Technical University of Denmark

Written examination: 22nd May 2024, 9 AM — 1 PM.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted (closed internet).

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

This exam only allows for electronic hand-in.
You hand in your answers at `https://eksamen.dtu.dk/`. To hand in your answers, write them in the file `answers.txt` (this file is available from the same place you downloaded this file). When you are done, upload the `answers.txt` file (and nothing else). Double-check that you uploaded the correct version of the file from your computer.

Do not change the format of `answers.txt`
The file is automatically parsed after hand-in. Do not change the file format of `answers.txt` to any other format such as `rtf`, `docx`, or `pdf`. Do not change the file structure. Only edit the portions of the file indicated by question marks.

Answers:

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

	Attribute description	Abbrev.
x_1	Input voltage (Volt)	VIN
x_2	Input frequency (Hertz)	FIN
x_3	Amplification factor [0.10-2.30]	AMP
x_4	Power consumption (Watt)	PWR
x_5	Eletrocmagnetic radiance (Watt per square meter)	EMR
y	Protection mode	PROT
y_r	Total distortion and noise	THDN

Table 1: The Electronic Device Quality dataset considered in this exam contains 900 measurements of an electronic device for amplifying an input signal. The classification task concerns the prediction of whether the device's protection mode is Activated ($y = 2$) or Deactivated ($y = 1$) , while the regression task concerns the prediction of the total distortion and noise introduced by the amplification device, y_r.

	x_1	x_2	x_3	x_4	x_5
1 pct.	-1.19	-1.6	-1.76	-0.7	-1.21
5 pct.	-1.19	-1.36	-1.57	-0.7	-1.21
25 pct.	-1.11	-0.78	-0.84	-0.68	-0.87
50 pct.	-0.08	-0.13	0.03	-0.5	-0.23
75 pct.	1.13	0.65	0.89	0.31	0.7
95 pct.	1.21	1.77	1.48	2.43	1.93
99 pct.	1.21	2.59	1.48	3.21	2.25

Table 2: Percentiles for all attributes in the Electronic Device Quality dataset described in Table 1.

Question 1. We consider the standardized version of the Electronic Device Quality dataset described in Table 1, i.e., each attribute has a mean of zero and a standard deviation of one.

Figure 2 shows a histogram plot of two unidentified attributes A and B from the Electronic Device Quality dataset. In Table 2, specific percentiles of all the attributes are listed.

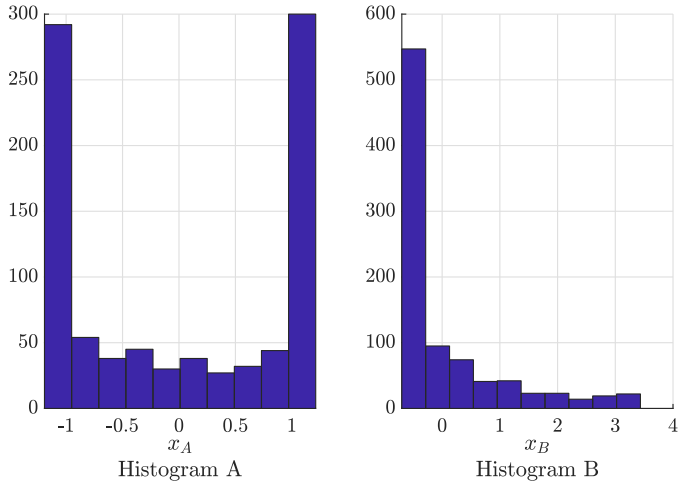

Figure 1: Histogram of two unidentified attributes (x_A and x_B) of the Electronic Device Quality dataset described in Table 1.

Which one of the following statements is true?

A. Histogram A corresponds to attribute x_3, and histogram B corresponds to attribute x_1.

B. Histogram A corresponds to attribute x_1, and histogram B corresponds to attribute x_4.

C. Histogram A corresponds to attribute x_4, and histogram B corresponds to attribute x_2.

D. Histogram A corresponds to attribute x_2, and histogram B corresponds to attribute x_1.

E. Don't know.

Solution 1. The correct answer is B.

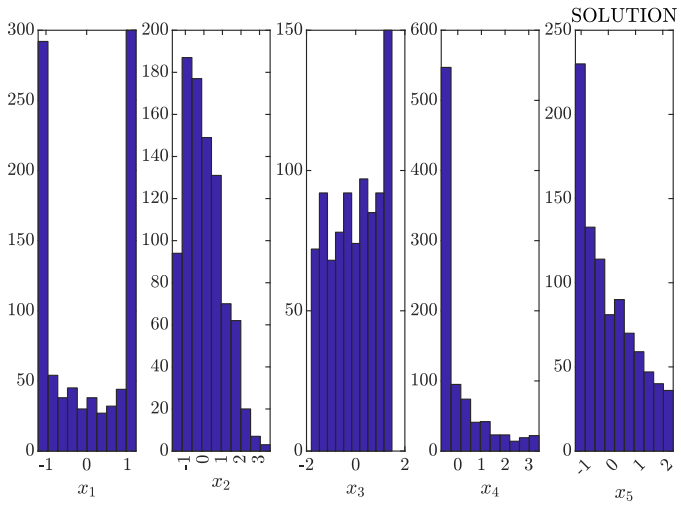

Figure 2: SOLUTION: Histogram of all attributes, ordered.

Question 2. Consider again the Electronic Device Quality dataset. The empirical covariance matrix of the first 4 attributes $x_1, \ldots, x_4$ is given by:

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} 1.0 & 0.036 & -0.077 & 0.643 \\ 0.036 & 1.0 & -0.03 & -0.029 \\ -0.077 & -0.03 & 1.0 & 0.459 \\ 0.643 & -0.029 & 0.459 & 1.0 \end{bmatrix}.$$

What is the empirical correlation of x_1 (VIN) and x_4 (PWR)?

A. ≈ 0.643

B. ≈ 0.8019

C. ≈ -0.8019

D. ≈ 0.4134

E. Don't know.

Solution 2. Recall the correlation is defined as

$$\text{cor}[x, y] = \frac{\text{cov}[x, y]}{\text{std}[x]\,\text{std}[y]}$$

Next, by definition the diagonal elements of the covariance matrix are estimates of the variance and the off-diagonal elements are estimates of the covariance, i.e. for $i \neq j$:

$$\hat{\Sigma}_{ii} = \text{Var}[x_i], \quad \hat{\Sigma}_{ij} = \text{cov}[x_i, x_j]$$

Therefore we get:

$$\text{cor}[x_i, y_j] = \frac{\hat{\Sigma}_{ij}}{\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}}}.$$

By simple insertion, we see option A is correct.

Question 3. Consider a PCA analysis of the Electronic Device Quality dataset in the $N \times M$ matrix, $\mathbf{X}$, where $N = 900$ is the number of observations and $M = 5$ is the number of attributes, as usual. We standardize the $\mathbf{X}$ matrix column-wise by subtracting the mean and dividing by the standard deviation to obtain $\tilde{\mathbf{X}}$. With this matrix, we perform principal component analysis via singular value decomposition, obtaining $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \tilde{\mathbf{X}}$. The three matrices are found to be:

$$\mathbf{U} = \begin{bmatrix} 0.000 & -0.041 & -0.032 & -0.045 & -0.003 \\ -0.048 & 0.044 & -0.030 & -0.016 & 0.058 \\ 0.010 & -0.055 & -0.026 & -0.027 & -0.054 \\ -0.017 & -0.039 & 0.013 & -0.021 & -0.037 \\ -0.065 & 0.018 & -0.030 & 0.038 & -0.016 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad (1)$$

$$\mathbf{\Sigma} = \begin{bmatrix} 49.164 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 31.401 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 29.678 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 12.436 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 7.837 \end{bmatrix} \quad (2)$$

$$\mathbf{V} = \begin{bmatrix} -0.47 & 0.52 & -0.24 & 0.54 & 0.40 \\ 0.00 & 0.44 & 0.90 & -0.04 & 0.02 \\ -0.31 & -0.73 & 0.37 & 0.41 & 0.27 \\ -0.58 & -0.04 & -0.01 & -0.73 & 0.36 \\ -0.60 & 0.01 & 0.01 & 0.08 & -0.80 \end{bmatrix} \quad (3)$$

With the SVD result in hand, the aim is to obtain a principal component representation with the fewest possible components retaining at least 90 % of the total variance.

For the first observation in $\tilde{\boldsymbol{X}}$, $\tilde{\boldsymbol{x}}_1 = \begin{bmatrix} 1.21 & 0.05 & -1.43 & -0.43 & -0.59 \end{bmatrix}^\top$, what would the coordinates, $\mathbf{b}_1$, be when $\tilde{\boldsymbol{x}}_1$ is projected onto the relevant subspace?

