Technical University of Denmark

Written examination: 28th May 2015, 9 AM - 1 PM. Page 1 of 12 pages.

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.

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

Please write your name and student number clearly and hand in the present page (page 1) as your answer of the written test. Other pages will not be considered.

Answers:

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

Name: _____

Student number: _____

HAND IN THIS PAGE ONLY

No.	Attribute description	Abbrev.
x_1	Student's sex (0: F, 1: M)	SEX
x_2	Student's age (15 to 22)	AGE
x_3	Mother's education (0: none, 1: 4th grade, 2: 5th to 9th grade, 3: secondary education or 4: higher education)	MEDU
x_4	Weekly study time in hours (1: 0 to 2, 2: 2 to 5, 3: 5 to 10, or 4: +10)	STUDYTIME
x_5	Number of past class failures (n if n from 1 to 3 else 4)	FAILURES
x_6	In romantic relationship (0: no, 1: yes)	ROMANTIC
x_7	Going out with friends (1: low to 5: high)	GOOUT
x_8	number of school absences (0 to 93 days)	ABSENCES
y	Final grade (0: low to 20: high)	GRADE

Table 1: Attributes of the *Student* dataset. The dataset includes 8 attributes $(x_1, \ldots, x_8)$ of 395 students and their final grade. The purpose is to evaluate how various factors (such as being in a romantic relationship) affects the final grade.

Question 1. Consider the *Student*[1] dataset of table 1. Which of the following statements are true?

A. The most suitable way to apply logistic regression to predict if y is greater than 12 is to first remove the binary features from the dataset

B. The most suitable method to predict the ROMANTIC variable for females is using association mining

C. MEDU is ordinal discrete

D. ABSENCES is interval but not ratio

E. Don't know.

Solution 1.

A There are no good reasons to remove binary features in order to apply logistic regression

B Predicting ROMANTIC for females is a classification task, possibly only on the female subset of the data

C MEDU is indeed ordinal discrete

D ABSENCES is ratio

[1]Dataset obtained from http://archive.ics.uci.edu/ml/datasets/Student+Performance. Notice the dataset has been pre-processed for this exam.

Question 2. A principal component analysis is carried out on the *Student* dataset based on the attributes $x_1, \cdots, x_8$ found in table 1. The data is standardized by (i) substracting the mean and (ii) dividing each column by it's standard deviation to obtain the standardized matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$ where

$$\boldsymbol{S} = \begin{bmatrix} 25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 23 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 22 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 18 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 17 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 15 \end{bmatrix}$$

$$\boldsymbol{V} = \begin{bmatrix} 0.1 & -0.6 & 0.2 & -0.1 & 0.3 & -0.5 & -0.1 & -0.5 \\ 0.5 & 0.2 & -0.0 & 0.1 & -0.1 & -0.7 & 0.0 & 0.4 \\ -0.3 & 0.0 & 0.7 & 0.1 & 0.1 & -0.0 & -0.5 & 0.4 \\ -0.3 & 0.6 & -0.1 & 0.4 & 0.1 & -0.3 & -0.4 & -0.5 \\ 0.6 & -0.1 & -0.2 & 0.0 & 0.1 & 0.4 & -0.7 & -0.0 \\ 0.3 & 0.4 & 0.3 & -0.3 & 0.7 & 0.1 & 0.2 & -0.1 \\ 0.3 & -0.1 & 0.3 & 0.8 & -0.0 & 0.2 & 0.3 & -0.1 \\ 0.3 & 0.3 & 0.5 & -0.3 & -0.6 & 0.1 & 0.0 & -0.4 \end{bmatrix}.$$

Notice the entries of the matrices have been rounded. Which one of the following statements is true?

A. The first four principal components accounts for more than 70% of the variance

B. An observation with a large projection onto the second principal component can be described as a *studious romantically involved female* (i.e. high STUDYTIME and ROMANTIC and a low value of SEX)

C. Since the variance is very similar for all 8 principal components this implies that any projection onto a single principal component will not be sufficient to predict the grade y.

D. The 3 principal components with the least variance account for less than 20% of the variance

E. Don't know.

Solution 2. Recall the variance of e.g. the first four components are

$$\text{var.} = \frac{\sum_{i=1}^{4} S_{ii}^2}{\sum_{j=1}^{8} S_{jj}^2}$$

Then the variance of the first four components is: 0.651 and the variance of the three last components: 0.246. We cannot say if any single principal component is sufficient to predict y by looking at the variances alone. This leaves option B which can be seen to be true by inspection.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9
o_1	0.00	3.85	4.51	4.39	4.08	3.97	2.18	3.29	5.48
o_2	3.85	0.00	2.19	3.46	3.66	3.93	3.15	3.47	4.11
o_3	4.51	2.19	0.00	3.70	4.30	4.83	3.86	4.48	4.19
o_4	4.39	3.46	3.70	0.00	1.21	3.09	4.12	3.22	3.72
o_5	4.08	3.66	4.30	1.21	0.00	2.62	4.30	2.99	4.32
o_6	3.97	3.93	4.83	3.09	2.62	0.00	4.15	1.29	3.38
o_7	2.18	3.15	3.86	4.12	4.30	4.15	0.00	3.16	4.33
o_8	3.29	3.47	4.48	3.22	2.99	1.29	3.16	0.00	3.26
o_9	5.48	4.11	4.19	3.72	4.32	3.38	4.33	3.26	0.00

Table 2: The pairwise Euclidian distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 9 observations from the *Student* dataset (recall $M = 8$). Each observation o_i corresponds to a row of the student matrix $\boldsymbol{X}$ of table 1 (the matrix has been normalized). The colors indicate classes such that the blue observations $\{o_1, o_2, o_3\}$ belongs to class C_1 (low GRADE), the red observations $\{o_4, o_5, o_6\}$ belongs to class C_2 (medium GRADE) and the black observations $\{o_7, o_8, o_9\}$ belongs to class C_3 (high GRADE).

Question 3. Consider the distances in table 2. The class labels C_1, C_2, C_3 (corresponding to $\{o_1, o_2, o_3\}$, $\{o_4, o_5, o_6\}$ and $\{o_7, o_8, o_9\}$) will be predicted using a k-nearest neighbour classifier based on the distances in table 2. Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 1-nearest classifier (i.e. $k = 1$). What is the error rate for the $N = 9$ observations?

A. error rate $= \frac{3}{9}$

B. error rate $= \frac{4}{9}$

C. error rate $= \frac{5}{9}$

D. error rate $= \frac{6}{9}$

E. Don't know.

Solution 3. The true accuracy is 0.444444444444 or 4/9. This is easy to see by going through table 2 and notice the "wrongly" classified observations are o_1, o_6, o_7, o_8 (they are paired as (o_1, o_7), (o_6, o_8), (o_7, o_1) and (o_8, o_6)).

Question 4. Consider the distances in table 2 and suppose we wish to apply mixture modelling and we use the normal density as the mixture distributions:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma) = (2\pi\sigma^2)^{-\frac{M}{2}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}}$$

Suppose we wish to compute the density at o_9 based on a mixture model of $K = 8$ components, the parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_8$ of each component is taken to be the position of the observations $o_1, \ldots, o_8$ and the components are weighted equally. Suppose we set $\sigma = 5$, what is the probability density at the *last* observation o_9?

A. $p(o_9) = \frac{1}{8(\pi 50)^4}\left(e^{\frac{-5.48^2}{50}} + e^{\frac{-4.11^2}{50}} + e^{\frac{-4.19^2}{50}} + e^{\frac{-3.72^2}{50}} + e^{\frac{-4.32^2}{50}} + e^{\frac{-3.38^2}{50}} + e^{\frac{-4.33^2}{50}} + e^{\frac{-3.26^2}{50}}\right)$

B. $p(o_9) = \frac{1}{(\pi 50)^8}\left(e^{\frac{-5.48}{50}} + e^{\frac{-4.11}{50}} + e^{\frac{-4.19}{50}} + e^{\frac{-3.72}{50}} + e^{\frac{-4.32}{50}} + e^{\frac{-3.38}{50}} + e^{\frac{-4.33}{50}} + e^{\frac{-3.26}{50}}\right)$

C. $p(o_9) = \frac{1}{8(\pi 50)^4}\left(e^{\frac{-3.85^2}{50}} + e^{\frac{-4.51^2}{50}} + e^{\frac{-4.39^2}{50}} + e^{\frac{-4.08^2}{50}} + e^{\frac{-3.97^2}{50}} + e^{\frac{-2.18^2}{50}} + e^{\frac{-3.29^2}{50}} + e^{\frac{-5.48^2}{50}}\right)$

D. $p(o_9) = \frac{1}{(\pi 50)^4}\exp\left(\frac{-5.48}{50} + \frac{-4.11}{50} + \frac{-4.19}{50} + \frac{-3.72}{50} + \frac{-4.32}{50} + \frac{-3.38}{50} + \frac{-4.33}{50} + \frac{-3.26}{50}\right)$

E. Don't know.

Solution 4. Options B and D are not properly normalized by the number of mixture components (D is also of the wrong functional form). Option C uses the wrong distances, namely the distances from observation o_1 to the other elements o_i. Accordingly option A is the correct answer.

Question 5. We wish to compute the *average relative KNN density* (a.r.d) of observation o_1 of table 2 using the distances given in the table. Letting $d(\boldsymbol{x}, \boldsymbol{y})$ denote the Euclidian distance metric the a.r.d. is defined as

$$\text{density}(\boldsymbol{x}, K) = \frac{1}{\frac{1}{K}\sum_{\boldsymbol{y} \in N(\boldsymbol{x}, K)} d(\boldsymbol{x}, \boldsymbol{y})}$$

$$\text{a.r.d}(\boldsymbol{x}, K) = \frac{\text{density}(\boldsymbol{x}, K)}{\frac{1}{K}\sum_{\boldsymbol{z} \in N(\boldsymbol{x}, K)} \text{density}(\boldsymbol{z}, K)},$$

$$N(\boldsymbol{x}, K) : \text{set of } K\text{-nearest neighbours of } \boldsymbol{x}.$$

What is the a.r.d. of observation o_1 using $K = 2$ nearest neighbours?

A. **a.r.d**$(\boldsymbol{x} = o_1, K = 2) \approx 0.868$

B. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.434$

C. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.569$

D. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.502$

E. Don't know.

Solution 5. The nearest neighbour of o_1 is o_7, o_8 and the nearest neighbours of o_7 is o_1, o_2 and for o_8 it is

o_5, o_6. The densities are

$$\text{density}(o_1, K=2) = 0.36563071298$$
$$\text{density}(o_7, K=2) = 0.375234521576$$
$$\text{density}(o_8, K=2) = 0.467289719626$$

from which it follows

$$\text{a.r.d.}(o_1, K=2) = \frac{0.36563071298}{\frac{1}{2}(0.375234521576 + 0.467289719626)} = 0.867941111008$$

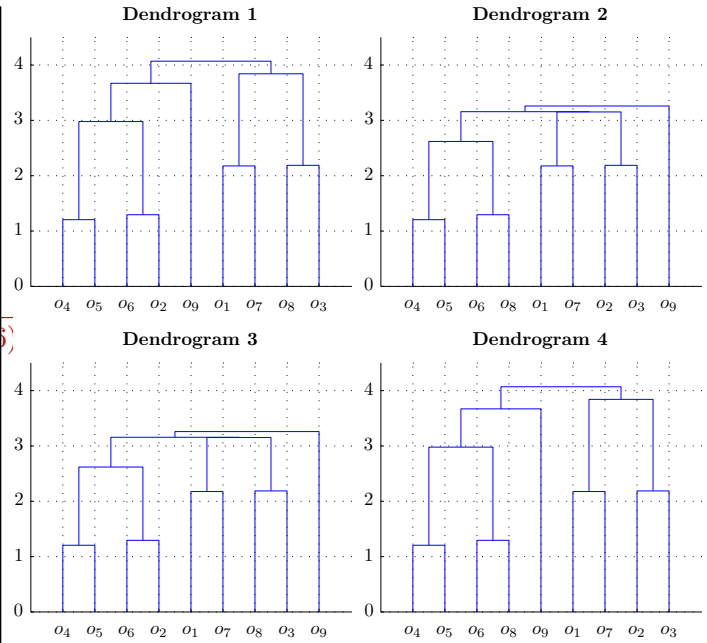

Figure 1: Proposed hierarchical clustering of the 9 observations considereded in table 2

Question 6. In table 2 is given the pairwise distances between 9 observations. A hierarchical clustering is applied to these nine observations using *minimum* linkage. Which of the dendrograms shown in fig. 1 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 6. The true answer is B, dendrogram 2. Considering the distances between o_2 and o_6 allow us to rule out dendrograms 1 and 3. Then considering the minimum distance between observation o_9 and o_8 is 3.26 allow us to rule out dendrogram 4. This leave only option *B*.

Question 7.

In table 2 is given the pairwise euclidian distances between 9 observations of the *Student* dataset of table 1. Suppose we wish to train a Gaussian Mixture-model (GMM) using the EM-algorithm with $K = 2$ clusters. The cluster centers $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are initialized to coincide with the location of observations o_1 and o_4 respectively, the covariance matrices are initialized to be the unit matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$ and the prior class-probability is selected to be $w_1 = 0.2$, $w_2 = 0.8$. The

EM algorithm is applied to compute the updated assignment of points to classes. Assuming the dimensionality of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ is $M = 8$, with what probability is point o_3 assigned to class C_1?

Hint: Notice the multivariate normal distribution for e.g. class C_1 becomes:

$$\mathcal{N}(o_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = \frac{1}{(2\pi)^{\frac{M}{2}}} \exp\left(\frac{-d(o_i, o_1)^2}{2}\right)$$

A. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0089$

B. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0347$

C. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0003$

D. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0013$

E. Don't know.

Solution 7. The probability can be computed by using Bayes rule

$$\begin{aligned} &p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \\ &\quad = \frac{\mathcal{N}(o_3; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) w_1}{\mathcal{N}(o_3; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) w_1 + \mathcal{N}(o_3; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) w_2} \\ &\quad = \frac{w_1 e^{\frac{-d(o_3, o_1)^2}{2}}}{w_1 e^{\frac{-d(o_3, o_1)^2}{2}} + w_2 e^{\frac{-d(o_3, o_4)^2}{2}}} \\ &\quad = 0.00891253249006 \end{aligned}$$

Question 8. In table 2 is given the pairwise euclidian distances between 9 observations from the *Student* dataset of table 1. Suppose the Euclidian norm of observations o_1 and o_2 is:

$$\|o_1\| = \sqrt{\sum_{k=1}^{M} x_{1k}^2} = 2.99, \quad \|o_2\| = \sqrt{\sum_{k=1}^{M} x_{2k}^2} = 2.26$$

What can be concluded about the extended Jaccard similarity of these two observations? (Hint: recall for vectors $\boldsymbol{x}$, $\boldsymbol{y}$ that $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{x}^\top \boldsymbol{y}$)

A. $\mathrm{EJ}(o_1, o_2) \approx -0.0523$

B. $\mathbf{EJ}(o_1, o_2) \approx -0.0268$

C. $\mathrm{EJ}(o_1, o_2) \approx -0.0261$

D. $\mathrm{EJ}(o_1, o_2) \approx -0.1052$

E. Don't know.

Solution 8. Notice the inner product can be recovered as

$$o_1^\top o_2 = \frac{\|o_1\|_2^2 + \|o_2\|_2^2 - d(o_1, o_2)^2}{2} = -0.3874$$

and the definition of the extended Jaccard similarity is

$$\mathrm{EJ}(o_1, o_2) = \frac{o_1^\top o_2}{\|o_1\|_2^2 + \|o_2\|_2^2 - o_1^\top o_2}$$

GRADE	less than 10	from 10 to 12	more than 12
SEX=F	55	44	88
SEX=M	75	59	74

Table 3: Number of students of the two sexes in the three classes of GRADE. For instance, there are 44 females with a grade between 10 and 12

Question 9.

Consider the *Student* dataset of table 1. Suppose the variable GRADE is divided into three classes depending on whether GRADE is (i) < 10 (ii) between $10-12$ (iii) > 12, thereby creating a 3-class classification problem. Suppose we attempt to train a decision tree and we initially consider a split on the variable SEX. If the number of students in the three classes of either sex is as listed in table 3, what is the *impurity gain* Δ of the split if the *Gini* impurity measure is used?

A. $\Delta \approx 0.00329$

B. $\Delta \approx 0.65548$

C. $\Delta \approx 0.64415$

D. $\Delta \approx 0.00497$

E. Don't know.

Solution 9. The relevant definitions can be found in section 4.3 of Tan et.al. We first need the frequencies for all students as well as for the males and females. Letting C_1, C_2 and C_3 denote the low, medium and high classes:

$$\begin{aligned}\text{Female: } & p(C_1|F) = 55/187, p(C_2|F) = 44/187, p(C_3|F) = 88/187\\ \text{Male: } & p(C_1|M) = 75/208, p(C_2|M) = 59/208, p(C_3|M) = 74/208\\ \text{All: } & p(C_1|A) = 130/395, p(C_2|A) = 103/395, p(C_3|A) = 162/395\end{aligned}$$

From this we can compute the impurity: $I(x) = 1 - \sum_i p(i|x)^2$

$$\begin{aligned}\text{Female: } & I(F) = 0.636678200692\\ \text{Male: } & I(M) = 0.662953032544\\ \text{All: } & I(A) = 0.655484697965.\end{aligned}$$

Then combining these we have

$$\begin{aligned}\Delta &= I(A) - (187/395)I(F) - (208/395)I(M)\\ &= 0.00497063644952.\end{aligned}$$

Question 10.

Consider the $N = 9$ students from table 2 and assume the data has been processed to the 9×6 binary matrix described in table 4. Suppose we only consider the first two features f_1, f_2 and train a Naïve-Bayes classifier to distinguish between class C_1, C_2 and C_3 based on only these two features. If an observation has $f_1 = 1, f_2 = 0$, what is the probability the observation belongs to class C_3 according to the Naive-Bayes classifier?

A. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.143$

B. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.133$

C. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.375$

D. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.125$

E. Don't know.

Solution 10. True answer is: 0.125. This can be found by computing the per-class probabilities

$$\begin{aligned}p(f_1 = 1|C_1) = 1,\ p(f_2 = 0|C_1) = 1/3\\ p(f_1 = 1|C_2) = 2/3,\ p(f_2 = 0|C_2) = 2/3\\ p(f_1 = 1|C_3) = 1/3,\ p(f_2 = 0|C_3) = 1/3\end{aligned}$$

The class label priors are the same $p(C_i) = \frac{1}{3}$ and so the Naive-Bayes estimate is

$$p_{NB}(C_3|f_1 = 1, f_2 = 0) = \frac{p(f_1 = 1|C_3)p(f_2 = 0|C_3)p(C_3)}{\sum_{i=1}^{3} p(f_1 = 1|C_i)p(f_2 = 0|C_i)p(C_i)} = \frac{1}{8}$$

	f_1	f_2	f_3	f_4	f_5	f_6
o_1	1	0	1	0	1	0
o_2	1	1	1	1	1	0
o_3	1	1	0	0	1	0
o_4	1	1	1	0	0	1
o_5	1	0	1	0	0	1
o_6	0	0	1	1	0	1
o_7	1	1	1	1	1	1
o_8	0	0	1	1	1	1
o_9	0	1	0	1	0	1

Table 4: Processed version of the $N = 9$ observations of table 2. For each student we record 6 features corresponding to $f_1 : SEX$, $f_2 : AGE$, $f_3 : MEDU$, $f_4 : STUDYTIME$, $f_5 : GOOUT$ and $f_6 : ABSENCES$. The features are binarized by thresholding at the mean value. The categories still reflect GRADE, i.e. the blue category C_1 (o_1, o_2, o_3) is low GRADE, the red category C_2 (o_4, o_5, o_6) is medium GRADE and the black category C_3 (o_7, o_8, o_9) is high GRADE.

Question 11. Suppose we consider the binary matrix in table 4 as a market-basket problem consisting of $N = 9$ "transactions" $o_1, \ldots, o_9$ and $M = 6$ "items" $f_1, \ldots, f_6$. Which of the following options represents all itemsets with support greater than 0.4?

