Technical University of Denmark

Written examination: December 16th 2020, 9:00–13:30.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: $4\frac{1}{2}$ 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.

When you hand in your answers, you have to upload two files:

1. Your answers to the multiple choice exam using the `answers.txt` file.
2. Your written full explanations of how you found the answer to each question not marked as E ("Don't know") either as a `.zip`-file (with `bmp`, `png`, `tiff` and `jpg` as allowed file formats, if you take photos of your handwritten explanations) or as a PDF file (if you write a document with your answers)[1].

Failing to timely upload both documents will count as not having handed in the exam.

Guessing on an answer is not allowed for the online exam, as each answer has to include an accompanying argumentation in writing for the answer.

Questions, where the answers in the `answers.txt` file (file 1) differ from the explanation (file 2) or where explanations are insufficient or unreadable, will be treated as "Don't know". Systematic discrepancy between the answers in the two hand-in files will potentially count as an attempt of cheating the exam.

Only in the exceptional case, where the exam submission system stops working, you should send your two files in a single email to `exam-02450@compute.dtu.dk`. Files sent to the email address will only be accepted in the rare event that the exam submission system fails.

In the event that there is an error in the exam, then you should contact Morten Mørup (`mmor@dtu.dk`) using your study/DTU email.

Answers:

1	2	3	4	5	6	7	8	9	10
C	B	C	A	C	D	A	C	C	B
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
C	D	B	C	C	A	D	B	A	B

[1] The *original* file format must be either zip or PDF.

21	22	23	24	25	26	27
D	A	B	C	C	D	C

No.	Attribute description	Abbrev.
x_1	Bill[2] length (millimeters)	Bill length
x_2	Bill depth (millimeters)	Bill depth
x_3	Flipper length (millimeters)	Flipper length
x_4	Body mass (grams)	Mass
x_5	Penguin sex (1=male, 2=female)	Sex
x_6	Study year (2007, 2008, or 2009)	Year
y		Species

Table 1: Description of the features of the Palmer Penguins dataset used in this exam. The dataset is collected in the Palmer Archipelago (Antarctica) and contains $M = 6$ attributes for three different species of penguins. The dataset has been pre-processed such that data objects with missing values have been removed. The binary attribute *Sex* is encoded as an integer, where a male penguin takes the value $x_5 = 1$ and a female penguin takes the value $x_5 = 2$. For classification, the objective is to predict the species, and the output variable y is taking values $y = 1$ (corresponding to an Adelie), $y = 2$ (corresponding to a Gentoo), and $y = 3$ (corresponding to a Chinstrap). After removing missing values, there are $N = 333$ observations in total.

Question 1. The main dataset used in this exam is the Palmer Penguins dataset[3] described in Table 1. We will consider the type of an attribute *as the highest level* it obtains in the type-hierarchy (nominal, ordinal, interval and ratio). Which one of the following statements is true about the types of the attributes $x_1, \ldots, x_6$ and the output y in the Palmer Penguins dataset?

A. All attributes except x_5 (*Sex*) are ratio.

B. x_5 (*Sex*) and x_6 (*Year*) are both ordinal.

C. x_5 (*Sex*) and y (*Species*) are both nominal.

D. x_4 (*Mass*) and x_6 (*Year*) are both interval.

E. Don't know.

Solution 1. The problem is solved by simply thinking about what the attributes represent and comparing them to the definition in the different types. Recall that

- Nominal is a type that only allow comparison (equal or different)
- Ordinal allows ordering (but not differences)
- Interval allows differences but no (physically well-defined) zero
- Ratio is a type with a zero with a well-defined meaning

With these definitions, we see that

x_1 **(*Bill length*)** is ratio

x_2 **(*Bill depth*)** is ratio

x_3 **(*Flipper length*)** is ratio

x_4 **(*Mass*)** is ratio

x_5 **(*Sex*)** is nominal

x_6 **(*Year*)** is interval

y **(*Species*)** is nominal

and therefore option C is correct.

[2] Bill is synonymous for beak.

[3] Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. `https://allisonhorst.github.io/palmerpenguins/`

	mean	$x_{p=25\%}$	$x_{p=50\%}$	$x_{p=75\%}$
Bill length	43.99	39.45	44.50	48.63
Bill depth	17.16	15.60	17.30	18.70
Flipper length	200.97	190.00	197.00	213.00
Mass	4207.06	3550.00	4050.00	4781.25

Table 2: Summary statistics of four attributes from the Palmer Penguins dataset. The column $x_{p=25\%}$ refers to the 25'th percentile of the given attribute, $x_{p=50\%}$ to the median and $x_{p=75\%}$ to the 75'th percentile.

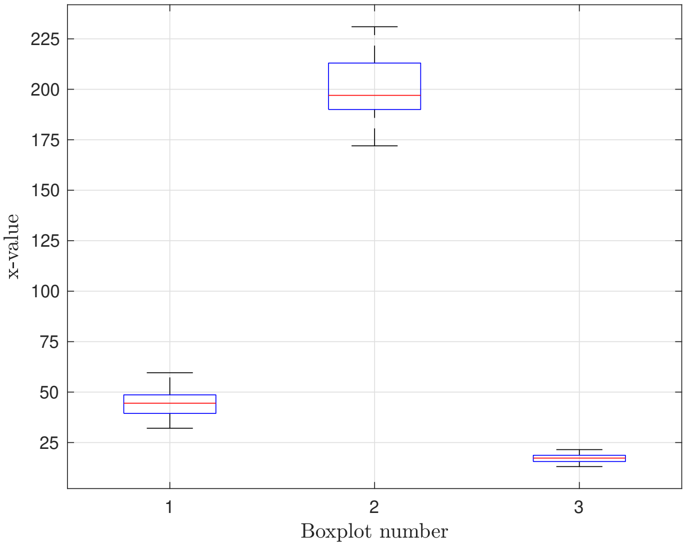


Figure 1: Boxplots corresponding to three of the variables with summary statistics indicated in Table 2 but not necessarily in that order.

Question 2. Table 2 contains the summary statistics of the four first attributes (x_1, x_2, x_3, x_4) from the Palmer Penguins dataset. Which boxplots in Figure 1 match which attributes?

A. Attribute *Bill length* corresponds to boxplot 1, *Bill depth* corresponds to boxplot 2, and *Flipper length* corresponds to boxplot 3.

B. Attribute *Bill length* corresponds to boxplot 1, *Flipper length* corresponds to boxplot 2, and *Bill depth* corresponds to boxplot 3.

C. Attribute *Flipper length* corresponds to boxplot 1, *Mass* corresponds to boxplot 2, and *Bill length* corresponds to boxplot 3.

D. Attribute *Bill depth* corresponds to boxplot 1, *Flipper length* corresponds to boxplot 2, and *Bill length* corresponds to boxplot 3.

E. Don't know.

Solution 2. We can read off the medians (red line) of the boxplots. We observe that

- Boxplot 1 has median between 25 and 50, and the only attribute with a median in this interval is *Bill length.*
- Boxplot 2 has median between 175 and 200, and the only attribute with a median in this interval is *Flipper length.*
- Boxplot 3 has median below 25, and the only attribute with a median below this value is *Flipper depth.*

Therefore B is the only correct answer.

