Technical University of Denmark

Disclaimer: The solutions provided in old exam sets are not written with pedagogical/didact clarity in mind. They are written so examiners can validate the answers and the steps needed to answer the questions.

Written examination: 23rd May 2023, 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.

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

No.	Attribute description	Abbrev.
x_1	Peak-to-average power ratio	P2AR
x_2	Periodicity	PER
x_3	Roughness	RO
x_4	Fluctuation	FLU
x_5	Sharpness	SHP
x_6	Loudness	LN
y	Sound class	
y_r	Percieved annoyance measure (PAM)	

Table 1: The dataset contains 564 audio recordings from two classes, Machine and Natural sounds. Machine sounds are produced by man-made machinery such as cars or chainsaws. Natural sounds result from natural phenomena such as rain or dogs barking. Additionally, a perceived annoyance measure (PAM), y_r, has been observed, quantifying the annoyance perceived by a human listener. Each recording has been preprocessed to extract 6 attributes representing the recording. The attributes are described in the table above. The classification task concerns the prediction of the sound type, y (Machine or Natural), while the regression task concerns the prediction of the percieved annoyance level, $y_r \in \mathbb{R}$.

Question 1. The main dataset used in this exam is the Sound Classification dataset[1] described in Table 1. Figure 1 showns the histograms of observations from the Sound Classification dataset. Which one of the following statements is true?

A. Attribute x_1 has an empirical mean of $\approx$ -3.72 and attribute x_6 has an empirical variance of $\approx$ 2.34.

B. Attribute x_2 has an empirical mean of $\approx$ 4.76 and attribute x_1 has an empirical variance of $\approx$ 0.12.

C. Attribute x_3 has an empirical mean of $\approx$ 27.43 and attribute x_1 has an empirical variance of $\approx$ 9.6.

D. Attribute x_4 has an empirical mean of $\approx$ 8.73 and attribute x_3 has an empirical variance of $\approx$ 0.27.

E. Don't know.

Solution 1. The correct answer is A from inspection and illustrated in Figure 2.

[1]Dataset obtained from `https://github.com/karolpiczak/ESC-50`

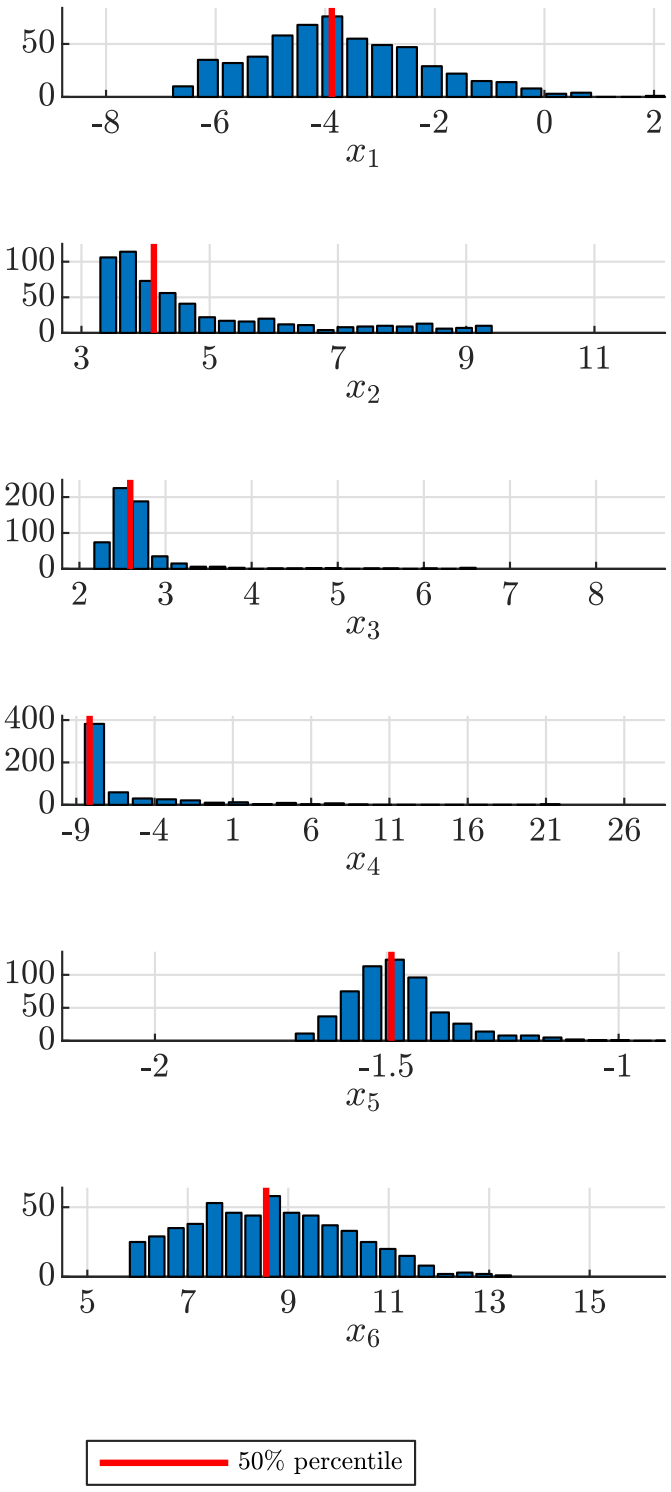

Figure 1: Histograms of observations from the Sound Classification dataset described in Table 1.

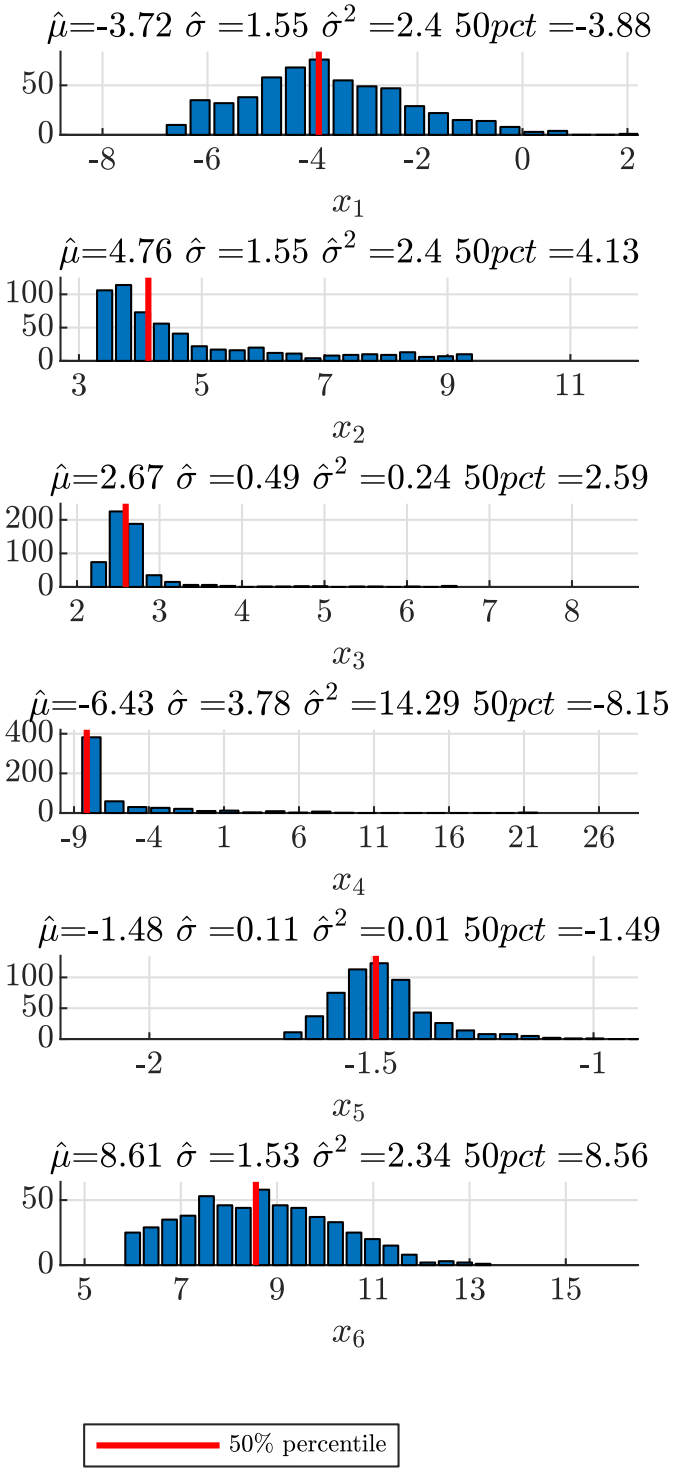

Figure 2: SOLUTION: Histograms of observations from the the Sound Classification dataset described in Table 1.

Question 2. The empirical correlation between the first 5 attributes in the Sound Classification dataset is given by the following correlation matrix

$$Corr = \begin{bmatrix} 1.0 & 0.48 & -0.14 & 0.15 & -0.06 \\ 0.48 & 1.0 & -0.14 & 0.19 & 0.19 \\ -0.14 & -0.14 & 1.0 & 0.12 & 0.0 \\ 0.15 & 0.19 & 0.12 & 1.0 & 0.26 \\ -0.06 & 0.19 & 0.0 & 0.26 & 1.0 \end{bmatrix},$$

where element $Corr[i, j]$ is the correlation between the two attributes x_i and x_j, e.g., $Corr[x_2, x_5] = 0.19$.

Additionally, the empirical standard deviation of x_1 is 1.55, and the empirical standard deviation of x_4 is 3.78.

Determine which one of the following statements is true.

A. The covariance between x_1 and x_4 is ≈ 0.88

B. The covariance between x_1 and x_4 is ≈ 0.36

C. The covariance between x_1 and x_4 is ≈ 1.36

D. The covariance between x_1 and x_4 is ≈ 0.03

E. Don't know.

Solution 2. The answer A is correct.