A. $\{f_1\}$, $\{f_3\}$, $\{f_6\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_1, f_3\}$, $\{f_4\}$, $\{f_5\}$, $\{f_6\}$, $\{f_3, f_6\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_1, f_2\}$, $\{f_3\}$, $\{f_1, f_3\}$, $\{f_4\}$, $\{f_3, f_4\}$, $\{f_5\}$, $\{f_1, f_5\}$, $\{f_3, f_5\}$, $\{f_6\}$, $\{f_3, f_6\}$, $\{f_4, f_6\}$

D. $\{f_1\}$, $\{f_2\}$, $\{f_1, f_2\}$, $\{f_3\}$, $\{f_1, f_3\}$, $\{f_2, f_3\}$, $\{f_1, f_2, f_3\}$, $\{f_4\}$, $\{f_2, f_4\}$, $\{f_3, f_4\}$, $\{f_5\}$, $\{f_1, f_5\}$, $\{f_2, f_5\}$, $\{f_1, f_2, f_5\}$, $\{f_3, f_5\}$, $\{f_1, f_3, f_5\}$, $\{f_4, f_5\}$, $\{f_3, f_4, f_5\}$, $\{f_6\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_3, f_6\}$, $\{f_1, f_3, f_6\}$, $\{f_4, f_6\}$, $\{f_3, f_4, f_6\}$

E. Don't know.

Solution 11. Recall by chapter 6.1 of Tan et al. the support count is the number of "transactions" containing a given set of items. The problem is then to find all subsets of items that occur in at least 4 of the 9 transactions. These are easily seen to be those in option C and no other.

Question 12.

We consider again the $N = 9$ students from table 4 as 6-dimensional binary vectors. Which one of the following statements is true regarding the Jaccard/cosine similarity and the simple matching coefficient?

A. $\text{SMC}(o_2, o_6) > \text{J}(o_2, o_7)$

B. $\text{SMC}(o_2, o_6) > \text{COS}(o_2, o_6)$

C. $\textbf{COS}(o_2, o_7) > \textbf{J}(o_2, o_7)$

D. $\text{COS}(o_2, o_6) > \text{COS}(o_2, o_7)$

E. Don't know.

Solution 12. It is easily verified only option C is correct by plugging in the following values:

$$\begin{aligned} \text{SMC}(o_2, o_6) &= 0.333333333333 \\ \text{J}(o_2, o_7) &= 0.833333333333 \\ \text{COS}(o_2, o_6) &= 0.516397779494 \\ \text{COS}(o_2, o_7) &= 0.912870929175 \end{aligned}$$

Question 13. Suppose we consider the binary matrix of table 4 as a market-basket problem consisting of $N = 9$ "transactions" $o_1, \ldots, o_9$ and $M = 6$ "items" $f_1, \ldots, f_6$. What is the *lift* of the rule $\{f_2, f_4, f_6\} \to \{f_1, f_5\}$ if the lift for a rule $A \to B$ is defined as

$$\text{Lift}(A \to B) = \frac{\text{Confidence}(A \to B)}{\text{Support}(B)}$$

A. $\textbf{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{9}{8}$

B. $\text{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{8}{9}$

C. $\text{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{6}{9}$

D. $\text{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{5}{9}$

E. Don't know.

Solution 13. the lift is 1.125. Recall the confidence is defined as (see chapter 6.1 of Tan et al)

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

All items f_1, f_2, f_4, f_5, f_6 occur only in 1 transaction, the items f_2, f_4, f_6 only in 2 transactions and finally f_1, f_5 only in 4 transactions. The lift is then:

$$\text{Lift}(A \to B) = \frac{\frac{1/9}{2/9}}{4/9} = \frac{9}{8}$$

Feature(s)	A_{train}	A_{test}
None	0.737	0.574
x_1	0.569	0.526
x_2	0.644	0.563
x_3	0.572	0.52
x_4	0.678	0.581
x_1, x_2	0.625	0.589
x_1, x_3	0.805	0.645
x_1, x_4	0.709	0.607
x_2, x_3	0.535	0.5
x_2, x_4	0.634	0.608
x_3, x_4	0.609	0.499
x_1, x_2, x_3	0.738	0.623
x_1, x_2, x_4	0.738	0.614
x_1, x_3, x_4	0.763	0.596
x_2, x_3, x_4	0.547	0.525
x_1, x_2, x_3, x_4	0.579	0.552

Table 5: The *accuracy* on a training set A_{train} and test set A_{test} of linear regression models (predictions are made by thresholding at 0.5) trained on different subsets of features of the Students dataset of table 1

Question 14. Consider the Students dataset of table 1 and suppose GRADE has been binarized to only take two values and we only consider the first four features x_1, x_2, x_3, x_4. Suppose we wish to examine which subset of these features gives the **highest** accuracy on a test set. In table 5 is shown how different combinations of features give rise to different values of the accuracy on a training and test set for a classifier. Which one of the following statements is true?

A. Forward and backward selection will select the same number of features.

B. Forward selection will select a model with higher accuracy on the test set than backward selection.

C. Backward selection will select *more* features than forward selection.

D. Backward selection will select *less* features than forward selection.

E. Don't know.

Solution 14. Firstly notice the column with the training set accuracy can be disregarded. Forward selection then first selects x_4, then x_2, x_4 and finally x_1, x_2, x_4 and then terminates with an accuracy on the test set of 0.614. Backward selection will first select x_1, x_2, x_3, then x_1, x_3 and terminate with an accuracy of 0.645 on the test set. Accordingly only option D is correct.

Question 15. Consider the attributes ROMANTIC, GOOUT, and GRADE of table 1. Assume each attribute has been binarized by thresholding at the median value giving the 3 binary attributes:

- RO = yes, no: *In a romantic relationship or not*
- GO = yes, no: *Going out in the evening or not*
- GR = high, low: *Has a high or low grade*

and there are thus $2^3 = 8$ possible outcomes. Suppose we are given the following information

$$\begin{aligned} p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}) &= 0.36 \\ p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{y}) &= 0.39 \\ p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{n}) &= 0.47 \\ p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{n}) &= 0.48 \\ \text{and} \quad p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{y}) &= 0.23 \\ p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{y}) &= 0.46 \\ p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{n}) &= 0.11 \\ p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{n}) &= 0.21 \end{aligned}$$

What is then the probability that a student goes out and has a romantic relationship given the student attains high grades?

A. $p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.45$

B. $p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.32$

C. $p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.18$

D. $\boldsymbol{p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.2}$

E. Don't know.

Solution 15. The problem can be solved by applying Bayes theorem:

$$\begin{aligned} p(\mathrm{GR}{=}\mathrm{h}) &= p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y})p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{y}) \\ &\quad + p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{n})p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{n}) \\ &\quad + p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{y})p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{y}) \\ &\quad + p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{n})p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{n}) \end{aligned}$$

$$\begin{aligned} &p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \\ &\quad = \frac{p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y})p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{y})}{p(\mathrm{GR}{=}\mathrm{h})} \\ &\quad \approx 0.2 \end{aligned}$$

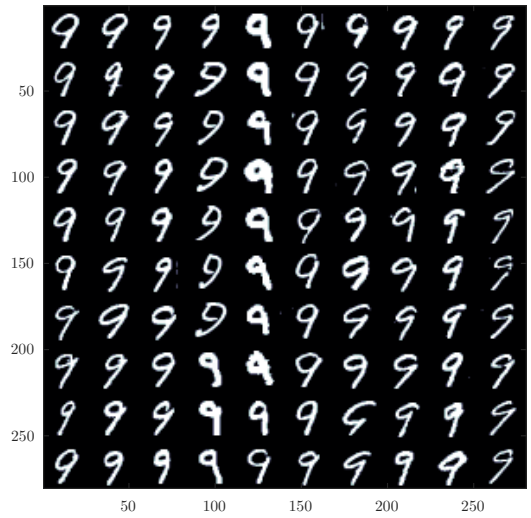

Figure 2: 100 images of the number 9 in a 10×10 grid

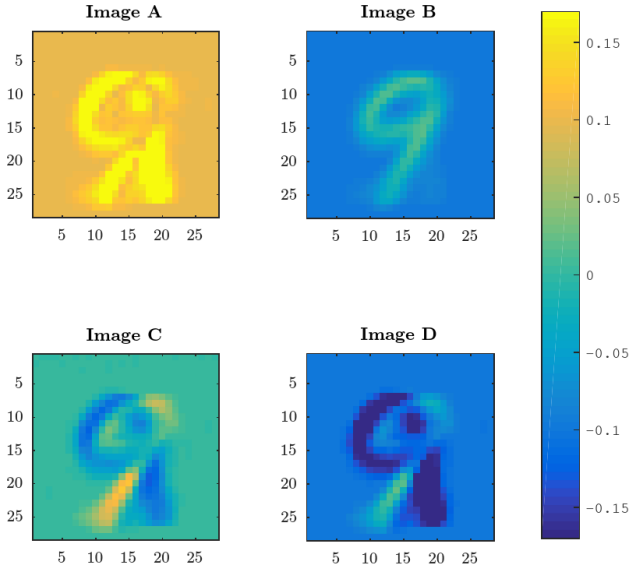

Figure 3: Four candidates for the first principal component of the image dataset of fig. 2. Notice the colorbar indicating coordinate magnitude

Question 16. In fig. 2 is shown a dataset[2] consisting of 100 images of the number 9 each of size 28×28 pixels. Each image is considered as a vector of $28^2 = 784$ coordinates and a principal component analysis is carried out as usual on the dataset. Which one of the four images, Image A, Image B, Image C or Image D in fig. 3 represents the first principal component reshapen as a 28×28 image and plotted using colors to indicate coordinate magnitude? (Hint: Use the colorbar)

A. Image A

B. Image B

C. Image C

D. Image D

E. Don't know.

Solution 16. Notice the instances of the number 9 only vary in pixel intensity in the middle region of the picture (or put in another way, most pixels along the boundary of the image remains zero). We can expect the variance along these boundary pixels to be zero, and this only leaves image C. Alternatively one can observe the large number of non-zero values implies only image C can be normalized.

[2]The numbers are a subset of the MNIST dataset obtained from http://yann.lecun.com/exdb/mnist/

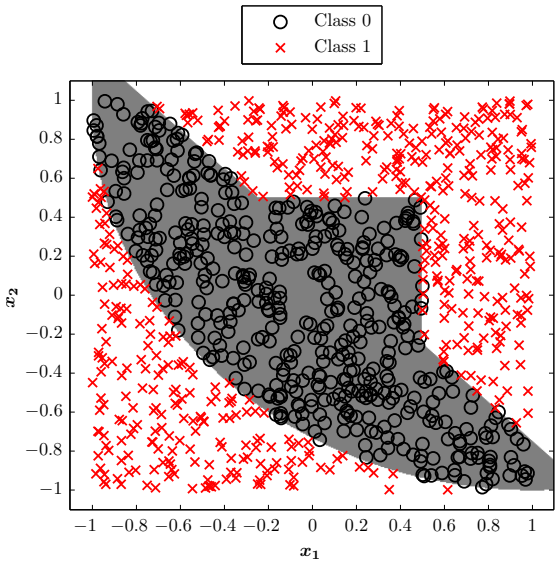

Figure 4: Two-class classification problem

Question 17. Suppose we wish to solve the two-class classification problem in fig. 4 using a classification tree of the form given in fig. 5. What rules, acting on the coordinates $\boldsymbol{x} = (x_1, x_2)$, should be assigned to the three internal nodes A, B and C of the tree to give rise to the indicated decision boundary?

A. $A : \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2, \quad B : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 < 2.25$
$C : \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$

B. $A : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2.25, \quad B : \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$
$C : \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2$

C. $A : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2.25, \quad B : \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2$
$C : \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$

D. $A : \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2, \quad B : \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$
$C : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2.25$

E. Don't know.

Solution 17. First consider the point $(0.4, 0.4)$ which should belong to the dark class 0. For this point $\|[0.4, 0.4] - [-1, -1]\|_1 = 1.4 + 1.4 = 2.8$ and so in option B and C it will be classified incorrectly leaving option A and D.

For option A, consider the point $(1, 1)$. Obviously node A will evaluate to false but node B will evaluate to $\|[1, 1] - [-1, -1]\|_1 = 2 + 2 = 4$ and so this node will evaluate to false and the point will incorrectly register as belonging to the dark class 0.

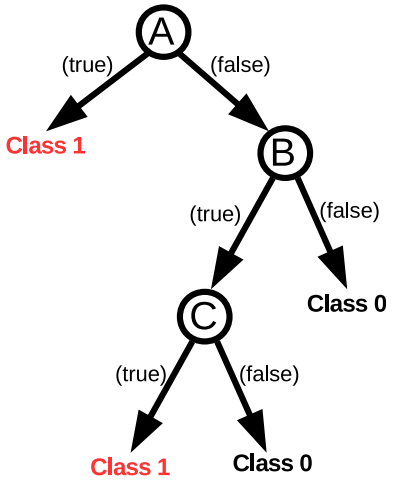

Figure 5: Decision tree with 3 nodes A, B and C

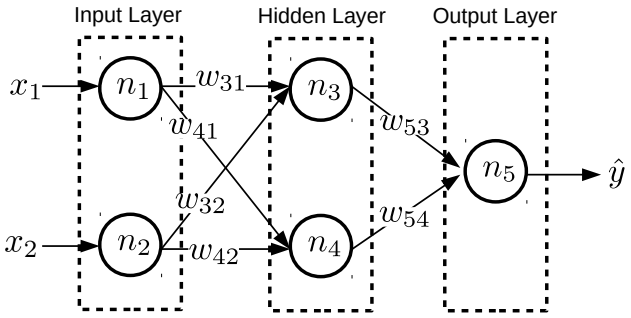

Figure 6: Simple neural network of 6 weights

Question 18. Consider the feedforward neural network shown in fig. 6. The network has no bias weights. Suppose the weights of the neural network after training are

$$w_{31} = 0.05, \quad w_{41} = 0, \quad w_{32} = 0.1,$$
$$w_{42} = -0.05, \quad w_{53} = 0.1, \quad w_{54} = -10$$

and the activation function of all five neurons n_1, n_2, n_3, n_4 and n_5 is the rectified linear unit

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \frac{1}{10}x & \text{otherwise.} \end{cases}$$

Suppose the network is evaluated on input $x_1 = 0.5$, $x_2 = 1$, what is the output?

A. $\hat{y} = 0.5125$

B. $\hat{y} = 0.05125$

C. $\hat{y} = -0.00375$

D. $\hat{y} = 0.0625$

E. Don't know.

Solution 18. To compute the output, initialize $n_1 = f(0.5) = 0.5, n_2 = f(1) = 1$. Then we can compute:

$$n_3 = f(n_1 * 0.05 + n_2 * 0.1) = f(1/8) = 1/8$$
$$n_4 = f(n_1 * 0 + n_2 * (-0.05)) = f(-1/20) = -1/200$$

Then for the output of the neural network we have

$$\hat{y} = n_5 = f(n_3 * 0.1 + n_4 * (-10)) = f(1/16) = 0.0625.$$

Question 19. Consider the classification problem given in fig. 4. Suppose the problem is solved using the following four classifiers

(1NN) A 1-nearest neighbour classifier

(TREE) A decision tree

(LREG) Logistic regression

(NNET) An artificial neural network with four hidden units

All classifiers are using only the two attributes x_1, x_2, corresponding to the position of each observation, as well as the class label. Which of the descriptions (1NN),(TREE),(LREG),(NNET) matches the boundariers of the four plots (Plot A, B, C and D) indicated in fig. 7?

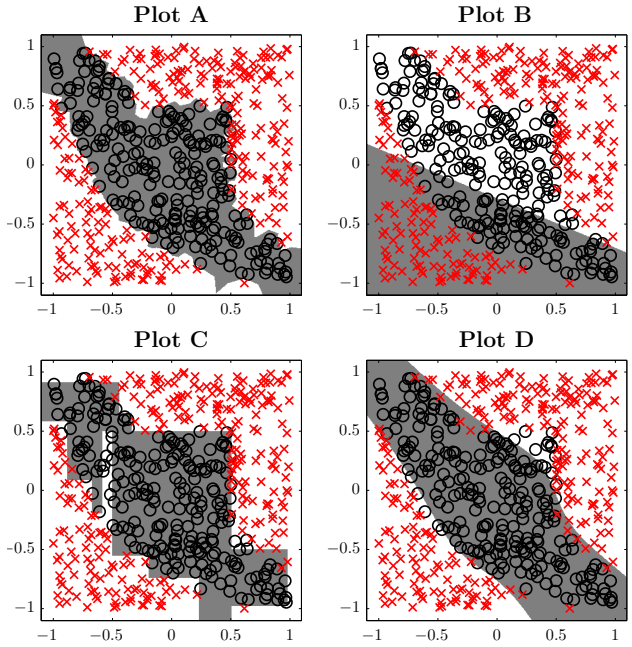

Figure 7: Four classifiers applied to a two-class classification problem

A. Plot A is 1NN, Plot B is LREG, Plot C is NNET, Plot D is TREE.

B. Plot A is 1NN, Plot B is LREG, Plot C is TREE, Plot D is NNET.

C. Plot A is 1NN, Plot B is NNET, Plot C is TREE, Plot D is LREG.

D. Plot A is LREG, Plot B is 1NN, Plot C is TREE, Plot D is NNET.

E. Don't know.

Solution 19. Plot A is a 1NN classifier (notice all points are correctly classified), B is the only classifier with a linear boundary and must be logistic regression and C has the "boxes" characteristic for a decision tree.

Question 20. Suppose the 2D dataset shown in fig. 8 was generated from a Gaussian mixture-model (GMM) with three components. Which of the following is the most likely equation of the density of the mixture model?

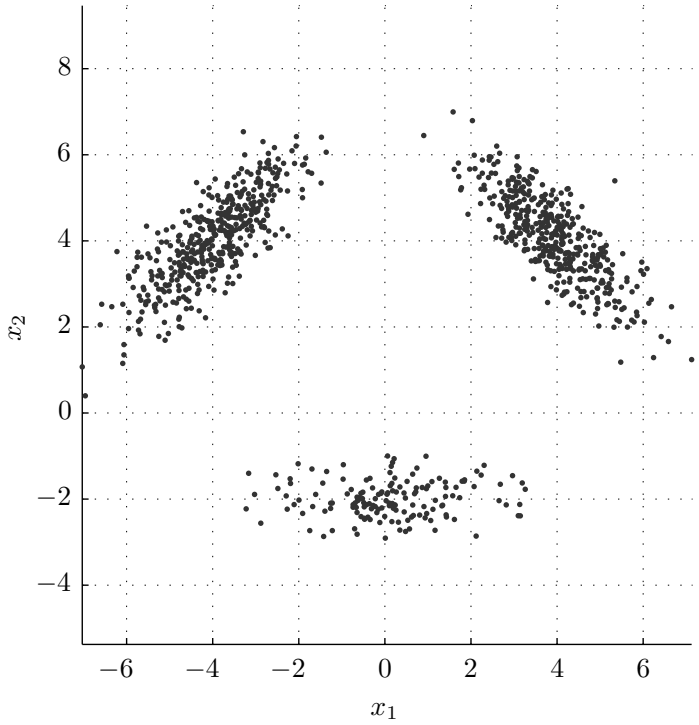

Figure 8: Scatter plot of observations

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.0 & -0.8 \\ -0.8 & 1.0 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.8 & 0.0 \\ 0.0 & 0.2 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix},$$
$$\boldsymbol{\mu}_1 = \begin{bmatrix} -4 \\ 4 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix},$$

A. The density is:

$$p(\boldsymbol{x}) = 0.15\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_3)$$

B. The density is:

$$p(\boldsymbol{x}) = 0.15\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_3)$$

C. The density is:

$$p(\boldsymbol{x}) = 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_1) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

D. The density is:

$$p(\boldsymbol{x}) = 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_3)$$

E. Don't know.

Solution 20. Options C and D can be ruled out because the densities cannot be weighted equally in the true mixture distribution. Then simply recall a covariance matrix with negative off-diagonal elements (such as $\boldsymbol{\Sigma}_1$) corresponds to a density slanted in the top-left to the bottom-right direction.

Question 21.