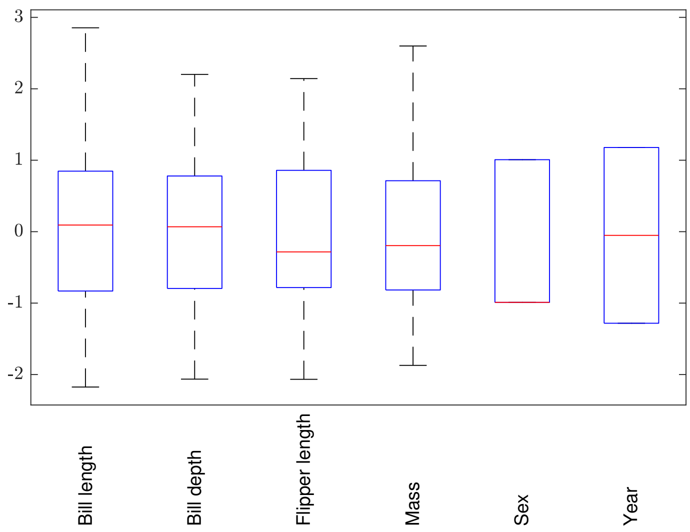


Figure 2: Boxplots of the six attributes (x_1, x_2, x_3, x_4, x_5, x_6) from the Palmer Penguins dataset. The attributes are standardized.

Question 3. Figure 2 shows boxplots of the six attributes (x_1, x_2, x_3, x_4, x_5, x_6) from the Palmer Penguins dataset. The attributes are standardized (i.e., the mean has been subtracted and the attributes divided by their standard deviations). Which one of the following statements about the original dataset can be concluded from the boxplots (i.e., based only on information regarding the data provided by Figure 2)?

A. The variance of *Bill length* is larger than the variance of *Flipper length*.

B. For *Flipper length* the mean and median values coincide.

C. There are more male penguins than female penguins.

D. *Bill length* and *Bill depth* have positive correlation.

E. Don't know.

Solution 3.

- A is false, since the boxplots are based on standardized data (with unit variance) and therefore it cannot inform about the variance of the original data (in fact bill length has smaller variance than flipper length).
- B is false, since you cannot read of the mean from a boxplot.
- C is true, since the boxplot shows that the median is close to -1. As males are encoded as 1 and female as 2, the negative median after standardization indicates that more had the lower value (in fact there are 168 males and 165 females).
- D is false, since boxplots do not inform about correlation, but only summarizes the individual feature (in fact it turns out that the two features have negative correlation).

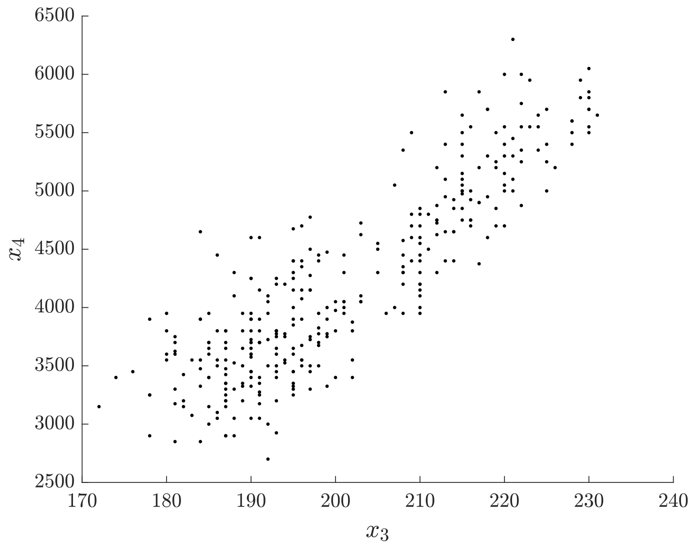


Figure 3: Black dots show attributes x_3 and x_4 of the Palmer Penguins dataset.

Question 4. Figure 3 shows a scatterplot for the two attributes x_3 (*Flipper length*) and x_4 (*Mass*) of the Palmer Penguins dataset. The two attributes have a positive correlation coefficient $\rho = 0.87$ and covariance $\text{cov}(x_3, x_4) = 9852$. Which of the following are the best estimates for the variance of x_3 and variance of x_4?

A. $\sigma^2_{x_3} = 196$ and $\sigma^2_{x_4} = 648025$

B. $\sigma^2_{x_3} = 38$ and $\sigma^2_{x_4} = 298$

C. $\sigma^2_{x_3} = 648025$ and $\sigma^2_{x_4} = 196$

D. $\sigma^2_{x_3} = 298$ and $\sigma^2_{x_4} = 38$

E. Don't know.

Solution 4. From Figure 3 we can observe that the spread of x_3 is smaller than the spread of x_4, i.e., $\sigma_{x_3} < \sigma_{x_4}$. This means that we can rule out answer C and D.

The correlation coefficient between x_3 and x_4 is defined as

$$\rho = \frac{\text{cov}(x_3, x_4)}{\sigma_{x_3}\sigma_{x_4}}.$$

If we use the number from A, we find that

$$\rho = \frac{9852}{\sqrt{196}\sqrt{648025}} \approx 0.87 \ ,$$

which means that this is the correct answer.

Question 5. A Principal Component Analysis (PCA) is carried out on the Palmer Penguins dataset in Table 1 based on the attributes x_1, x_2, x_3, x_4. The data is standardized by (i) substracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized data matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.45 & -0.60 & -0.64 & 0.15 \\ -0.40 & -0.80 & 0.43 & -0.16 \\ 0.58 & -0.01 & 0.24 & -0.78 \\ 0.55 & -0.08 & 0.59 & 0.58 \end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix} 30.19 & 0.0 & 0.0 & 0.0 \\ 0.0 & 16.08 & 0.0 & 0.0 \\ 0.0 & 0.0 & 11.07 & 0.0 \\ 0.0 & 0.0 & 0.0 & 5.98 \end{bmatrix}.$$

Which one of the following statements is *correct*?

A. The first principal component accounts for less than 50 percent of the variance.

B. The first two principal components account for more than 90 percent of the variance.

C. The first three principal components account for more than 95 percent of the variance.

D. The last principal component accounts for more than 3 percent of the variance.

E. Don't know.

Solution 5. The correct answer is D. To see this, recall the variance explained by a given component k of the PCA is given by

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

where M is the number of attributes in the dataset being analyzed. The values of σ_k can be read off as entry $\sigma_k = S_{kk}$ where $\boldsymbol{S}$ is the diagonal matrix of the SVD computed above. We therefore find the variance explained by components x_1, x_2, x_3 is:

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2} = 0.9731.$$

Question 6. Consider again the principal component analysis described in Question 5. Recall that the d'th column of $\boldsymbol{V}$ defines the d'th principal component direction. Which one of the following statements is *correct*?

A. Any single observation with a high negative projection onto the second principal component will in general have high negative values for all four features x_1, x_2, x_3, x_4.

B. The first column vector of $\boldsymbol{V}$ is longer than the second column vector of $\boldsymbol{V}$ measured by the Euclidean norm (2-norm).

C. The first principal component primarily separates penguins with relatively short *Flipper length* and high *Body mass* from penguins with relatively long *Flipper length* and low *Body mass.*

D. The fourth principal component primarily separates penguins with relatively short *Flipper length* and high *Body mass* from penguins with relatively long *Flipper length* and low *Body mass.*

E. Don't know.

Solution 6.

- A is false: since all the coordinates in the second PC are negative, an observation with high negative values for all features will have a high positive projection onto the second principal component.
- B is false: since all column vectors have unit length.
- C is false: The first PC has nearly the same positive value for both *Flipper length* (x_3) and *Body mass* (x_4). Therefore, flipping the values of x_3 and x_4 would nearly give the same projection.
- D is true: The forth PC has a large negative coefficient for flipper length (x_3) and a large positive coefficient for body mass (x_4). Therefore, the forth PC primarily separates penguins with relatively low x_3 (short flipper length) and high x_4 (high body mass) from penguins with relatively high x_3 (long flipper length) and low x_4 (low body mass).