Recall, that

$$Corr[x_i, x_j] = \frac{Cov[x_1, x_2]}{\sqrt{\sigma_i^2 \sigma_j^2}}$$

thus

$$Cov[x_i, x_j] = Corr[x_i, x_j] \sqrt{\sigma_i^2 \sigma_j^2}$$

. Specifically, with the provided values, $Cov[x_1, x_4] = 0.15 \times \sqrt{1.55^2 \times 3.78^2}$

For completeness (not needed to answer the question), the full covariance matrix is

$$\Sigma = \begin{bmatrix} 2.4 & 1.15 & -0.11 & 0.88 & -0.01 \\ 1.15 & 2.4 & -0.11 & 1.11 & 0.03 \\ -0.11 & -0.11 & 0.24 & 0.22 & 0.0 \\ 0.88 & 1.11 & 0.22 & 14.29 & 0.11 \\ -0.01 & 0.03 & 0.0 & 0.11 & 0.01 \end{bmatrix},$$

which is based on the following standard deviations

$$\sigma = \begin{bmatrix} 1.55 & 1.55 & 0.49 & 3.78 & 0.11 \end{bmatrix}.$$

Question 3. A Principal Component Analysis (PCA) is carried out on the Sound Classification dataset in Table 1 based on the attributes x_1, x_2, x_3, x_4 and 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 (SVD) is then carried out on the standardized data matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

$$V = \begin{bmatrix} -0.5939 & 0.2906 & v_{1,3} & 0.0621 & 0.6652 \\ -0.6521 & 0.0759 & 0.0004 & 0.3813 & v_{2,5} \\ 0.2028 & -0.5105 & -0.7036 & 0.4508 & 0.0010 \\ -0.3696 & -0.5414 & -0.1781 & -0.7244 & -0.1173 \\ -0.2102 & -0.5967 & 0.5973 & 0.3503 & 0.3467 \end{bmatrix} \tag{1}$$

$$S = \begin{bmatrix} 30.3832 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & \sigma_{2,2} & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 22.7730 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 19.7263 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 16.0724 \end{bmatrix} \tag{2}$$

It is noted that certain elements are missing in the SVD, specifically $v_{1,3}, v_{2,5}$ and $\sigma_{2,2}$. Additionally, $\left\|\tilde{\boldsymbol{X}}\right\|_F^2 = 2814.8909$.

Determine which one of the following statements is correct.

A. $v_{1,3} \approx -0.1399$, $v_{2,5} \approx -0.2116$
$\sigma_{2,2} \approx 5.1903$

B. $v_{1,3} \approx -0.3413$, $v_{2,5} \approx -0.6508$
$\sigma_{2,2} \approx 26.9387$

C. $v_{1,3} \approx 0.3425$, $v_{2,5} \approx -0.6506$
$\sigma_{2,2} \approx 26.9387$

D. $v_{1,3} \approx -1.8385$, $v_{2,5} \approx -0.1629$
$\sigma_{2,2} \approx 5.1903$

E. Don't know.

Solution 3. The correct answer is B.

With $\boldsymbol{v}_{:,a}$ denoting the first collum with missing element and $\boldsymbol{v}_{:,b}$ the second collum with a missing element.

B: Correct; all vectors $\boldsymbol{v}_{:,j}$ orthonormal (orthogonal and norm one).
D: Incorrect. $\boldsymbol{v}_{:,a}$ and $\boldsymbol{v}_{:,b}$ are NOT orthogoal to each other; $\boldsymbol{v}_{:,a}$ not length 1 (easy to check; see below)
A: Incorrect. $\boldsymbol{v}_{:,a}$ are orthogoal to the rest; $\boldsymbol{v}_{:,b}$ NOT length 1 (easy to check via length).
C: Incorrect. $\boldsymbol{v}_{:,b}$ is NOT orthogonal to the known eigenvectors, $\boldsymbol{v}_{:,a}$ is unconstrained. Both have norm $\approx$ 1. Requires a check of orthogonality.

The complete SVD is given by:

$$V = \begin{bmatrix} -0.5939 & 0.2906 & -0.3413 & 0.0621 & 0.6652 \\ -0.6521 & 0.0759 & 0.0004 & 0.3813 & -0.6508 \\ 0.2028 & -0.5105 & -0.7036 & 0.4508 & 0.001 \\ -0.3696 & -0.5414 & -0.1781 & -0.7244 & -0.1173 \\ -0.2102 & -0.5967 & 0.5973 & 0.3503 & 0.3467 \end{bmatrix}$$

$$S = \begin{bmatrix} 30.3832 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 26.9387 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 22.773 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 19.7263 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 16.0724 \end{bmatrix}$$

The norms (used to quickly exclude certain options) :

$$B : \begin{bmatrix} 0.9999 & 1.0 & 1.0 & 1.0 & 1.0 \end{bmatrix}$$
$$D : \begin{bmatrix} 0.9999 & 1.0 & 2.0649 & 1.0 & 0.7766 \end{bmatrix}$$
$$A : \begin{bmatrix} 0.9999 & 1.0 & 0.9503 & 1.0 & 0.7882 \end{bmatrix}$$
$$C : \begin{bmatrix} 0.9999 & 1.0 & 1.0004 & 1.0 & 0.9999 \end{bmatrix}$$

For completeness, we provide the pairwise inner products between the eigenvectors corresponding to the var-ious solutions (**not** required to solve the problem):

$$B : \begin{bmatrix} 0.9999 & -0.0001 & 0.0 & 0.0 & 0.0 \\ -0.0001 & 1.0 & 0.0001 & 0.0 & 0.0 \\ 0.0 & 0.0001 & 1.0 & 0.0 & -0.0 \\ 0.0 & 0.0 & 0.0 & 0.9999 & 0.0 \\ 0.0 & 0.0 & -0.0 & 0.0 & 1.0 \end{bmatrix}$$

$$D : \begin{bmatrix} 0.9999 & -0.0001 & 0.8892 & 0.0 & -0.3181 \\ -0.0001 & 1.0 & -0.435 & 0.0 & 0.0371 \\ 0.8892 & -0.435 & 4.2637 & -0.093 & -0.9958 \\ 0.0 & 0.0 & -0.093 & 0.9999 & 0.186 \\ -0.3181 & 0.0371 & -0.9958 & 0.186 & 0.603 \end{bmatrix}$$

$$A : \begin{bmatrix} 0.9999 & -0.0001 & -0.1196 & 0.0 & -0.2863 \\ -0.0001 & 1.0 & 0.0586 & 0.0 & 0.0334 \\ -0.1196 & 0.0586 & 0.9031 & 0.0125 & 0.1341 \\ 0.0 & 0.0 & 0.0125 & 0.9999 & 0.1675 \\ -0.2863 & 0.0334 & 0.1341 & 0.1675 & 0.6213 \end{bmatrix}$$

$$C : \begin{bmatrix} 0.9999 & -0.0001 & -0.4061 & 0.0 & -0.0001 \\ -0.0001 & 1.0 & 0.1988 & 0.0 & 0.0 \\ -0.4061 & 0.1988 & 1.0008 & 0.0425 & 0.4548 \\ 0.0 & 0.0 & 0.0425 & 0.9999 & 0.0001 \\ -0.0001 & 0.0 & 0.4548 & 0.0001 & 0.9997 \end{bmatrix}$$

Question 4. Consider again the PCA for the Sound Classification dataset, in particular the SVD of $\tilde{\boldsymbol{X}}$ in Equation (1). Which one of the following statements is true?

A. An observation with a low value of x_1 **(P2AR)**, a high value of x_2 **(PER)**, a low value of x_3 **(RO)**, a high value of x_4 **(FLU)**, and a high value of x_5 **(SHP)** will typically have a positive value of the projection onto principal component number 1.

B. An observation with a high value of x_2 **(PER)**, a high value of x_3 **(RO)**, a low value of x_4 **(FLU)**, and a high value of x_5 **(SHP)** will typically have a negative value of the projection onto principal component number 4.

C. An observation with a low value of x_1 **(P2AR)**, a high value of x_3 **(RO)**, a high value of x_4 **(FLU)**, and a high value of x_5 **(SHP)** will typically have a positive value of the projection onto principal component number 2.

D. An observation with a high value of x_1 (P2AR), a high value of x_2 (PER), a low value of x_3 (RO), a high value of x_4 (FLU), and a high value of x_5 (SHP) will typically have a negative value of the projection onto principal component number 1.

E. Don't know.

Solution 4. The correct answer is D. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_1$ (i.e. column one of $\boldsymbol{V}$) is

$$b_1 = \boldsymbol{x}^\top \boldsymbol{v}_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} \begin{bmatrix} -0.5939 \\ -0.6521 \\ 0.2028 \\ -0.3696 \\ -0.2102 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and negative, this occurs if x_1, x_2, x_3, x_4, x_5 has large magnitude and the sign convention given in option D.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	1.3	4.1	3.8	4.5	2.4	3.2	2.7	3.0	3.9
o_2	1.3	0.0	3.2	3.1	4.7	2.3	2.6	2.2	2.7	4.2
o_3	4.1	3.2	0.0	0.4	4.9	2.7	1.1	1.6	2.4	4.8
o_4	3.8	3.1	0.4	0.0	4.6	2.5	0.9	1.3	2.1	4.5
o_5	4.5	4.7	4.9	4.6	0.0	3.1	4.4	3.7	2.8	2.3
o_6	2.4	2.3	2.7	2.5	3.1	0.0	1.8	1.2	0.9	2.8
o_7	3.2	2.6	1.1	0.9	4.4	1.8	0.0	1.0	1.7	4.1
o_8	2.7	2.2	1.6	1.3	3.7	1.2	1.0	0.0	1.1	3.6
o_9	3.0	2.7	2.4	2.1	2.8	0.9	1.7	1.1	0.0	2.9
o_{10}	3.9	4.2	4.8	4.5	2.3	2.8	4.1	3.6	2.9	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 Sound Classification dataset (recall that $M = 6$). 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 $\{o_1, o_2, o_3, o_4\}$ belong to class C_1 (corresponding to Machine), and the red observations $\{o_5, o_6, o_7, o_8, o_9, o_{10}\}$ belong to class C_2 (corresponding to Natural). To avoid single features to dominate, the dataset was standardized by subtracting the mean and dividing by the standard deviation.

Question 5. Consider the distances in Table 2 based on 10 observations from the Sound Classification dataset. 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 2. 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. 0.732

B. 0.769

C. 0.469

D. 0.640

E. Don't know.

Solution 5.

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:

$$N_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = \{o_2, o_6, o_8\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = 0.469$$

Therefore option C is correct.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$K=1$	0	0	1	0	0
$K=3$	1	1	1	1	1
$K=4$	1	1	1	0	0

Table 3: Error rates corresponding to a specific inner fold, j, and model complexity, K, in **outer fold 1**.

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
$K=1$	0	0	1	0	1
$K=3$	0	0	1	0	0
$K=4$	0	0	1	0	0

Table 4: Error rates corresponding to a specific inner fold, j, and model complexity, K, in **outer fold 2**.