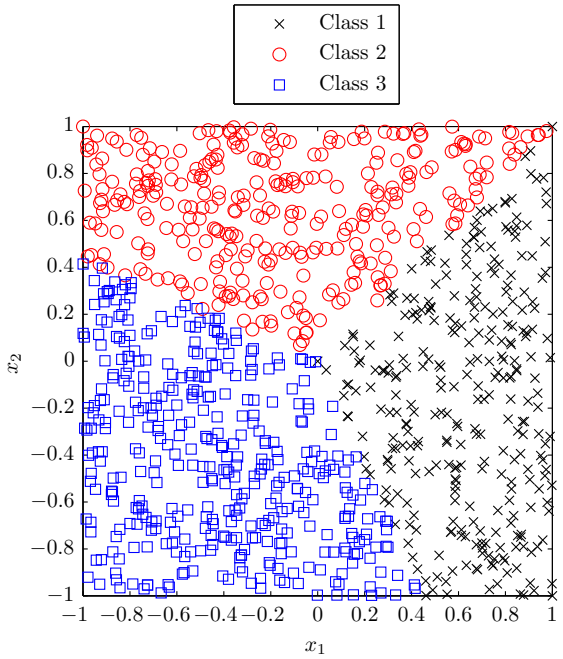

Figure 9: Observations labelled with the most probable class

Consider a multinomial regression classifier for a three-class problem where for each point $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$ we compute the class-probability by first computing the intermediate values

$$y_1 = \boldsymbol{w}_1^\top \boldsymbol{x}, \quad y_2 = \boldsymbol{w}_2^\top \boldsymbol{x}, \quad y_3 = \boldsymbol{w}_3^\top \boldsymbol{x}$$

and then combine these to the per-class probability

$$P(\hat{y} = k) = \frac{e^{y_k}}{e^{y_1} + e^{y_2} + e^{y_3}}$$

A dataset of $N = 1000$ points where each point is labelled according to the maximum class-probability is shown in fig. 9. Which setting of the weights was used? (Hint: consider points (x_1, x_2) where either x_1 or x_2 is zero)

A. $\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \boldsymbol{w}_3 = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}$

B. $\boldsymbol{w}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \boldsymbol{w}_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \boldsymbol{w}_3 = \begin{bmatrix} 0.8 \\ 0.8 \end{bmatrix}$

C. $\boldsymbol{w}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \boldsymbol{w}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{w}_3 = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}$

D. $\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \boldsymbol{w}_3 = \begin{bmatrix} -0.8 \\ -0.8 \end{bmatrix}$

E. Don't know.

Solution 21. Consider for instance $x_1 = -1$ and $x_2 = 0$. Then we have for the four settings of weights:

$$\begin{aligned} A &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & -0.3 \end{bmatrix} \\ B &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -0.8 \end{bmatrix} \\ C &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -0.3 \end{bmatrix} \\ D &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0.8 \end{bmatrix} \end{aligned}$$

Next, since the multinomial regression function preserves order we need only consider the maximal value. Accordingly A is classified to class 2, B and C to class 1 and only D correctly to class 3.

Question 22. Consider a two-dimensional data set consisting of $N = 8$ observations shown in fig. 10. The dataset contains three classes indicated by the black crosses (class 1), red circles (class 2) and blue squares (class 3). In the figure, the decision boundaries for four K-nearest neighbor classifiers (KNN) are shown. Which of the plots correspond to the $K = 3$ nearest-neighbour classifier assuming ties are broken by assigning to the *nearest* neighbour's class?

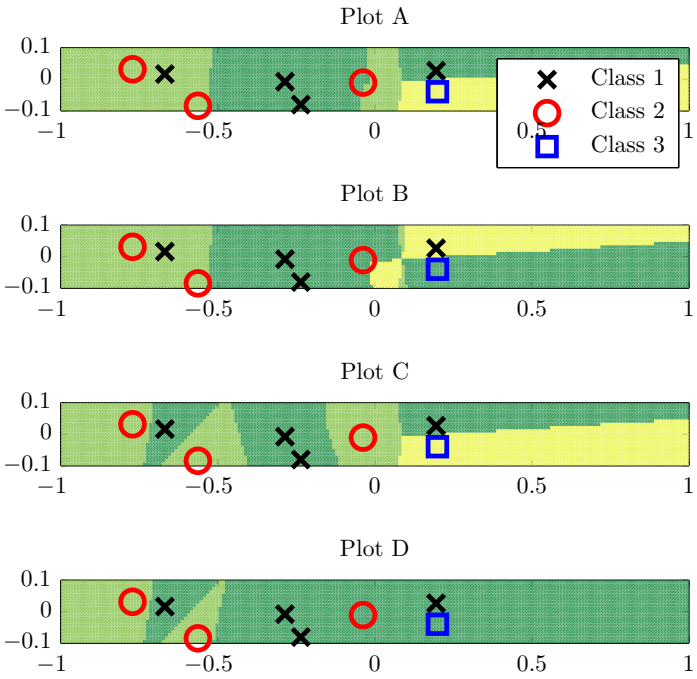

Figure 10: Decision boundaries for four KNN classifiers.

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 22. Lets focus on the blue square (Class 3). The three nearest neighbours at the blue square represents all three classes and (since ties are broken by assigning to the nearest class) the blue square should belong to it's own class. Evidently this rules out D (only two classes) and B (why should the blue class extend so far to the left?).

Consider then plot C. The left-most black cross and red circle has the same 3-nearest neighbours consisting of one black cross and two red circles, accordingly they should be in the same class. This rules out option C leaving only A.

X	1	3	4	6	7	8	13	15

Table 6: Simple 1-dimensional dataset comprised of $N = 8$ observations.

Question 23. Consider the 1-dimensional data set comprised of $N = 8$ observations shown in table 6. Which one of the following clusterings corresponds to converged state of a K-means algorithm using standard Euclidian distances?

A. {1, 3, 4}, {6, 7, 8}, {13, 15}

B. {1}, {3, 4, 6}, {7, 8}, {13, 15}

C. {1, 3, 4}, {6, 7}, {8, 13, 15}

D. {1, 3, 4}, {6, 7}, {8, 13}, {15}

E. Don't know.

Solution 23. The problem can be solved by explicit calculation, however it is easier solved by drawing the points on a paper and ruling out the clusterings that look the most "odd". For instance:

- For B notice the second cluster has mean 4.33 and the third has mean 7.5. Accordingly $x = 6$ is in the wrong cluster.
- For partition C notice the two last clusters have mean 6.5 and 12 and so $x = 8$ is in the wrong cluster.
- For options D the two last clusters has mean 10.5 and 15 and accordingly $x = 13$ is in the wrong cluster.

It is easy to check the first option has converged.

Question 24. Consider a dataset comprised of two classes as shown in fig. 11. For each observations i, there is an associated value y_i, $i = 1, \ldots, n$, and the curves indicate the density of each class. The two classes are composed of a comparable number of observations. I.e. most of the black observations (the *negative* class) has y-values between -2 and 2 and most of the red observations (the *positive* class) has y values between 2 and 10.

By thresholding at different levels θ, i.e. assign each observation i to class 0 (the predicted negative class) if $y_i \leq \theta$ and otherwise to class 1 (the predicted positive class), one obtains different values of the TP and FN (true positives and false negatives) which in turn allow

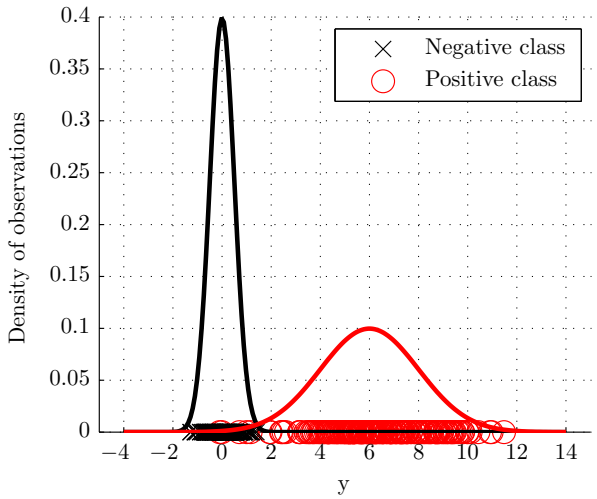

Figure 11: A simple two-class problem. Colors indicate the two classes and the curves indicates the density of each class. Each "point" is simply the value of y_i for observation i.

us to compute the TPR (true positive rate) curve. Which of the true positive rate (TPR) curves A,B,C or D shown in fig. 12 corresponds to the two-class problem of fig. 11?

A. Figure A

B. Figure B

C. Figure C

D. Figure D

E. Don't know.

Solution 24. Recall the true positive rate is defined as

$$\text{TPR} = \frac{\sharp\text{true positives}}{\sharp\text{total positives}}$$

Accordingly we only need to consider the positive class (red circles). Put in a different way, for a given threshold θ we have:

$$\text{TPR} = \frac{\sharp\text{red circles to the right of } \theta}{\sharp\text{red circles}}$$

so if we for instance select $\theta = 2$, the true positive rate should be about 0.95, ruling out all but option C. Alternatively the other options can be ruled out since they reflect changes in the area where we only have black squares (the negative class).

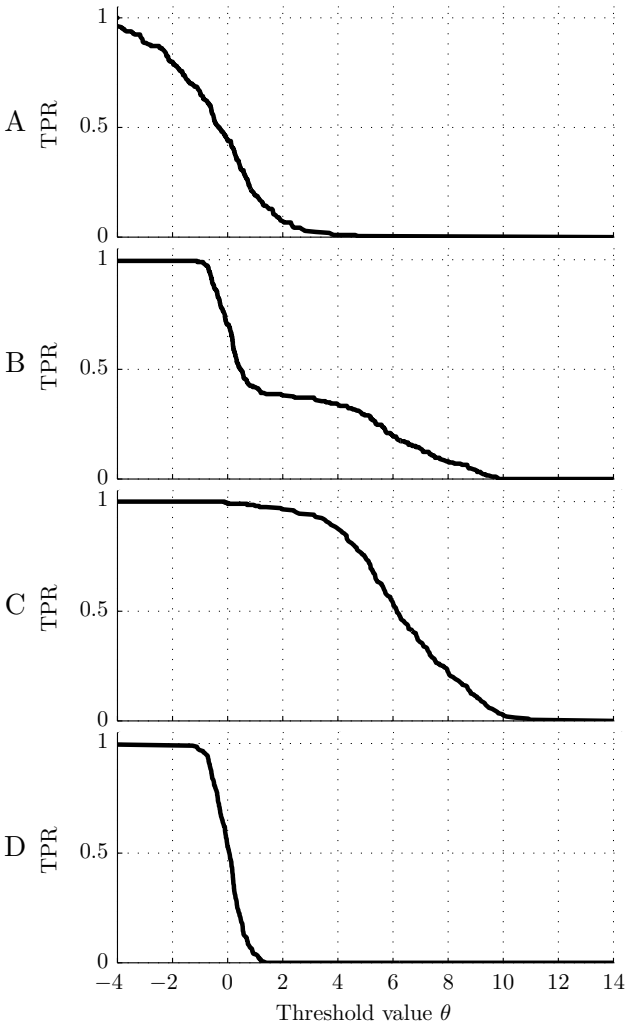

Figure 12: Proposed TPR (true positive rate) curves

Question 25. Suppose for a given problem the true positive rate (TPR) as a function of the threshold θ is as shown in fig. 12(D). Suppose we consider predictions made at a threshold of $\theta = 0$, and suppose we are told that the number of true positives at $\theta = 0$ is $TP = 113$ and the TPR at this threshold is $TPR = 0.55$, what is the (approximate) total number of observations in the positive class?

A. Actual positives: 252

B. Actual positives: 205

C. Actual positives: 159

D. Actual positives: 420

E. Don't know.

Solution 25. This is easily calculated by observing

$$\text{TPR} = \frac{\text{TP}}{\sharp\text{observations in the positive class}}$$

so

$$\sharp\text{positives} = \frac{113}{0.55} \approx 205.45$$

Question 26. Suppose a neural network classifier is applied to a small binary classification problem of only $N = 4$ observations shown in table 7. We attempt to improve the performance by applying AdaBoost (the version in the lecture notes, chapter 15). AdaBoost works by first sampling a new dataset D_1 with replacement, then training a classifier C_1 on D_1 and proceeding with the subsequent steps of the AdaBoost algorithm.

i	x_i	y_i	$C_1(x_i)$
1	50	1	1
2	22	1	0
3	20	0	1
4	76	0	0

Table 7: True values y_i and predictions $C_1(x_i) = \hat{y}_i$ for a neural network classifier C_1 trained on the (subsampled) dataset D_1 (see text) in an AdaBoost iteration.

Suppose in the first iteration of the AdaBoost algorithm a dataset D_1 is selected and the classifier C_1 trained on D_1. The predictions of the classifier C_1 is given in table 7. If AdaBoost is applied for $k = 1$

rounds of boosting what is the resulting (approximate) value for the weights $\boldsymbol{w}$?

A. $\boldsymbol{w} = \begin{bmatrix}0.072 & 0.428 & 0.428 & 0.072\end{bmatrix}$

B. $\boldsymbol{w} = \begin{bmatrix}0.019 & 0.481 & 0.481 & 0.019\end{bmatrix}$

C. $\boldsymbol{w} = \begin{bmatrix}0.250 & 0.250 & 0.250 & 0.250\end{bmatrix}$

D. $\boldsymbol{w} = \begin{bmatrix}0.130 & 0.370 & 0.370 & 0.130\end{bmatrix}$

E. Don't know.

Solution 26. The classifier C_1 classifies observation 2 and 3 incorrectly. From the lecture notes we have for a classifier C_i that

$$\varepsilon_i = \left[\sum_{j=1}^{N} w_j I\left(C_i(\boldsymbol{x}_j) \neq y_j\right)\right]$$
$$\alpha_i = \frac{1}{2}\log\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

and accordingly $\varepsilon_1 = \frac{1}{4} \times 2 = \frac{1}{2}$. This gives

$$\alpha_1 = \frac{1}{2}\log\frac{1-\frac{1}{2}}{\frac{1}{2}} = \frac{1}{2}\log 1$$

and so for $\boldsymbol{w}$ we get

$$\boldsymbol{w} \propto \begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}$$

and normalizing:

$$\boldsymbol{w} = \frac{1}{4}\begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}$$

accordingly option C is correct.

Question 27. Consider a neural network model applied to a dataset of $N = 1000$ observations. Suppose we wish to select both the optimal number of hidden units of the network as well as estimate the generalization error. To simplify the problem, we only consider 4 possible number of hidden units

$$n_{\text{hidden}} = 2, 4, 6, 8.$$

We opt for a two-level cross-validation strategy in which we use an inner loop of K_2-fold cross-validation to estimate the optimal number of units and an outer loop of K_1 fold cross-validation to estimate the generalization error. That is, for each of the K_1 outer folds, the dataset is divided into a validation set $D_{\text{validation}}$ and the remainder D_{CV} is used for the K_2-cross-validation to select the optimal number of neurons for this outer fold. Then, having estimated the optimal number of neurons for this outer fold, we train a new model on D_{CV} and use it to predict the values in $D_{\text{validation}}$ in order to estimate the generalization error. The full generalization error is obtained as the average of the K_1 outer folds.

Suppose we select $K_1 = 5$ and $K_2 = 10$. How many times in total must we *train* a neural network model?

A. 205

B. 200

C. 210

D. 55

E. Don't know.

Solution 27. This can easily be obtained noting for each of the K_1 outer folds we must both (i) train K_2 models on the $L = 4$ different settings of the number of hidden units (ii) train a single new model to estimate the generalization error for this fold. Accoringly the number of trained models is

$$K_1(K_2 L + 1) = 5(10 \cdot 4 + 1) = 205$$

Technical University of Denmark

Written examination: 16th December 2015, 9 AM - 1 PM. Page 1 of 17 pages.

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.

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

Please write your name and student number clearly. Only the present page (page 1) gives your answers to the written test. Other pages will not be considered.

Answers:

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

Name: _____

Student number: _____

ONLY THIS PAGE IS USED FOR THE EVALUATION.
ALL 17 PAGES MUST BE HANDED IN.

No.	Attribute description	Abbrev.
x_1	Age	AGE
x_2	Blood pressure	BP
x_3	Blood glucose random	BGR
x_4	Blood urea	BU
x_5	Serum creatinine	SC
x_6	Hemoglobin	HEMO
y	Chronic kidney disease (0: no 1: yes)	CKD

Table 1: Attributes of the *Chronic Kidney Disease* dataset. The dataset includes 6 attributes $(x_1, \ldots, x_6)$ of 294 person and whether they have a chronic kidney disease or not.

Question 1. Consider the *Chronic Kidney Disease* dataset with attributes given in Table 1[1]. Notice, the dataset has been pre-processed for this exam and only some of the attributes in the original data are presently considered whereas persons with missing data have been removed. A boxplot of the attributes are given in Figure 1. Which of the following statements is true?

A. Regression methods are more suitable than classification methods to predict the output y for this data.

B. BP is ratio whereas y is nominal.

C. All the attributes appear to be normal distributed.

D. From the boxplots it can be seen that there are many outliers in the data that have to be removed.

E. Don't know.

Solution 1.

A As y is nominal classification approaches are more well suited than regression methods.

B Indeed the BP attribute is ratio as zero would constitute an absence of what is being measured. As y is one or zero indicating if the person has a chronic kidney disease or not this is nominal. Thus, this answer option is correct.

C Not all attributes appear to be normal distributed. For instance BP seems to have the 50th and 75th percentiles coincide which would not be the case had the attribute been normal distributed.

[1]Dataset obtained from http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

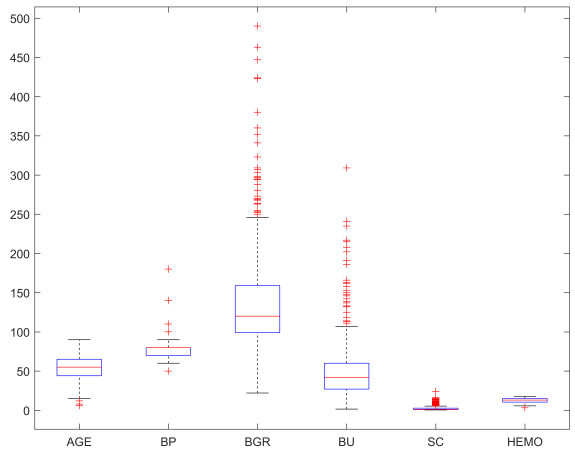

Figure 1: Boxplots of the six attributes of the Chronic Kidney Disease data.

D Even though outliers are indicated in red in the boxplot this is only because they are outside the 1.5 times interquartile range, but there is no reason to believe these values should not be correct.

Question 2. A principal component analysis is carried out on the *Chronic Kidney Disease* dataset based on the attributes $x_1, \ldots, x_6$ found in Table 1. We standardize the data, i.e. by subtracting the mean from each attribute and dividing each attribute by its standard deviation to form the standardized data matrix $\tilde{\boldsymbol{X}}$ of size 294 *subjects* $\times$ 6 *attributes* and apply a singular value decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$ where

$$\boldsymbol{S} = \begin{bmatrix} 27.9 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 18.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 15.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 14.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 11.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 8.4 \end{bmatrix}$$

$$\boldsymbol{V} = \begin{bmatrix} 0.26 & -0.57 & 0.40 & 0.66 & -0.07 & 0.01 \\ 0.29 & -0.19 & -0.89 & 0.25 & -0.12 & -0.05 \\ 0.27 & -0.62 & 0.06 & -0.70 & -0.21 & -0.00 \\ 0.51 & 0.34 & 0.16 & -0.02 & -0.22 & -0.74 \\ 0.50 & 0.37 & 0.11 & 0.01 & -0.42 & 0.66 \\ -0.51 & -0.04 & -0.01 & 0.09 & -0.85 & -0.13 \end{bmatrix}.$$

We note that the entries of the matrices above have been rounded. Which one of the following statements is true?