A. $\mathbf{b}_1^\top = [0.021, -1.298, -0.953, -0.560]$

B. Part of the first row in the matrix $\tilde{\mathbf{X}}\mathbf{V}$.

C. $\mathbf{b}_1^\top = [0, -0.048, 0.010]$

D. $\mathbf{b}_1^\top = [-0.466, -0.522, -0.243, -0.536]$

E. Don't know.

Solution 3. From Σ, we see that the cummulated variance explained by the components are 0.53, 0,75, 0.95, 0.98 and 1. Thus 3 components are sufficient (i.e. option A and D are clearly wrong). The principal component representation of $\mathbf{b_1}$ is available as the 3 first elements of the first row of matrices $\mathbf{U\Sigma}$ or $\tilde{\mathbf{X}}\mathbf{V}$.

Question 4. Consider again the PCA analysis for the Electronic Device Quality dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (3). Which one of the following statements is true?

A. An observation with a low value of x_1 **(VIN)**, a low value of x_3 **(AMP)**, a low value of x_4 **(PWR)**, and a high value of x_5 **(EMR)** will typically have a positive value of the projection onto principal component number 5.

B. An observation with a high value of x_1 **(VIN)**, a high value of x_3 **(AMP)**, a high value of x_4 **(PWR)**, and a high value of x_5 **(EMR)** will typically have a positive value of the projection onto principal component number 1.

C. An observation with a high value of x_1 **(VIN)**, a high value of x_2 **(FIN)**, and a low value of x_3 **(AMP)** will typically have a negative value of the projection onto principal component number 2.

D. An observation with a low value of x_1 (VIN), a high value of x_2 (FIN), and a high value of x_3 (AMP) will typically have a positive value of the projection onto principal component number 3.

E. Don't know.

Solution 4. The correct answer is D. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_3$ (i.e. column three of $\boldsymbol{V}$) is

$$b_3 = \boldsymbol{x}^\top \boldsymbol{v}_3 = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \begin{bmatrix} -0.24 \\ 0.9 \\ 0.37 \\ -0.01 \\ 0.01 \end{bmatrix}$$

(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 positive, this occurs if x_1, x_2, x_3 has large magnitude and the sign convention given in option D.

Question 5. Consider the application of Kernel Density Estimation (KDE) to a synthetic 1D problem based on three observed data points $\boldsymbol{X} = \{-0.7, 1.1, 1.8\}$. Two possible kernel widths $\lambda = \{0.5, 1\}$ are considered with the final choice to be made based on leave-one-out cross-validation on $\boldsymbol{X}$.

Determine which one of the following test points is *most* likely to be a potential outlier using the KDE approach.

A. $x^* = -0.9$

B. $x^* = -0.3$

C. $x^* = 0.6$

D. $x^* = 1.2$

E. Don't know.

Solution 5. First, we need to find the optimal σ using LOOCV for the given dataset $\boldsymbol{X}$. Therefore, we estimate the density for each point x_i of the training set using the KDE as

$$p(x_i) = \frac{1}{N-1}\sum_{j\neq i}\mathcal{N}(x_i|x_j,\sigma^2)$$
$$= \frac{1}{N-1}\sum_{j\neq i}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-x_j)^2}{2\sigma^2}\right)$$

and then the likelihood as $\frac{1}{N}\sum_{i=1}^{N} p(x_i)$.

For the point $x_1 = -0.7$ the density is estimated as $p(x_1) = \frac{1}{2}(\mathcal{N}(x_1|x_2,\sigma_1^2) + \mathcal{N}(x_1|x_3,\sigma_1^2)) \approx 0.15$
For the point $x_2 = 1.1$ the density is estimated as $p(x_2) = \frac{1}{2}(\mathcal{N}(x_2|x_1,\sigma_1^2) + \mathcal{N}(x_2|x_3,\sigma_1^2)) \approx 0.15$
For the point $x_3 = 1.8$ the density is estimated as $p(x_3) = \frac{1}{2}(\mathcal{N}(x_3|x_1,\sigma_1^2) + \mathcal{N}(x_3|x_2,\sigma_1^2)) \approx 0$

This gives an estimation for the likelihood $\frac{1}{N}\sum_{n=1}^{N} p(x_n) \approx 0.1$ given $\sigma_1 = 0.5$. Similarly we estimate the likelihood for $\sigma_2 = 1$.

For the point $x_1 = -0.7$ the density is estimated as $p(x_1) = \frac{1}{2}(\mathcal{N}(x_1|x_2,\sigma_2^2) + \mathcal{N}(x_1|x_3,\sigma_2^2)) \approx 0.2$
For the point $x_2 = 1.1$ the density is estimated as $p(x_1) = \frac{1}{2}(\mathcal{N}(x_2|x_1,\sigma_2^2) + \mathcal{N}(x_2|x_3,\sigma_2^2)) \approx 0.16$
For the point $x_3 = 1.8$ the density is estimated as $p(x_1) = \frac{1}{2}(\mathcal{N}(x_3|x_1,\sigma_2^2) + \mathcal{N}(x_3|x_2,\sigma_2^2)) \approx 0.05$

which gives an estimation for the likelihood 0.14. Therefore, the optimal $sigma_* = 1$, and using this one we can estimate the density for the candidates: $p(x = -0.9) = 0.15$, $p(x = -0.3) = 0.19$, $p(x = 0.6) = 0.24$, $p(x = 1.2) = 0.27$. So the potential outlier is $x = -0.9$.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	3.7	3.4	2.3	1.8	1.6	2.4	2.6	2.5	4.3
o_2	3.7	0.0	3.2	5.1	3.5	2.9	4.4	4.6	4.0	7.2
o_3	3.4	3.2	0.0	4.1	2.6	4.0	2.7	5.3	2.9	6.1
o_4	2.3	5.1	4.1	0.0	3.7	3.2	1.7	2.7	1.9	2.6
o_5	1.8	3.5	2.6	3.7	0.0	2.8	3.2	4.4	3.4	5.6
o_6	1.6	2.9	4.0	3.2	2.8	0.0	3.4	2.1	3.1	5.3
o_7	2.4	4.4	2.7	1.7	3.2	3.4	0.0	3.8	0.8	4.1
o_8	2.6	4.6	5.3	2.7	4.4	2.1	3.8	0.0	3.5	4.0
o_9	2.5	4.0	2.9	1.9	3.4	3.1	0.8	3.5	0.0	4.4
o_{10}	4.3	7.2	6.1	2.6	5.6	5.3	4.1	4.0	4.4	0.0

Table 3: 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 Electronic Device Quality 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 $\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$ belong to class C_1 (corresponding to Deactivated), and the red observations $\{o_8, o_9, o_{10}\}$ belong to class C_2 (corresponding to Activated). To avoid single features to dominate, the dataset was standardized by subtracting the mean and dividing by the standard deviation.

Question 6. To examine if observation o_1 may be an outlier, we will calculate the K-nearest neighborhood density using only the observations and distances in Table 3. For an observation o_i, recall the density is computed using the set of K nearest neighbors of observation o_i excluding the i'th observation itself, $N_{\boldsymbol{X}_{\setminus i}}(o_i, K)$, and is denoted by $\text{density}_{\boldsymbol{X}_{\setminus i}}(o_i, K)$. What is the density for observation o_1 for $K = 3$ nearest neighbors?

A. 0.625

B. 0.462

C. 1.139

D. 0.526

E. Don't know.

Solution 6.

The density is given as:

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

So to solve the problem, we only need to plug in the values. We find that the $k = 3$ neighborhood of o_1 and density is:

$$N_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = \{o_6, o_5, o_4\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = 0.526$$

Therefore option D is correct.

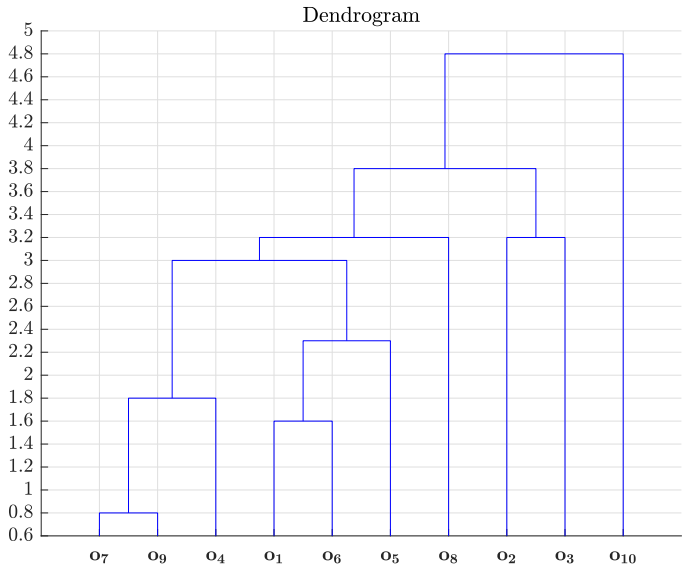

Figure 3: Dendrogram.