Question 6. We consider the generalization error of K-nearest neighbour (KNN) classifiers with $K = [1, 3, 4]$ neighbors based on the observations in Table 2.

We apply two-layer cross-validation (cf. Algorithm 6 in the lecture notes) to estimate the generalization error with 2 outer and 5 inner folds. Table 3 and Table 4 provides partial results of the associated two-layer cross validation. It is noted that with other things being equal, a model with the lowest value of K is preferred in this context.

The specific split of the dataset in the two outer folds is also known:

- Outer fold $i = 1$ training set: $o_1,\ o_2,\ o_4,\ o_8,\ o_9$.
- Outer fold $i = 2$ training set: $o_3,\ o_5,\ o_6,\ o_7,\ o_{10}$.

Determine the error rates on the test set in each of the 2 outer folds based on the information in Table 3, Table 4 and Table 2.

A. $E^{test}_{i=1} = 0.1$ and $E^{test}_{i=2} = 0.3$

B. $E^{test}_{i=1} = 0.2$ and $E^{test}_{i=2} = 0.2$

C. $E^{test}_{i=1} = 0.8$ and $E^{test}_{i=2} = 0.4$

D. $E^{test}_{i=1} = 0.2$ **and** $E^{test}_{i=2} = 0.6$

E. Don't know.

Solution 6.

The correct answer is D.

From the available error rates in the inner fold (in the tables) we first select K in the two outer folds to

be [1, 3] based on the minimum error rate across the inner folds.

Secondly, we note that only half the observations should be used as a training set in each fold.

The error rate is computed as usual for a KNN classifier with the selected values of K.

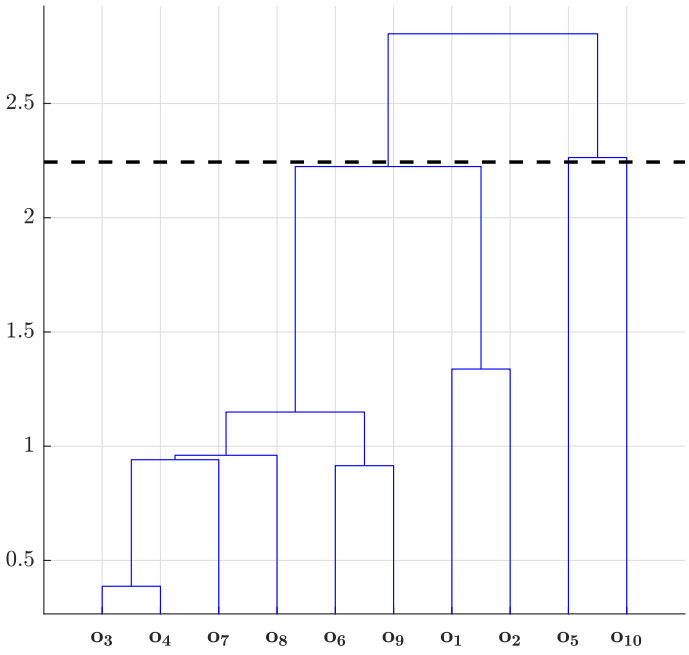

Figure 3: A dendrogram with a cutoff indicated by the dotted line, thereby generating 3 clusters.

Question 7. Consider the dendrogram shown in Figure 3 which is based on the observations and distances in Table 2 using single linkage.

Suppose we apply a cutoff (indicated by the black line in Figure 3) 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. What is the Rand index of the two clusterings?

A. $\mathrm{R}(Z, Q) \approx 0.428$

B. $\mathbf{R}(Z, Q) \approx 0.444$

C. $\mathrm{R}(Z, Q) \approx 0.48$

D. $\mathrm{R}(Z, Q) \approx 0.411$

E. Don't know.

Solution 7. To compute $\mathrm{R}[Z, Q]$, note Z is the clustering corresponding to the colors in Table 2 and Q the clustering obtained by cutting the dendrogram in Figure 3 given as:

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

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

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

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

$$S = 12, D = 8$$

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

Question 8. Consider the ten observations in Table 2 and denote the partition implied by the ground truth classification by Z. We also consider an alternative partition Q, where all ten observations are placed in the same cluster. Let $R(Z, Q)$ denote the Rand index and $J(Z, Q)$ the Jaccard similarity between Z and Q. Which one of the following statements is true?

A. $R(Z, Q) > J(Z, Q)$

B. $R(Z, Q) = J(Z, Q)$

C. $R(Z, Q) < J(Z, Q)$

D. $R(Z, Q)$ and $J(Z, Q)$ are not defined in this case

E. Don't know.

Solution 8. The Rand index between Z and Q is defined as

$$R(Q, P) = \frac{S + D}{\frac{1}{2}N(N-1)}.$$

and the Jaccard similarity as

$$J(Q, P) = \frac{S}{\frac{1}{2}N(N-1) - D}$$

where D denotes the number of times that Z and Q agree that two distinct observations are not in the same cluster. Since Q only contains one cluster we have that $D = 0$ and therefore $R(Z, Q) = J(Z, Q)$. This can also be confirmed numerically in the given case.

Question 9. We consider a synthetic 1-dimensional dataset with the following $N = 5$ observations:

$$\boldsymbol{X} = \begin{bmatrix} 2 \\ 5 \\ 8 \\ 12 \\ 13 \end{bmatrix}$$

Suppose we want to perform K-means clustering with $K = 2$ centroids. The centroids are initialized at locations $\mu_1 = 4$ and $\mu_2 = 10$. What will be the total cost (sum of squared distances), E, after the first iteration of K-means?

A. $E \approx 18.5$

B. $E \approx 9.0$

C. $E \approx 3.7$

D. $E \approx 299.75$

E. Don't know.

Solution 9.

The correct answer is A.

The squared distances from the centroids to all points is:

$$\begin{bmatrix} 4 & 1 & 16 & 64 & 81 \\ 64 & 25 & 4 & 4 & 9 \end{bmatrix}$$

We partition the points as follows:

$$Z = \begin{bmatrix} 1 & 1 & 2 & 2 & 2 \end{bmatrix}$$

The centroids are then updated as the mean of the assiged observations.

$$\mu = \begin{bmatrix} 3.5 & 11.0 \end{bmatrix}$$

The squared distance from the new centroid to the observations:

$$\begin{bmatrix} 2.25 & 2.25 & 20.25 & 72.25 & 90.25 \\ 81.0 & 36.0 & 9.0 & 1.0 & 4.0 \end{bmatrix}$$

The squared distances to the nearest centroid is:

$$\begin{bmatrix} 2.25 & 2.25 & 9.0 & 1.0 & 4.0 \end{bmatrix}$$

Summing the distances yields the result: $E = 18.5$.

$p(\hat{x}_2, \hat{x}_4, y)$	$y = Machine$	$y = Natural$
$\hat{x}_2 = 0,\ \hat{x}_4 = 0$	0.18	0.08
$\hat{x}_2 = 0,\ \hat{x}_4 = 1$	0.17	0.07
$\hat{x}_2 = 1,\ \hat{x}_4 = 0$	0.08	0.16
$\hat{x}_2 = 1,\ \hat{x}_4 = 1$	0.1	0.16

Table 5: Probability of observing particular values of $\hat{x}_2$, $\hat{x}_4$, and y.

Question 10. Consider the Sound Classification dataset from Table 1. Suppose the attributes have been binarized such that $\hat{x}_2 = 0$ corresponds to $x_2 \leq -0.405$ (and otherwise $\hat{x}_2 = 1$) and $\hat{x}_4 = 0$ corresponds to $x_4 \leq -0.456$ (and otherwise $\hat{x}_4 = 1$). Suppose the probability for each of the configurations of $\hat{x}_2$, $\hat{x}_4$ and y are as given in Table 5.

What is then the probability y corresponds to *Machine* given an observation has $\hat{x}_2 = 1$?

A. $p(y = Machine \mid \hat{x}_2 = 1) = 0.18$

B. $p(y = Machine \mid \hat{x}_2 = 1) = 0.34$

C. $p(y = Machine \mid \hat{x}_2 = 1) = 0.36$

D. $p(y = Machine \mid \hat{x}_2 = 1) = 0.261$

E. Don't know.

Solution 10. Recall the (generic) marginalization rule which gives:

$$\sum_{x_b} p(x_a, x_b, y) = p(x_a, y)$$

Next we marginalize out y to get the probability $p(x_a)$

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

From this we can obtain the conditional probability using the product rule $p(y|x_a) = p(x_a, y)/p(x_a)$.

In our case we obtain the values:

$$p(\hat{x}_2 = 1, y = Machine) = 0.18$$

and $p(\hat{x}_2 = 1) = 0.5$

Thus, we see see answer C is correct using Bayes theorem

Question 11. We wish to predict which of the two classes an observation $\boldsymbol{x}$ belongs to in the Sound Classification dataset described in Table 1. To accomplish this we apply a Naïve-Bayes classifier where we model each of the $M = 6$ features using a 1-dimensional normal distribution. The classifier will be used in an embedded setting where model prediction speed is paramount. Therefore, consider a single model evaluation:

$$p(y = Machine|\boldsymbol{x}).$$

What is the minimum number of evaluations of the normal density function $\mathcal{N}(x|\mu, \sigma^2)$ we have to perform to compute this quantity?

A. 14

B. 21

C. 12

D. 18

E. Don't know.

Solution 11. Recall the formula for Naïve-Bayes is

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

The total number of evaluations is equal to the total number of evaluations of terms $p(x_k|y)$. Note that once we have evaluated the denominator, we will also have evaluated the numerator since they share the same terms $p(x_k|y)$, cutting down on computations. The total number of evaluations is therefore simply CM where C is the number of classes and therefore answer C is correct.

	M_A correct	M_A wrong
M_B correct	416	42
M_B wrong	38	68
M_C correct	68	38
M_C wrong	42	416

Table 6: Outcome of cross-validation with paired test observations. The values indicate the number of times where two models both classify correctly, both classify incorrectly (i.e., both wrong), or one classifies correctly and the other one incorrectly.

Question 12. Consider the comparison of a classification model, M_A, with two alternative models, M_B and M_C, using McNemar's test as defined in Section 11.3.3 of the lecture notes.
The test is based on the output of three-fold cross-validation on the Sound Classification dataset with the relevant results of the cross-validation shown in Table 6.
The comparison between M_A and M_B results in a p-value (p_{AB}) and the difference in accuracy ($\hat{\theta}_{AB}$).
Similarly, the comparison between M_A and M_C results in a p-value (p_{AC}) and the difference in accuracy ($\hat{\theta}_{AC}$).
Determine which one of the following statements is correct.
Hint: The problem can be solved without evaluating complicated mathematical functions.