A. The first two principal components account for more than 60 % of the variance in the data.

B. The last principal component accounts for less than 2 % of the variance in the data.

C. The first principal component accounts for more than 50 % of the variance in the data.

D. The performed principal component analysis will mainly be driven by BGR as this attribute has the largest variance.

E. Don't know.

Solution 2. Recall that the variance of the first k components are

$$\text{var.} = \frac{\sum_{i=1}^{k} S_{ii}^2}{\sum_{j=1}^{6} S_{jj}^2}$$

Thus, the first two principal components account for $\frac{27.9^2+18^2}{27.9^2+18^2+15.8^2+14.5^2+11.1^2+8.4^2} = 0.6278$, whereas the last principal component accounts for $\frac{8.4^2}{27.9^2+18^2+15.8^2+14.5^2+11.1^2+8.4^2} = 0.0402$, and the first principal component for $\frac{27.9^2}{27.9^2+18^2+15.8^2+14.5^2+11.1^2+8.4^2} = 0.4433$ of the variance. Since the data has been standardized each attribute is given equal importance in the PCA. However, had the data not been standardized the analysis would be highly driven by BGR and it would be expected the first principal component would mainly capture variance along the direction of BGR.

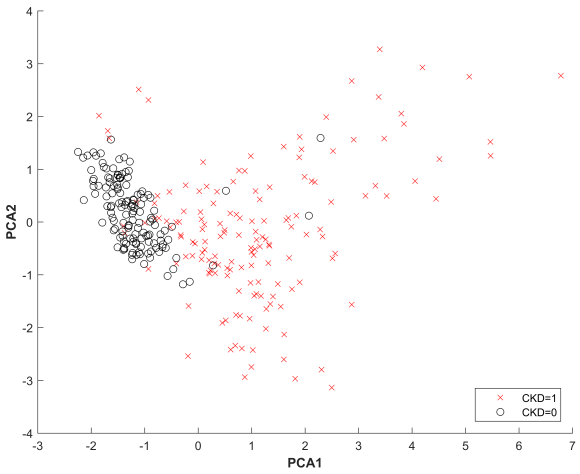

Figure 2: PCA of the chronic kidney disease data.

Question 3. In Figure 2 is given the data projected onto the first two principal components where observations corresponding to $CKD = 1$ is marked by red crosses (x) whereas observations corrresponding to $CKD = 0$ are makred by black circles (○). Which statement is correct?

A. The observations projected onto the first principal component direction can be found as the first column of the matrix $\tilde{\boldsymbol{X}}\boldsymbol{S}^\top$.

B. It seems to be more likely to have chronic kidney disease if a person has relative low values of AGE, BP, BGR, BU, SC and a relatively high value of $HEMO$.

C. Principal component three appears to mainly be discriminating old people with low blood pressure from young people with high blood pressure.

D. Principal component analysis identifies features that are optimal for discriminating between persons with chronic kidney disease from persons not having chronic kidney disease.

E. Don't know.

Solution 3. The data projected onto the first PCA direction is given by $X\boldsymbol{v}_1$. According to the data projected onto the first PCA given in figure fig. 2 it appears that high and not low values of AGE, BP, BGR, BU, SC and low not high values of $HEMO$ would result in a large projection onto PCA which is where most red crosses indicating observations with chronic kidney disease are found. Indeed it appears as if principal

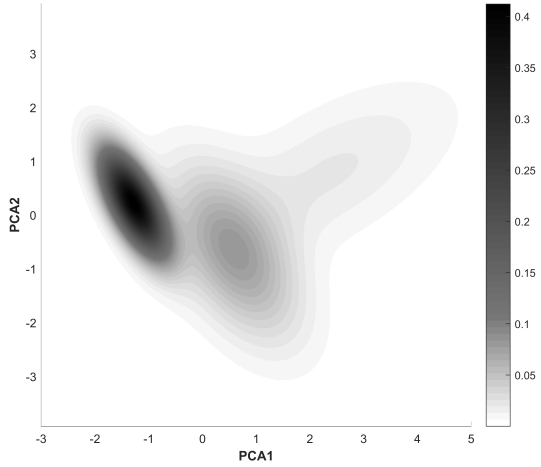

Figure 3: The density of a GMM fitted to the data projected onto the first two principal components given in Figure 2.

component three is mainly discriminating old people with low blood pressure from young people with high blood pressure as the first and second coefficients with largest magnitude pertain to x_1 and x_2 are 0.40 and -0.89 thus having opposing signs. PCA is optimized for accounting for variance and not for discrimination persons with and without chronic kidney disease.

Question 4. A Gaussian Mixture model is fitted to the data projected onto the first two principal components using K=3 mixture components. The estimated density is shown in Figure 3. Which of the following expressions corresponds to the estimated density?

A.

$$p(\boldsymbol{x}) = 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & -0.44 \\ -0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & -0.16 \\ -0.16 & 0.39 \end{bmatrix}).$$

B.

$$p(\boldsymbol{x}) = 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & -1.06 \\ -1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & 0.44 \\ 0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & 0.16 \\ 0.16 & 0.39 \end{bmatrix}).$$

C.

$$p(\boldsymbol{x}) = 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.16 & -0.16 \\ -0.16 & 0.39 \end{bmatrix}) + 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.64 & -0.44 \\ -0.44 & 1.13 \end{bmatrix}).$$

D.

$$p(\boldsymbol{x}) = 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & 0.44 \\ 0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & 0.16 \\ 0.16 & 0.39 \end{bmatrix}).$$

E. Don't know.

Solution 4. The density is:

$$p(x) = 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & -0.44 \\ -0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & -0.16 \\ -0.16 & 0.39 \end{bmatrix}).$$

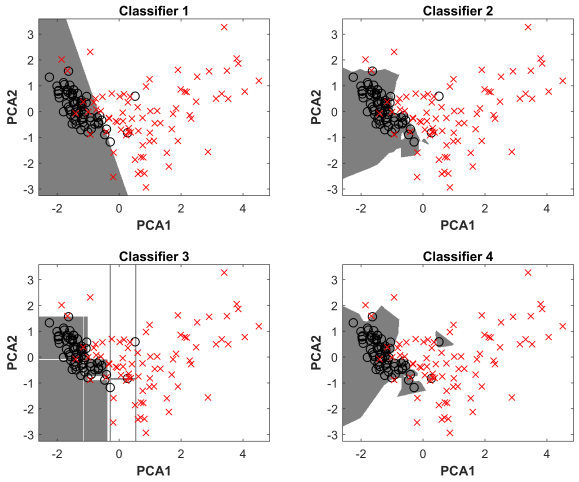

Figure 4: Four classifiers trained on half of the data using PCA1 and PCA2 as features for the classifier.

This can be observed sa the cluster with center at $\begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}$ has positive covariance between PCA1 and PCA2 whereas the remaining two clusters have negative covariance. While the first cluster also have the lowest density values, i.e. lowest value of the mixing proportions. This is only the case for this density.

Question 5. Using the data projected onto the first two principal components four different classifiers are trained on half of the data and their decision boundaries given in Figure 4. Which one of the following statements is true?

A. Classifier 1 is a decision tree, Classifier 2 is a one-nearest-neighbour classifier, Classifier 3 is a logistic regression model, and Classifier 4 is an artificial neural network with 5 hidden units.

B. Classifier 1 is a logistic regression, Classifier 2 is a one-nearest-neighbour classifier, Classifier 3 is a decision tree, and Classifier 4 is an artificial neural network with 5 hidden units.

C. Classifier 1 is a three-nearest-neighbour classifier, Classifier 2 is a decision tree, Classifier 3 is a logistic regression, and Classifier 4 is an artificial neural network with 5 hidden units.

D. Classifier 1 is a logistic regression, Classifier 2 is a three-nearest-neighbour classifier, Classifier 3 is a decision tree, and Classifier 4 is a one-nearest neighbor classifier.

E. Don't know.

Solution 5. Classifier 1 is a logistic regression as the boundary is defined by a straight line, Classifier 2 is a three-nearest neighbor classifier since single observations are not surrounded by a decision boundary for their classes, Classifier 3 is a decision tree which is observed from the horizontal and vertical lines, and Classifier 4 is a one-nearest-neighbour classifier as it can be observed that single observations are surrounded by decision boundaries pertaining to them .

Feature(s)	Training ErrorRate	Test ErrorRate
x_1	0.3537	0.3061
x_2	0.4286	0.4422
x_3	0.3605	0.2517
x_4	0.2993	0.3061
x_1 and x_2	0.3265	0.3401
x_1 and x_3	0.3401	0.2653
x_1 and x_4	0.2517	0.2381
x_2 and x_3	0.2857	0.2653
x_2 and x_4	0.2245	0.2449
x_3 and x_4	0.1701	0.1497
x_1 and x_2 and x_3	0.2653	0.2449
x_1 and x_2 and x_4	0.2041	0.2313
x_1 and x_3 and x_4	0.1701	0.1429
x_2 and x_3 and x_4	0.1769	0.1565
x_1 and x_2 and x_3 and x_4	0.1701	0.1633

Table 2: Error rate for the training and test set when using a logistic regression to predict kidney disease based only on the four attributes x_1, x_2, x_3 and x_4.

Question 6. A logistic regression classifier is trained using only combinations of x_1, x_2, x_3 and x_4 and the error rate on the training and test data where half the data again is used for training and the other half for testing is given in Table 2. Which one of the following statements is correct?

A. Forward selection will select the feature set x_3 and x_4.

B. Forward selection will select the feature set x_1 and x_3 and x_4.

C. Backward selection will select the features set x_1 and x_2 and x_3 and x_4.

D. Backward selection will select the feature set x_3 and x_4.

E. Don't know.

Solution 6. Using forward selection, x_3 will initially be selected according to the test error rate, next x_4 and subsequently x_1. As the error rate does not improve by selecting x_2 the forward selection procedure will terminate choosing x_1, x_3, x_4. Backward selection will also terminate at this features set.

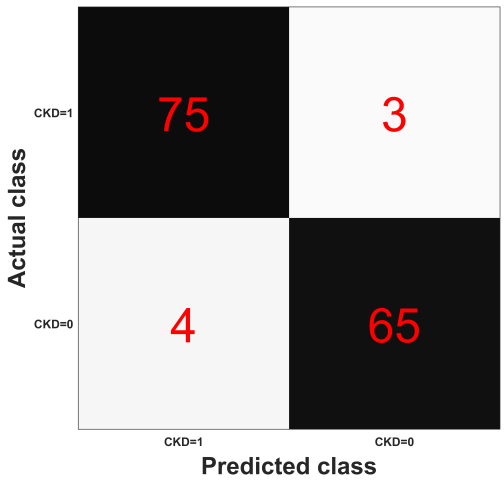

Figure 5: The confusion matrix of a logistic regression classifier evaluated on the test data.

Question 7. The confusion matrix of a logistic regression classifier evaluated on the test data is given in Figure 5. To generate the confusion matrix data has been split in two, half of the data is used for training the classifier and the other half for testing. We will presently consider CKD=1 as the positive class and CKD=0 as the negative class of the classifier. Which one of the following statements is true?

A. The accuracy of the classifier is 7/147 and the error rate is 140/147.

B. The Precision of the classifier is 75/78.

C. The used procedure corresponds to two-fold cross-validation.

D. The used procedure corresponds to the holdout method.

E. Don't know.

Solution 7. The accuracy is 140/147 and the error rate 7/147, not the reverse. The precision of the classifier is 75/(75+4), see also p. 297. The used procedure corresponds to the hold-out method where 50% of the data has been hold out and not two fold-cross validation as only one and not two models are trained.

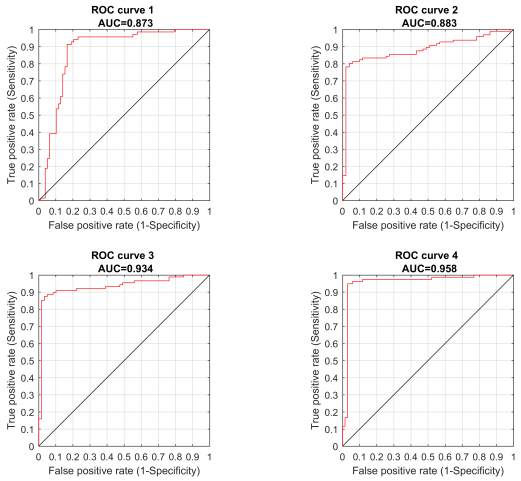

Figure 6: The Receiver Operator Characteristic (ROC) curve for four different classifiers.

Question 8. The performance of four different classifiers are given in terms of their Receiver Operator Characteristic (ROC) curve in Figure 6. One of the classifiers correspond to the classifier with confusion matrix given in Figure 5, which one?

A. ROC curve 1.

B. ROC curve 2.

C. ROC curve 3.

D. ROC curve 4.

E. Don't know.

Solution 8. According to the confusion matrix there exist a threshold in which the true positive rate TPR=75/(75+3)=0.9615 and false positive rate FPR=4/(4+65)=0.0580. This only holds for ROC curve 4.

	O1	O2	O3	O4	O5	O6	O7
O1	0	69	55	117	50	326	36
O2	69	0	36	128	104	303	85
O3	55	36	0	129	94	314	78
O4	117	128	129	0	85	220	91
O5	50	104	94	85	0	303	23
O6	326	303	314	220	303	0	307
O7	36	85	78	91	23	307	0

Table 3: Pairwise Euclidean distance, i.e $d(Oa, Ob) = ||\boldsymbol{x}_a - \boldsymbol{x}_b||_2 = \sqrt{\sum_m (x_{am} - x_{bm})^2}$, between the first four subjects with and first three subjects without chronic kidney disease respectively. Red observations (i.e., O1, O2, O3, and O4) correspond to the four subjects having chronic kidney disease (CKD=1) whereas black observations (i.e., O5, O6, O7) correspond to the four subjects without (CKD=0).

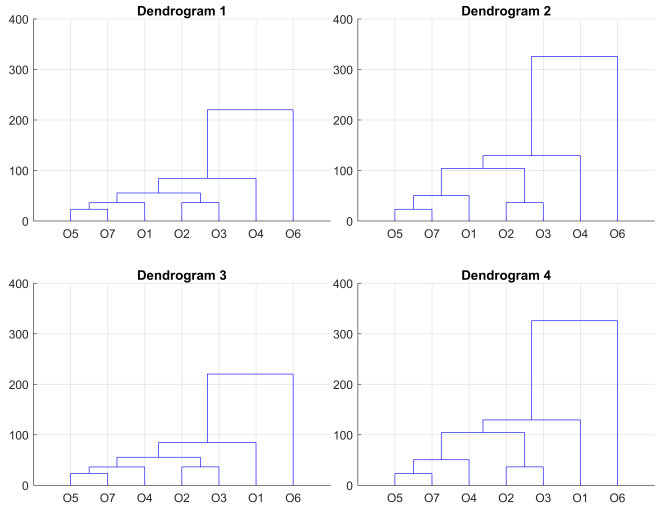

Figure 7: Hierarchical clustering of the seven observations considered in Table 3.

Question 9. In Table 3 is given the pairwise distances between the first four subjects with and the first three subjects without chronic kidney disease. A hierarchical clustering is used to cluster these seven observations using complete (i.e., maximum) linkage. Which one of the dendrograms given in Figure 7 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 9. In complete linkage the observations that is the furthest between the clusters define the level in

which they merge. Initially, O5 and O7 will merge at 23 and O2 and O3 at 36. Next O1 will merge with O5, O7 at 50 and subsequently O2, O3 will merge with O1, O5, O7 at 104, then O4 with O1, O2, O3, O5, O7 at 129, and finally 06 will merge with O1, O2, O3, O4, O5, O7 at 326. This corresponds to dendrogram 2.

Question 10. In order to predict if an observation corresponds to a subject having chronic kidney disease or not we will use a k-nearest neighbor (KNN) classifier based on the Euclidean distance between the seven observations given in Table 3. We will use leave-one-out cross-validation for the KNN in order to classify whether the seven considered observations constitute subjects with or without chronic kidney disease (CKD=1 given in red, i.e. observation O1, O2, O3, O4) and (CKD=0 given in black, i.e. observation O5, O6, O7) using a five-nearest neighbor classifier, i.e. $K = 5$. The analysis will be based only on the data given in Table 3. Which one of the following statements is *correct*?

A. The error rate of the classifier will be 3/7.

B. The error rate of the classifier will be 4/7.

C. All subjects without chronic kidney disease will be correctly classified.

D. All subjects will be correctly classified.

E. Don't know.

Solution 10. All subjects with chronic kidney disease, i.e. O1, O2, O3, and O4 will be correctly classified as there neighbors will be the other observations except observation O6. As there are only two observations that do not have chronic kidney disease available when using leave-one-out the majority will be having chronic kidney disease thus observation O5, O6 and O7 will be miss-classified.

Question 11. We suspect that observation O6 in Table 3 is an outlier. Which procedure would *not* indicate that O6 is the strongest candidate of being an outlier of the seven observations O1–O7?

A. Using the inverse average distance to K-nearest neighbor as density with K=1.

B. Using the average relative density with K=3.

C. Hierarchical clustering using single linkage when inspecting the top split in the dendrogram.

D. Using the density obtained from a Gaussian Mixture Model (GMM) having two clusters with the first cluster mean fixed at observation O1 and the second cluster mean fixed at observation O6 (i.e., the mean values are not updated during the M-step).

E. Don't know.

Solution 11. Using the inverse average distance to K-nearest neighbours as density measure, average relative density, and hierarchical clustering will all clearly point to observation O6 being an outlier. However, the Gaussian Mixture Model with two components, one centered on observation O1 and one centered on observation O6 would create a covariance matrix for the second cluster that would be very small therefore providing a very high density of observation O6 that would therefore not be indicated to be an outlier from the density value.

No.	Attribute description	Abbrev.
x_1	Red blood cells	RBC
x_2	Pus cell	PC
x_3	Pus cell clumps	PCC
x_4	Hypertension	HTN
x_5	Diabetes mellitus	DM
x_6	Coronary artery disease	CAD
x_7	Pedal edema	PE
y	Chronic kidney disease (0: no 1: yes)	CKD

Table 4: Binary attributes of the *Chronic Kidney Disease* dataset. The dataset includes seven attributes $(x_1, \ldots, x_7)$ of 232 persons and whether they have a chronic kidney disease or not.

Question 12. We will now consider some of the Binary attributes also available in the original Chronic Kidney Disease dataset[2]. These binary attributes are given in Table 4. Using these attributes it is found that:

- 56.27 % of the subjects have chronic kidney disease (CKD=1).
- 49.46 % of the subjects with chronic kidney disease (CKD=1) have coronary artery disease (CAD=1).
- 0.7 % of the subjects without chronic kidney disease (CKD=0) have coronary artery disease (CAD=1).

According to this data what is the probability that a subject that has coronary artery disease also has chronic kidney disease?

A. 27.83 %

B. 56.27 %

C. 87.89 %

D. 98.91 %

E. Don't know.

Solution 12. Using Bayes theorem we have :

$$P(CKD=1|CAD=1) = \frac{P(CAD=1|CKD=1)P(CKD=1)}{P(CAD=1)}$$
$$= \frac{P(CAD=1|CKD=1)P(CKD=1)}{P(CAD=1|CKD=1)P(CKD=1)+P(CAD=1|CKD=0)P(CKD=0)}$$
$$= \frac{0.4946 \cdot 0.5627}{0.4946 \cdot 0.5627 + 0.007 \cdot (1-0.5627)}$$
$$= 0.9891 \approx 98.91\%$$

[2]Dataset obtained from http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

	RBC	PC	PCC	HTN	DM	CAD	PE
O_1	0	0	0	0	1	0	0
O_2	0	1	1	1	0	0	1
O_3	0	0	0	0	0	0	0
O_4	0	1	0	0	1	0	1
O_5	0	1	1	1	1	0	0
O_6	1	1	1	1	1	0	0
O_7	1	1	1	1	1	0	1
O_8	0	1	1	1	1	1	1
O_9	0	1	0	1	0	0	0
O_{10}	1	1	0	0	0	1	0
O_{11}	0	0	0	0	1	0	0
O_{12}	0	0	0	0	0	0	0
O_{13}	0	0	0	0	0	0	0
O_{14}	0	0	0	0	0	0	0
O_{15}	0	0	0	0	0	0	0

Table 5: For each observation there are $M = 7$ binary features and $N = 15$ observations $O_1, \ldots, O_{15}$ belonging to two categories (i.e., CKD=1 for $O_1, \ldots, O_9$ and CKD=0 for $O_{10}, \ldots, O_{15}$).

Question 13. We consider the fifteen subjects given in Table 5. We will consider this data set a market basket problem in which the fifteen subjects have various combinations of the seven items denoted RBC, PC, PCC, HTN, DM, CAD, PE. Which one of the proposed solutions below includes *all* the frequent itemsets with support of more than 30 %?

