatrix}$

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). This rules out all but option B.

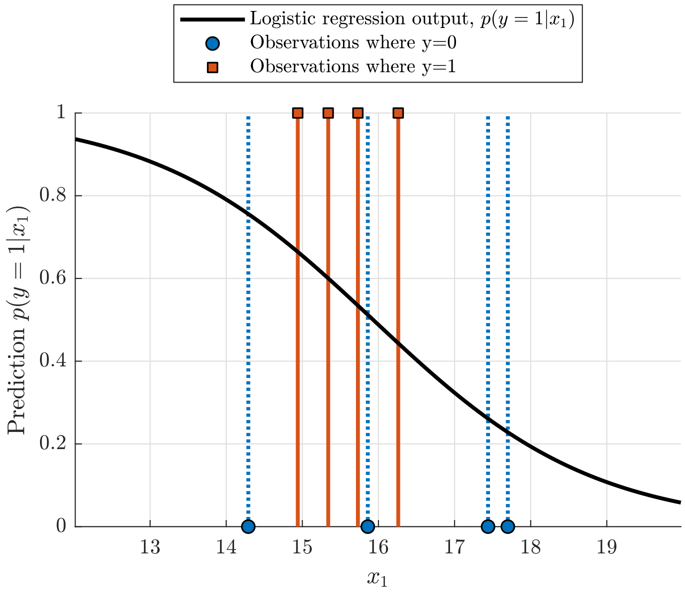

Figure 14: Output of a logistic regression classifier trained on 8 observations from the dataset.

Question 27. Consider again the Cruise ship dataset. To simplify the setup further, we select just 8 observations and train a logistic regression classifier using only the feature x_1 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term such that x_1 becomes the second coordinate). To be consistent with the lecture notes, we label the output as $y=0$ (corresponding to an old ship model) and $y=1$ (corresponding to a new ship model).

In Figure 14 is shown the predicted probability that an observation belongs to the positive class, $p(y=1|x_1)$. What are the weights?

A. $\boldsymbol{w} = \begin{bmatrix} 4.22 \\ 0.85 \end{bmatrix}$

B. $\boldsymbol{w} = \begin{bmatrix} 0.0 \\ -0.99 \end{bmatrix}$

C. $\boldsymbol{w} = \begin{bmatrix} 0.0 \\ -1.83 \end{bmatrix}$

D. $\boldsymbol{w} = \begin{bmatrix} 10.98 \\ -0.69 \end{bmatrix}$

E. Don't know.

Solution 27. The solution is easily found by simply computing the predicted $\hat{y} = p(y=1|x_1)$-value for an appropriate choice of x_1. Notice that

$$p(y=1|x_1) = \sigma(\tilde{\boldsymbol{x}}_1^T \boldsymbol{w})$$

If we select $x_1 = 16$ and select the weights as in option D we find $p(y=1|x_1) = 0.49$, in good agreement with the figure. On the other hand, for the weights in option A we obtain $\hat{y} = 1$, for C that $\hat{y} = 0$ and finally for B that $\hat{y} = 0$. We can therefore conclude that D is correct.