A. $p_{AB} = p_{AC}$ **and** $\hat{\theta}_{AB} = -\hat{\theta}_{AC}$**.**

B. $p_{AB} \neq p_{AC}$ and $\hat{\theta}_{AB} = \hat{\theta}_{AC}$.

C. $p_{AB} = p_{AC}$ and $\hat{\theta}_{AB} = \hat{\theta}_{AC}$.

D. $p_{AB} \neq p_{AC}$ and $\hat{\theta}_{AB} = -\hat{\theta}_{AC}$.

E. Don't know.

Solution 12.
The correct answer is A.
We identify the contingency tables (here for A and B):

n_{11} : Both classifiers are correct
n_{12} : A is correct, B is wrong
n_{21} : A is wrong, B is correct
n_{22} : Both classifiers are wrong

We then exploit the symmertries and note that n is the same for the two cases and $n_{12}^{AB} = n_{21}^{AC}$.

$\hat{\theta}$: The difference in accuracy is given by

$$\hat{\theta} = (n_{12} - n_{21})/n$$

i.e. we see that $\theta_{AB} = -\theta_{AC}$.

p-value: The p-value for McNemar's test is given by:

$$p = 2cdf_{binom}(m = \min\{n_{12}, n_{21}\} \mid \theta = \frac{1}{2}, N = n_{12}+n_{21})$$

i.e. it only depends on n_{12} and n_{21}. Since n_{12} and n_{21} only enter as the minimum value and in a sum, the p-values must be the same.

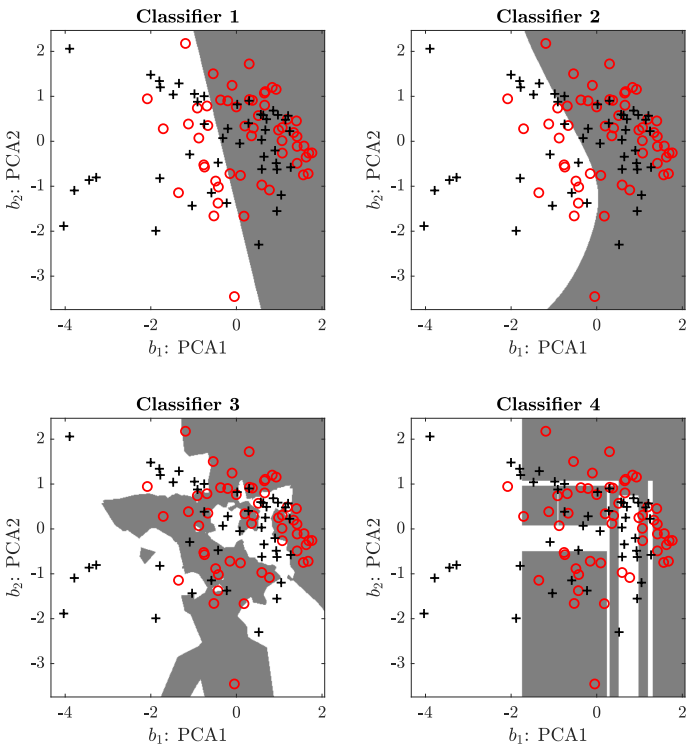

Figure 4: Decision boundaries for four different classifiers trained on the Sound Classification dataset when projected onto the first two principal components.

Question 13. Consider a subset of the Sound Classification dataset (described in Table 1) after it has been projected onto the first two principal components, thereby giving rise to a two-dimensional dataset with each observation having two coordinates, b_1 and b_2.

We will consider the following four classifiers:

LR: Logistic regression

ANN: Artificial neural network with 5 hidden units

CT: Classification tree with regular axis-aligned splits

KNN: K-nearest neighbours with $K = 3$

Suppose the classifiers are trained on the two-dimensional dataset and the decision boundary for each of the four classifiers is given in Figure 4. Which one of the following statements is correct?

A. Classifier 1 corresponds to LR, Classifier 2 corresponds to ANN, Classifier 3 corresponds to KNN, Classifier 4 corresponds to CT.

B. Classifier 1 corresponds to **ANN**,
Classifier 2 corresponds to **KNN**,
Classifier 3 corresponds to **LR**,
Classifier 4 corresponds to **CT**.

C. Classifier 1 corresponds to **KNN**,
Classifier 2 corresponds to **CT**,
Classifier 3 corresponds to **ANN**,
Classifier 4 corresponds to **LR**.

D. Classifier 1 corresponds to **LR**,
Classifier 2 corresponds to **KNN**,
Classifier 3 corresponds to **ANN**,
Classifier 4 corresponds to **CT**.

E. Don't know.

Solution 13. To solve this problem, we have to use our intuition about what the typical decision boundaries for the different methods look like:

- A KNN method will have decision boundaries dictated by the nearest neighbors. That is, points (x, y) where the nearest K neighbors are in one class must be in the same class and therefore the boundaries will be fairly complex and respect the data distribution well.
- A decision tree has axis aligned splits, therefore the boundaries must be vertical or horizontal
- A multivariate regression model must have linear boundaries
- An artificial neural network with few hidden units can have some non-linearity, but otherwise have boundaries of limited complexity and consisting of relatively simple shapes

It is easy to see this rules out all but option A.

Question 14. Determine which one of the following statements is always true about two-layer cross-validation with K_1 outer folds, K_2 inner folds, S model complexities, and a dataset of size N (cf. Algorithm 6 in the lecture notes).

A. The total number of times each observation is used for testing is $K_1 \times (K_2 \times S + 1)$.

B. The total number of times each observation is used for training is $(K_1 - 1) \times ((K_2 - 1) \times S + 1)$**.**

C. The total number of models needed to be trained is $K_2 \times S \times (K_1 + 1)$.

D. The size of the test set in each of the K_1 outer folds is always greater than the test set in the K_2 inner folds.

E. Don't know.

Solution 14. The correct answer is B. The solution is found by carefully tracing the two-layer cross-validation loops (in Algorithm 6) and counting the number of times a point is used to train a model.

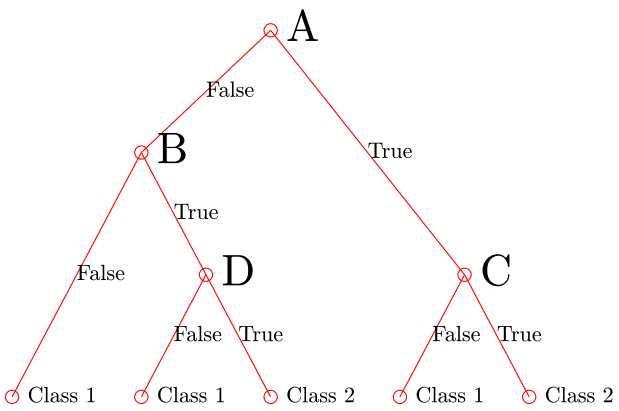

Figure 5: Classification tree.

Question 15. Consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 5 resulting in a partition into classes indicated by the colors/markers in Figure 6. What is the correct rule assignment to the nodes in the decision tree?

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

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

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

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

E. Don't know.

Solution 15.

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.

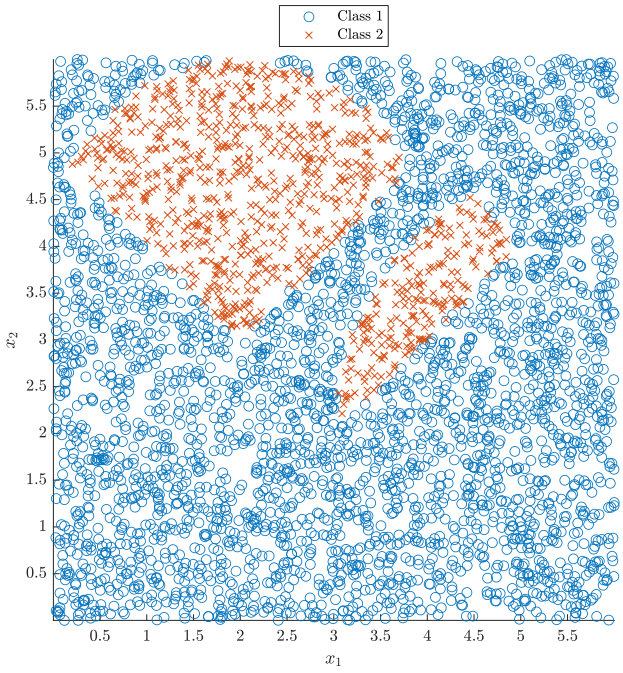

Figure 6: Classification boundary.

The resulting decision boundaries for each of the options are shown in Figure 7 and it follows answer B is correct.

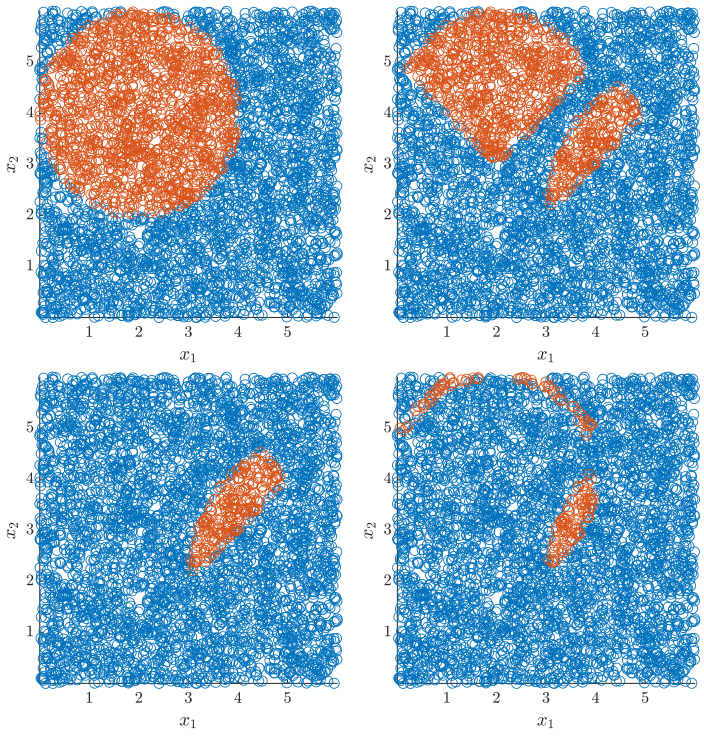

Figure 7: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

Question 16. Consider standard linear regression (no regularization) with the prediction rule $f(\boldsymbol{x}_i; \boldsymbol{w}) = \tilde{\boldsymbol{x}}_i^\top \boldsymbol{w}$. Here $\tilde{\boldsymbol{x}}_i^\top = \begin{bmatrix} 1 & \boldsymbol{x}_i^\top \end{bmatrix}$ is the i'th observation, $\boldsymbol{x}_i$, concatenated with a constant with a value of 1. Suppose, the weights, $\boldsymbol{w} = \begin{bmatrix} w_0 & w_1 & \dots & w_M \end{bmatrix}^\top$ (including the intercept term w_0), were learned by predicting y_r from $\tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{\boldsymbol{x}}_1 & \tilde{\boldsymbol{x}}_2 & \dots & \tilde{\boldsymbol{x}}_N \end{bmatrix}^\top$ (N is the number of observations) and minimizing the sum-of-squares error function, as usual.