A. {PC}, {PCC}, {HTN}, and {DM}.

B. {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, and {PCC, HTN}.

C. {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, {PC, DM}, and {PCC, HTN}.

D. {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, {PC, DM}, {PCC, HTN}, {PC, PCC, HTN}.

E. Don't know.

Solution 13. Support of more than 30 % implies that there has to at least be 5 observations in an itemset (i.e. 33% of observations). This is the case for the itemsets: {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, {PC, DM}, {PCC, HTN}, {PC, PCC, HTN}.

Question 14. What is the support and confidence for the decision rule {RBC, PC} → {CAD}?

A. The support is 1/15 and the confidence is 1/3.

B. The support is 1/3 and the confidence is 1/15.

C. The support is 1/3 and the confidence is 1/2.

D. The support is 1/3 and the confidence is 2/15.

E. Don't know.

Solution 14. The support is given by the support of the itemset {RBC, PC, CAD} which is 1/15. The confidence is given as the number of transactions with {RBC, PC, CAD} divided by the number of transactions with {RBC, PC} which is 1/3.

Question 15. We will use a one-nearest neighbor classifier to classify observations as having chronic kidney disease (CKD=1) or not having chronic kidney disease (CKD=0) respectively using the data in Table 5. For the one-nearest neighbor we will use as distance measure $d(O_x, O_y) = 1/SMC(O_x, O_y)$, i.e. high $SMC(O_x, O_y)$ value implies a low $d(O_x, O_y)$ value. If several observations are equally close we will use majority voting to determine the class. Which one of the following statements is correct?

A. The first observation O_1 will be *correctly* classified.

B. The third observation O_3 will be *correctly* classified.

C. The fifth observation O_5 will be *correctly* classified.

D. The twelfth observation O_{12} will be *incorrectly* classified.

E. Don't know.

Solution 15. Recalling that the simple matching coefficient between two binary observations is $SMC(O_x, O_y) = \frac{f_{00}(O_x,O_y)+f_{11}(O_x,O_y)}{M}$ we have that the inverse of the SMC similarity is $SMC^{-1}(O_x, O_y) = \frac{M}{f_{11}(O_x,O_y)+f_{00}(O_x,O_y)}$. Thus:
O_1 is closest to O_{11} and will be miss-classified
O_3 is equally close to O_{12}, O_{13}, O_{14} and will be miss-classified
O_5 is closest to O_6 and will be correctly classified.

Question 16. Nine of the fifteen observations in Table 5 have chronic kidney disease (i.e., O_1–O_9 given in red) whereas six of the observations do not have chronic kidney disease (i.e., O_{10}–O_{15} given in black). We would like to predict whether a subject has chronic kidney disease or not using the data in Table 5 and the attributes RBC, PC, DM, and CAD. We will apply a Naïve Bayes classifier that assumes independence between the four attributes. Given that a subject has these four attributes (i.e., $RBC = 1$, $PC = 1$, $DM = 1$, and $CAD = 1$) what is the probability that the person has chronic kidney disease, i.e., what is $P(CKD = 1|RBC = 1, PC = 1, DM = 1, CAD = 1)$ according to the Naïve Bayes classifier?

A. 2.56 %

B. 96.14 %

C. 98.03 %

D. 100 %

E. Don't know.

Solution 16. According to the Naïve Bayes classifier we have

$$P(CKD = 1|RBC = 1, PC = 1, DM = 1, CAD = 1) = \frac{\begin{pmatrix} P(RBC = 1|CKD = 1)\times \\ P(PC = 1|CKD = 1)\times \\ P(DM = 1|CKD = 1)\times \\ P(CAD = 1|CKD = 1)\times \\ P(CKD = 1) \end{pmatrix}}{\begin{pmatrix} P(RBC = 1|CKD = 1)\times \\ P(PC = 1|CKD = 1)\times \\ P(DM = 1|CKD = 1)\times \\ P(CAD = 1|CKD = 1)\times \\ P(CKD = 1) \end{pmatrix} + \begin{pmatrix} P(RBC = 1|CKD = 0)\times \\ P(PC = 1|CKD = 0)\times \\ P(DM = 1|CKD = 0)\times \\ P(CAD = 1|CKD = 0)\times \\ P(CKD = 0) \end{pmatrix}}$$

$$= \frac{2/9 \cdot 7/9 \cdot 6/9 \cdot 1/9 \cdot 9/15}{2/9 \cdot 7/9 \cdot 6/9 \cdot 1/9 \cdot 9/15 + 1/6 \cdot 1/6 \cdot 1/6 \cdot 1/6 \cdot 6/15} = 0.9614.$$

Question 17. We will consider a Bayes classifier using the attributes RBC, PC, and DM in Table 5 (i.e., we no longer consider the attribute CAD). What is $P(CKD = 1|RBC = 1, PC = 1, DM = 1)$ according to a Bayes classifier (i.e. we are no longer imposing independence as in the Naïve Bayes classifier)?

A. 26.67 %

B. 97.07 %

C. 98.03 %

D. 100 %

E. Don't know.

Solution 17. According to the Bayes classifier we have

$$P(CKD = 1|RBC = 1, PC = 1, DM = 1) = \frac{\left(\begin{array}{r} P(RBC = 1, PC = 1, DM = 1|CKD = 1)\times \\ P(CKD = 1) \end{array}\right)}{\left(\begin{array}{r} P(PRBC = 1, PC = 1, DM = 1|CKD = 1)\times \\ P(CKD = 1) \end{array}\right) + \left(\begin{array}{r} P(RBC = 1, PC = 1, DM = 1|CKD = 0)\times \\ P(CKD = 0) \end{array}\right)} = \frac{2/9 \cdot 9/15}{2/9 \cdot 9/15 + 0/6 \cdot 6/15} = 1$$

Question 18. We will use the data in Table 5 to build a decision tree. We will consider splitting at the root of the tree according to the attribute PC (i.e., according to whether $PC = 0$ or $PC = 1$). What is the purity gain, Δ, of splitting according to PC using the classification error as impurity measure $I(t)$, (i.e., $I(t) = 1 - \max_i[p(i|t)]$)?

A. 2/15

B. 1/5

C. 4/15

D. 2/5

E. Don't know.

Solution 18. The purity gain is given by $\Delta = I(Parent) - (N_{Left}/N \cdot I(left) + N_{Rigth}/N \cdot I(Right))$ where $I = 1 - max_c p(c|j)$. At the root of the tree we have:
$I(Parent) = 1 - 9/15 = 6/15$. For the attribute conditions we obtain:
PC: $N_{Left}/N \cdot I(left) + N_{Right}/N \cdot I(Right) = 7/15 \cdot (1 - 5/7) + 8/15 \cdot (1 - 7/8) = 3/15$ Thus, the purity gain is given as $\Delta = 6/15 - 3/15 = 3/15$.

Question 19. We cluster the binary data considering only the attribute RBC according to K-means using euclidean distance with two clusters having centroids in 0 and 1 respectively. The purity of the clustering is given as:

$$Purity = \sum_{i=1}^{K} \frac{m_i}{m} p_i, \text{ where } p_i = \max_j m_{ij}/m_i$$

where m_i is the number of observation in cluster i, m the total number of observations in all clusters and m_{ij} the number of observations in cluster i of class j. What is the Purity of the clustering?

A. Purity=1/5

B. Purity=3/5

C. Purity=2/3

D. Purity=12/15

E. Don't know.

Solution 19. As cluster 1 containing the observations having $RBC = 0$ has $O_1, O_2, O_3, O_4, 0_5, O_8, O_9, O_{11}, O_{12}, O_{13}, O_{14}, O_{15}$ with seven chronic kidney disease observations, we have that $p_1 = \max\{7/12, 5/12\} = 7/12$. Whereas cluster 2 containing the observations having $RBC = 1$ are O_6, O_7, O_{10} with two chronic kidney disease observations, we have that $p_1 = \max\{2/3, 1/3\} = 2/3$. Thus $Purity = 12/15 \cdot 7/12 + 3/15 \cdot 2/3 = 9/15 = 3/5$

Question 20. Using all 232 observations of the binary data in Table 4 we would like to investigate the generalization performance of logistic regression using two-level cross-validation where we in the outer folds use leave-one-out cross validation and in the inner folds use five-fold cross-validation. In our inner cross-validation we determine the optimal feature combination from all potential combinations of the seven attributes (RBC, PC, PCC, HTN, DM, CAD, and PE) providing $2^7 = 128$ different logistic regression models. Once we from the inner folds have determined the optimal feature combination we train a model using this feature combination on all the training data from the outer fold and estimate the generalization error on the test data of the outer fold. Which one of the following statements is correct?

A. In total 29696 logistic regression models will be trained.

B. In total 148480 logistic regression models will be trained.

C. In total 148712 logistic regression models will be trained.

D. In total 178176 logistic regression models will be trained.

E. Don't know.

Solution 20. As we use leave-one-out-cross validation in the outer fold we have 232 splits into training and testing. For each of these splits we have 5 folds where we evaluate for each of these 5 folds 128 different model settings which gives 640 models. Once we have found the optimal setting we train one additional model on all the training data which provide one additional model to fit, i.e. a total of 641 models fitted 232 times (leave-one-cross validation provides 232 folds) which gives $641 \cdot 232 = 148712$ models. endsolution

Question 21. In Figure 8 is given a boxplot of an attribute denoted A. Which of the following statements regarding this attribute is correct?

A. The mean value of the attribute A is 0.

B. The mode of the attribute A is 0.

C. The range of the attribute A is 5.

D. The attribute A appears to be normal distributed.

E. Don't know.

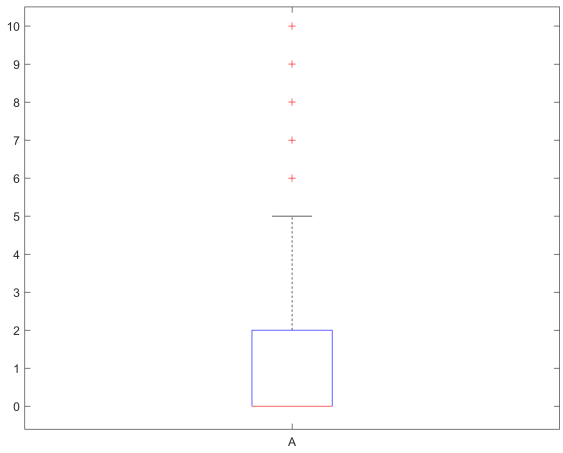

Figure 8: Boxplot of an attribute denoted A.

Solution 21. The mean value of the attribute A is greater than zero as the smallest value of A is zero and there are several values larger than zero, thus the mean will also be larger than zero. The mode is zero as the 50th percentile is located at 0 with at least 50% of the observations therefore taking the value 0. The range of A is 10 with smallest value of zero and largest of 10. The attribute A does not appear to be normally distributed as the distribution is highly asymmetric/skewed.

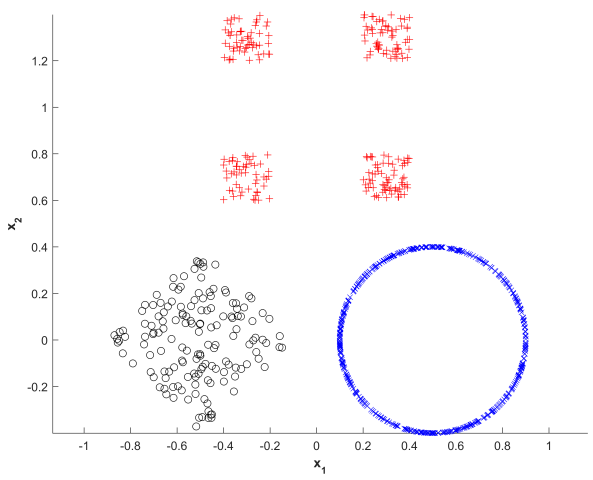

Figure 9: A dataset with two attributes x_1 and x_2 and three clusters given by red plus (+), black circles ($\circ$) and blue crosses (x).

Question 22. We will use the decision tree given in Figure 10 to attempt to separate the observations into the three classes (i.e. red pluses, black circles, and blue crosses) given in Figure 9. Which one of the following choices for the two decisions A, and B in the decision tree would be the most well suited to separate the three classes?

A. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_2 < 0.4$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \|_1 < 1$

B. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_2 < 0.4$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \|_2 < 0.5$

C. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_\infty < 0.5$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \|_1 < 0.5$

D. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_1 < 0.5$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \|_2 < 0.5$

E. Don't know.

Solution 22. The three classes would be well separated by the decisions

A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_\infty < 0.5$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \|_1 < 05$
Decision A would capture the red pluses within a square box centered at (0,1) and radius 0.5. Decision B will separate the black circles from blue croses as the black circles are well contained within a diamond shape with radius 0.5 centered at (-0.5, 0) forming the 1-norm.

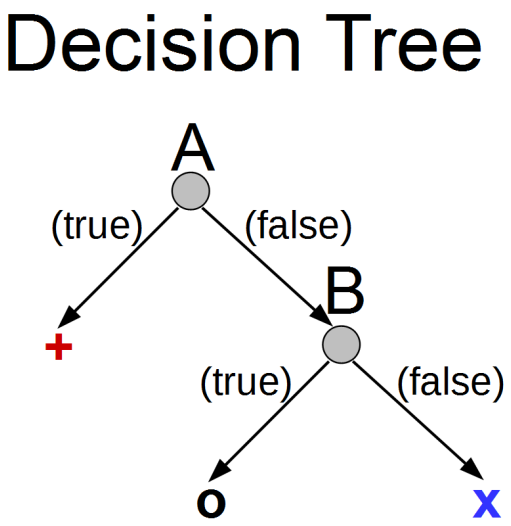

Figure 10: A decision tree with two decisions denoted A and B resulting in a separation into the three clusters given in Figure 9.

Question 23. We again consider the data given in Figure 9 containing the three classes given by red pluses (+), black circles (∘) and blue crosses(x). Which one of the following clustering approaches is most suited to separate the data into the three classes?

A. Well-separated.

B. Center-based.

C. Hierarchical clustering using single linkage.

D. Density-based.

E. Don't know.

Solution 23. For Well-separated we have: Each point is closer to all points in its cluster than any point in another cluster.
Center-based: Each point is closer to the center of its cluster than to the center of any other cluster.
Hierarchical clustering using single linkage: Clusters are merged according to their minimal distance between two points.
Density-based: Clusters are regions of high density separated by regions of low density.
The center-based definition would adequately separate the three classes into clusters as they would all be closer to their center than the center of other clusters. Well-separated would not work since for instance the lower right red observations are closer to some of the blue observations. Contiguity based would at first seem reasonable but it would fail when merging the lower right red cluster as this is closer to the some blue observations. Density based requires clusters be regions of high density separated by regions of low density. However, the red cluster itself contains low-density regions and would therefore not be adequately defined.

Question 24. Which one of the following statements regarding ensemble methods is correct?

A. Ensemble methods aim at combining weak classifiers that each are as similar to each other as possible in terms of how they make predictions.

B. Random Forest corresponds to combining multiple decision trees where each decision tree is trained using features selected according to the AdaBoost sampling procedure.

C. In each Bagging round the same observation can occur more than once in the training set.

D. Bagging puts more emphasis on miss-classified observations.

E. Don't know.

Solution 24. Ensemble methods aim at combining weak classifiers that are independent of each other and not as similar to each other as possible. Random forest randomly sub-samples when training each decision tree however it generally considers more than one attribute for each tree. Bagging uses sampling with replacement, thus, the same observation can occur more than once in the training set. However, Bagging does not put more emphasis on miss-classified observations - this is the strategy of Boosting.

X	2	4	8	11	15	19	20	27	30	31

Table 6: Simple 1-dimensional dataset with $N = 10$ observations.

Question 25. Consider the 1-dimensional data set having $N = 10$ observations shown in table 6. We will cluster the data using K-means based on Euclidean distance and initialized with centroids positioned at the first three observations, i.e. cluster one is located at $x_1 = 2$, cluster two at $x_2 = 4$ and cluster three at $x_3 = 8$. We will use the basic K-means algorithm described in the book and lecture slides. What will be the converged solution of the K-means procedure? (Note: This exercise cannot be solved using computer implementations of K-means as they may use update schemes that are not the same as the basic algorithm given in the book and lecture slides.)

A. {2, 4}, {8, 11, 15}, {19, 20, 27, 30, 31}.

B. {2, 4}, {8, 11, 15, 19, 20}, {27, 30, 31}.

C. {2, 4, 8}, {11, 15, 19, 20}, {27, 30, 31}.

D. {2, 4, 8, 11},{15, 19, 20},{ 27, 30, 31}.

E. Don't know.

Solution 25. Initially only cluster 3 will be closest to the remaining observations and its centroid updated to 20.1250. Subsequently, cluster 2 will be closer to 8 and 11 and cluster 2 and cluster 3 therefore updated to: Cluster 1: 3, cluster 2: 7.66, cluster 3: 23.66
Subsequently the centroids will be updated to: Cluster 1: 3, cluster 2: 11.33 cluster 3: 25.40
After which no more change in assignment will occur and the procedure therefore converge.

Question 26. Which one of the following approaches will not be sensitive to the initialization of the parameters of the model?

A. K-means with five clusters.

B. An Artificial Neural Network (ANN) with 3 hidden units.

C. An Artificial Neural Network (ANN) with 10 hidden units.

D. A one cluster Gaussian Mixture Model (GMM).

E. Don't know.

Solution 26. K-means and artificial neural networks are prone to local minima as is Gaussian Mixture Models. However, when there are only one cluster the expectation step will always be giving probability one of all observations belonging to the same cluster and therefore the GMM with one component will converge to the same solution regardless of initialization.

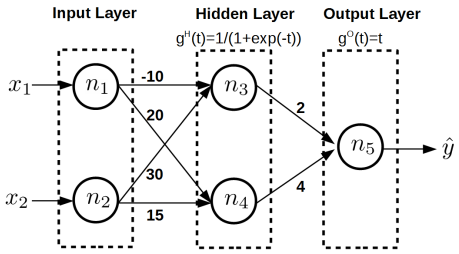

Figure 11: A neural network.

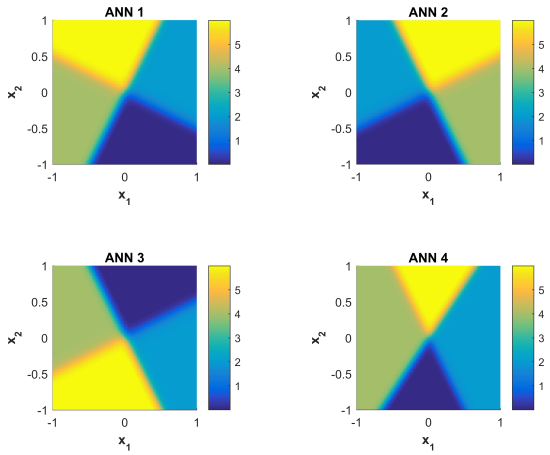

Figure 12: The predictions made by four different neural networks, denoted ANN 1, ANN 2, ANN 3, and ANN 4.

Question 27. Consider the artificial neural network in Figure 11 that has a logistic function as non-linearity in the hidden layer and a linear function in the output layer. I.e., the transfer function for n_3 and n_4 is $g^H(t) = 1/(1 + exp(-t))$ whereas the transfer function for n_5 is $g^O(t) = t$. There are no biases in the network, i.e. the bias for all neurons are 0. Which one of the four artificial networks in Figure 12 corresponds to the ANN in Figure 11?

A. ANN 1

B. ANN 2

C. ANN 3

D. ANN 4

E. Don't know.

Solution 27. Propagating $x_1 = -1$ and $x_2 = -1$ we propagate something that is almost zero through each hidden unit to obtain an output close to zero. This is only the case for ANN 2.

Technical University of Denmark

Written examination: 27th May 2016, 9 AM - 1 PM. Page 1 of 12 pages.

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.

You must *either* use the electronic file or the form on this page to hand in your answers *but not both*. **We strongly encourage that you hand in your answers digitally using the electronic file**. If you hand in using the form on this page, please write your name and student number clearly.