Question 7. Based on the principal component analysis described in Question 5, Figure 4 shows 2D scatter plots for the Palmer Penguin dataset projected onto different combinations of the principal components. The class labels y (penguin species *Adelie, Gentoo, Chinstrap*) are indicated in the plots.

Consider a new observation $\boldsymbol{x}$ for which the four features x_1, x_2, x_3, x_4 are standardized (by subtracting the mean from each column and dividing by the standard deviations), resulting in the point $\tilde{\boldsymbol{x}}$ with coordinates:

$$\tilde{x}_1 = -1,\ \tilde{x}_2 = -1,\ \tilde{x}_3 = -1,\ \tilde{x}_4 = 1.$$

Consider the point projected onto the principal components. Furthermore, consider a k-nearest neighbor classification with $k = 1$ using the Euclidean distance measure and using the projected dataset shown in the individual 2D plots as the training data. In which one of the following 2D projections (shown in Figure 4) will the 1-nearest neighbor classifier classify the point $\tilde{\boldsymbol{x}}$ as a *Chinstrap*?

A. Data projected onto PC1 and PC4.

B. Data projected onto PC2 and PC4.

C. Data projected onto PC2 and PC3.

D. Data projected onto PC3 and PC4.

E. Don't know.

Solution 7. The point $\tilde{\boldsymbol{x}}^T = [-1, -1, -1, 1]$ is projected to the vector with coordinates

$$\tilde{\boldsymbol{x}}^T \boldsymbol{V} = [-0.08, 1.33, 0.56, 1.37]^T$$

which can be seen to only being closest to a '+' in one figure (PC1 vs. PC4), which is illustrated in Figure 5.

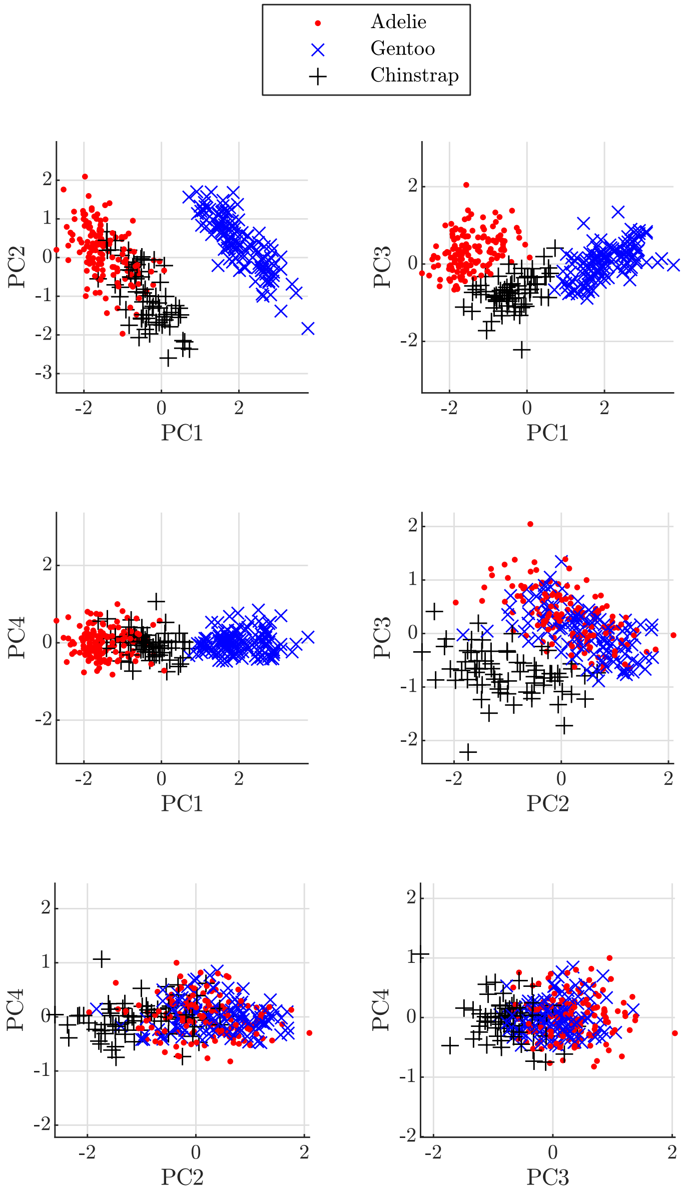


Figure 4: Projection of the Palmer Penguins dataset onto different combinations of two principals components obtained from the principal component analysis described in question 5. The three species of penguins are indicated with an *Adelie* as a red dot, a *Gentoo* as a blue 'x' and a *Chinstrap* as a black '+'.

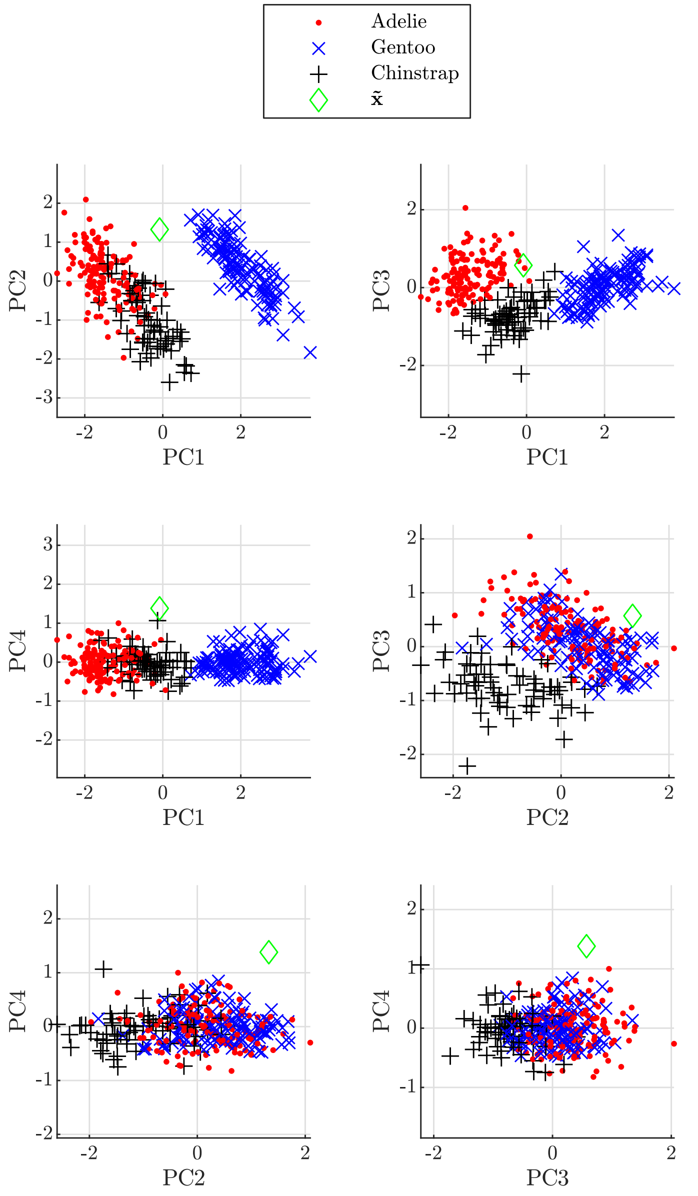
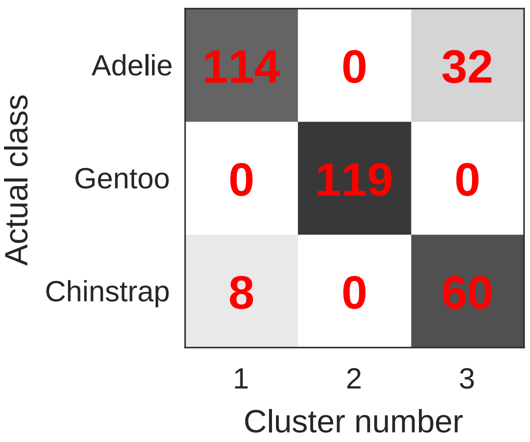


Figure 5: The point $\tilde{\boldsymbol{x}}$ (green diamond) project onto the different combinations of two principals components.