What solution, $\boldsymbol{w}_{\text{scaled}}$, do you obtain if you scale all elements in $\tilde{\boldsymbol{X}}$ by a factor of $1/\beta$, with $\beta \in \mathbb{R}$ being a constant, before learning the weights?

A. $\boldsymbol{w}_{\text{scaled}} = \frac{1}{\beta}\boldsymbol{w}$

B. $\boldsymbol{w}_{\text{scaled}} = \boldsymbol{w}^\beta$

C. $\boldsymbol{w}_{\text{scaled}} = \beta\, \boldsymbol{w}$

D. $\boldsymbol{w}_{\text{scaled}} = \beta^2 \boldsymbol{w}$

E. Don't know.

Solution 16. The correct answer is C.

Formally, the original solution, $\boldsymbol{w}$, is found by

$$\boldsymbol{w} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{y}_r$$

.

The new solution, $\boldsymbol{w}_{\text{scaled}}$, is found by scaling $\tilde{\boldsymbol{X}}$ by $1/\beta$, i.e., elementwise multiplication of each element (also the bias term) in $\tilde{\boldsymbol{X}}$ with $1/\beta$, i.e scaling by $1/\beta$

$$\begin{aligned}
\boldsymbol{w}_{\text{scaled}} &= ((\frac{1}{\beta}\tilde{\boldsymbol{X}})^\top (\frac{1}{\beta}\tilde{\boldsymbol{X}}))^{-1} (\frac{1}{\beta}\tilde{\boldsymbol{X}})^\top \boldsymbol{y} && (3)\\
&= \beta^2 (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} (\frac{1}{\beta}\tilde{\boldsymbol{X}})^\top \boldsymbol{y} && (4)\\
&= \frac{1}{\beta}\beta^2 (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^\top \boldsymbol{y} && (5)\\
&= \beta \hat{\mathbf{w}} && (6)
\end{aligned}$$

Note: The problem does not require the formal derivation shown above. It can e.g. be solved by fitting a model on some data, X, and then scaling X and refitting the model to see what happens to the weights.

Question 17. Suppose a neural network is trained to classify sounds. As part of training the network, we wish to select between eight different ways to transform the attributes (i.e., $S = 8$ models) and estimate the generalization error of the optimal choice. In the outer loop we opt for $K_1 = 3$-fold cross-validation, and in the inner $K_2 = 5$-fold cross-validation. The time taken to *train* a single model is 20 minutes, and this can be assumed constant for each fold. If the time taken to test a model is negligible, what is the total time required for the 2-level cross-validation procedure?

A. 360 minutes

B. 2460 minutes

C. 4920 minutes

D. 2400 minutes

E. Don't know.

Solution 17. Going over the 2-level cross-validation algorithm we see the total number of models to be *trained* is:

$$K_1(K_2S + 1) = 123$$

Multiplying by the time taken to train a single model we obtain a total training time of 2460 minutes and therefore answer B is correct.

Question 18. Suppose a logistic regression model has been trained on a two-dimensional dataset based on two attributes from the the Sound Classification dataset. Now consider four test observations:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$$
$$\boldsymbol{x}_2 = \begin{bmatrix} -1.0 & 0.6 \end{bmatrix}^\top$$
$$\boldsymbol{x}_3 = \begin{bmatrix} -0.5 & -0.5 \end{bmatrix}^\top$$
$$\boldsymbol{x}_4 = \begin{bmatrix} -0.91 & -0.16 \end{bmatrix}^\top$$

Additionally, the following probabilities are known

$$p(y = Machine \mid \boldsymbol{x}_1) = 0.73$$
$$p(y = Machine \mid \boldsymbol{x}_2) = 0.47$$
$$p(y = Machine \mid \boldsymbol{x}_3) = 0.73$$

Determine which one of the following statements is true.

Hint: The problem can be solved without complicated numerical computation.

A. $p(y = Machine \mid \boldsymbol{x}_4) < p(y = Machine \mid \boldsymbol{x}_2)$

B. $p(y = Machine \mid \boldsymbol{x}_4) = p(y = Machine \mid \boldsymbol{x}_1)$

C. $p(y = Machine \mid \boldsymbol{x}_4) > p(y = Machine \mid \boldsymbol{x}_2)$

D. $p(y = Machine \mid \boldsymbol{x}_4) > p(y = Machine \mid \boldsymbol{x}_1)$

E. Don't know.

Solution 18. The correct answer is C.

The easieist approach is to argue based on the knowlegde of the decision boundries and location of the points.

Initially, we note that $p(y = Machine \mid o) = 1 - p(y = Natural \mid o)$. We know that the isoprobablity lines are straight line for LR, and since o_1 and o_3 have the same probability and together they define the probablity line of 0.73.

All other isoprobability lines must be parallel to this for an LR curve. We note that observation o_2 lies on the 0.47 isoprobability line.

We need to figure out where $o4$ lies releative to various isoprobability lines. A quick sketch or inspection demonstartes that o_4 lies such that $p(y = Machine \mid \boldsymbol{x}_4) > p(y = Machine \mid \boldsymbol{x}_2)$.

For completeness, we enumerate all the relevant probabilities:

$p(y = Machine \mid o_1) = 0.73, p(y = Natural \mid o_1) = 0.27$

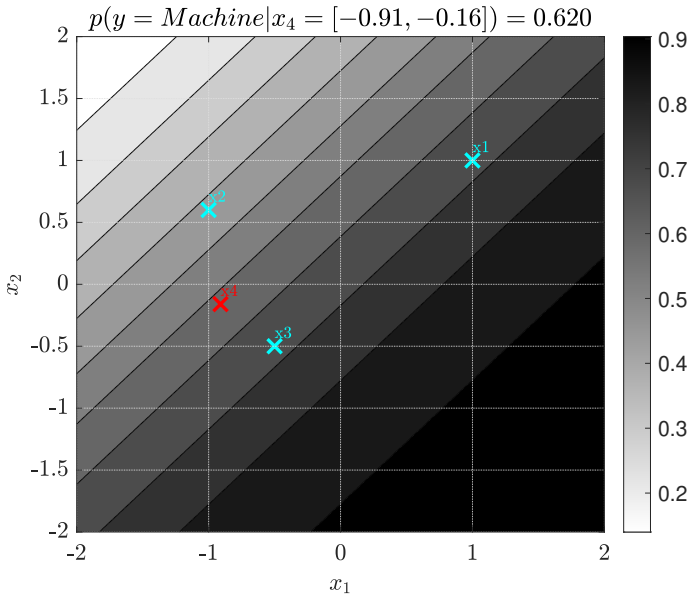

Figure 8: **Solution**

$p(y = Machine \mid o_2) = 0.47, p(y = Natural \mid o_2) = 0.53$

$p(y = Machine \mid o_3) = 0.73, p(y = Natural \mid o_3) = 0.27$

$p(y = Machine \mid o_4) = 0.62, p(y = Natural \mid o_4) = 0.38$

An alternative but cumbersome approach would be to nummerically solve for w and then insert o_3 to nummerical determine $p(y \mid o_3)$.

	$i=1$	$i=2$	$i=3$	$i=4$
x_i	-0.5	0.39	1.19	-1.08
$y_{r,i}$	-0.86	-0.61	1.37	0.1

Table 7: A subset of the Sound Classification dataset.

Question 19. Consider once again the Sound Classification dataset, but this time we limit ourselves to $N = 4$ observations and a single attribute from the full dataset with the i'th observation denoted by x_i. In particular, we consider transformations of the i'th observation, x_i, resulting in $\tilde{\boldsymbol{x}}_i$ (a column vector).

The goal is to predict y_r, and to achieve this; we will apply ridge regression with the prediction rule $f(\tilde{\boldsymbol{x}}_i; w_0, \boldsymbol{w}) = w_0 + \tilde{\boldsymbol{x}}_i^\top \boldsymbol{w}$. Ridge regression determines the constant offset, w_0, and weights $\boldsymbol{w} = \begin{bmatrix} w_1 & w_2 & \dots & w_M \end{bmatrix}^\top$ (M is the dimensionality of $\tilde{\boldsymbol{x}}_i$), by minimizing a cost function of the form:

$$E_\lambda = \sum_{i=1}^{N} (y_{r,i} - f(\tilde{\boldsymbol{x}}_i; w_0, \boldsymbol{w}))^2 + \lambda \|\boldsymbol{w}\|_2^2$$

Suppose w_0 and $\boldsymbol{w}$ were learned based on the dataset in Table 7 using a specific transformation of x. The weigths were learned with $\lambda = 0.25$ resulting in $E_{\lambda=0.25} \approx 0.2$, $w_0 = 0.0$ and $\boldsymbol{w} = \begin{bmatrix} 0.39 & 0.77 \end{bmatrix}^\top$.

Standardization was performed on the data matrix with the transformed observations, $\tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{\boldsymbol{x}}_1 & \tilde{\boldsymbol{x}}_2 & \tilde{\boldsymbol{x}}_3 & \tilde{\boldsymbol{x}}_4 \end{bmatrix}^\top$, prior to learning the parameters by subtracting the column-wise mean and dividing by the unbiased estimate of the column-wise standard deviation.

Which one of the following transformations was used when learning w_0 and $\boldsymbol{w}$?

A. $\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & x_i^3 \end{bmatrix}^\top$

B. $\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & sin(x_i) & x_i^2 \end{bmatrix}^\top$

C. $\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & sin(x_i) \end{bmatrix}^\top$

D. $\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & x_i^2 \end{bmatrix}^\top$

E. Don't know.

Solution 19. The correct answer is D.