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

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

Answers:

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

Name: _____

Student number: _____

No.	Attribute description	Abbrev.
x_1	Room temperature	Temperature
x_2	Room humidity	Humidity
x_3	Light intensity in room	Light
x_4	CO2 concentration in room	CO2
x_5	Feature-transformed variable	HumidityRatio
y	Is the room occupied?	Occupancy

Table 1: Attributes of the *Occupancy* dataset. The dataset includes 5 attributes $(x_1, \ldots, x_5)$ of 8143 measurements of rooms made over time as well as a binary variable indicating if the room is occupied or not, y. The purpose is to predict if the room is occupied based on the other observations.

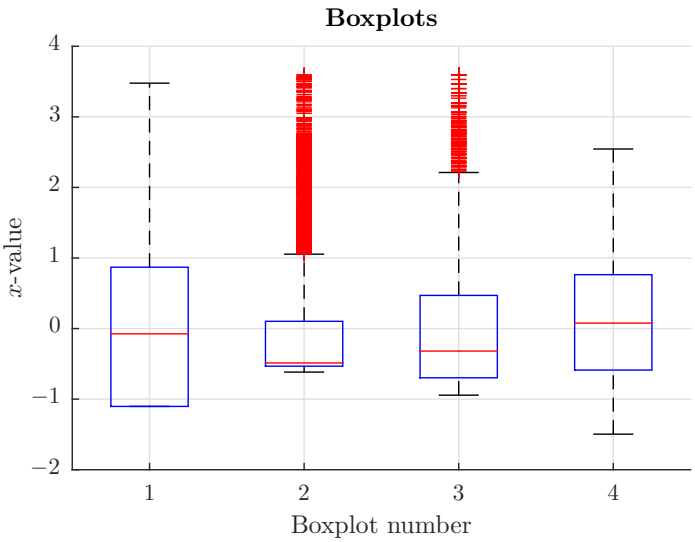

Figure 1: Boxplots corresponding to the variables x_1, x_2, x_3, x_4 of Figure 2 but not necessarily in that order. Notice the features have been standardized.

Question 1. In Figure 2 and Figure 1 are shown histograms and boxplots of the *Occupancy* dataset[1] based on the attributes $x_1, \cdots, x_4$ found in Table 1. The dataset has been standardized for this question.

[1]Dataset obtained from `http://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+`. Notice the dataset has been pre-processed for this exam.

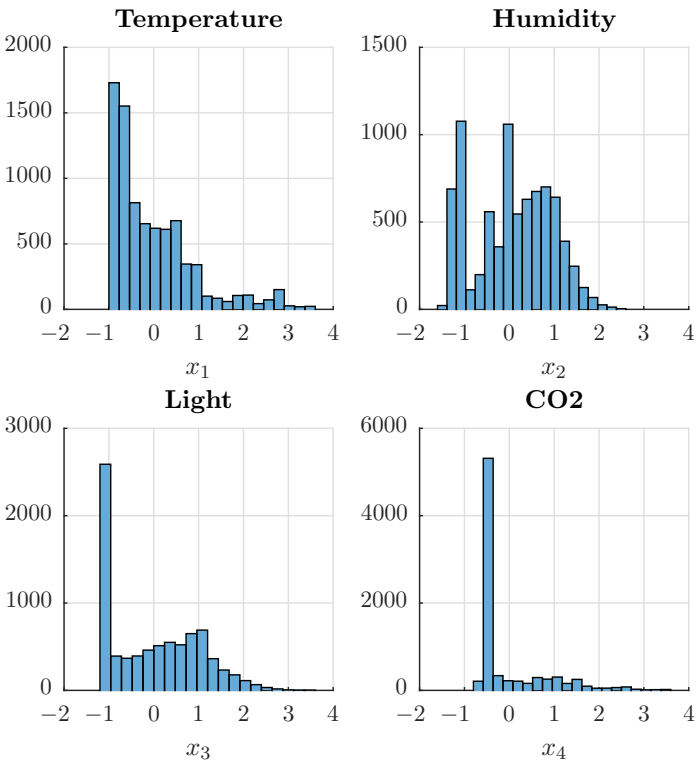

Figure 2: Plot of observations x_1, x_2, x_3, x_4 of the *Occupancy* dataset of Table 1 as histograms. Notice the features have been standardized.

Which histograms x_1, x_2, x_3, x_4 match which boxplots?

A. Boxplot 1 is x_3, Boxplot 2 is x_4, Boxplot 3 is x_1 and Boxplot 4 is x_2

B. Boxplot 1 is x_4, Boxplot 2 is x_1, Boxplot 3 is x_2 and Boxplot 4 is x_3

C. Boxplot 1 is x_3, Boxplot 2 is x_1, Boxplot 3 is x_4 and Boxplot 4 is x_2

D. Boxplot 1 is x_3, Boxplot 2 is x_2, Boxplot 3 is x_1 and Boxplot 4 is x_4

E. Don't know.

Solution 1. The two histograms x_3, x_4 have a minimum cutoff (indicated by the large spikes for low x-values) and will therefore correspond to boxplots 1 and 2 which have no lower whiskers. By considering the median values one can then determine that x_3 corresponds to boxplot 1 and boxplot 3 must correspond to x_1.

Question 2. A principal component analysis is carried out on the *Occupancy* dataset based on the attributes $x_1, \cdots, x_5$ found in Table 1. The data is standardized by (i) substracting the mean and (ii) dividing each column by it's standard deviation to obtain the standardized matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized matrix to obtain the decomposition $\boldsymbol{USV}^\top = \tilde{\boldsymbol{X}}$ where

$$\boldsymbol{S} = \begin{bmatrix} 149 & 0 & 0 & 0 & 0 \\ 0 & 118 & 0 & 0 & 0 \\ 0 & 0 & 53 & 0 & 0 \\ 0 & 0 & 0 & 42 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{bmatrix} \tag{1}$$

$$\boldsymbol{V} = \begin{bmatrix} -0.3 & -0.5 & 0.7 & 0.2 & 0.2 \\ -0.4 & 0.6 & -0.0 & 0.2 & 0.7 \\ -0.4 & -0.4 & -0.7 & 0.4 & -0.0 \\ -0.6 & -0.1 & -0.1 & -0.8 & 0.1 \\ -0.5 & 0.4 & 0.2 & 0.2 & -0.7 \end{bmatrix}. \tag{2}$$

Notice the entries of the matrices have been rounded. Which one of the following statements is true?

A. The three principal components with the least variance account for less than 10% of the variance

B. The first principal component accounts for more than 60% of the variance

C. The two principal components with the least variance account for less than 4% of the variance

D. The first two principal components account for more than 85% of the variance

E. Don't know.

Solution 2. Recall the variance of a given component is

$$\text{var.} = \frac{S_{ii}^2}{\sum_{j=1}^5 S_{jj}^2}$$

Then the variance of the three last components is 0.113, the variance of the first components is: 0.545, the variance of the last two components is 0.044, and the variance of the first two principal components 0.887.

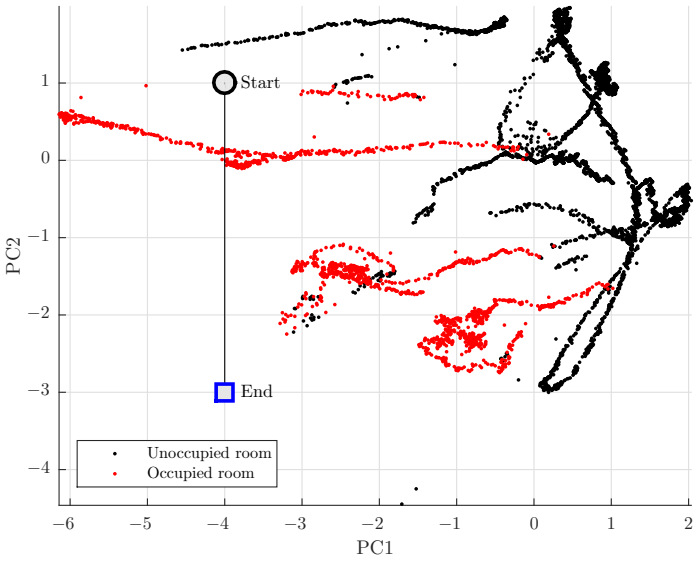

Figure 3: Plot of the 8143 observations from the *Occupancy* dataset of Table 1 projected onto the first two principal directions.

Question 3. Consider again the *Occupancy* dataset of Table 1. A plot of each observation plotted onto the two first principal directions given in Equation (2) is shown in Figure 3. As seen in the plot, the observations are made successively over time and therefore the room measurements form "trajectories". Suppose a rooms measurements starts at the black circle and ends a few hours later at the blue square. Which of the following statements best describes the development of the measurements?

A. The room temperature increases, the room humidity drops and the room becomes lighter

B. The room temperature drops, the humidity increases and the room room becomes darker

C. The room temperature drops, the room humidity drops and the room becomes darker

D. The room temperature increases, the room humidity increases and the room becomes lighter

E. Don't know.

Solution 3. If we compute the difference between the points projected onto the second component we obtain:

$$-3v_2 - 1v_2 = \begin{bmatrix} 2.14 \\ -2.30 \\ 1.78 \\ 0.48 \\ -1.65 \end{bmatrix}$$

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9
o_1	0.00	4.84	0.50	4.11	1.07	4.10	4.71	4.70	4.93
o_2	4.84	0.00	4.40	5.96	4.12	2.01	5.36	3.59	3.02
o_3	0.50	4.40	0.00	4.07	0.72	3.75	4.66	4.48	4.64
o_4	4.11	5.96	4.07	0.00	4.48	4.69	2.44	3.68	4.15
o_5	1.07	4.12	0.72	4.48	0.00	3.54	4.96	4.62	4.71
o_6	4.10	2.01	3.75	4.69	3.54	0.00	3.72	2.23	1.95
o_7	4.71	5.36	4.66	2.44	4.96	3.72	0.00	2.03	2.73
o_8	4.70	3.59	4.48	3.68	4.62	2.23	2.03	0.00	0.73
o_9	4.93	3.02	4.64	4.15	4.71	1.95	2.73	0.73	0.00

Table 2: The pairwise Euclidian distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 9 observations from the *Occupancy* dataset (recall $M = 5$). Each observation o_i corresponds to a row of the occupancy matrix $\boldsymbol{X}$ of Table 1 (the data has been standardized). The colors indicate classes such that the black observations $\{o_1, o_2, o_3, o_4, o_5\}$ belongs to class C_1 (unoccupied) and the red observations $\{o_6, o_7, o_8, o_9\}$ belongs to class C_2 (Occupied).

thus we see the temperature goes up, the humidity drops and the light goes up, therefore option A is correct.

Question 4. Consider the distances in Table 2. The class labels C_1, C_2 (corresponding to $\{o_1, o_2, o_3, o_4, o_5\}$ and $\{o_6, o_7, o_8, o_9\}$) will be predicted using a k-nearest neighbour classifier based on the distances given in Table 2. Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 3-nearest neighbour classifier (i.e. $k = 3$). What is the error rate computed for all $N = 9$ observations?

A. error rate $= \frac{1}{9}$

B. error rate $= \frac{2}{9}$

C. error rate $= \frac{3}{9}$

D. error rate $= \frac{4}{9}$

E. Don't know.

Solution 4. The true error rate is 0.22 or 2/9. This is easy to see by going through Table 2 and notice the "wrongly" classified observations are o_2, o_4 which are closer to two observations in the red class than the observations in the black class.

Question 5. Consider the distances in Table 2 and suppose we wish to apply mixture modelling and we use the normal density as the mixture distributions[2]:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma) = (2\pi\sigma^2)^{-\frac{M}{2}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}}.$$

Suppose we wish to compute the density at o_9 based on a mixture model of $K = 8$ components, the parameters $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_8$ of each component is taken to be the position of the observations $o_1, \ldots, o_8$ and the components are weighted equally. Suppose we set $\sigma = 2$, what is the probability density at the *last* observation o_9?

A. $p(o_9) = \frac{1}{9(\pi 8)^5}\left(e^{\frac{-4.93}{8}} + e^{\frac{-3.02}{8}} + e^{\frac{-4.64}{8}} + e^{\frac{-4.15}{8}} + e^{\frac{-4.71}{8}} + e^{\frac{-1.95}{8}} + e^{\frac{-2.73}{8}} + e^{\frac{-0.73}{8}}\right)$

B. $p(o_9) = \frac{1}{9(\pi 8)^{\frac{5}{2}}}\left(e^{\frac{-4.93^2}{8}} + e^{\frac{-3.02^2}{8}} + e^{\frac{-4.64^2}{8}} + e^{\frac{-4.15^2}{8}} + e^{\frac{-4.71^2}{8}} + e^{\frac{-1.95^2}{8}} + e^{\frac{-2.73^2}{8}} + e^{\frac{-0.73^2}{8}}\right)$

C. $p(o_9) = \frac{1}{8(\pi 8)^{\frac{5}{2}}} \exp\left(\frac{-4.93}{8} + \frac{-3.02}{8} + \frac{-4.64}{8} + \frac{-4.15}{8} + \frac{-4.71}{8} + \frac{-1.95}{8} + \frac{-2.73}{8} + \frac{-0.73}{8}\right)$

D. $p(o_9) = \frac{1}{8(\pi 8)^{\frac{5}{2}}}\left(e^{\frac{-4.93^2}{8}} + e^{\frac{-3.02^2}{8}} + e^{\frac{-4.64^2}{8}} + e^{\frac{-4.15^2}{8}} + e^{\frac{-4.71^2}{8}} + e^{\frac{-1.95^2}{8}} + e^{\frac{-2.73^2}{8}} + e^{\frac{-0.73^2}{8}}\right)$

E. Don't know.

[2] Remember that $\exp(x) = e^x$.

Solution 5. Options A and B are not properly normalized by the number of mixture components. Option C does not use the squared distances. Accordingly option D is the correct answer.

Question 6. We wish to compute the *average relative KNN density* (a.r.d) of observation o_9 from the *Occupancy* dataset described in Table 1 using the distances given in Table 2. Letting $d(\boldsymbol{x}, \boldsymbol{y})$ denote the Euclidian distance metric the a.r.d. is defined as

$$\text{density}(\boldsymbol{x}, K) = \frac{1}{\frac{1}{K}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},K)} d(\boldsymbol{x}, \boldsymbol{y})}$$

$$\text{a.r.d}(\boldsymbol{x}, K) = \frac{\text{density}(\boldsymbol{x}, K)}{\frac{1}{K}\sum_{\boldsymbol{z}\in N(\boldsymbol{x},K)} \text{density}(\boldsymbol{z}, K)},$$

$$N(\boldsymbol{x}, K) : \text{set of } K\text{-nearest neighbours of } \boldsymbol{x}.$$

What is the a.r.d. of observation o_9 using $K = 2$ nearest neighbours?

A. $\text{a.r.d}(\boldsymbol{x} = o_9, K = 2) \approx 2.428$

B. $\text{a.r.d}(\boldsymbol{x} = o_9, K = 2) \approx 0.399$

C. $\text{a.r.d}(\boldsymbol{x} = o_9, K = 2) \approx 1.214$

D. $\text{a.r.d}(\boldsymbol{x} = o_9, K = 2) \approx 1.618$

E. Don't know.

Solution 6. The nearest neighbour of o_9 is o_6, o_8 and the nearest neighbours of o_6 is o_2, o_9 and for o_8 it is o_7, o_9. The densities are

$$\text{density}(o_9, K = 2) = 0.746268656716$$
$$\text{density}(o_6, K = 2) = 0.505050505051$$
$$\text{density}(o_8, K = 2) = 0.724637681159$$

from which it follows

$$\text{a.r.d.}(o_9, K = 2) = \frac{0.7463}{\frac{1}{2}(0.5051 + 0.7246)} = 1.2138$$

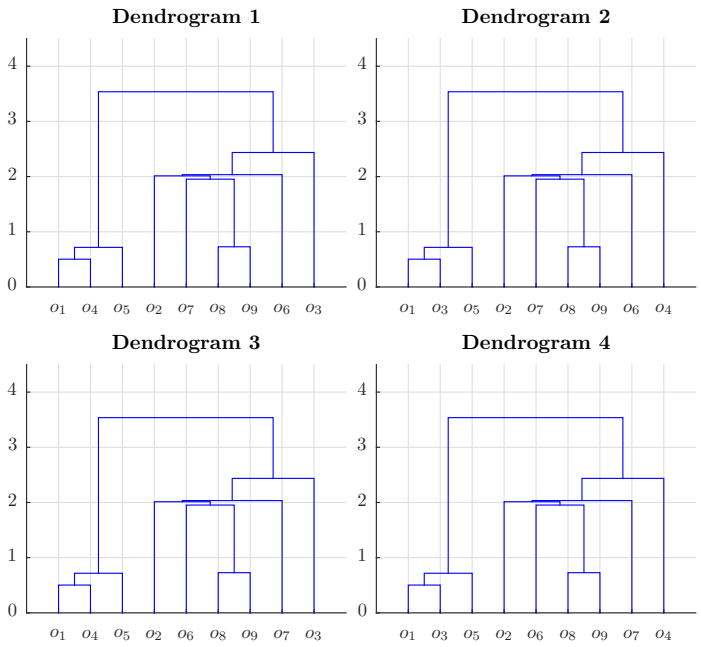

Figure 4: Proposed hierarchical clustering of the 9 observations considered in Table 2

Question 7. A hierarchical clustering is applied to the 9 observations in Table 2 using *minimum* linkage. Which of the dendrograms shown in Figure 4 corresponds to the clustering?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

Solution 7. The correct answer is D, dendrogram 4. o_8 and o_9 are grouped together in all diagrams. Since the distance from o_6 to o_8 is lower than the distance from o_7 to o_6 then o_6 should link to o_8, o_9 before o_7. This allows us to rule out dendrogram 1 and dendrogram 2.

Finally, o_3 and o_1 should clearly link together allowing us to rule out dendrogram 3. This leaves only option D.

Question 8. In Table 2 is given the pairwise euclidian distances between 9 observations from the *Occupancy* dataset of Table 1. Suppose the Euclidian norm of observations o_2 and o_3 is:

$$\|o_2\| = \sqrt{\sum_{k=1}^{M} x_{1k}^2} = 3.04, \quad \|o_3\| = \sqrt{\sum_{k=1}^{M} x_{2k}^2} = 1.5$$

Split nr.	Splitting rule	$y=0$	$y=1$
Split 1	Temperature <20	45	1
	$20 \leq$ Temperature ≤ 22	47	66
	$22<$ Temperature	8	33
Split 2	Temperature <21	76	20
	$21 \leq$ Temperature ≤ 22	16	47
	$22<$ Temperature	8	33
Split 3	Temperature <19.5	25	0
	$19.5 \leq$ Temperature ≤ 21	55	23
	$21<$ Temperature	20	77

Table 3: Three potential splits of a subset of the *Occupancy* dataset based on the variable Temperature. Each split is a three-way split where the dataset is divided into three sets. For instance, in the second set of split 1 (corresponding to $20 \leq$ Temperature ≤ 22), there are 47 observations of an unoccupied room ($y=0$) and 66 observations of an occupied room ($y=1$).

What can be concluded about the Cosine similarity of these two observations? (Hint: recall for vectors $\boldsymbol{x}$, $\boldsymbol{y}$ that $\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{x}^\top \boldsymbol{y}$)

A. $\cos(o_2, o_3) \approx 0.7127$

B. $\cos(o_2, o_3) \approx 0.4314$

C. $\cos(o_2, o_3) \approx -0.8712$

D. $\cos(o_2, o_3) \approx -0.8628$

E. Don't know.

Solution 8. Notice the inner product can be recovered as

$$o_2^\top o_3 = \frac{\|o_2\|_2^2 + \|o_3\|_2^2 - d(o_2, o_3)^2}{2} = -3.9342$$

and the definition of Cosine similarity is

$$\cos(o_2, o_3) = \frac{o_2^\top o_3}{\|o_2\|_2 \|o_3\|_2}$$

Thus the true answer is -0.862763157895

Question 9.

Consider a subset of the *Occupancy* dataset of Table 1 and suppose we wish to predict the occupied status y using a decision tree build using Hunts algorithm. Hunt's algorithm consider potential splits and select the one with the greatest purity gain Δ. In Table 3 is indicated three potential splits using the Temperature variable (Split 1 to 3) where in each case we consider a three-way split. Suppose the number of observations in the unoccupied $y=0$ and occupied $y=1$ class is as given in Table 3, what will Hunt's algorithm do if *classification error* is used as impurity measure?