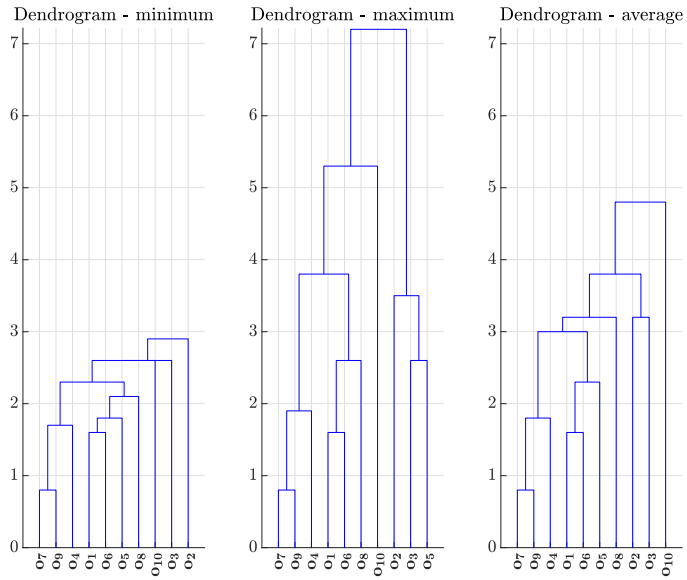

Figure 4: Dendrogram.

Question 7. Consider the observations and the pairwise distances in Table 3 along with the dendrogram in Figure 3.

Determine which linkage function was used to produce the dendrogram.

A. Minimum linkage.

B. Average linkage.

C. Maximum linkage.

D. The linkage function can not be determined from the available information.

E. Don't know.

Solution 7. The correct answer is Average linkage.

The solution is found easiest by considering the heigth at which a group of 1 and a group merges (all linkage functions will merge singletons at the same height.)

All the options are illustrated in **??**.

Question 8. Consider a Gaussian mixture model (GMM) with three components that define the density from which we draw random samples to create a dataset. We compute the principal components of this dataset. The corresponding directions are shown in Fig. 5. Which one of the following GMMs was used to generate the data?

A.

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4.9 \\ -2.6 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.47 & 1.3 \\ 1.3 & 1.54 \end{bmatrix} \quad w_1 = 0.34$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} -0.04 \\ 2.6 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 5.3 & -0.1 \\ -0.1 & 0.32 \end{bmatrix} \quad w_2 = 0.12$$
$$\boldsymbol{\mu}_3 = \begin{bmatrix} -1.17 \\ -1.13 \end{bmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.84 & 0.89 \\ 0.89 & 4.4 \end{bmatrix} \quad w_3 = 0.54$$

B.

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4.9 \\ -2.6 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.84 & 0.89 \\ 0.89 & 4.4 \end{bmatrix} \quad w_1 = 0.12$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} -0.04 \\ 2.6 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.47 & 1.3 \\ 1.3 & 1.54 \end{bmatrix} \quad w_2 = 0.34$$
$$\boldsymbol{\mu}_3 = \begin{bmatrix} -1.17 \\ -1.13 \end{bmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 5.3 & -0.1 \\ -0.1 & 0.32 \end{bmatrix} \quad w_3 = 0.54$$

C.

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4.9 \\ -2.6 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 5.3 & -0.1 \\ -0.1 & 0.32 \end{bmatrix} \quad w_1 = 0.34$$
$$\boldsymbol{\mu}_2 = \begin{bmatrix} -0.04 \\ 2.6 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.47 & 1.3 \\ 1.3 & 1.54 \end{bmatrix} \quad w_2 = 0.12$$
$$\boldsymbol{\mu}_3 = \begin{bmatrix} -1.17 \\ -1.13 \end{bmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.84 & 0.89 \\ 0.89 & 4.4 \end{bmatrix} \quad w_3 = 0.54$$

D.

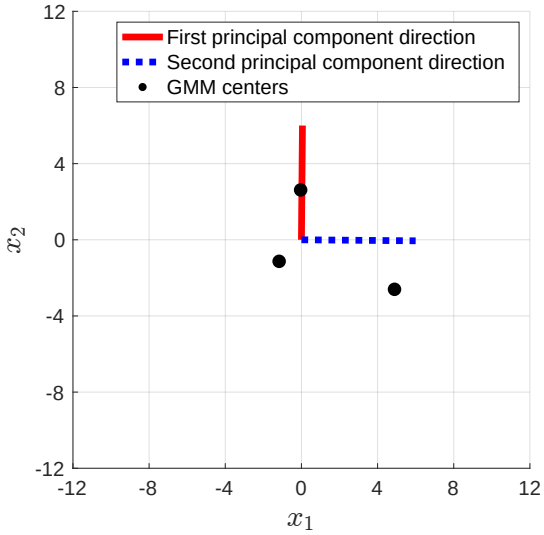

Figure 5: The two principal component directions illustrated along with the centers of the components in the Gaussian mixture model. The lines indicating the principal component directions have arbitrary lengths (for illustrative purposes).

Solution 8. In Fig. 6 we see that only the Option 4 generates a dataset for which the PCA result is similar to the desired one.

Note that it is not necessary to generate datasets from the associated GMMs, and then compute the PCA to find the correct answer. We can just observe that the GMM of Option 4 generates more data ($\sim$ 88%) with components 2 and 3 along the vertical axis which implies that the first principal component will capture this variance.

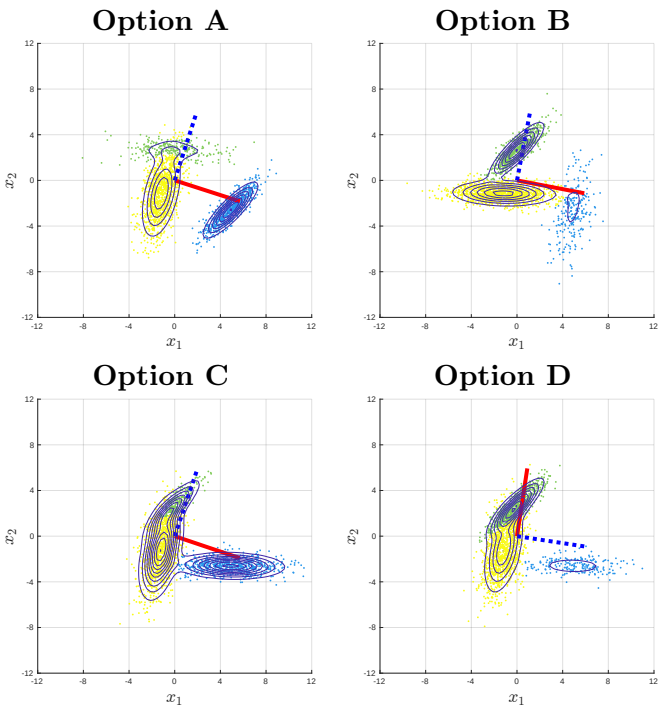

Figure 6: The densities of the Gaussian mixture models together with generated samples and the PCA results.

Question 9. Consider again the Electronic Device Dataset. Suppose we wish to predict the class label y using a decision tree model. To improve performance, we apply AdaBoost (the specific variant described in Chapter 17 of the lecture notes). We will use the full dataset for training, and it will be denoted by $\mathcal{D}$. We run $T = 3$ boosting rounds, and suppose that the algorithm proceeds as follows:

- The first classifier, f_1, has a weighted error, $\epsilon_{t=1}$, of 0.50 when evaluated on $\mathcal{D}$.
- The second classifier, f_2, has a weighted error, $\epsilon_{t=2}$, of 0.75 when evaluated on $\mathcal{D}$.
- The third classifier, f_3, has a weighted error, $\epsilon_{t=3}$, of 0.50 when evaluated on $\mathcal{D}$.

What is the *accuracy* of the combined classifier (majority voting classifier) produced by AdaBoost when evaluated on $\mathcal{D}$?

A. ≈ 0.25

B. ≈ 0.50

C. ≈ 0.63

D. ≈ 0.75

E. Don't know.

Solution 9. AdaBoost starts with uniform weights $w_i(1) = \frac{1}{N}$ and therefore the weighted error on all data for f_1 is $\epsilon_1 = 1 - 0.5 = 0.5$. This means that the importance of classifier f_1 is $\alpha_1 = \frac{1}{2}\log\frac{1-0.5}{0.5} = 0$ and that the weights for the second boosting round are also uniform, as $w_i(1)e^{-\alpha_1} = w_i(1)e^{-\alpha_1} = w_i(1)$.

With uniform weights in the second round, we find that the weighted error on all data for f_2 is $\epsilon_2 = 1 - 0.75 = 0.25$. The importance of classifier f_2 is $\alpha_2 = \frac{1}{2}\log\frac{1-0.75}{0.75} \approx 0.55$.

The majority voting classifier of AdaBoost (Algorithm 7 in the book) is defined as follows

$$f^*(\boldsymbol{x}) = \arg\max_{y=1,2} \sum_{t=1}^{T} \alpha_t \delta_{f_t(\boldsymbol{x}),y} \, .$$

The first and third classifier has importance $\alpha_1 = 0$ and $\alpha_3 = 0$, which means they will not contribute to the sum over classifies. Therefore, we can write the classifies as

$$f^*(\boldsymbol{x}) = \arg\max_{y=1,2} \alpha_2 \delta_{f_2(\boldsymbol{x}),y} \, .$$

i.e. we will obtain a accuracy of 0.75 provided by f_2.