Actual class	1	2	3
Adelie	114	0	32
Gentoo	0	119	0
Chinstrap	8	0	60

Figure 6: Confusion matrix obtained by comparing the actual classes in the Palmer Penguins dataset with three cluster obtained by the k-means algorithm for $k = 3$.

Question 8. The k-means algorithm with $k = 3$ was applied to the Palmer Penguins dataset. Figure 6 shows the confusion matrix, where the cluster assignment obtained by the k-means algorithm is compared to the ground truth class label.

We wish to compare the quality of the k-means clustering, Z, to the ground truth clustering, Q. What is the Rand index (Rand similarity) between Z and Q?

A. $R(Z, Q) = 0.72$

B. $R(Z, Q) = 0.79$

C. $R(Z, Q) = 0.87$

D. $R(Z, Q) = 0.96$

E. Don't know.

Solution 8. First note that the confusion matrix is equivalent to the counting matrix $\boldsymbol{n}$ that can be used to calculate the Rand similarity.

We can calculate the number of times Z and Q agrees two observations are S or are not in D the same cluster by

$$\begin{aligned} S &= \sum_{k=1}^{3}\sum_{m=1}^{3} \frac{n_{km}(n_{km}-1)}{2} \\ &= \frac{114\cdot 113 + 32\cdot 31 + 119\cdot 118 + 8\cdot 7 + 60\cdot 59}{2} \\ &= 15756 \end{aligned}$$

and

$$\begin{aligned} D &= \frac{N(N-1)}{2} - \sum_{k=1}^{K} \frac{n_k^Z(n_k^Z-1)}{2} - \sum_{m=1}^{M} \frac{n_m^Q(n_m^Q-1)}{2} + S \\ &= \frac{333\cdot 332}{2} - \frac{146\cdot 145 + 119\cdot 118 + 68\cdot 67}{2} \\ &\quad - \frac{122\cdot 121 + 119\cdot 118 + 92\cdot 91}{2} + 15756 \\ &= 32562. \end{aligned}$$

We then find the Rand similarity to be

$$R(Z, Q) = \frac{S + D}{\frac{1}{2}N(N-1)} = \frac{15756 + 32562}{\frac{1}{2}\cdot 333\cdot 332} = 0.87.$$

Question 9. Which one of the following machine-learning models is not trained by fitting any parameters?

A. Linear regression applied to data with one attribute.

B. Neural network with no hidden layers.

C. K nearest neighbor classification using the $K = 1$ nearest neighbour classification rule.

D. Multinomial regression applied to data with three output classes.

E. Don't know.

Solution 9. KNN is instance based with no learned parameters when K is fixed.

Question 10. Three of the following actions will typically reduce the amount of over-fitting, and one of them will typically increase it. Which option will typically *increase* the amount of over-fitting?

A. Reduce the number of attributes.

B. Reduce the amount of training data.

C. Select a less complex model.

D. Add model regularisation.

E. Don't know.

Solution 10. Reducing the amount of training data is the only option that will typically increase the amount of over-fitting. The other choices will typically decrease over-fitting.

$p(\hat{x}_1, \hat{x}_2 \vert y)$	$y = 1$	$y = 2$	$y = 3$
$\hat{x}_1 = 0, \hat{x}_2 = 0$	0.23	0.15	0.07
$\hat{x}_1 = 0, \hat{x}_2 = 1$	0.75	0	0.01
$\hat{x}_1 = 1, \hat{x}_2 = 0$	0	0.85	0.16
$\hat{x}_1 = 1, \hat{x}_2 = 1$	0.02	0	0.76

Table 3: Probability of observing particular values of $\hat{x}_1$ and $\hat{x}_2$ conditional on y.

Question 11. Consider the Palmer Penguins dataset from Table 1. We wish to predict the species based on the attributes *Bill length* and *Bill depth* using a Bayes classifier. Suppose the attributes have been binarized such that $\hat{x}_1 = 0$ corresponds to $x_1 \leq 44.5$ (and otherwise $\hat{x}_1 = 1$) and $\hat{x}_2 = 0$ corresponds to $x_2 \leq 17.3$ (and otherwise $\hat{x}_2 = 1$). Suppose the probability for each of the configurations of $\hat{x}_1$ and $\hat{x}_2$ conditional on the species y are as given in Table 3 and the prior probability of the species are

$$p(y = 1) = 0.44, \; p(y = 2) = 0.36, \; p(y = 3) = 0.20.$$

Using this, what is then the probability an observation is *Chinstrap* ($y = 3$) given that $\hat{x}_1 = 1$ and $\hat{x}_2 = 0$?

A. $p(y = 3|\hat{x}_1 = 1, \hat{x}_2 = 0) = 0.033$

B. $p(y = 3|\hat{x}_1 = 1, \hat{x}_2 = 0) = 0.085$

C. $p(y = 3|\hat{x}_1 = 1, \hat{x}_2 = 0) = 0.095$

D. $p(y = 3|\hat{x}_1 = 1, \hat{x}_2 = 0) = 0.158$

E. Don't know.

Solution 11. The problem is solved by a simple application of Bayes' theorem:

$$p(y = 3|\tilde{x}_1 = 1, \tilde{x}_2 = 0)$$
$$= \frac{p(\tilde{x}_1 = 1, \tilde{x}_2 = 0|y = 3)p(y = 3)}{\sum_{k=1}^{3} p(\tilde{x}_1 = 1, \tilde{x}_2 = 0|y = k)p(y = k)}$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_1 = 1, \tilde{x}_2 = 0|y)$ in Table 3. Inserting the values we see option D is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0	725	800	150	1000	525	600	500	400	850
o_2	725	0	75	575	275	1250	1325	226	325	125
o_3	800	75	0	650	200	1325	1400	300	400	51
o_4	150	575	650	0	850	675	750	350	250	700
o_5	1000	275	200	850	0	1525	1600	500	600	150
o_6	525	1250	1325	675	1525	0	75	1025	925	1375
o_7	600	1325	1400	750	1600	75	0	1100	1000	1450
o_8	500	226	300	350	500	1025	1100	0	100	350
o_9	400	325	400	250	600	925	1000	100	0	450
o_{10}	850	125	51	700	150	1375	1450	350	450	0

Table 4: The pairwise Euclidean 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 Palmer Penguins dataset (recall that $M = 4$). 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 $\{o_1, \; o_2, \; o_3, \; o_4, \; o_5\}$ belong to class C_1 (corresponding to an Adelie), the red observations $\{o_6, \; o_7\}$ belong to class C_2 (corresponding to a Gentoo), and the blue observations $\{o_8, \; o_9, \; o_{10}\}$ belong to class C_3 (corresponding to a Chinstrap). The distances are rounded to integers.

Question 12. Table 4 shows the pairwise Euclidean distances between 10 observations $o_1, o_2, \ldots, o_{10}$ from the Palmer Penguins dataset. To examine if observation o_2 may be an outlier, we will calculate the average relative density using the Euclidean distance based on the observations given in Table 4 only. We recall that the KNN density and average relative density (ard) for the observation $\boldsymbol{x}_i$ are given by:

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

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

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors. What is the average relative density for observation o_2 for $K = 2$ nearest neighbors?

A. 0.01

B. 0.37

C. 0.68

D. 0.73

E. Don't know.

Solution 12.