A. Hunt's algorithm will select split 1 over split 2

B. Hunt's algorithm will select split 2 over split 3

C. Hunt's algorithm will select split 1 over split 3

D. Hunt's algorithm will select split 3 over split 2

E. Don't know.

Solution 9. The relevant definitions can be found in section 4.3 of Tan et.al. We need to compute the purity gain for each of the three splits. There are $n=200$ observations. Then

$$I_0 = 1 - \frac{1}{2} = \frac{1}{2}$$

And we compute:

$$\Delta_1 = I_0 - \frac{46}{n}(1 - \frac{45}{46}) - \frac{113}{n}(1 - \frac{66}{113}) - \frac{41}{n}(1 - \frac{33}{41}) = 0.22$$
$$\Delta_2 = I_0 - \frac{96}{n}(1 - \frac{76}{96}) - \frac{63}{n}(1 - \frac{47}{63}) - \frac{41}{n}(1 - \frac{33}{41}) = 0.28$$
$$\Delta_3 = I_0 - \frac{25}{n}(1 - \frac{25}{25}) - \frac{78}{n}(1 - \frac{55}{78}) - \frac{97}{n}(1 - \frac{77}{97}) = 0.285$$

	f_1	f_2	f_3	f_4	f_5
o_1	0	1	1	0	1
o_2	0	0	1	0	0
o_3	1	0	0	0	1
o_4	1	0	0	1	1
o_5	1	0	0	1	0
o_6	1	1	0	1	1
o_7	1	0	1	0	0
o_8	1	0	1	1	1
o_9	0	1	1	1	1
o_{10}	1	0	1	1	0
o_{11}	0	1	1	0	0

Table 4: Processed version of the $N = 11$ observations of Table 2. For each observation we binarize the features by thresholding at the median to produce the binary features $f_1, \ldots, f_5$. The categories indicated by the color still indicate occupancy status y, i.e. the black category $(o_1, o_2, o_3, o_4, o_5, o_6)$ corresponds to $y = 0$ and the red category $(o_7, o_8, o_9, o_{10}, o_{11})$ to $y = 1$.

Question 10.

Consider the $N = 11$ observations from Table 2 and assume the data has been processed to the 11×5 binary matrix shown in Table 4. Suppose we consider the first three features f_1, f_2, f_3 and train a Naïve-Bayes classifier to distinguish between unoccupied and occupied rooms $y = 0$ and $y = 1$ based on only these three features. If an observation has $f_1 = 0, f_2 = 1, f_3 = 1$, what is the probability that the room is occupied, $y = 1$, according to the Naive-Bayes classifier?

A. $p_{NB}(y = 1|f_1 = 0, f_2 = 1, f_3 = 1) = 0.730$

B. $p_{NB}(y = 1|f_1 = 0, f_2 = 1, f_3 = 1) = 0.783$

C. $p_{NB}(y = 1|f_1 = 0, f_2 = 1, f_3 = 1) = 0.812$

D. $p_{NB}(y = 1|f_1 = 0, f_2 = 1, f_3 = 1) = 0.764$

E. Don't know.

Solution 10. True answer is: 0.783. This can be found by computing the per-class probabilities

$$p(f_1 = 0|y = 0) = \frac{1}{3}, \; p(f_1 = 0|y = 1) = \frac{2}{5}$$
$$p(f_2 = 1|y = 0) = \frac{1}{3}, \; p(f_2 = 1|y = 1) = \frac{2}{5}$$
$$p(f_3 = 1|y = 0) = \frac{1}{3}, \; p(f_3 = 1|y = 1) = 1$$

The class prior is $p(y = 0) = \frac{6}{11}$ and so the Naive-Bayes estimate is

$$p_{NB}(y = 1|f_1 = 0, f_2 = 1, f_3 = 1) = \frac{x_1}{x_0 + x_1}$$
$$\approx 0.7826$$

where

$$x_0 = p(f_1 = 0|y = 0)p(f_2 = 1|y = 0)p(f_3 = 1|y = 0)p(y = 0)$$
$$= 0.0202$$
$$x_1 = p(f_1 = 0|y = 1)p(f_2 = 1|y = 1)p(f_3 = 1|y = 1)p(y = 1)$$
$$= 0.0727$$

Question 11. Suppose we consider the binary matrix in Table 4 as a market-basket problem consisting of $N = 11$ "transactions" $o_1, \ldots, o_{11}$ and $M = 5$ "items" $f_1, \ldots, f_5$. Which of the following options represents all itemsets with support greater than 0.32?

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

B. $\{f_1\}$, $\{f_3\}$, $\{f_4\}$, $\{f_5\}$, $\{f_1, f_4\}$

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

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

E. Don't know.

Solution 11. Recall by chapter 6.1 of Tan et al. the support count is the number of "transactions" containing a given set of items. The problem is then to find all subsets of items that occur in at least 4 of the 11 transactions. These are easily seen to be those in option C and no other.

Question 12. We again consider the binary matrix of Table 4 as a market-basket problem consisting of $N = 11$ "transactions" $o_1, \ldots, o_{11}$ and $M = 5$ "items" $f_1, \ldots, f_5$. Which of the following rules has the highest *confidence*?

A. $\{f_3, f_4\} \to \{f_5\}$

B. $\{f_1, f_5\} \to \{f_4\}$

C. $\{f_1, f_4\} \to \{f_5\}$

D. $\{f_2, f_4\} \to \{f_1\}$

E. Don't know.

Solution 12. Recall the confidence is defined as (see chapter 6.1 of Tan et al)

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

We can then compute the confidence of the three rules as, respectively,

$$c(\{f_3, f_4\} \to \{f_5\}) = \frac{2}{3} = 0.666666666667$$
$$c(\{f_1, f_5\} \to \{f_4\}) = \frac{3}{4} = 0.75$$
$$c(\{f_1, f_4\} \to \{f_5\}) = \frac{3}{5} = 0.6$$
$$c(\{f_2, f_4\} \to \{f_1\}) = \frac{1}{2} = 0.5$$

Question 13.

We consider the $N = 11$ observations from Table 4 as 5-dimensional binary vectors. Which one of the following statements is true regarding the Jaccard/cosine similarity and the simple matching coefficient?

A. $\mathrm{COS}(o_1, o_2) > \mathrm{SMC}(o_1, o_2)$

B. $\mathbf{COS}(o_1, o_2) > \mathbf{COS}(o_1, o_3)$

C. $\mathrm{J}(o_1, o_3) > \mathrm{SMC}(o_1, o_2)$

D. $\mathrm{J}(o_1, o_3) > \mathrm{COS}(o_1, o_3)$

E. Don't know.

Solution 13. It is easily verified only option B is correct by plugging in the following values:

$$\mathrm{SMC}(o_1, o_2) = 0.6$$
$$\mathrm{J}(o_1, o_3) = 0.25$$
$$\mathrm{COS}(o_1, o_2) = 0.57735026919$$
$$\mathrm{COS}(o_1, o_3) = 0.408248290464$$

Feature(s)	E_{train}	E_{test}
None	0.711	0.9
x_1	0.657	0.622
x_2	0.648	0.721
x_3	0.584	0.446
x_4	0.604	0.645
x_1, x_2	0.568	0.574
x_1, x_3	0.465	0.311
x_1, x_4	0.42	0.503
x_2, x_3	0.448	0.428
x_2, x_4	0.421	0.515
x_3, x_4	0.338	0.458
x_1, x_2, x_3	0.275	0.324
x_1, x_2, x_4	0.273	0.534
x_1, x_3, x_4	0.221	0.314
x_2, x_3, x_4	0.182	0.391
x_1, x_2, x_3, x_4	0.139	0.641

Table 5: The *error rate* on a training set E_{train} and test set E_{test} for a classification model trained on different subsets of features of the *Occupancy* dataset of Table 1

Question 14. Consider the *Occupancy* dataset of Table 1 and suppose we only consider the first four features x_1, x_2, x_3, x_4. Suppose we wish to examine which subset of these features can be expected to give the optimal generalization error. In Table 5 is shown how different combinations of features give rise to different error rates on a training and a test set for a classifier. Which one of the following statements is true?

A. Forward and backward selection will select the same number of features

B. Forward selection will select a better model (measured by the generalization error) than backward selection

C. Backward selection will select *more* features than forward selection

D. Backward selection will select *less* features than forward selection

E. Don't know.

Solution 14. Firstly, notice the column with the training set error rates can be disregarded. Forward selection then first selects x_3, then x_1, x_3 and then terminates. Backward selection will first select x_1, x_3, x_4, then x_1, x_3 and then terminates. Accordingly, option A is correct.

No.	Attribute description
x_1	Species (Oak, pine, ...)
x_2	Year planted (e.g. 1946)
x_3	Tree height (in feet)
x_4	Tree quality score $(1, 2, ..., 5)$
y	Expected selling price

Table 6: Attributes of the *Trees* dataset. The dataset includes 4 attributes $(x_1, \ldots, x_4)$ of 1306 trees in a forrest.

Question 15. Consider the first two attributes of Table 1 and suppose they have been binarized by thresholding at the mean value to produce the binary attributes g_1, g_2. Suppose we are told that $p(y = 1) = 0.5$ and that

$$
\begin{aligned}
P(g_1 = 0, g_2 = 0 | y = 0) &= 0.23 \\
P(g_1 = 0, g_2 = 1 | y = 0) &= 0.40 \\
P(g_1 = 1, g_2 = 0 | y = 0) &= 0.28 \\
P(g_1 = 1, g_2 = 1 | y = 0) &= 0.09 \\
P(g_1 = 0, g_2 = 0 | y = 1) &= 0.01 \\
P(g_1 = 0, g_2 = 1 | y = 1) &= 0.03 \\
P(g_1 = 1, g_2 = 0 | y = 1) &= 0.46 \\
P(g_1 = 1, g_2 = 1 | y = 1) &= 0.50
\end{aligned}
$$

What is then the probability that a room is humid given that it is occupied?

A. $p(g_2 = 1 | y = 1) \approx 0.53$

B. $p(g_2 = 1 | y = 1) \approx 0.51$

C. $p(g_2 = 1 | y = 1) \approx 0.265$

D. $p(g_2 = 1 | y = 1) \approx 0.245$

E. Don't know.

Solution 15. This question can be solved by only using the sum rule. Since

$$
\begin{aligned}
p(g_2 = 1 | y = 1) = p(g_1 = 0, g_2 = 1 | y = 1) \\
+ p(g_1 = 1, g_2 = 1 | y = 1)
\end{aligned}
$$

option A is correct.

Question 16. In Table 6 is shown a dataset where each observation corresponds to a tree. Which of the

following statements is true?

A. x_2 **is interval and** x_3 **is ratio**

B. x_2 is ratio and x_1 is nominal

C. x_1, x_4 are ordinal

D. Considered pairwise, all variables in the dataset can be expected to have little correlation

E. Don't know.

Solution 16. Height, year planted and selling price can all be expected to be correlated ruling out option D. Species is nominal, tree quality score is nomimal, year is interval and tree height and selling price are ratio. This rules out all options except option 1.

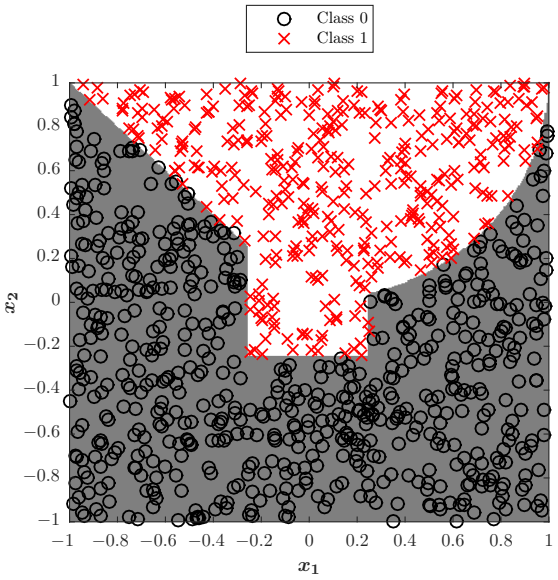

Figure 5: Two-class classification problem

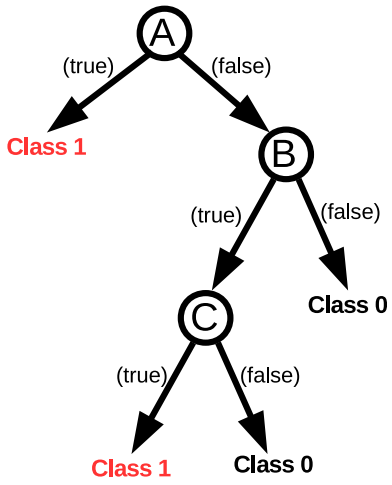

Figure 6: Decision tree with 3 nodes A, B and C

Question 17. Suppose we wish to solve the two-class classification problem in Figure 5 using a classification tree of the form given in Figure 6. What rules, acting on the coordinates $\boldsymbol{x} = (x_1, x_2)$, should be assigned to the three internal nodes A, B and C of the tree to give rise to the indicated decision boundary?

A. $A : \|\boldsymbol{x}\|_\infty < \frac{1}{4}$, $B : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2$
$C : \left\|\boldsymbol{x} - \begin{bmatrix}0\\1\end{bmatrix}\right\|_2 < 1$

B. $A : \left\|\boldsymbol{x} - \begin{bmatrix}0\\1\end{bmatrix}\right\|_2 < 1$, $B : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2$
$C : \|\boldsymbol{x}\|_\infty < \frac{1}{4}$

C. $A : \left\|\boldsymbol{x} - \begin{bmatrix}0\\1\end{bmatrix}\right\|_2 < 1$, $B : \|\boldsymbol{x}\|_\infty < \frac{1}{4}$
$C : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2$

D. $A : \|\boldsymbol{x}\|_\infty < \frac{1}{4}$, $B : \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 < 2$
$C : \left\|\boldsymbol{x} - \begin{bmatrix}0\\1\end{bmatrix}\right\|_2 < 1$

E. Don't know.

Solution 17. First consider the point $(0, -0.2)$ which should belong to the red class 1. This point will be classified incorrectly according to option B and C. For option D, consider the point $(0, 1)$ which will also be classified incorrectly. This leaves option *A* which is correct.

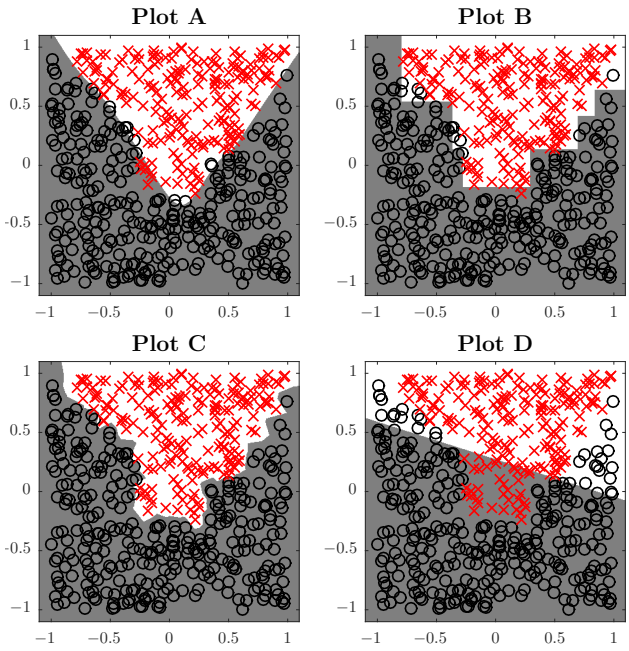

Figure 7: Four classifiers applied to a two-class classification problem

Question 18. Consider the classification problem given in Figure 5. Suppose the problem is solved using the following four classifiers

(1NN) A 1-nearest neighbour classifier

(TREE) A decision tree

(LREG) Logistic regression

(NNET) An artificial neural network with four hidden units

All classifiers are using only the two attributes x_1, x_2, corresponding to the position of each observation, as well as the class label. Which of the descriptions (1NN),(TREE),(LREG),(NNET) matches the boundariers of the four plots (Plot A, B, C and D) indicated in Figure 7?

A. Plot A is 1NN, Plot B is TREE, Plot C is NNET, Plot D is LREG

B. Plot A is LREG, Plot B is TREE, Plot C is 1NN, Plot D is NNET

C. Plot A is NNET, Plot B is TREE, Plot C is LREG, Plot D is 1NN

D. Plot A is NNET, Plot B is TREE, Plot C is 1NN, Plot D is LREG

E. Don't know.

Solution 18. Plot C is a 1NN classifier (notice all points are correctly classified), D is the only classifier with a linear boundary and must be logistic regression and B has the "boxes" characteristic for a decision tree.

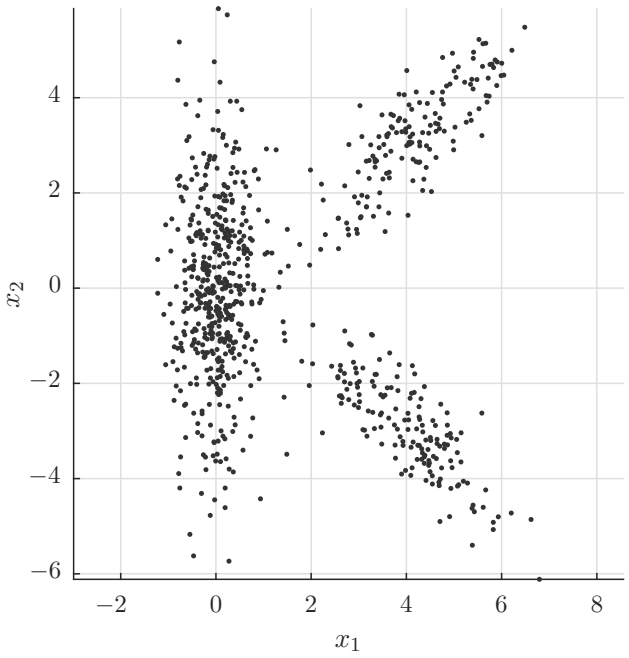

Figure 8: Scatter plot of observations generated from a Gaussian mixture-model

Question 19. Suppose the 2D dataset shown in Figure 8 was generated from a Gaussian mixture-model (GMM) with three components. Which of the following is the most likely equation of the density of the mixture model?

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.0 & -0.8 \\ -0.8 & 1.0 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.2 & 0.0 \\ 0.0 & 3.5 \end{bmatrix},$$
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix} 4 \\ -3 \end{bmatrix},$$

A. The density is:

$$p(\boldsymbol{x}) = 0.6\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.2\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.2\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_3)$$

B. The density is:

$$p(\boldsymbol{x}) = 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.25\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_1) + 0.25\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_3)$$

C. The density is:

$$p(\boldsymbol{x}) = 0.5\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_3) + 0.25\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.25\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$$

D. The density is:

$$p(\boldsymbol{x}) = 0.6\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_3) + 0.2\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_1) + 0.2\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_2)$$

E. Don't know.

Solution 19. Focusing on the axis aligned "cigar" to the left we have that $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_3$ must go together leaving only option A and D. The upper-right cigar goes from south-west to north-east indicating $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_2$ must go together. By process of elimination, this leaves option D.

Question 20. Consider a 1D GMM mixture model

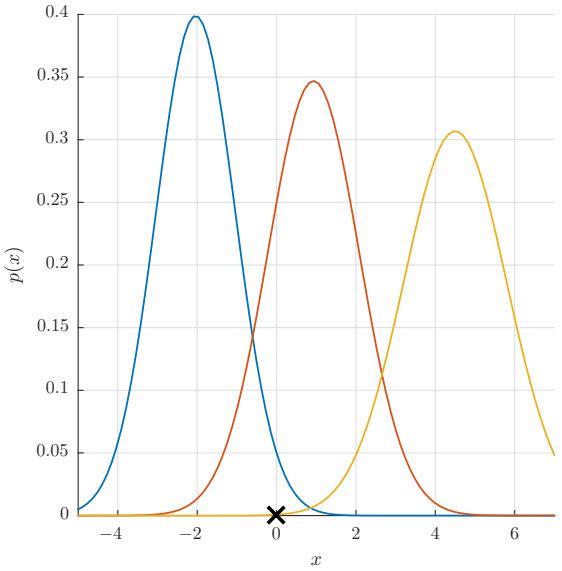

Figure 9: Mixture components in a GMM mixture model with $K = 3$

where each of the $K = 3$ (Gaussian) mixture components are illustrated in Figure 9 as the colored curves and the figure also shows a new observation indicated by the cross. Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step (i.e. assuming the mixture components has the means and variances indicated by Figure 9 and equal weights). According to the EM algorithm, what is the (approximate) probability the black cross is assigned to the blue (left-most) mixture component?