We notice that option is clearly wrong as it would require a 3 dimensional weigth vector, w.

This leaves three options for which we need to compute the various terms of the loss:

The regularization term is constant for all options, i.e. $\lambda \|\boldsymbol{w}\|_2^2 = 0.186$.

The squared loss is relatively easy to compute as $w_0 = 0$, thus

$$\sum_{i=1}^{N} (y_i - f([x\, x^2]; w_0, \boldsymbol{w}))^2 = 0.016$$

$$\sum_{i=1}^{N} (y_i - f([x\, x^3]; w_0, \boldsymbol{w}))^2 = 1.867$$

$$\sum_{i=1}^{N} (y_i - f([x\, sin(x)]; w_0, \boldsymbol{w}))^2 = 2.033$$

Combined we get

$\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & x_i^2 \end{bmatrix}^\top$: $0.016 + 0.186 = 0.2025$

$\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & x_i^3 \end{bmatrix}^\top$: $1.867 + 0.186 = 1.9258$

$\tilde{\boldsymbol{x}}_i = \begin{bmatrix} x_i & sin(x_i) \end{bmatrix}^\top$: $2.033 + 0.186 = 2.076$

i.e. only option D matchees the desired value of $E(\lambda)$.

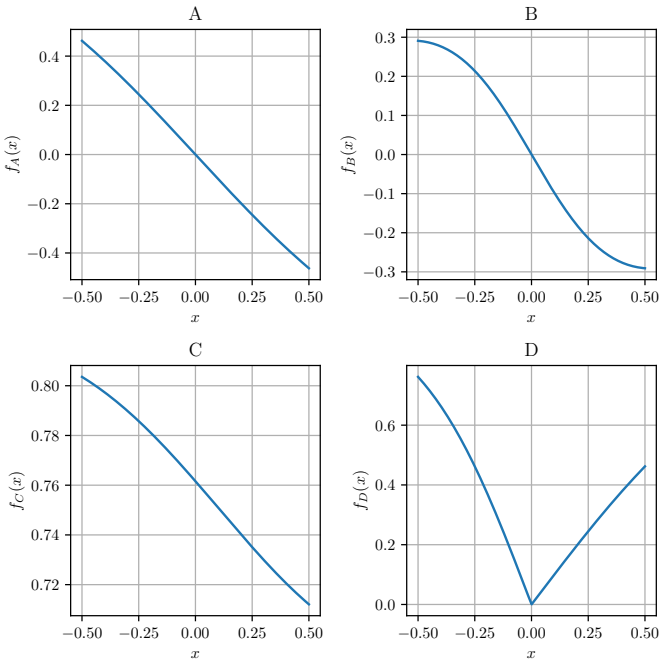

Figure 9: The output of a neural network $f(x)$ as a function of the input x for four different choices of the activation function $h^{(1)}$ for the hidden layer.

Question 20. Consider a two-layer neural network $f : \mathbb{R} \to \mathbb{R}$ for regression with two hidden units (i.e. $\boldsymbol{z}^{(1)} \in \mathbb{R}^2$) and of the form

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

where $\tilde{\boldsymbol{x}} = \begin{bmatrix} 1 & x \end{bmatrix}$, $\tilde{\boldsymbol{z}}^{(1)} = \begin{bmatrix} 1 & z_1^{(1)} & z_2^{(1)} \end{bmatrix}$, $h^{(1)}(x)$ is the activation function for the hidden layer that is applied elementwise, and $h^{(2)}(x)$ is the activation function of the output layer.

Assume that $h^{(2)}(x)$ is the hyperbolic tangent activation function and the network's weights are given by

$$\boldsymbol{W}^{(1)} = \begin{bmatrix} 0 & 0 \\ 1 & -2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{W}^{(2)} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Figure 9 shows $f(x)$ for four different choices of the activation function $h^{(1)}$. Which one of the following statements is correct?

A. For f_A, $h^{(1)}$ is the **rectified linear unit**
For f_B, $h^{(1)}$ is the **identity function**
For f_C, $h^{(1)}$ is the **hyperbolic tangent**
For f_D, $h^{(1)}$ is the **logistic sigmoid**

B. For f_A, $h^{(1)}$ is the **identity function**
For f_B, $h^{(1)}$ is the **logistic sigmoid**
For f_C, $h^{(1)}$ is the **hyperbolic tangent**
For f_D, $h^{(1)}$ is the **rectified linear unit**

C. For f_A, $h^{(1)}$ is the identity function
For f_B, $h^{(1)}$ is the hyperbolic tangent
For f_C, $h^{(1)}$ is the logistic sigmoid
For f_D, $h^{(1)}$ is the rectified linear unit

D. For f_A, $h^{(1)}$ is the **hyperbolic tangent**
For f_B, $h^{(1)}$ is the **identity function**
For f_C, $h^{(1)}$ is the **logistic sigmoid**
For f_D, $h^{(1)}$ is the **rectified linear unit**

E. Don't know.

Solution 20. First, let $\mathrm{id}(x) = x$ denote the identity function, tanh denote hyperbolic tangent, σ denote the logistic sigmoid function, and $\mathrm{ReLu}(x) = \max(0, 1)$ denote the rectified linear unit. We note that $\mathrm{id}(0) = \tanh(0) = \mathrm{ReLU}(0) = 0$, while $\sigma(0) = 0.5$.

For $x = 0$ we have that $\tilde{\boldsymbol{x}}^\intercal \boldsymbol{W}^{(1)} = \begin{bmatrix} 0 & 0 \end{bmatrix}^\intercal$. For $h^{(1)} = \sigma$, this means that $z^{(1)} = \begin{bmatrix} \sigma(0) & \sigma(0) \end{bmatrix}^\intercal = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}^\intercal$ and therefore $f(0) = \tanh(0.5+0.5) \approx 0.76$. For id, tanh and ReLU, we have $z^{(1)} = \begin{bmatrix} 0 & 0 \end{bmatrix}^\intercal$ and consequently $f(x) = \tanh(0 + 0) = 0$. Looking at the curves in Figure 9 for $x = 0$, this means that f_C uses the logistic sigmoid.

We will now evaluate the remaining networks in $x = 0.5$ and first find that

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

This gives us that the following possible values of $f(0.5)$:

$$f(0.5) = \tanh(\mathrm{id}(0.5) + \mathrm{id}(-1)) \approx \tanh(-0.5) \approx -0.46$$
$$f(0.5) = \tanh(\tanh(0.5) + \tanh(-1)) \approx \tanh(-0.30) \approx -0.29$$
$$f(0.5) = \tanh(\mathrm{ReLU}(0.5) + \mathrm{ReLU}(-1)) \approx \tanh(0.5) \approx 0.46$$

Comparing these function values to Figure 9, we see that f_A uses the **identity function**, f_B, $h^{(1)}$ uses **hyperbolic tangent** and f_D uses the **rectified linear unit** for the activation function of the hidden layer.

We could also calculate that

$$f(0.5) = \tanh(\sigma(0.5) + \sigma(-1)) \approx \tanh(0.89) \approx 0.71$$

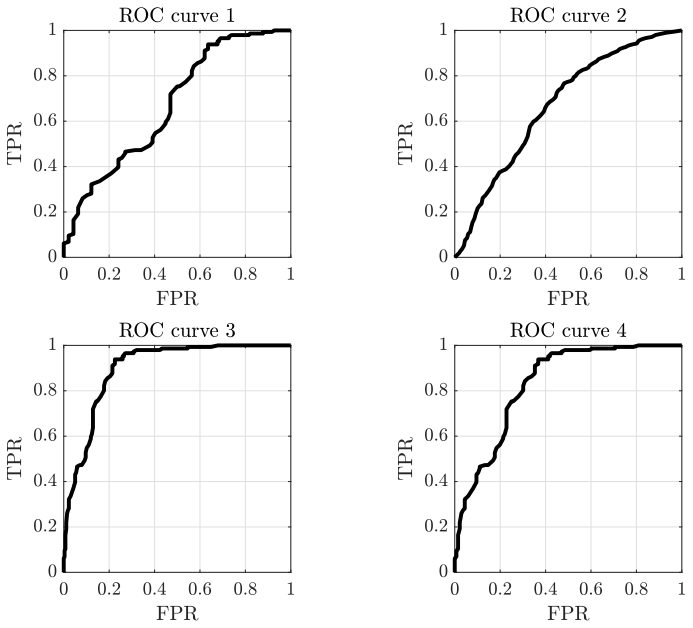

Figure 10: Candidate ROC curves for the classifier.

Question 21. We wish to predict whether an observation from the Sound Classification dataset (see Table 1) belongs to the Machine class (or not). To accomplish this, we fit a logistic regression model to the dataset, and for each observation $\boldsymbol{x}_i, y_i$ obtain a class-probability prediction $\hat{y}_i \in [0, 1]$. We threshold the class-probability at different values θ thereby obtaining, for each value of θ, the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These are plotted as functions of θ in Figure 11. Which of the receiver-operator characteristic (ROC) plots shown in Figure 10 corresponds to these graphs?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

Solution 21. From the TP curve (left-most value) we get that the total number of positive-class observations are $P = 146$ and from the TN curve (right-most value) we get $N = 136$. The simplest approach is to compute a point on the ROC curve. Most values will do, however we choose the point corresponding to $\theta = 0.5$, at which the number of false positives is $FP = 42$ and

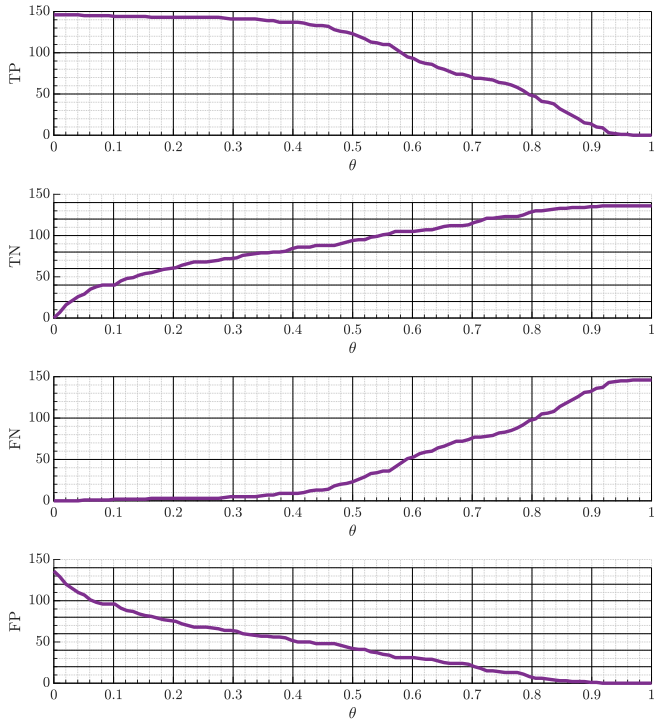

Figure 11: TP, TN, FN, and FP as functions of the threshold value θ.

true positives is $TP = 123$. We therefore see that the following point must lie on the ROC curve:

$$(\text{fpr}, \text{tpr}) = (\frac{FP}{N}, \frac{TP}{P}) = (0.31, 0.84)$$

this rules out all options except D.