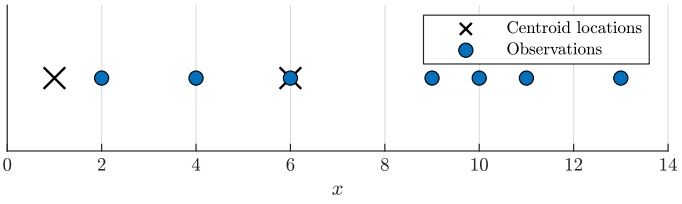

Figure 7: A small 1-dimensional dataset and initial values of centroids.

Question 10. Consider a small dataset comprised of $N = 7$ one-dimensional observations shown as the filled circles in Figure 7.

Suppose a k-means algorithm is applied to the dataset with $K = 2$ and using Euclidean distances. We will assume the location of the centroids are initialized to the values indicated by the crosses in Figure 7. After initialization, the k-means algorithm is evaluated for one step, comprised of assigning observations to centroids and updating the location of the centroids. After the first step, what will be the new location of the centroids?

A. $\mu_1 = 3$, and $\mu_2 = \frac{49}{5}$.

B. $\mu_1 = 4$, and $\mu_2 = \frac{43}{4}$.

C. $\mu_1 = 2$, and $\mu_2 = \frac{53}{6}$.

D. $\mu_1 = \frac{22}{3}$, and $\mu_2 = 11$.

E. Don't know.

Solution 10. The location of the observations and centroids is first read from Figure 7. When this is done, each observation is assigned to the nearest centroid. The observations are thereby partitioned into the two clusters $\{2\}$, $\{4, 6, 9, 10, 11, 13\}$.

The new location of the centroids are simply the mean of the observation in each of these two sets. Doing this, we see C is the correct answer.

Question 11. Consider a regularized linear regression model trained using $\lambda = 2.23$ on the Electronic Device Quality dataset using the five attributes in addition to second-order polynomial terms as input, i.e.

$$\tilde{\mathbf{x}} = [1, x_1, x_2, \ldots, x_5, x_1^2, x_2^2, \ldots, x_5^2].$$

The model performs poorly on a held-out test set and has been confirmed to have very high bias and very low variance.

Which one of the following actions is most likely to improve the generalization performance on unseen data?

A. Increasing the amount of regularization applied to the model.

B. Remove the second-order polynomial transformations from the input vector.

C. Decrease the amount of regularization applied to the model.

D. Train on the original training dataset with $\lambda = 2.23$ but only test on half the observations in the test set.

E. Don't know.

Solution 11. Correct answer is C as the model seem to underfit, i.e., it is too simple.
A: Increasing the amount of regularization applied to the model: This will lead even more bias and lower variance B: Remove the second order polynomial transformations from the input vector.: Reduces complexity C: Decrease the amount of regularization applied to the model: This will lead less bias and higher variance, i.e. the correct thing to do. variance D: Keep the orginal training dataset but only test on half the observations in the test set: The model's generalizatione error is not influenced by the size of the test set (only the estimate of it). will stil

Question 12. Suppose Hunt's algorithm is used to compute the purity gain for a decision tree (DT) classifier trained on the Electronic Device Quality dataset using a hold-out cross-validation procedure. The purity gain is defined as $\Delta = I(r) - \sum_{k=1}^{K} \frac{N(v_k)}{N(r)} I(v_k)$, where $I(z)$ is the impurity measure evaluated at the root, r, or a specific branch, v_k.

At the root, the training set contains 400 examples from class 1 and 100 from class 2. A particular split along x_4 results in two branches such that:

- Branch v_1 has 175 observations from class 1 and 75 observations from class 2.
- Branch v_2 has 225 observations from class 1 and 25 observations from class 2.

Determine which one of the following expressions correctly computes the purity gain for the provided split when using the *entropy* impurity measure.

A. $\Delta = -\sum_{c=1}^{2} p(c \mid r) \log_2 p(c \mid r) - \sum_{c=1}^{2} p(c \mid v_1) \log_2 p(c \mid v_1)$

B. $\Delta = \sum_{c=1}^{2} \left(-p(c \mid r) \log_2 p(c \mid r) + \frac{1}{2} p(c \mid v_1) \log_2 p(c \mid v_1) + \frac{1}{2} p(c \mid v_2) \log_2 p(c \mid v_2)\right)$

C. $\Delta = -\sum_{c=1}^{2} p(c \mid r) \log_2 p(c \mid r) + \sum_{k=1}^{2} \sum_{c=1}^{2} p(c \mid v_k) \log_2 p(c \mid v_k)$

D. $\Delta = -\sum_{k=1}^{2} \sum_{c=1}^{2} p(c \mid v_k) \log_2 p(c \mid v_k)$

E. Don't know.

Solution 12. The correct answer is B.

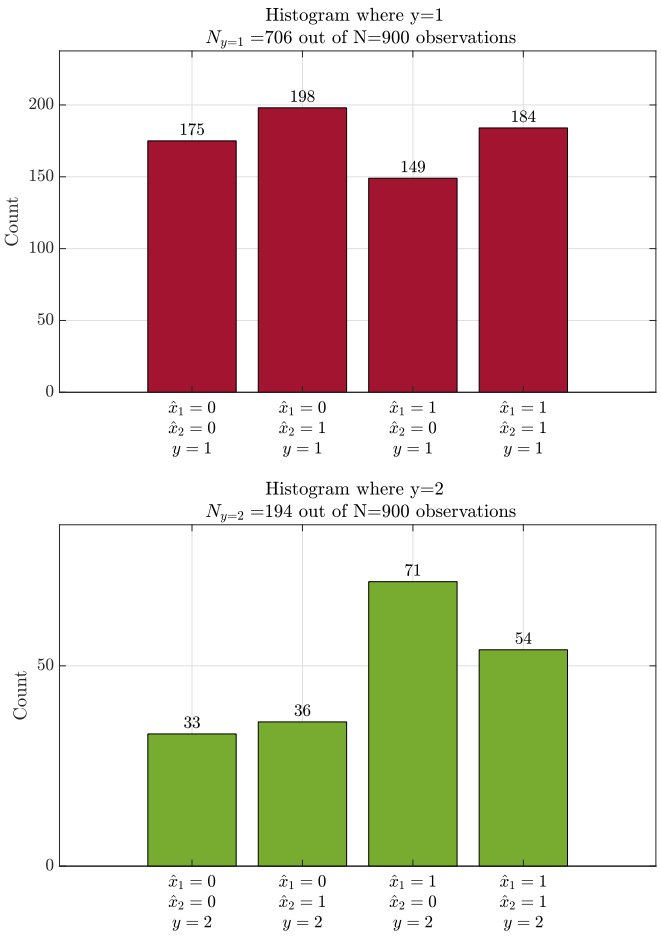

Figure 8: Histogram of the binarized version of the Electronic Device Quality dataset described in Table 1.

Question 13. Consider the Electronic Device Quality dataset from Table 1. The attributes have been binarized to produce new features denoted $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_M$. A summary of the binarised dataset is given in Figure 8 for a subset of the (binarized) attributes and label, namely $\hat{x}_1$, $\hat{x}_2$ and y.

Determine the probability of $y = 1$ given an observation has $\hat{x}_1 = 0$ and $\hat{x}_2 = 1$ using Bayes theorem (without applying the Naïve Bayes assumption).

A. 0.846

B. 0.664

C. 0.714

D. 0.28

E. Don't know.

Solution 13. The correct answer is A.

A: The key aspect of the question is to map from the counts to a joint (or conditional distribution). The answer is found using $p(y = 1|x_1 = 0, x_2 = 1) = p(x_1, x_2, y)/p(x_1, x_2)$, where $p(x_1 = 1, x_2 = 0, y = 1) = 0.220$ is found by enumerating the counts from the table (and normalizing by N). $p(x_1, x_2) = 0.260$ is found by marginalizing out y from the joint.
C: Incorrect, usign wrong joint (wrong y)
B: Incorrect, use joint and prior in nummerator
D: Incorrect, using just the conditional $p(x1, x2|y)$

Question 14. Consider clustering $N = 10$ observations using two clustering algorithms. The resulting cluster assignments is indicated by:

$$Q = [? \quad ? \quad ? \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1]$$
$$Z = [? \quad ? \quad ? \quad 2 \quad 3 \quad 3 \quad 4 \quad 2 \quad 1 \quad 1]$$

The elements in the vectors indicate the cluster index for a specific observation with "?" representing a missing/unknown cluster assignment.

We use the *Rand index* to compare the two results. It is known that the total number of times Q and Z agree that two observations are *not* in the same cluster is $D = 20$. Additionally, it is known that the total number of times Q and Z agree that two observations are in the same cluster is $S = 7$. Which one of the following options is the correct combination of missing assignments?