To solve the problem, first observe the $k = 2$ neighborhood of o_2 and density is:

$$N_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = \{o_3, o_{10}\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = 0.01$$

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

$$N_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = \{o_{10}, o_2\}, \quad N_{\boldsymbol{X}_{\setminus 10}}(\boldsymbol{x}_{10}) = \{o_3, o_2\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = 0.016, \text{density}_{\boldsymbol{X}_{\setminus 10}}(\boldsymbol{x}_{10}) = 0.011.$$

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

Question 13. Again, consider the pairwise Euclidean distances between the 10 observations $o_1, o_2, \dots, o_{10}$ from the Palmer Penguins dataset in Table 4. Consider the two clusters

$$C_2 = \{o_6, o_7\}$$
$$C_3 = \{o_8, o_9, o_{10}\}$$

What is the distance between C_2 and C_3 using average linkage as the linkage function?

A. $d(C_2, C_3) \approx 1108.3$

B. $d(C_2, C_3) \approx 1145.8$

C. $d(C_2, C_3) \approx 1183.3$

D. $d(C_2, C_3) \approx 1450.0$

E. Don't know.

Solution 13. We have that the average linkage is given by

$$\begin{aligned} d(C_2, C_3) &= \frac{\sum_{x \in C_1, y \in C_2} ||x - y||_2}{|C_1||C_2|} \\ &= \frac{1025 + 925 + 1375 + 1100 + 1000 + 1450}{2 \cdot 3} \end{aligned}$$

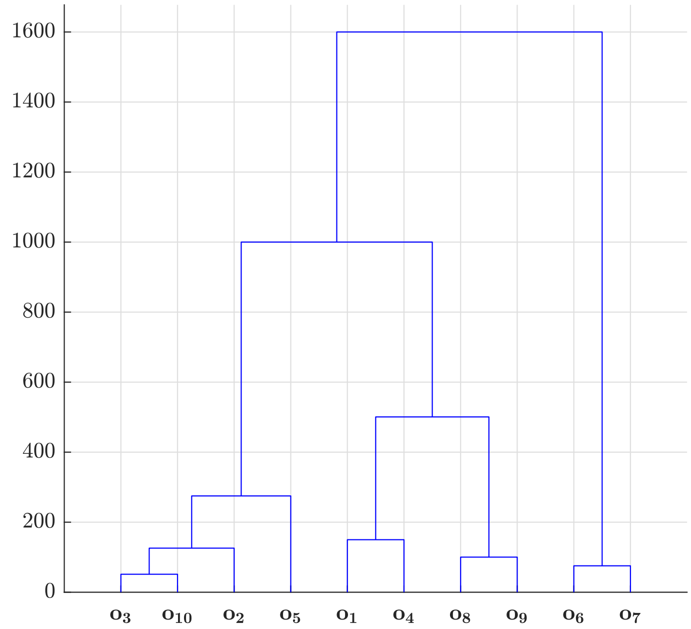


Figure 7: Hierarchical clustering of the 10 observations in Table 4.

Question 14. A hierarchical clustering is applied to the 10 observations in Table 4 using *maximum* linkage, and the result is shown in Figure 7. Which one of the following set of clusters *cannot* be obtained from the dendrogram by applying a valid cutoff?

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

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

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

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

E. Don't know.

Solution 14.

- A can be achieved by a cutoff at 400.
- B can be achieved by a cutoff at 200.
- C cannot be achieved, as o_2 cannot be split from $\{o_3, o_{10}, o_2\}$ without also splitting $\{o_1, o_4\}$. This is hence the correct answer.
- D can be achieved by by a cutoff at 125.
- Don't know.

	f_1	f_2	f_3	f_4	x_5
o_{11}	0	0	1	0	1
o_{12}	1	0	1	0	1
o_{13}	0	1	0	0	1
o_{14}	1	0	1	1	1
o_{15}	1	0	0	0	1
o_{16}	0	0	0	1	1
o_{17}	1	0	1	0	2
o_{18}	0	1	0	0	2
o_{19}	0	0	1	0	2
o_{20}	1	0	0	0	2

Table 5: Binarized version of 10 observations from the Palmer Penguins 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 binary feature x_5 indicates the sex of the penguin.

Question 15. Table 5 shows $N = 10$ observations from the Palmer Penguins dataset. The data is processed to produce four new binary features such that $f_i = 1$ corresponds to a value x_i greater than the median. Here, we make the following conditional independence assumption about the binary features

$$p(f_1, f_2, f_3, f_4|x_5) = p(f_1|x_5)p(f_2, f_3|x_5)p(f_4|x_5).$$

Using this assumption and Bayes' rule, we obtain the classifier

$$p(x_5|f_1, f_2, f_3, f_4) = \frac{p(f_1|x_5)p(f_2, f_3|x_5)p(f_4|x_5)p(x_5)}{\sum_{k=1}^{2} p(f_1|x_5 = k)p(f_2, f_3|x_5 = k)p(f_4|x_5 = k)p(x_5 = k)}.$$

Consider a new observations o_{21} with the following values for the binary features:

	f_1	f_2	f_3	f_4
o_{21}	0	0	1	0

What is the probability that o_{21} is classified as *male* ($x_5 = 1$) using the classifier above?

A. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{1}{10}$

B. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{1}{5}$

C. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{1}{2}$

D. $p(x_5 = 1|f_1 = 0, f_2 = 0, f_3 = 1, f_4 = 0) = \frac{10}{19}$

E. Don't know.

Solution 15. From Table 5 we find that

- $p(x_5 = 1) = \frac{6}{10}$
- $p(f_1 = 0 | x_5 = 1) = \frac{3}{6}$
- $p(f_2 = 0, f_3 = 1 | x_5 = 1) = \frac{3}{6}$
- $p(f_4 = 0 | x_5 = 1) = \frac{4}{6}$
- $p(x_5 = 2) = \frac{4}{10}$
- $p(f_1 = 0 | x_5 = 2) = \frac{2}{4}$
- $p(f_2 = 0, f_3 = 1 | x_5 = 2) = \frac{2}{4}$
- $p(f_4 = 0 | x_5 = 2) = \frac{4}{4}$

Therefore we find that

$$\begin{aligned} p(x_5 = 1 | f_1 = 0, f_2 = 1, f_3 = 1, f_4 = 0) \\ = \frac{\frac{3}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{6}{10}}{\frac{3}{6} \cdot \frac{3}{6} \cdot \frac{4}{6} \cdot \frac{6}{10} + \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{4} \cdot \frac{4}{10}} = \frac{1}{2} \end{aligned}$$

Question 16. We again consider the Palmer Penguins dataset from Table 1 and the $N = 10$ observations we already encountered in Table 5. Recall that, the data is processed to produce four new binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[4], and we thereby arrive at the $N \times M = 10 \times 4$ binary matrix in Table 5. In this exercise, we do not consider attribute x_5.

Then the matrix can be considered as representing $N = 10$ transactions $o_{11}, o_{12}, \dots, o_{20}$ and $M = 4$ items $f_1, f_2, \dots, f_4$. Which of the following options represents all (non-empty) itemsets with support greater than 0.25 (and only itemsets with support greater than 0.25)?

A. $\{f_1\}$, $\{f_3\}$, $\{f_1, f_3\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_1, f_3\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_1, f_3\}$, $\{f_1, f_4\}$, $\{f_3, f_4\}$, $\{f_1, f_3, f_4\}$

D. $\{f_1\}$, $\{f_3\}$

E. Don't know.

Solution 16. Recall the support of an itemset is the number of rows containing all items in the itemset divided by the total number of rows. Therefore, to have a support of 0.25, an itemset needs to be contained in 3 rows. It is easy to see this rules out all options except *A*.