Question 22. In a market basket problem, we consider two item sets X and Y. Assume that the support of X is $\frac{3}{5}$, the support of Y is $\frac{8}{15}$, and the confidence of the association rule $X \to Y$ is $\frac{1}{6}$. What is the confidence of $Y \to X$?

A. $\text{conf}(Y \to X) = \frac{1}{10}$

B. $\text{conf}(Y \to X) = \frac{4}{27}$

C. $\text{conf}(Y \to X) = \frac{1}{6}$

D. conf$(Y \to X) = \frac{3}{16}$

E. Don't know.

Solution 22. Option D is correct Using the definition of confidence, we have that

$$\text{supp}(X \cup Y) = \text{conf}(X \to Y)\,\text{supp}(X) = \frac{1}{6}\frac{3}{5} = \frac{1}{10}.$$

Then we can calculate the support of $Y \to X$

$$\text{conf}(Y \to X) = \frac{\text{supp}(Y \cup X)}{\text{supp}(Y)} = \frac{\frac{1}{10}}{\frac{8}{15}} = \frac{3}{16}$$

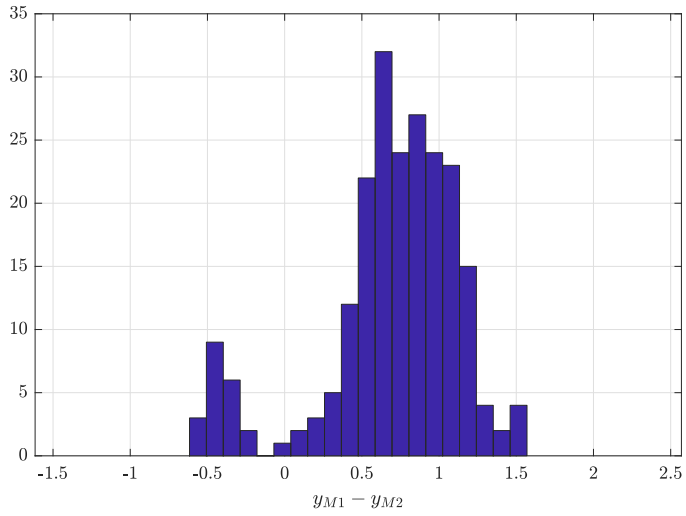

Figure 12: Histogram of $y_{M1} - y_{M2}$ based on all test observations.

Question 23. Consider a statistical procedure to determine if there is a difference between two regression models, M1 and M2, using a paired test as described in Section 11.3.6 of the lecture notes.

The distribution of the paired differences in predictions between the two models, $y_{M1} - y_{M2}$, using a 5-fold cross validation setup is illustrated as a histogram in Figure 12.

Determine which combination of estimated difference ($\hat{z}$), confidence interval (CI) and p-value is the only plausible answer.

A. $\hat{z} = 0.69$, $CI = [0.63, 0.75]$, **p-value**< 0.05

B. $\hat{z} = -1.05$, $CI = [-1.29, -0.81]$, p-value< 0.05

C. $\hat{z} = 0.63$, $CI = [0.57, 0.69]$, p-value< -0.05

D. $\hat{z} = 0.76$, $CI = [0.59, 0.76]$, p-value< 0.05

E. Don't know.

Solution 23. Option A is correct (estimate, CI and p-value all feasible and teh estimate, 0.69, matches the mean as observed from the histogram.

B) The estimate of -1.29 does not align with the mean as observed from the histogram (approx. 0.6-0.8)
C) A p-value less than zero is not meaningful! The estimate might be meaningful,
D) The CI is not centered around the estimate, which would be the case in a standard paired test as described in section 11.3.6 of the book.

	b_1	b_2	b_3	b_4	b_5	b_6
o_1	1	1	0	1	0	0
o_2	1	0	1	1	0	0
o_3	0	0	1	1	1	1
o_4	0	1	1	1	1	1
o_5	1	1	0	1	1	0
o_6	1	1	0	1	1	1
o_7	0	1	1	1	0	1
o_8	0	1	0	1	1	1
o_9	0	1	0	1	1	1
o_{10}	1	1	1	1	1	0

Table 8: Binarized version of the Sound Classification dataset. Each of the features b_i are obtained by taking a feature x_i and letting $b_i = 1$ correspond to a value x_i greater than the median (otherwise $b_i = 0$). The colors indicate classes such that the black observations $\{o_1,\ o_2,\ o_3,\ o_4\}$ belong to class C_1 (corresponding to Machine), and the red observations $\{o_5,\ o_6,\ o_7,\ o_8,\ o_9,\ o_{10}\}$ belong to class C_2 (corresponding to Natural).

Question 24. Consider again the Sound Classification dataset. Suppose we wish to predict the class label y using a decision tree model, and to improve performance, we wish to apply AdaBoost. We apply AdaBoost (as described in Algorithm 7 of the lecture notes) to the binarized version of the dataset in Table 8. Suppose that the classifier learned on dataset D_1 in the first round of boosting is

$$f_1(b_1, b_2, b_3, b_4, b_5, b_6) = \begin{cases} C_1, & \text{if } b_3 = 1 \text{ and } b_4 = 1 \\ C_2, & \text{otherwise.} \end{cases}$$

What is the importance of the classifier f_1?

A. $\alpha_1 \approx 0.84$

B. $\alpha_1 \approx 0.42$

C. $\alpha_1 \approx 0.70$

D. $\alpha_1 \approx 0.30$

E. Don't know.

Solution 24. Option B is correct

From table 8, we see that o_1, o_7 and o_{10} are miss classified by f_1. Since we have $N = 10$ observations in out dataset, this means that

$$\epsilon_1 = \frac{1}{10} \cdot 3,$$

and following Algorithm 7, we find the importance of the classifier to be

$$\alpha_1 = \frac{1}{2} \log \frac{1 - \frac{3}{10}}{\frac{3}{10}} = \frac{1}{2} \log \frac{7}{3} = \log \frac{\sqrt{21}}{3} \approx 0.42\,.$$

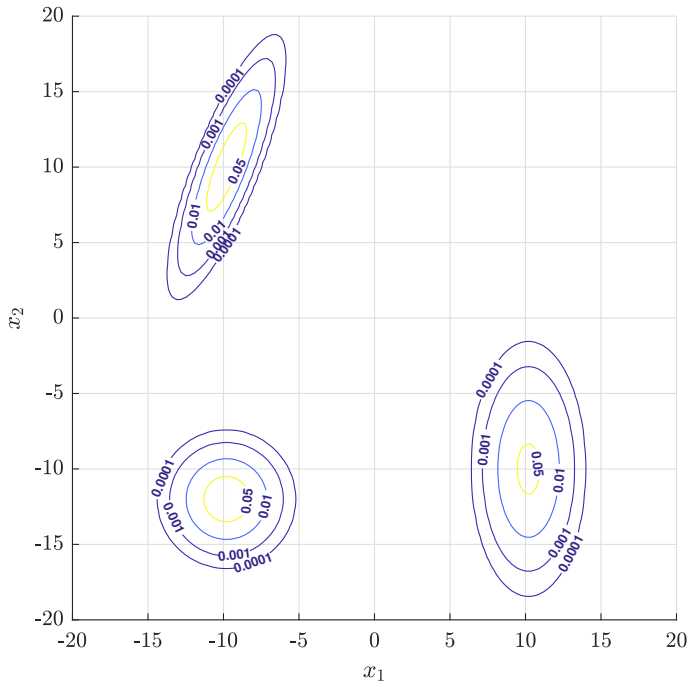

Figure 13: Contour lines corresponding to the density of the three individual components in the Gaussian Mixture Model.

Question 25. A Gaussian Mixture Model with $K = 3$ components has been trained on a centered dataset (i.e., the column-wise mean has been subtracted) with two attributes. The model is defined as

$$p(\boldsymbol{x}) = \sum_{k=1}^{3} w_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \,.$$

The contour lines of the densities corresponding to the individual mixture components are shown in Figure 13.

The component weights and covariance matrices are:

$$w_1 = 0.5, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.1 & 2.0 \\ 2.0 & 5.5 \end{bmatrix}$$

$$w_2 = 0.49, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.1 & 0.0 \\ 0.0 & 5.5 \end{bmatrix}$$

$$w_3 = 0.01, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{bmatrix}$$

Determine which one of the following options corresponds to the first principal component direction, $\boldsymbol{v}_1$,

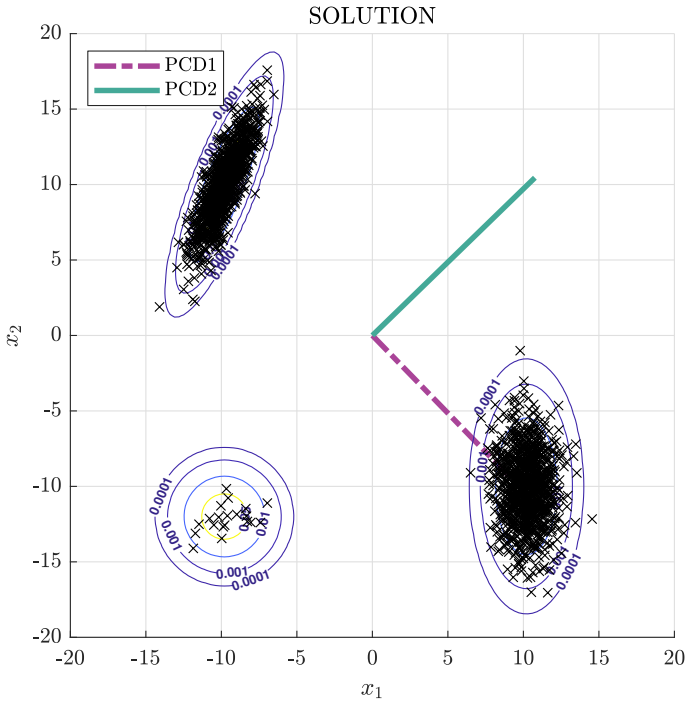

Figure 14: SOLUTION: Contour lines corresponding to the density of the three individual components in the Gaussian Mixture Model.

and the second principal component direction, $\mathbf{v}_2$, of the dataset modeled by $p(\mathbf{x})$.