A.

$$Q = [3 \quad 1 \quad 3 \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1]$$
$$Z = [1 \quad 3 \quad 1 \quad 2 \quad 3 \quad 3 \quad 4 \quad 2 \quad 1 \quad 1]$$

B.

$$Q = [1 \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1]$$
$$Z = [4 \quad 1 \quad 3 \quad 2 \quad 3 \quad 3 \quad 4 \quad 2 \quad 1 \quad 1]$$

C.

$$Q = [1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1]$$
$$Z = [2 \quad 3 \quad 2 \quad 2 \quad 3 \quad 3 \quad 4 \quad 2 \quad 1 \quad 1]$$

D.

$$Q = [1 \quad 2 \quad 2 \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 3 \quad 1]$$
$$Z = [1 \quad 3 \quad 4 \quad 2 \quad 3 \quad 3 \quad 4 \quad 2 \quad 1 \quad 1]$$

E. Don't know.

Solution 14. Using the Rand Index formula we can compute the $S = R(A, B) \cdot 0.5 \cdot N \cdot (N-1) - D = 7$. Or alternatively if S is given instead of R, that $R(A, B) = \frac{S+D}{0.5 \cdot N \cdot (N-1)} = 0.60$

So we need to find which combination gives a matrix $\boldsymbol{n}$ that results to $S = 7$. Since a part in each assignment remains constant, we can pre-compute the matrix $\boldsymbol{n}' = \begin{bmatrix} 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$, which remains constant in all cases, and hence, we only need to add the residual information using the missing assignments.

Option A: $\boldsymbol{n}' + \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \end{bmatrix} \Rightarrow S = 5$ Wrong

Option B: $\boldsymbol{n}' + \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow S = 3$ Wrong

Option C: $\boldsymbol{n}' + \begin{bmatrix} 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow S = 7$ Correct

Option D: $\boldsymbol{n}' + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow S = 4$ Wrong

i	z_i^A	z_i^B	z_i^C
1	17	21	17
2	21	21	22
3	19	23	18
4	20	21	20
5	22	22	53

Table 4: Squared error from three regression models for $n = 5$ observations.

Question 15. Consider comparing three regression models (A, B, and C) using hold-out cross-validation and the setup presented in Section 11.3.6 of the lecture notes. The squared error on each of the five observations for all three models is shown in Table 4.

Determine which one of the following statements is correct.

A. Model A's losses are always lower or equal to Model B's losses, thus a paired test on the difference in performance between the models is guaranteed to show that that Model A is better than Model B with a p-value < 0.05.

B. The confidence interval for the estimated difference in performance between Model A and Model C is $\approx [17.0, 53.0]$

C. The losses of Model C seems far from being normally distributed due to the outlier ($z_5^C = 53$), so one should be careful with any application and interpretation of the usual paired test.

D. When comparing Model A's and Model B's losses using a paired test, the degrees of freedom parameter ν in the test will be $\nu = (n-1) \times n = 20$.

E. Don't know.

Solution 15. The solution is C.
A: Even though the losses are lower, the test will not necessarily show a low p-value. That depends on the number of data points and the variance of the data. In the case here the p-value is ≈ 0.12.
B: The CI is in the other direction.
C: The outlier make the data far from normally distributed. Furthermore, there are very few data points.
D: For a paired test there are only $N = 5 - 1 = 4$ degrees of freedom.

i	1	2	3	4	5
$y_{r,i} - f(\hat{\mathbf{x}}_i, \hat{\mathbf{w}})$	0.00162	-0.00756	-0.0161	0.0248	-0.00274

Table 5: Regression result.

Question 16. Consider a regularized linear regression model on the (non-standardized) attribute, y_r, using three of the standardized attributes x_3, x_4, x_5. We use a standard regression function defined as $f(\hat{\mathbf{x}}, \hat{\mathbf{w}}) = \hat{\mathbf{x}}^\top \hat{\mathbf{w}}$, where $\hat{\mathbf{w}}$ is the weight vector including the intercept term as the first element and $\hat{\mathbf{x}}^\top$ contains the standardized attributes augmented with a constant for the intercept term, i.e., $\hat{\mathbf{x}} = [1, \tilde{\mathbf{x}}]^\top$.

Consider an error function for the regularized linear regression problem defined as

$$E(\hat{\mathbf{w}}) = \frac{1}{N} \sum_{i=1}^{N} L(y_{r,i}, f(\hat{\mathbf{x}}_i, \hat{\mathbf{w}})) + \lambda \|\mathbf{w}\|_2^2 ,$$

where the function $L(y_{r,i}, \hat{y}_{r,i})$ is chosen to be the *squared* difference between the actual observation, $y_{r,i}$, and the prediction, $\hat{y}_{r,i}$. $\mathbf{w}$ is the weight vector *excluding* the intercept term.

The model parameters are estimated based on five observations with the relevant information shown in Table 5, resulting in $\hat{\mathbf{w}} = [0.0318, 0.04, 0.000197, w_5]^\top$ and $E(\hat{\mathbf{w}}) \approx 0.00025757$. $\lambda = 0.02$ was used in the estimation.

Determine which value of w_5 was estimated.

A. $w_5 \approx -0.043$

B. $w_5 \approx 0.018$

C. $w_5 \approx 0.22$

D. $w_5 \approx 0.60$

E. Don't know.

Solution 16. The correct answer is A.

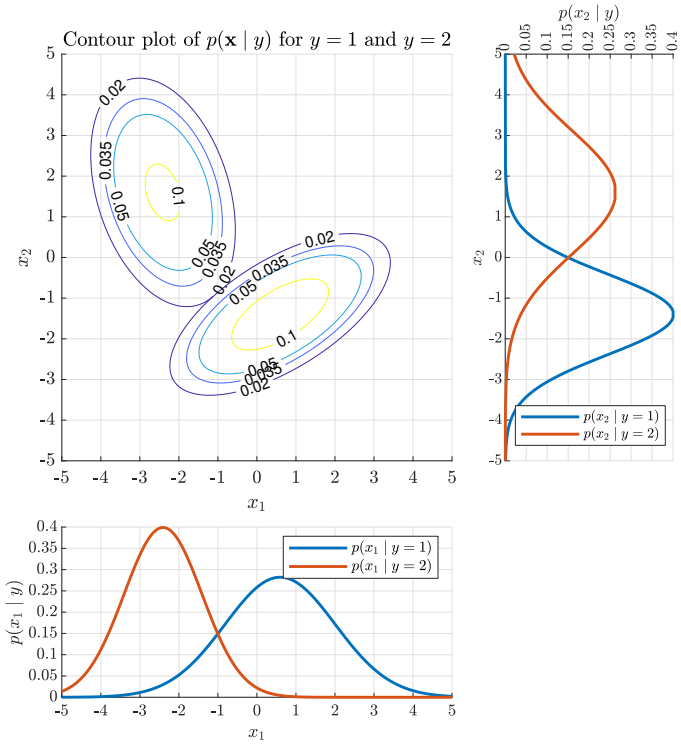

Figure 9: Conditional distributions related to a Naïve Bayes classifier.

Question 17. Consider the task of classifying the label y from two continuous attributes, namely x_1 and x_2 using Bayes rule.

We use a normal distribution to model the class conditional distribution of $\mathbf{x} = [x_1, x_2]^\top$, i.e. in general $p(\mathbf{x} \mid \boldsymbol{\mu}_{y=k}, \boldsymbol{\Sigma}_{y=k})$.

The specific distributions are illustrated in Figure 9. Additionally, it is noted that $p(y = 1) = 0.784$ and $p(y = 2) = 0.216$.

Determine the value of $p(y = 1 \mid \mathbf{x} = [-1, 0]^\top)$ using a Naïve Bayes classifier.

A. ≈ 0.675

B. ≈ 0.784

C. ≈ 0.864

D. ≈ 0.5

E. Don't know.

Solution 17. Recall the formula for Naïve-Bayes is

$$p(y|\boldsymbol{x}) = \frac{p(y)\prod_{k=1}^{M} p(x_k|y)}{\sum_{y'} p(y')\prod_{k=1}^{M} p(x_k|y')}.$$

Thus, we focus on on $p(y \mid x)$ in Figure 9.

We note that $p(y = 1\boldsymbol{x}_1 = -1)$ =p(y=2 $\boldsymbol{x}_1 = -1$) =p(y=1 $\boldsymbol{x}_2 = 0$) =p(y=1 $\boldsymbol{x}_2 = 0$) ≈ 0.15

I.e. plugging into the equation reveils that the result is given by the prior, $p(y = 1)$ as all other terms cancels.

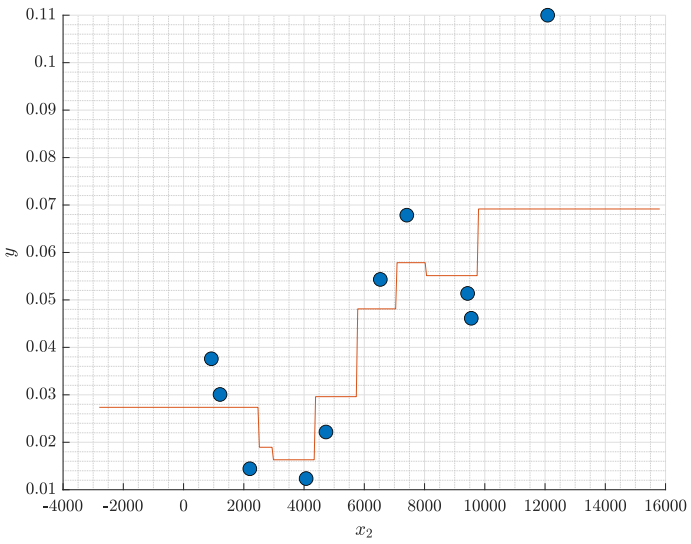

Figure 10: KNN regression model in which the red line is fitted to a small 1-dimensional dataset.