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

Question 17. We again consider the binary matrix from Table 5 as a market basket problem consisting of $N = 10$ transactions $o_{11}, \ldots, o_{20}$ and $M = 4$ items $f_1, \ldots, f_4$.

What is the *confidence* of the rule $\{f_1, f_4\} \to \{f_3\}$?

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

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

C. The confidence is $\frac{7}{10}$

D. The confidence is 1

E. Don't know.

Solution 17. The confidence of the rule is computed as

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

Question 18. Consider classifying the Palmer Penguins dataset according to the species attribute y. The dataset contains 146 observations for $y = 1$ (*Adelie*), 119 observation for $y = 2$ (*Gentoo*) and 68 observation for $y = 3$ (*Chinstrap*).

During training of a decision tree, the classification error has been used as impurity measure

$$\text{classError}(v) = 1 - \max_c p(c|v)$$

where $p(c|v)$ denotes the fraction of observations belonging to class c at a given node v. The tree is constructed by Hunts algorithm using the purity gain

$$\Delta = \text{classError}(parent) - \sum_{k=1}^{2} \frac{N(v_k)}{N} \text{classError}(v_k)$$

where N is the total number of observations at the parent node and $N(v_k)$ is the number of observations associated with the k^{th} child node, v_k.

After the first split you learn that the left node contains all the *Adelie* and *Gentoo* observations while the right node contains all the *Chinstrap* observations. What is the purity gain for the split?

A. The purity gain is $\dfrac{1}{5}$.

B. The purity gain is $\dfrac{68}{333}$**.**

C. The purity gain is $\dfrac{68}{265}$.

D. The purity gain is $\dfrac{265}{333}$.

E. Don't know.

Solution 18. With I denoting the class error, we obtain:

- $I(parent) = 1 - \frac{146}{333}$
- $I(left) = 1 - \frac{146}{146+119} = 1 - \frac{146}{265}$
- $I(right) = 1 - \frac{68}{68} = 0$
- $\Delta = \left(1 - \frac{146}{333}\right) - \frac{265}{333} \cdot \left(1 - \frac{146}{265}\right) - 0 = 1 - \frac{265}{333} = \frac{68}{333}$

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

Table 6: For each of the $N = 7$ observations (first column), the table indicates the true class labels y^{true} (second column) and the predicted outputs of the AdaBoost classifier (third column) for the first round of boosting $t = 1$.

Question 19. Consider again the Palmer Penguins dataset of Table 1. Suppose we limit ourselves to $N = 7$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_1 and x_2. We use a KNN classification model ($K = 3$) to this dataset and apply AdaBoost to improve the performance. After the first round of boosting, we obtain predictions and the true class labels as tabulated in Table 6. Given this information, how will the AdaBoost update the weights $\boldsymbol{w}$?

A. $\begin{bmatrix}0.083 & 0.083 & 0.5 & 0.083 & 0.083 & 0.083 & 0.083\end{bmatrix}$

B. $\begin{bmatrix}0.165 & 0.165 & 0.008 & 0.165 & 0.165 & 0.165 & 0.165\end{bmatrix}$

C. $\begin{bmatrix}0.152 & 0.152 & 0.085 & 0.152 & 0.152 & 0.152 & 0.152\end{bmatrix}$

D. $\begin{bmatrix}0.01 & 0.01 & 0.937 & 0.01 & 0.01 & 0.01 & 0.01\end{bmatrix}$

E. Don't know.

Solution 19.

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

$$\{y_3\}$$

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

$$\epsilon_{t=1} = \sum_i w_i(t)(1 - \delta_{f_t(x_i),y_i}) = 0.143$$

From this, we compute α_t as

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} = 0.896$$

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

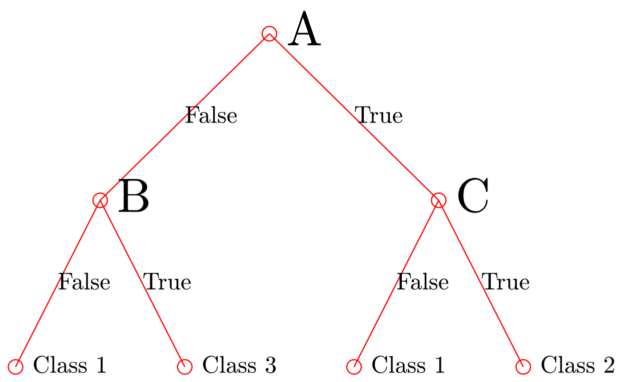


Figure 8: Example classification tree.

Question 20. We consider an artificial dataset of $N = 4000$ observations. The dataset is classified according to a decision tree of the form shown in Figure 8 resulting in a partition into classes indicated by the colors/markers in Figure 9. What is the correct rule assignment to the nodes in the decision tree?

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

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

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

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

E. Don't know.

Solution 20.

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

The resulting decision boundaries for each of the options are shown in Figure 10 and it follows answer B is correct.

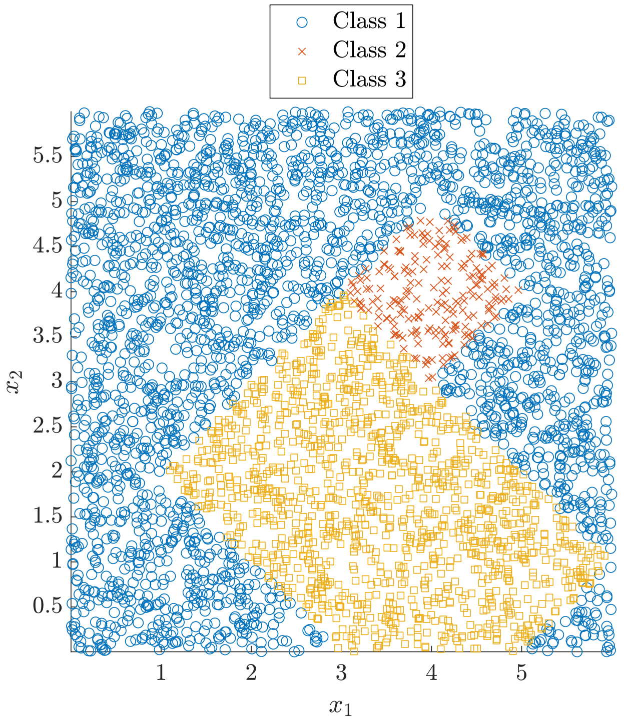


Figure 9: Classification boundaries.