A. $\boldsymbol{v}_1 \approx \begin{bmatrix} -0.3 & -1.0 \end{bmatrix}^\top, \boldsymbol{v}_2 \approx \begin{bmatrix} -1.0 & 0.3 \end{bmatrix}^\top$

B. $\boldsymbol{v}_1 \approx \begin{bmatrix} -0.3 & -1.0 \end{bmatrix}^\top, \boldsymbol{v}_2 \approx \begin{bmatrix} 0.0 & 1.0 \end{bmatrix}^\top$

C. $\boldsymbol{v}_1 \approx \begin{bmatrix} 0.7 & -0.7 \end{bmatrix}^\top, \boldsymbol{v}_2 \approx \begin{bmatrix} 0.7 & 0.7 \end{bmatrix}^\top$

D. $\boldsymbol{v}_1 \approx \begin{bmatrix} 0.7 & 0.7 \end{bmatrix}^\top, \boldsymbol{v}_2 \approx \begin{bmatrix} 0.7 & -0.7 \end{bmatrix}^\top$

E. Don't know.

Solution 25.
Option C is correct.
First, the components are identified to establish their weight from the shape parameters. Secondly, we realise that one of the weights is very small and would therefore not be associated with many observations. Hence, the PCDs are mainly determined by the two components with large weights as illustrated in Figure 14.

Question 26. Consider again the Sound Classification dataset in Table 1. We apply backward selection to find an interpretable linear regression model which uses a subset of the $M = 6$ attributes to predict the Percieved annoyance measure (PAM) y_r. Recall backward selection chooses models based on the test error as determined by cross-validation, and in our case we use the hold-out method to generate a single test/training split.

Suppose backward selection ends up selecting the attributes x_1, x_4, x_5, and x_6, what is the minimal number of models which were *tested* in order to obtain this result?

A. 15 models

B. 19 models

C. 16 models

D. 10 models

E. Don't know.

Solution 26.

The correct answer is C.

Since we use backward selection, we first have to evaluate a single model with all features.

Then we evaluated all models with a single missing feature giving an additional M models.

One of these models were selected and variable selection proceeded at the next level where an additonal $M - 1$ models were evaluated.

This continues until all possible models with an additional missing feature has a higher cost than the currently best model i.e. none were selected and the method terminated.

For 4 features this gives

$$1 + (M) + (M - 1) + (M - 2)$$

,

or specifically with $M = 6$

$$1 + 6 + 5 + 4 = 16$$

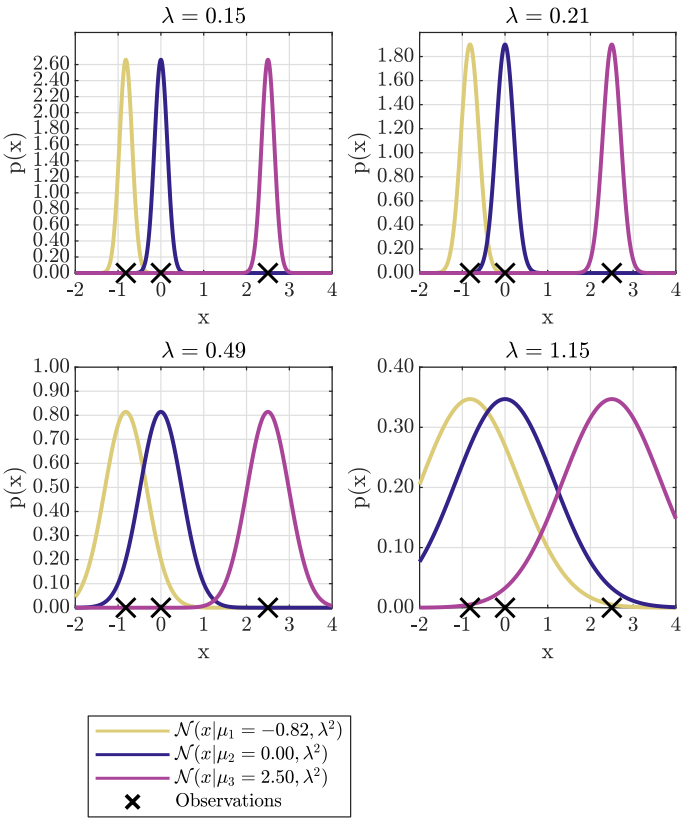

Figure 15: KDE applied to a 1-dimensional dataset using a Gaussian kernel with kernel widths $\lambda = \begin{bmatrix}0.15 & 0.21 & 0.49 & 1.15\end{bmatrix}$.

Question 27. A small 1-dimensional dataset with $N = 3$ observations has been subsampled from the Sound Classification dataset, specifically:

$$\boldsymbol{X} = \begin{bmatrix} -0.82 \\ 0.0 \\ 2.5 \end{bmatrix}$$

A kernel density estimator (KDE) has been fitted to the small dataset with different λ using the standard Gaussian kernel (i.e., the individual Gaussian components in the KDE have variance $\sigma^2 = \lambda^2$). The resulting densities are illustrated in Figure 15.

Consider a leave-one-out (LOO) procedure for computing the generalization error for different λ as in Section 20.1.1 of the lecture notes. The test log likelihood for the whole dataset is defined as

$$\mathcal{L}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} log\left[\sum_{j\neq i}\frac{1}{N-1}\mathcal{N}(x_i|x_j,\lambda^2)\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N} log\left[\sum_{j\neq i}\frac{1}{N-1}M_{ij}\right] = \frac{1}{N}\sum_{i=1}^{N} L_i$$

i.e. L_i corresponds to the log density, $\log p(x)$, of the KDE constructed without x_i and evaluated at x_i.

Determine which value of λ in Figure 15 results in $L = \begin{bmatrix}-2.3 & -2.3 & -13.91\end{bmatrix}$.

A. $\lambda = 1.15$

B. $\lambda = 0.15$

C. $\lambda = 0.21$

D. $\lambda = 0.49$

E. Don't know.

Solution 27. The correct answer is D, i.e., $\lambda = 0.49$.

Focusing on the correct answer, we obtain the answer by the following procedure. $\tilde{p}(x)$ denotes the KDE when an observation is removed.:

Step 1) Leave out $x = -0.82$, i.e., component 1 centered at $x = \mu_1 = -0.82$ (yellow curve): We evaluate the resulting two-component (component 2 (blue) and component 3 (purpule)) KDE at the held out observation $x = -0.82$.

$$\mathcal{N}(x = -0.82 \mid \mu_2, \lambda^2) \approx 0.2 \text{ (blue curve at } -0.82)$$

$$\mathcal{N}(x = -0.82 \mid \mu_3, \lambda^2) \approx 0 \text{ (purple curve at } -0.82)$$

Thus $\tilde{p}(x = -0.82) \approx (0.2 + 0)/2 = 0.1$ and $L_1 = \log \tilde{p}(x) = -2.3$.

Note: The observant student* should/*could end here as only* $\lambda = 0.49$ *gives* the correct *values of L.

Step 2) Leave out component 2 centered at $x = \mu_2 = 0.0$ (blue curve): We evaluate the resulting two-component (component 1 (yellow) and component 3 (purpule)) KDE at the held out observation $x = 0.0$.

$$\mathcal{N}(x = 0.0 \mid \mu_1, \lambda^2) \approx 0.2 \text{ (yellow curve at } 0.0)$$

$$\mathcal{N}(x = 0.0 \mid \mu_3, \lambda^2) \approx 0 \text{ (purple curve at } 0.0)$$

Thus $\tilde{p}(x = 0.0) \approx (0.2 + 0)/2 = 0.1$ and $L_2 = \log \tilde{p}(x) = -2.3$.

Step 3) Leave out $x = 2.5$, i.e. component 3 centered at $x = \mu_2 = 2.5$ (purple curve): We evaluate the resulting two-component (component 1 (yellow) and component 2 (blue)) KDE at the held out observation $x = 2.5$.

$$\mathcal{N}(x = 2.5 \mid \mu_1, \lambda^2) \approx 0 \text{ (not exactly 0; yellow curve)}$$

$$\mathcal{N}(x = 2.5 \mid \mu_2, \lambda^2) \approx 0 \text{ (not exactly 0; blue curve)}$$

Thus $\tilde{p}(x = 0.0) \approx (0+0)/2 \approx 0.0$ (but not exatcly 0!). Thus $L_3 << L1$ and $L2$ is very low relative to $L1$ and L_2 (it is not $-\infty$ as we know $\tilde{p}(x)$ is not exactly zero).

Overall the values of L_1 and L_2 leaves only D as the correct answer.

The full (exact) evaluations of $\tilde{p}(x)$ and L are given below for completeness:

$$\lambda = 0.15 : \tilde{p}(x) = \begin{bmatrix}0 & 0 & 0\end{bmatrix}$$
$$\lambda = 0.21 : \tilde{p}(x) = \begin{bmatrix}0 & 0 & 0\end{bmatrix}$$
$$\lambda = 0.49 : \tilde{p}(x) = \begin{bmatrix}0.1 & 0.1 & 0.0\end{bmatrix}$$
$$\lambda = 1.15 : \tilde{p}(x) = \begin{bmatrix}0.14 & 0.15 & 0.02\end{bmatrix}$$

$$\lambda = 0.15 : L = \log \tilde{p}(x) = \begin{bmatrix}-14.66 & -14.66 & -138.6\end{bmatrix}$$
$$\lambda = 0.21 : L = \log \tilde{p}(x) = \begin{bmatrix}-7.68 & -7.68 & -70.91\end{bmatrix}$$
$$\lambda = 0.49 : L = \log \tilde{p}(x) = \begin{bmatrix}-2.3 & -2.3 & -13.91\end{bmatrix}$$
$$\lambda = 1.15 : L = \log \tilde{p}(x) = \begin{bmatrix}-1.99 & -1.89 & -3.96\end{bmatrix}$$

For completeness, we report the full $\mathcal{L}(\lambda)$ for the four options (not required to answer the question).

$$E^{test} = \begin{bmatrix}-55.97274350665651\\ -28.754385811304935\\ -6.170686733754928\\ -2.6134008381902736\end{bmatrix}$$