Question 18. Suppose a K-nearest neighbors regression model is fitted to a small 1-dimensional dataset with $N = 10$ observations. The predicted response is shown in Figure 10. How many neighbors (i.e. K) was used?

A. $K = 5$

B. $K = 2$

C. $K = 3$

D. $K = 4$

E. Don't know.

Solution 18.

The problem could be solved by using the definition of the KNN regression model and test various points, but it is much quicker solved using an intuitive argument. The KNN regression model consist of a series of steps, and the important information is where the discontinuities occur. If $K = 1$, the y-value has to pass through the training observations. On the other hand, if we consider the y-value at the right-most end of the x-axis, we note it is quite large consistent with it being computed using the two left-most observations, but not consisting with including the third observation from the right (or additional observations). Hence, we conclude that $K = 2$.

Question 19. Consider again the Electronic Device Quality dataset in Table 1. We apply forward or backward selection to find an interpretable linear regression model which uses a subset of the $M = 5$ attributes to predict the y_r attribute.

Recall that backward and forward selection chooses models based on the test error as determined by cross-validation. In our case we use the hold-out method to generate a single test/training split.

Suppose backward selection ends up selecting the attributes x_1 and x_5.

Determine which one of the following statements is correct.

A. Forward selection is garanteed to also select x_1 and x_5.

B. Backward selection will always result in fewer attributes being selected compared to forward selection.

C. Backward selection requires *testing* 13 models if the procedure selects x_1 and x_5.

D. Forward selection requires *training* 13 models if the procedure selects two attributes.

E. Don't know

Solution 19. The correct answer is D.
Forward: $1 + M + (M-1) + (M-2) = 13$
Backward: $1 + M + (M-1) + (M-2) + (M-3) = 15$

Question 20. Suppose we fit three different logistic regression models (A, B and C) on the Electronic Device Quality dataset.

Model A uses x_1.

Model B uses x_1 and x_2.

Model C uses all five attributes $x_1, \ldots, x_5$.

We standardize the input attributes and add an intercept column with ones in front of the attribute columns.

The estimated parameters/weigths are:

$$\mathbf{w}^A = [-1.332, 0.378]^\top$$

$$\mathbf{w}^B = [-1.346, 0.387, -0.222]^\top$$

$$\mathbf{w}^C = [-15.913, 9.028, -0.273, 18.852, 2.133, -5.583]^\top$$

where the first element in each vector corresponds to the intercept/bias term.

We consider the (already standardized) observation $\boldsymbol{x}_3 = [1.209, -0.303, -1.070, -0.459, 0.013]^\top$ and the prediction $\hat{y}_i = \sigma(\mathbf{x}_i^\top \mathbf{w})$ where $\sigma(z)$ is the logistic sigmoid function.

Determine which one of the following statements is correct.

A. Model A's prediction is $\hat{y}_3^A \approx -0.87$

B. Model A's prediction is $\hat{y}_3^A \approx 0.15$

C. Model B's prediction is $\hat{y}_3^B \approx 0.31$

D. It is not possible to make a prediction using $\mathbf{w}^C$ as the dimension of $\mathbf{w}^C$ does not match the input.

E. Don't know.

Solution 20.
$\hat{y}_3^B = \sigma(-1.346 \cdot 1 + 0.387 \cdot 1.209 + (-0.222) \cdot (-0.303)) \approx 0.31$.

DT		Predicted	
		Activated	Deactivated
Actual	Activated	373	110
	Deactivated	82	335

ANN		Predicted	
		Activated	Deactivated
Actual	Activated	355	128
	Deactivated	64	353

Table 6: Confusion matrices.

Question 21. Table 6 shows the confusion matrices related to a neural network (ANN) classifier and a decision tree (DT) classifier trained on a particular instance of the Electronic Device Quality dataset to predict y (PROT). The dataset has been resampled to almost balance the number of observations from the two classes, but the number of observations is still $N = 900$.

We define $y = 1$ to mean a *negative* outcome (i.e., PROT is Deactivated), and $y = 2$ to be a *positive* outcome (i.e., PROT is Activated).

On the particular instance of the dataset, the goal is to have low error rate and simultaneously avoid misclassifying a device as having its protection mode Deactivated when it is actually Activated.

Determine which one of the following statements is true.

A. It is not possible to determine which classification model is better based on the provided information.

B. The DT is preferred over the ANN because the DT's precision is higher than the ANN's precision.

C. The DT is preferred over the ANN because the DT's false positive rate is lower than the ANN's false positive rate.

D. The DT is preferred over the ANN because the DT's recall is higher than the ANN's recall.

E. Don't know.

Solution 21. The correct answer is D.

The accuracy of the two models is the same, i.e., we can not differentiate based on this; however, it is possible to differentiate based on the recall which

is the metric we are looking to maksimise here since it involoves minimising FNs — the number of times we incorrectly predict Deactivated while actually Activated.

For DT we have:

$ACC = (TP + TN)/(TP + FN + FP + TN) = 0.79$

$FPR = FP/(FP + TN) = 0.772256728778468$

$TPR = Recall = TP/(TP + FN) = 0.77$

$Precision = TP/(TP + FP) = 0.82$

For ANN we have:

$ACC = (TP + TN)/(TP + FN + FP + TN) = 0.79$

$FPR = FP/(FP + TN) = 0.7349896480331263$

$TPR = Recall = TP/(TP + FN) = 0.73$

$Precision = TP/(TP + FP) = 0.85$

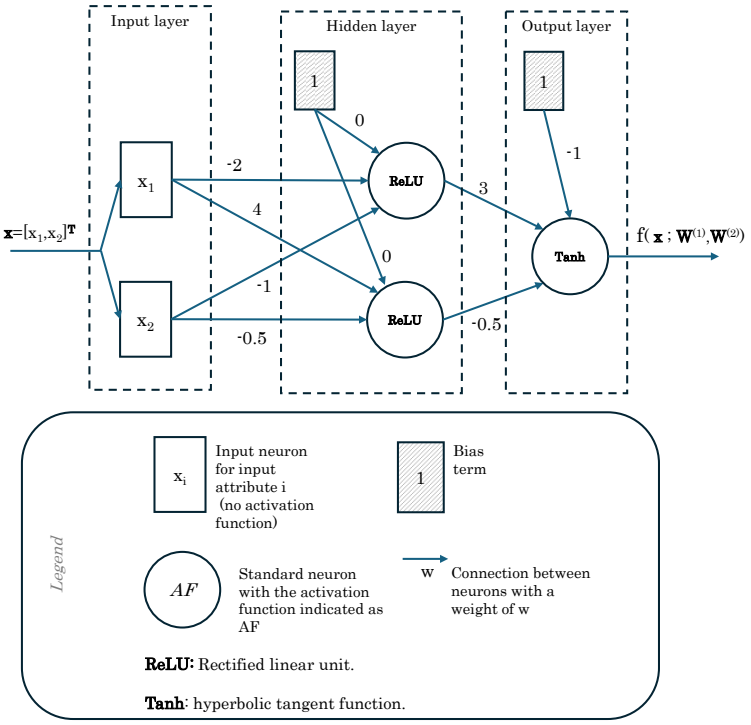

Figure 11: ANN schematic.

Question 22. Consider a two layer neural network $f : \mathbb{R}^2 \to \mathbb{R}$ for regression with two hidden units (i.e. $\boldsymbol{z}^{(1)} \in \mathbb{R}^2$) and of the form

$$\boldsymbol{z}^{(1)} = h^{(1)}((\boldsymbol{W}^{(1)})^\top \tilde{\boldsymbol{x}}),$$
$$f(\boldsymbol{x}; \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}) = h^{(2)}((\boldsymbol{W}^{(2)})^\top \tilde{\boldsymbol{z}}^{(1)}),$$

with a schematic provided in Figure 11.

Given an input, $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top = \begin{bmatrix} 0 & -1 \end{bmatrix}^\top$ determine the corresponding output.