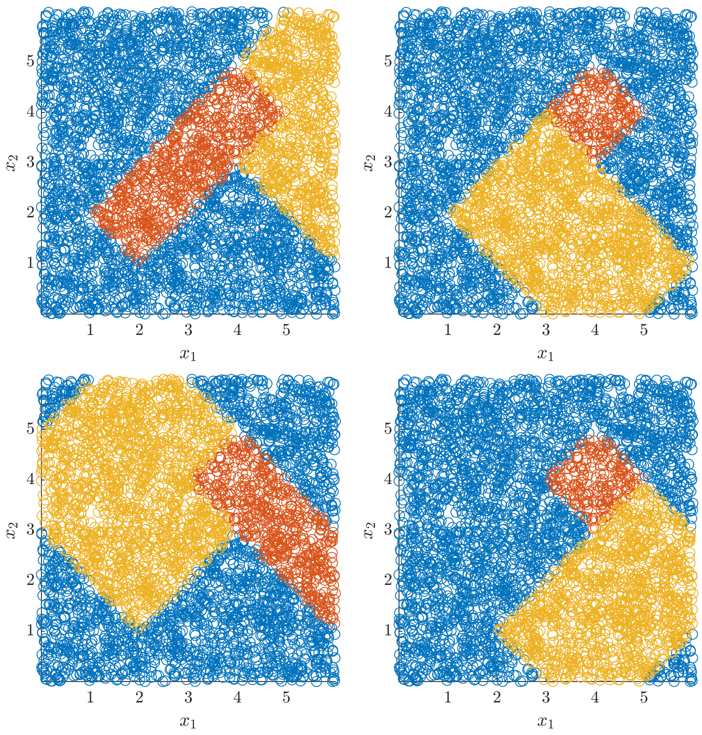


Figure 10: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

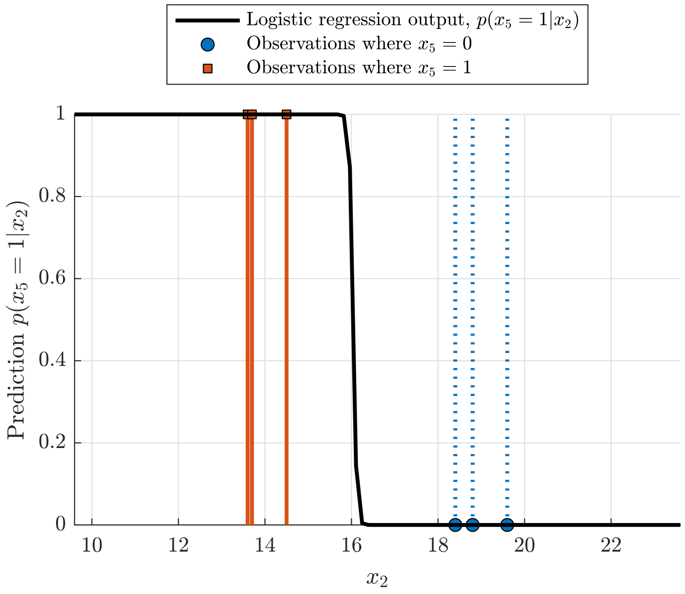


Figure 11: Output of a logistic regression classifier trained on 6 observations from the dataset.

Question 21. Consider again the Palmer Penguins dataset. To simplify the setup further, we select just 6 observations and train a logistic regression classifier using only the feature x_2 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term). As output, we want to predict the Penguin sex (x_5). To be consistent with the lecture notes, we label the output as $x_5 = 0$ (corresponding to *male*) and $x_5 = 1$ (corresponding to *female*).

In Figure 11 is shown the predicted output probability of an observation belonging to the positive class, $p(x_5 = 1|x_2)$. What are the weights?

A. $\boldsymbol{w} = \begin{bmatrix} 423.49 \\ 48.16 \end{bmatrix}$

B. $\boldsymbol{w} = \begin{bmatrix} 0.0 \\ -46.21 \end{bmatrix}$

C. $\boldsymbol{w} = \begin{bmatrix} 0.0 \\ -27.89 \end{bmatrix}$

D. $\boldsymbol{w} = \begin{bmatrix} 418.94 \\ -26.12 \end{bmatrix}$

E. Don't know.

Solution 21. The solution is easily found by simply computing the predicted $\hat{x}_5 = p(x_5 = 1|x_2)$-value for an appropriate choice of x_2. Notice that

$$p(x_5 = 1|x_2) = \sigma(\tilde{\boldsymbol{x}}_2^T \boldsymbol{w})$$

If we select $x_2 = 16$ and select the weights as in option D we find $\hat{x}_5 = p(x_5 = 1|x_2) = 0.735$, in good agreement with the figure. On the other hand, for the weights in option A we obtain $\hat{x}_5 = 1$, for C that $\hat{x}_5 = 0$ and finally for B that $\hat{x}_5 = 0$. We can therefore conclude that D is correct.

Question 22. We fit a GMM to a single feature x_2 from the Palmer Penguins dataset. Recall the density of a 1D GMM is

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

and suppose that the identified values of the mixture weights are

$$w_1 = 0.13, \ w_2 = 0.55, \ w_3 = 0.32$$

and the parameters of the multivariate normal densities:

$$\mu_1 = 18.347, \ \mu_2 = 14.997, \ \mu_3 = 18.421$$
$$\sigma_1 = 1.2193, \ \sigma_2 = 0.986, \ \sigma_3 = 1.1354.$$

According to the GMM, what is the probability an observation at $x_0 = 15.38$ is assigned to cluster $k = 2$?

A. 0.975

B. 0.389

C. 0.213

D. 0.042

E. Don't know.

Solution 22. Recall γ_{ik} is the posterior probability that observation i is assigned to mixture component 2 which can easily be obtained using Bayes' theorem. We see that:

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

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

$$p(x_i|z_{i1} = 1) = 0.017$$
$$p(x_i|z_{i2} = 1) = 0.375$$
$$p(x_i|z_{i3} = 1) = 0.01$$

Combining these with the class-assignment probabilities we obtain:

$$\gamma_{i,2} = 0.975$$

and conclude the solution is A.

Fold	M_1/M_2	$M_1/\overline{M}_2$	$\overline{M}_1/M_2$	$\overline{M}_1/\overline{M}_2$
1	86	8	7	10
2	65	15	11	20
3	79	5	17	10

Table 7: Outcome of cross-validation. Rows are combination of outcomes of the two models.

Question 23. We will consider the Palmer Penguins dataset and two models for predicting the class label y. Specifically, let M_1 be a $K = 1$ nearest neighbor classification model and M_2 a $K = 5$ nearest neighbor classification model. To compare them statistically, we perform $K = 3$ fold cross-validation, and for each fold we record the number of observations where both models are correct (as M_1/M_2), M_1 is correct and M_2 wrong (as $M_1/\overline{M}_2$), and so on. The outcome can be found in Table 7.

We want to test if the two classifiers perform differently. The null hypothesis is that the models have the same performance. Given the values of the binomial cumulative distribution function in Table 8 and assuming that the null hypothesis is true, what is the p-value from McNemar's test (rounded to two decimals)?

A. 0.23

B. 0.45

C. 0.84

D. 0.90

E. Don't know.

Solution 23. Since the cross-validation folds are non-overlapping, we can easily find by summing columns 2 and 3 in Table 7.

- n_1 the total number of times that M_1 is correct and M_2 is incorrect, and
- n_2 the total number of times that M_1 is incorrect and M_2 is correct.

We find that $n_1 = 28$ and $n_2 = 35$. We can calculate the p-value from the CDF of the binomial distribution as

$$p = 2\text{cdf}_{\text{binom}}(m = \min\{n_1, n_2\}|\theta = 1/2, N = n_1 + n_2)$$
$$= 2\text{cdf}_{\text{binom}}(m = 28|\theta = 1/2, N = 63) \approx 2 \cdot 0.225 \approx 0.45$$

Therefore, B is correct.

	$m = 14$	$m = 21$	$m = 28$	$m = 35$
$N = 30$	0.428	0.992	1.000	1.000
$N = 33$	0.243	0.960	1.000	1.000
$N = 36$	0.121	0.879	1.000	1.000
$N = 39$	0.054	0.739	0.998	1.000
$N = 42$	0.022	0.561	0.990	1.000
$N = 45$	0.008	0.383	0.964	1.000
$N = 48$	0.003	0.235	0.903	1.000
$N = 51$	0.001	0.131	0.799	0.998
$N = 54$	0.000	0.067	0.658	0.990
$N = 57$	0.000	0.031	0.500	0.969
$N = 60$	0.000	0.014	0.349	0.922
$N = 63$	0.000	0.006	0.225	0.843
$N = 66$	0.000	0.002	0.134	0.731
$N = 69$	0.000	0.001	0.074	0.595
$N = 72$	0.000	0.000	0.038	0.453
$N = 75$	0.000	0.000	0.018	0.322
$N = 78$	0.000	0.000	0.008	0.214
$N = 81$	0.000	0.000	0.004	0.133
$N = 84$	0.000	0.000	0.001	0.078
$N = 87$	0.000	0.000	0.001	0.043

Table 8: Values of the binomial cumulative distribution function $\text{cdf}_{\text{binom}}(m|N, \theta = \frac{1}{2})$ for different values of the number of successes m and the number of trials N.