A. ≈ -1.00

B. ≈ -0.85

C. ≈ 0.94

D. ≈ 1.75

E. Don't know.

Solution 22. We have $\tilde{\boldsymbol{x}} = \begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix}^\intercal$, $\tilde{\boldsymbol{z}}^{(1)} = \begin{bmatrix} 1 & z_1^{(1)} & z_2^{(1)} \end{bmatrix}$, $h^{(1)}(x) = \max(0, x)$ and $h^{(2)}(x) = tanh(x)$ is the activation function for the hidden layer (rectified linear unit) and output layer that is applied elementwise. We id the weights from the figure

$$\boldsymbol{W}^{(1)} = \begin{bmatrix} 0 & 0 \\ -2 & 4 \\ -1 & -0.5 \end{bmatrix}$$

and

$$\boldsymbol{W}^{(2)} = \begin{bmatrix} -1 & 3 & -0.5 \end{bmatrix}^\intercal.$$

i.e. $y = tanh([1; max(0, W1' * mtx)]' * W2) = 0.94$

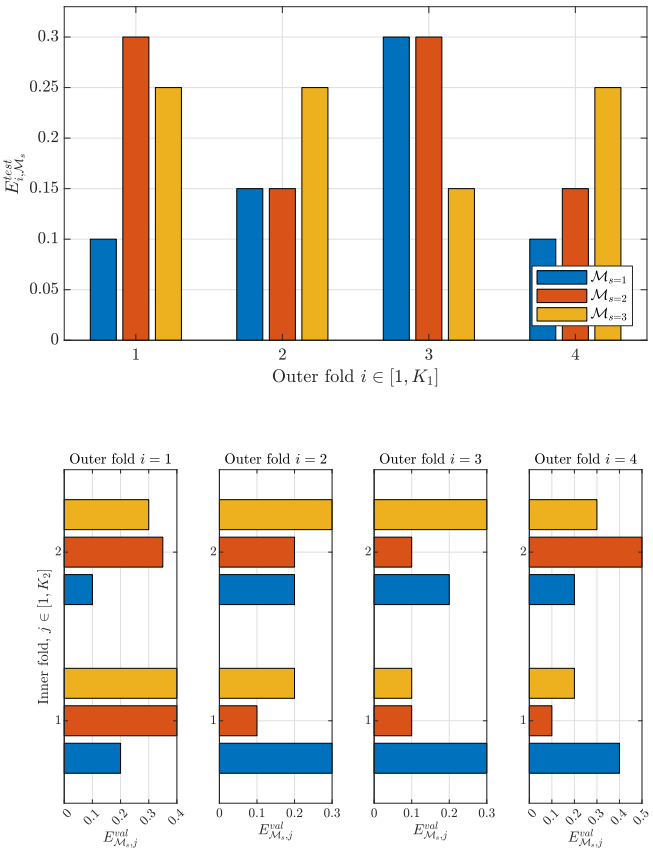

Figure 12: Result of running two-layer cross-validation.

Question 23. Consider evaluating three regression models ($\mathcal{M}_{s=1}, \mathcal{M}_{s=2}$ and $\mathcal{M}_{s=3}$), based on a dataset with 1200 observations. We use two-layer cross-validation (c.f. Algorithm 6 in the lecture notes) with $K_1 = 4$ outer folds and $K_2 = 2$ inner folds. The error is computed using the mean squared error.

The result of the procedure is shown in Figure 12. $E^{test}_{i,\mathcal{M}_s}$ denotes the performance of the models with complexities $s = \{1, 2, 3\}$ on the test set in the outer fold. Hence, we do not just evaluate the optimal model, $\mathcal{M}_{s^*}$, chosen based on the result of the inner fold, but compute the error for all three models on the test set in each of the outer folds.

Determine the value of the estimated generalization error $\hat{E}^{gen}$.

A. $\hat{E}^{gen} \approx 0.2$

B. $\hat{E}^{gen} \approx 0.3$

C. $\hat{E}^{gen} \approx 0.1$

D. $\hat{E}^{gen} \approx 0.25$

E. Don't know.

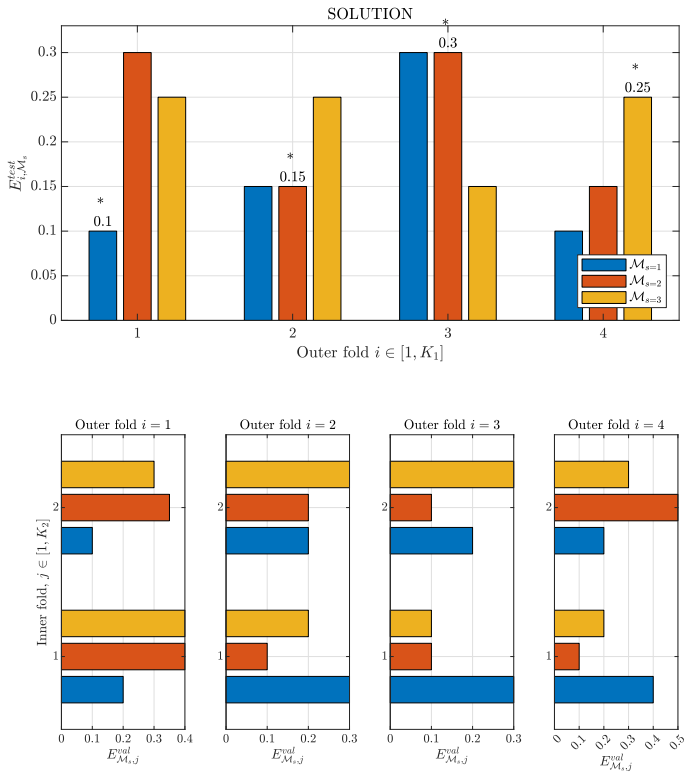

Figure 13: SOLUTION

Solution 23. The correct answer is A. See Figure 13 where the selected models and their performance on the outer fold is illustrated. The generalization error is the average of the indicated values. The optimal models is selected based on the average performance on the 2 inner folds.

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

Table 7: Binarized version of the Electronic Device Quality 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 $\{o_1,\ o_2,\ o_3,\ o_4,\ o_5,\ o_6,\ o_7\}$ belong to class C_1 (corresponding to Deactivated), and the red observations $\{o_8,\ o_9,\ o_{10}\}$ belong to class C_2 (corresponding to Activated).

Question 24. We consider the binary matrix from Table 7 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 $\{f_2, f_3, f_4\} \rightarrow \{f_1, f_5\}$?

A. The confidence is $\frac{3}{10}$

B. The confidence is 1

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

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

E. Don't know.

Solution 24. The confidence of the rule is computed as

$$\frac{\text{support}(\{f_2, f_3, f_4\} \cup \{f_1, f_5\})}{\text{support}(\{f_2, f_3, f_4\})} = \frac{\frac{1}{10}}{\frac{1}{10}} = 1.$$

Therefore, answer B is correct.

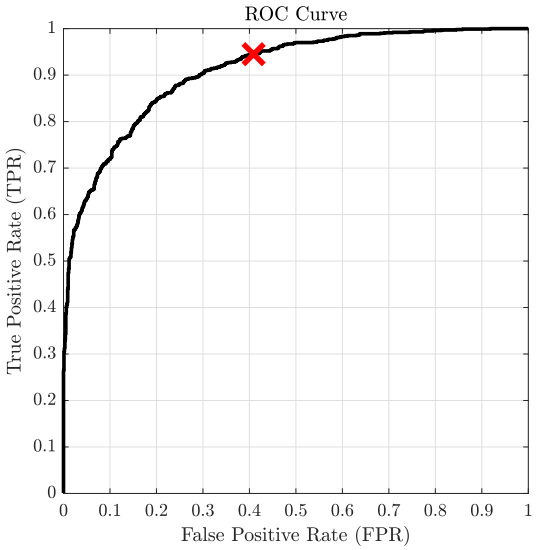

Figure 14: ROC.

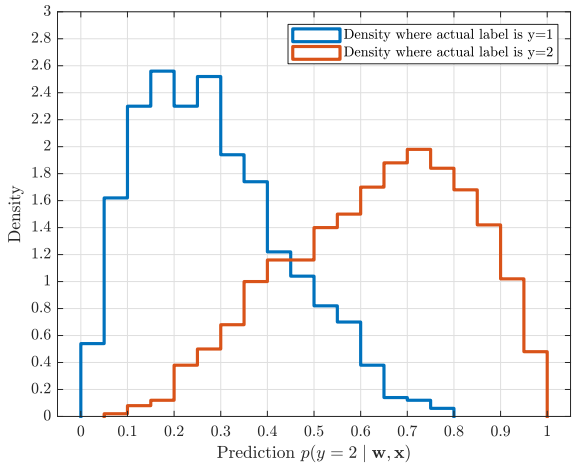

Figure 15: Density of predicted probabilities.

Question 25. Consider the ROC curve shown in Figure 14 originating from evaluating a logistic regression model.

Based on the density of the predictions shown in Figure 15, determine which threshold corresponds to the (FPR, TPR)-point indicated by the red cross in Figure 14.

A. ≈ 0.01

B. ≈ 0.1

C. ≈ 0.3

D. ≈ 0.7

E. Don't know.

Solution 25.

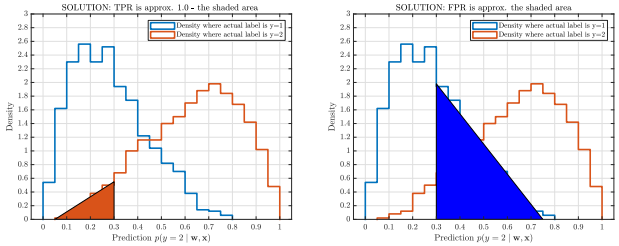

Figure 16: Illustration of the relevant areas.

The correct answer is C.

A given threshold results in a certain point on the ROC curve, i.e. a calculation of (FPR, TPR) values.

The plot shows a density of predicted probabilities, i.e. each curve/histogram integrates to 1.

Suppose we set a threshold of $\theta = 0.3$.

The $FPR = FP/(FP + TN)$ is computed as the area under under the $p(y = 1|w, x)$-curve (blue) in $[\theta, 1]$, as indicated in Figure 16. Here we approximate with a triangle, so the area is $(W \times H)/2 = (0.4 \times 2)/2 = 0.4$

The $TPR = TP/(TP + FN)$ is computed as the area under under the $p(y = 2|w, x)$-curve (brown/yellow) in $[\theta, 1]$ - or here as one minus the area $[0, \theta]$ as indicated in Figure 16, i.e. using a triangle $\approx .1 - (0.25 \times 0.5)/2 \approx 0.94 - 0.95$.

These values roughly match the values indicated by the red cross, i.e. $(0.41, 0.945)$

The exact corresponding values of θ, FPR and TPR:

$$\begin{bmatrix} \theta & 0.3000 & 0.0100 & 0.1000 & 0.7000 \\ FPR & 0.4090 & 0.9990 & 0.8930 & 0.0090 \\ TPR & 0.9450 & 1.0000 & 0.9990 & 0.4220 \end{bmatrix}$$

Question 26. Consider two classifiers evaluated on the Electronic Device Quality dataset with ($N = 900$) observations:

> Model A: A logistic regression model with the complexity controlled by the regularization parameter λ.
>
> Model B: A neural network with the complexity controlled by the number of hidden layers.

The models were compared using on two-layer cross-validation (Algorithm 6 in the lecture notes) with $K_1 = 10$ outer folds and $K_2 = 3$ inner folds, and a correct/ideal application of McNemar's test.

McNemar's test resulted in an estimated difference in accuracy of $\hat{\theta} = -0.097$ along with an undisclosed confidence interval and p-value.

McNemar's test relies on a 2×2 matched-pair matrix, $\left[\begin{smallmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{smallmatrix}\right]$, where

- n_{11} : Number of times both classifiers are correct
- n_{12} : Number of times A is correct and B is wrong
- n_{21} : Number of times A is wrong and B is correct
- n_{22} : Number of times both classifiers are wrong

Determine which one of the following matrices was used in the estimation.

A. $\begin{bmatrix} 707 & 13 \\ 100 & 80 \end{bmatrix}$

B. $\begin{bmatrix} 71 & 1 \\ 10 & 8 \end{bmatrix}$

C. $\begin{bmatrix} 236 & 4 \\ 33 & 27 \end{bmatrix}$

D. $\begin{bmatrix} 77 & 104 \\ 17 & 702 \end{bmatrix}$

E. Don't know.

Solution 26.

The correct answer is A.

We realise that McNemar's test (correctly) operates on the outer layer and on all $N = 900$ paired observations, i.e., the total number of elements in n should be N. E.g. it would not make sense to do McNemar's on e.g. an average accuracy per outer fold or a subset of the samples in folds.

A: Correct

D: Incorrect, elements sum to 900 but incorrect $\hat{\theta}$ (negated)

B: Incorrect, correct-ish $\hat{\theta}$, but elements in **n** does not sum to N ($N/K1$)

C: Incorrect, correct $\hat{\theta}$, but elements in **n** does not sum to N ($N/K2$)

INTERNAL Details only for debugging:

$n1 = \begin{bmatrix} 707 & 13 \\ 100 & 80 \end{bmatrix}$ $n2 = \begin{bmatrix} 77 & 104 \\ 17 & 702 \end{bmatrix}$

$n3 = \begin{bmatrix} 71 & 1 \\ 10 & 8 \end{bmatrix}$ $n4 = \begin{bmatrix} 236 & 4 \\ 33 & 27 \end{bmatrix}$

$$N = \begin{bmatrix} 900 & 900 & 90 & 300 \end{bmatrix}$$

$$\hat{\theta} = \begin{bmatrix} -0.097 & 0.097 & -0.1 & -0.097 \end{bmatrix}$$

$$\theta_L = \begin{bmatrix} -0.119 & 0.074 & -0.169 & -0.135 \end{bmatrix}$$

$$\theta_U = \begin{bmatrix} -0.074 & 0.12 & -0.031 & -0.058 \end{bmatrix}$$

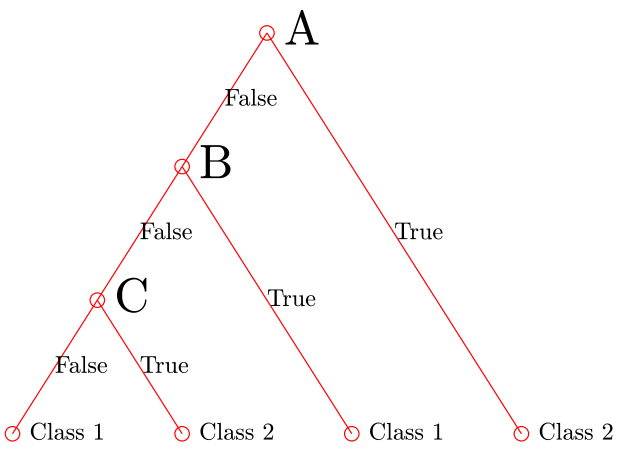

Figure 17: Example classification tree.

Question 27. We consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 17 resulting in a partition into classes indicated by the colors/markers in Figure 18. What is the correct rule assignment to the nodes in the decision tree?

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

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

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

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

E. Don't know.

Solution 27.

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.

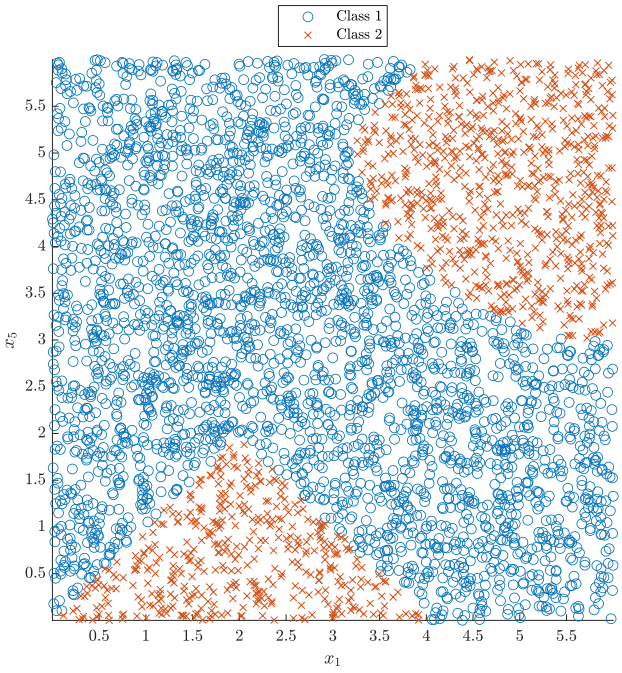

Figure 18: Classification boundary.

The resulting decision boundaries for each of the options are shown in Figure 19 and it follows answer D is correct.

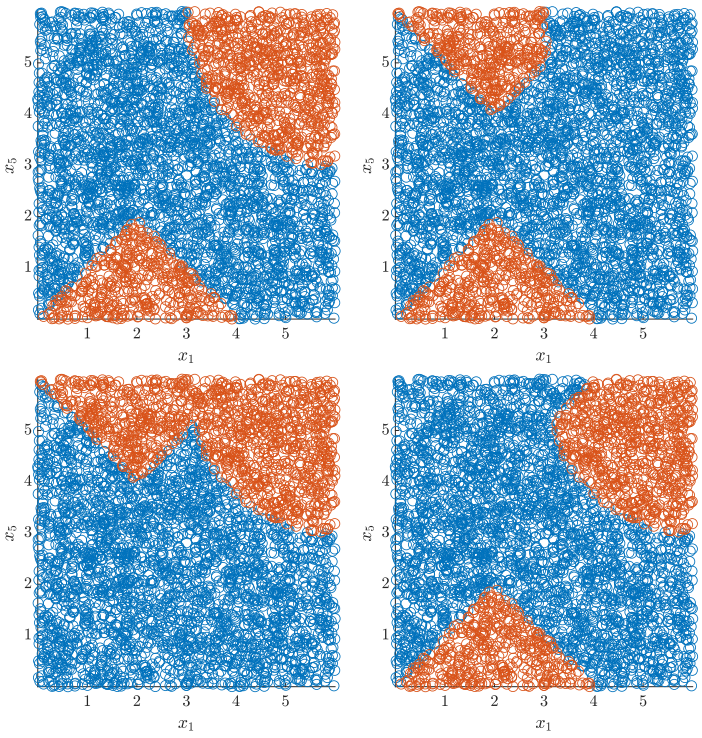

Figure 19: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)