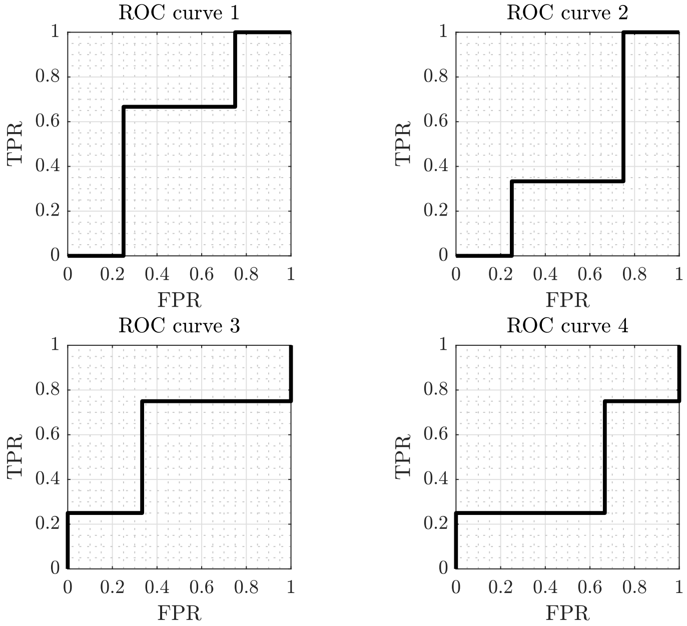


Figure 12: Proposed ROC curves for the neural network classifier with predictions/true class labels given in Table 9

y	1	0	0	1	1	0	1
$\hat{y}$	0.01	0.05	0.14	0.3	0.31	0.36	0.91

Table 9: Small binary classification dataset of $N = 7$ observations along with the predicted class probability $\hat{y}$.

Question 24. A neural network classifier is trained to distinguish between two classes $y \in \{0, 1\}$ in a small dataset consisting of $N = 7$ observations. Suppose the true class label y and predicted probability an observation belongs to class 1, $\hat{y}$, is as given in Table 9.

To evaluate the classifier, we will use the *area under curve* (AUC) of the *reciever operator characteristic* (ROC) curve. In Figure 12 is given four proposed ROC curves, which one of the curves corresponds to the classifier?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

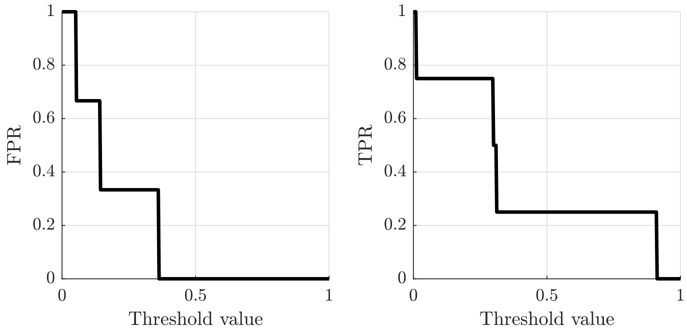


Figure 13: TPR, FPR curves for the classifier.

Solution 24. To compute the AUC, we need to compute the false positive rate (FPR) and true positive rate (TPR) for particular choices of threshold value $\hat{y}$. To compute e.g. the TPR, one assumes every observation predicted to belong to class 1 with a probability higher than $\hat{y}$ is actually assigned to class one. We then divide the total number of observations belonging to class one *and which are predicted to belong to class 1* with the number of observations in the *positive* class.

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

This procedure is then repeated for different threshold values to obtain the curves shown in Figure 13. The ROC curve is then obtained by plotting these two curves against each other. I.e. for each threshold value, the point

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

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

Question 25. We will consider an artificial neural network (ANN) trained on the Palmer Penguins dataset described in Table 1 to predict the class label y based on the first four attributes $x_1, \ldots, x_4$. The neural network has a single hidden layer containing $n_h = 6$ units, and will use the softmax activation function (specifically, we will use the over-parameterized softmax function described in section 15.3.2 *(Neural networks for multi-class classification)* of the lecture notes) to predict the class label y since it is a multi-class problem. For the hidden layer we will use the tanh non-linear activation function. How many parameters has to be trained to fit the neural network?

A. The network contains 36 parameters

B. The network contains 42 parameters

C. The network contains 51 parameters

D. The network contains 72 parameters

E. Don't know.

Solution 25. Each hidden unit has as many input unit weights as there are features $M = 4$ plus one (the bias), therefore they contribute with

$$(M + 1)n_h$$

weights. The softmax is computed deterministically from $C = 3$ units (as many as there are classes in the dataset), and each has as many weights as there are hidden units plus one (the bias):

$$(n_h + 1)C$$

Adding these two numbers together gives option C.

Question 26. Suppose that you have a deep neural network that can binary classify whether an image contains a penguin or not. If an image contains a penguin, the network will correctly classify it as a penguin with probability 97%. If an image does not contain a penguin, the network will classify it as a penguin with probability 3%. You apply the classifier to a dataset where 1% of the images contain a penguin. What is the probability that a random image from this dataset contains a penguin given that it is classified as a penguin?

A. ≈ 0.97

B. ≈ 0.75

C. ≈ 0.50

D. ≈ 0.25

E. Don't know.

Solution 26. Let A denote the event that the classifier classifies a picture as a penguin, and let B denote the event that a picture contains a penguin. From the text we are told

- $P(A|B) = 0.97$
- $P(A|\bar{B}) = 0.03$
- $P(B) = 0.01$
- $P(\bar{B}) = 1 - P(B)$

Using Bayes rule, we find that

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\bar{B})P(\bar{B})} \\ &= \frac{0.97 \cdot 0.01}{0.97 \cdot 0.01 + 0.03 \cdot (1 - 0.01)} \\ &\approx 0.25 \end{aligned}$$

Question 27. We use a multinomial regression model to predict the species (y) from the six attributes $x_1, \dots, x_6$ for the Palmer Penguins dataset in Table 1. We apply least squares regularization (i.e., L_2 regularization) to the weights of the multinomial regression model and use two-level cross-validation to select the optimal value of the regularization constant $\lambda \in \{0.001, 0.01, 0.1, 1.0\}$. In the outer fold, we use 3-fold cross-validation, and in the inner fold, we use leave-one-out cross-validation. With this setup and $N = 333$ observations $o_1, \dots, o_{333}$ in the dataset, how many times is o_1 used for training a model?

A. 999

B. 1768

C. 1770

D. 2667

E. Don't know.

Solution 27. Consider a single observation o_0 in the Palmer Penguins dataset:

- In the outer fold, we have three training sets $\{D_1^{\text{par}}, D_2^{\text{par}}, D_3^{\text{par}}\}$ each of size 222.
- o_0 is part of two of the training sets $\{D_1^{\text{par}}, D_2^{\text{par}}, D_3^{\text{par}}\}$.
- In the inner fold, we split D_i^{par} into training sets $\{D_1^{\text{train}}, \dots, D_{222}^{\text{train}}\}$. If $o_2 \in D_i^{\text{par}}$ then o_2 is part of 221 of these training sets.
- We train 4 models on each of the training sets D_j^{train}.
- Finally, we also train a single model for the optimal choice of λ on each of D_i^{par}.

This means that o_0 is used $2 \cdot (221 \cdot 4 + 1) = 1770$ times for training a model.