A. 0.05

B. 0.17

C. 0.25

D. 0.02

E. Don't know.

Solution 20. The probability of the black cross under each of the three mixture components can be read of as approximately $p(x_0|\mu_1, \sigma_1) \approx 0.05$, $p(x_0|\mu_2, \sigma_2) \approx 0.25$, $p(x_0|\mu_3, \sigma_3) \approx 0$. Since they are weighted equally the assignment to the left-most component is

$$p(z = 1|x_0) = \frac{\frac{1}{3}p(x_0|\mu_1, \sigma_1)}{\sum_{i=1}^{3} \frac{1}{3}p(x_0|\mu_i, \sigma_i)} \approx \frac{0.05}{0.05 + 0.25} = 0.17$$

Question 21.

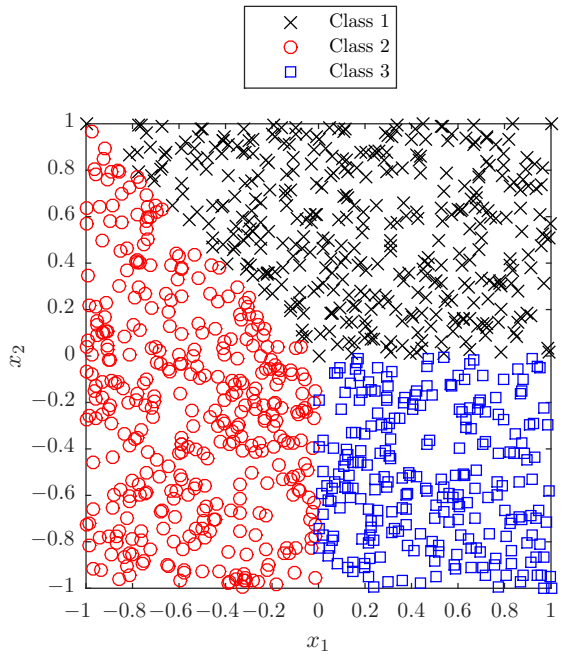

Figure 10: Observations labelled with the most probable class

Consider a multinomial regression classifier for a three-class problem where for each point $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$ we compute the class-probability using the softmax function

$$P(\hat{y} = k) = \frac{e^{\boldsymbol{w}_k^\top \boldsymbol{x}}}{e^{\boldsymbol{w}_1^\top \boldsymbol{x}} + e^{\boldsymbol{w}_2^\top \boldsymbol{x}} + e^{\boldsymbol{w}_3^\top \boldsymbol{x}}}.$$

A dataset of $N = 1000$ points where each point is labelled according to the maximum class-probability is shown in Figure 10. Which setting of the weights was used?

A. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

B. $w_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

C. $w_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

D. $w_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $w_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

E. Don't know.

Solution 21. Consider for instance the point $\boldsymbol{x}$ where $x_1 = 0$ and $x_2 = 1$. Then, letting $y_k = \boldsymbol{w}_k^T \boldsymbol{x}$, we obtain:

$$A: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & -1 \end{bmatrix}$$
$$B: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & -1 & 1 \end{bmatrix}$$
$$C: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \end{bmatrix}$$
$$D: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 \end{bmatrix}$$

Next, since the multinomial regression function preserves order we need only consider the maximal value. Accordingly the point $\boldsymbol{x}$ is only classified to the correct class 1 for option C.

Question 22. Consider a two-dimensional data set consisting of $N = 9$ observations shown in Figure 11. The dataset contains three classes indicated by the black crosses (class 1), red circles (class 2) and blue squares (class 3). In the figure, the decision boundaries for four K-nearest neighbor classifiers (KNN) are indicated by shades of gray. Which of the plots correspond to the $K = 5$ nearest-neighbour classifier assuming ties are broken assigning the observation to the *nearest* of the classes which are *tied*? (That is, if for a given observation $\boldsymbol{x}$, the 5 nearest-neighbours contains two observations from class A and two observations from class B, then compute the distance from $\boldsymbol{x}$ to all four observations and select the class where the distance is the smallest).

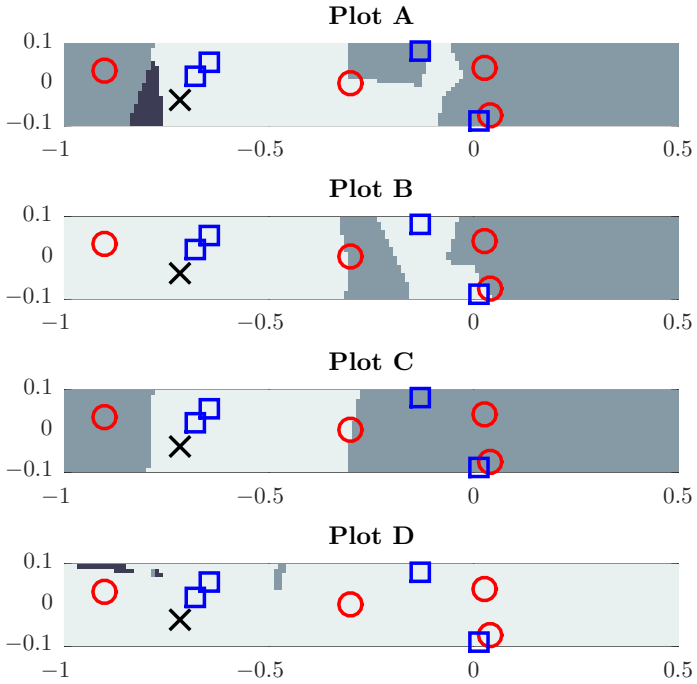

Figure 11: Decision boundaries for four KNN classifiers. The dataset contains three classes indicated by the black crosses (class 1), red circles (class 2) and blue squares (class 3).

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 22. The far left and right parts of the plot must all be assigned to the same class (corresponding to the red circle) because of the tie-breaking rule. In addition, no class can correspond to the black cross because there is only one black cross and 3 classes.

This leaves option C and D. However the two blue squares must also be assigned to their own class due to tie-breaking and so only option C is plausible.

X	1	3	4	6	7	8	13	15	16	17

Table 7: A 1-dimensional dataset of $N = 10$ observations.

Question 23. Consider the 1-dimensional data set comprised of $N = 10$ observations shown in Table 7. Which one of the following clusterings corresponds to a converged state of a K-means algorithm using standard Euclidian distances?

A. {1, 3}, {4, 6, 7}, {8, 13, 15, 16}, {17}

B. {1}, {3, 4, 6}, {7, 8}, {13, 15, 16, 17}

C. {1, 3, 4}, {6, 7, 8}, {13, 15, 16, 17}

D. {1, 3, 4}, {6, 7}, {8, 13}, {15}, {16, 17}

E. Don't know.

Solution 23. The problem can be solved by explicit calculation, however it is easier solved by drawing the points on a paper and ruling out the clusterings that look the most "odd". For instance:

- For option A cluster 2 has mean 5.66 and cluster 3 has mean 13 thus $x = 8$ is in the wrong cluster
- For option B cluster 2 has mean 4.33 and cluster 3 has mean 7.5 thus $x = 6$ is in the wrong cluster
- For option D cluster 2 has mean 6.5 and cluster 3 mean 10.5 so $x = 8$ is in the wrong cluster

It is easy to check the third option has converged.

Question 24. Consider a similarity measure $s(\boldsymbol{x}, \boldsymbol{y})$ defined for two vectors $\boldsymbol{x}, \boldsymbol{y}$:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2 - (\boldsymbol{x}^T\boldsymbol{y})^2}{\|\boldsymbol{x}\|^2\|\boldsymbol{y}\|^2}} \tag{3}$$

where $\|\cdot\|$ is the Euclidian norm. For this problem, we will say $s(\boldsymbol{x}, \boldsymbol{y})$ is translation invariant if for all numbers β: $s(\boldsymbol{x} + \beta, \boldsymbol{y}) = s(\boldsymbol{x}, \boldsymbol{y})$ and scale invariant if for all numbers $\alpha > 0$: $s(\alpha\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{x}, \boldsymbol{y})$. Suppose we apply the similarity measure in Equation (3) to the *Occupancy* dataset in Table 1, which of the following statements are true?

A. s is scale invariant

B. s is translation invariant

C. s is both translation and scale invariant

D. s is neither translation or scale invariant

E. Don't know.

Solution 24. The first option is correct since the measure is obviously scale invariant. The measure is easily seen not to be translation invariant.

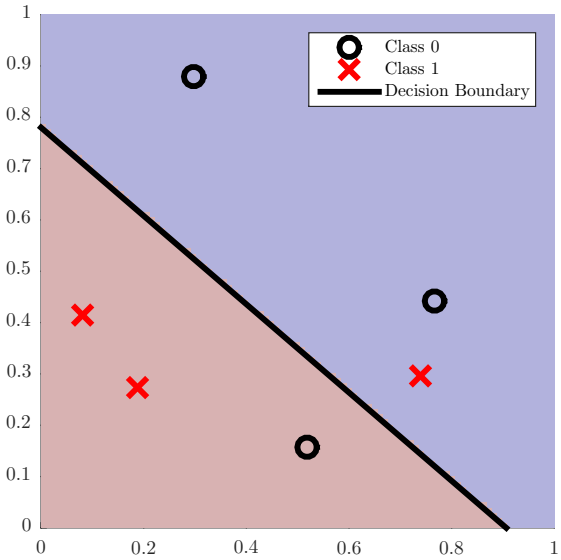

Figure 12: A binary classification problem and the decision boundary obtained by logistic regression. Observations left of the boundary are classified as belonging to the positive class 1 (red crosses) and observations right of the boundary to the negative class 0 (black circles)

Question 25. We wish to apply a logistic regression model to the binary classification problem shown in Figure 12. We attempt to improve the performance by applying AdaBoost (the version in *the lecture notes*, chapter 15. AdaBoost works by first sampling a new dataset with replacement, then training a classifier on the dataset and then proceeding with the subsequent steps of the AdaBoost algorithm.

Suppose in the first iteration of the AdaBoost algorithm the classification boundary of the trained classifier is as indicated by the black line (i.e. observations left of the black line are classified as in the positive class). What is the resulting (rounded) value for the updated weights $\boldsymbol{w}$?

A. $\boldsymbol{w} = [0.026 \quad 0.447 \quad 0.026 \quad 0.026 \quad 0.026 \quad 0.447]$

B. $\boldsymbol{w} = [0.125 \quad 0.250 \quad 0.125 \quad 0.125 \quad 0.125 \quad 0.250]$

C. $\boldsymbol{w} = [0.235 \quad 0.029 \quad 0.235 \quad 0.235 \quad 0.235 \quad 0.029]$

D. $\boldsymbol{w} = [0.120 \quad 0.260 \quad 0.120 \quad 0.120 \quad 0.120 \quad 0.260]$

E. Don't know.

Solution 25. The classifier classifies 2 out of $N = 6$ observations incorrectly. We have:

$$\varepsilon_i = \left[\sum_{j=1}^{N} w_j I\left(\hat{y}_j \neq y_j\right)\right]$$
$$\alpha_i = \frac{1}{2}\log\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

and accordingly $\varepsilon_1 = \frac{1}{N} \times 2 = \frac{1}{3}$. This gives

$$\alpha_1 = \frac{1}{2}\log\frac{1-\frac{1}{3}}{\frac{1}{3}} = \frac{1}{2}\log 2$$

and so for $\boldsymbol{w}$ we get

$$\boldsymbol{w} \propto \begin{bmatrix} e^{-\alpha_1} & e^{\alpha_1} & e^{-\alpha_1} & e^{-\alpha_1} & e^{-\alpha_1} & e^{\alpha_1} \end{bmatrix}$$

Simplifying by moving $\frac{1}{\sqrt{2}}$ outside the vector:

$$\boldsymbol{w} \propto \begin{bmatrix} 1 & 2 & 1 & 1 & 1 & 2 \end{bmatrix}$$

and normalizing:

$$\boldsymbol{w} = \frac{1}{8}\begin{bmatrix} 1 & 2 & 1 & 1 & 1 & 2 \end{bmatrix}$$

accordingly option B is correct.

Question 26. We again consider the logistic regression classifier in Figure 12. Recal the black line indicates the decision boundary obtained by thresholding at 0.5 when trained on a small 2-class dataset composed of a negative class (black circles) and a positive class (red crosses) and that the observations to the left of the boundary are classified as in the positive class and to the right of the boundary in the negative class. What is the AUC score of the logistic regression model?

A. $\frac{2}{3}$

B. $\frac{8}{9}$

C. $\frac{5}{9}$

D. $\frac{3}{4}$

E. Don't know.

Solution 26. The AUC is the area under the curve obtained by plotting the true positive rate (TPR) against the false positive rate (FPR) obtained by thresholding the logistic regression model at different levels (i.e. translating the decision boundary from the far right (everything in the positive class) to the far left (everything in the negative class). The curve can be seen in Figure 13. It begins in (1,1), then the false positive rate drops to $\frac{1}{3}$, then the true positive rate drops to $\frac{2}{3}$, then the false positive rate drops to 0 and then the true positive rate drops to 0. Computing the area under the curve gives $\frac{2}{3} \times \frac{1}{3} + \frac{2}{3} = \frac{8}{9}$.

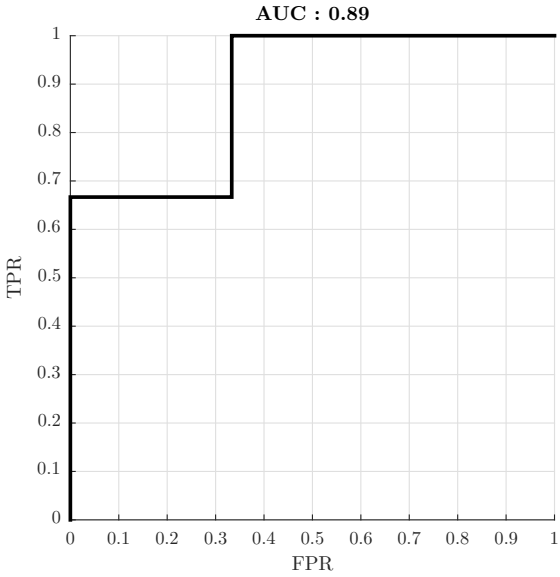

Figure 13: AUC scores computed by plotting the TPR and FPR of Figure 12 against each other. The rates are obtained by translating the decicion boundary horizontally starting from the far right (everything is in the positive class).

Question 27. Consider a classification tree model applied to a dataset of $N = 1000$ observations. Suppose we wish to both select the optimal pruning level and estimate the generalization error of the classification tree model by cross-validation. To simplify the problem, we only consider 3 possible pruning levels:

$$3, 4, 5.$$

We opt for a two-level cross-validation strategy in which we use an inner loop of K_2-fold cross-validation to estimate the optimal pruning level and an outer loop of K_1 fold cross-validation to estimate the generalization error. That is, for each of the K_1 outer folds, the dataset is divided into a validation set and a parameter estimation set on which K_2-fold cross-validation is used to select the optimal pruning level for this outer fold.

Suppose we have a computational budget such that we can only *train* a maximum of 100 models. Which of the following cross-validation strategies train the *most* models while still staying within our budget of 100 trained models?

A. $K_1 = 6, K_2 = 5$

B. $K_1 = 3, K_2 = 11$

C. $K_1 = 14, K_2 = 2$

D. $K_1 = 4, K_2 = 9$

E. Don't know.

Solution 27. This can easily be obtained noting for each of the K_1 outer folds we must both (i) train K_2 models on the $L = 3$ different settings of pruning level (ii) train a single new model to estimate the generalization error for this fold. Accordingly the number of trained models is

$$K_1(K_2L + 1).$$

This gives for each of the options:

$$96,\ 102,\ 98,\ 112$$

and so option C is the option which allows us to train the most models within our computational budget.

Technical University of Denmark

Written examination: 16 December 2016, 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.

You must either use the electronic file or the form on this page to hand in your answers but not both. **We strongly encourage that you hand in your answers digitally using the electronic file.** If you hand in using the form on this page, please write your name and student number clearly.

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

No.	Attribute description	Abbrev.
x_1	Area	A
x_2	Perimeter	P
x_3	Length of kernel	L
x_4	Width of kernel	W
y	Seed type	

Table 1: The attributes of the Seeds data set taken from http://archive.ics.uci.edu/ml/datasets/seeds. The output is given by the type of seed, i.e. y=1 corresponds to Kama, y=2 corresponds to Rosa, and y=3 corresponds to Canadian.

Question 1. We will consider the data of wheat kernels based on 70 observations of each class of three seed types, i.e., Kama, Rosa, and Canadian. The original data contains seven attributes, however, we presently only consider four of these attributes given in Table 1. Considering the attributes described in the table and visualized using boxplots in Figure 1 which one of the following statements is *correct*?

A. All the attributes x_1, x_2, x_3, and x_4 are continuous and ratio.

B. The output variable y is ordinal.

C. Rosa and Canadian do not appear to differ in terms of area (A).

D. The observations pertaining to Kama appear to contain clear outliers that must be removed.

E. Don't know.

Solution 1. As zero means absence of the attribute for all the attributes, i.e. zero area means no area etc. and it makes sense to talk about an attribute value being twice as large etc. than another attribute value whereas all the values are continuous, the first answer is correct. y is nominal indicating class, but not ordinal we in general cannot argue that one type of seed is better/higher than another, only whether a seed is different or not from another. Indeed it appears from the boxplot that Rosa and Canadian differ i terms of Area such that all Rosa seeds have larger area than Canadian seeds. Finally, although boxplots indicate observations that fall beyond the whiskers these observations should not be removed unless there are strong justifications for doing so which the boxplot does not provide reasons for.

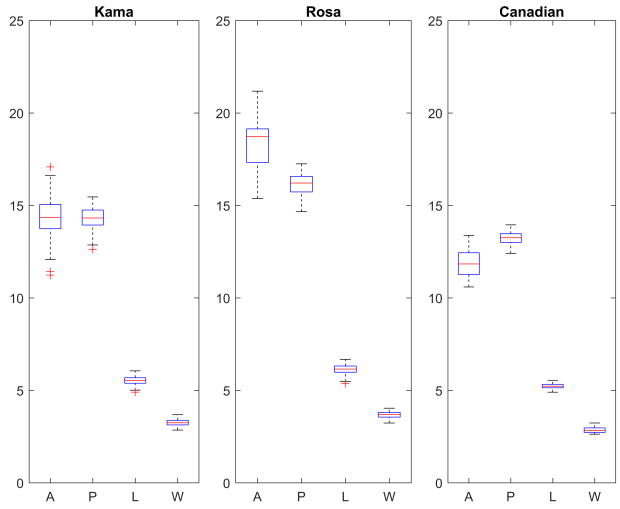

Figure 1: Boxplot of the data visualized separately for each of the three types of seeds; Kama, Rosa, and Canadian.

Question 2. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_4, forming the standardized matrix $\tilde{\boldsymbol{X}}$, resulting in the following $\boldsymbol{S}$ and $\boldsymbol{V}$ matrices obtained from a singular value decomposition:

$$\boldsymbol{S} = \begin{bmatrix} 28.4 & 0 & 0 & 0 \\ 0 & 5.5 & 0 & 0 \\ 0 & 0 & 1.2 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix},$$

$$\boldsymbol{V} = \begin{bmatrix} -0.51 & 0.11 & -0.39 & -0.76 \\ -0.51 & -0.13 & -0.58 & 0.62 \\ -0.49 & -0.69 & 0.53 & -0.05 \\ -0.49 & 0.71 & 0.47 & 0.19 \end{bmatrix}.$$

Which one of the following statements is *correct*?

A. The first principal component accounts for more than 95 % of the variance.

B. The two first principal components account for more than 99.9 % of the variance.

C. The fourth principal component accounts for more than 0.05% of the variance.

D. The attributes are not correlated as the data has been standardized.

E. Don't know.

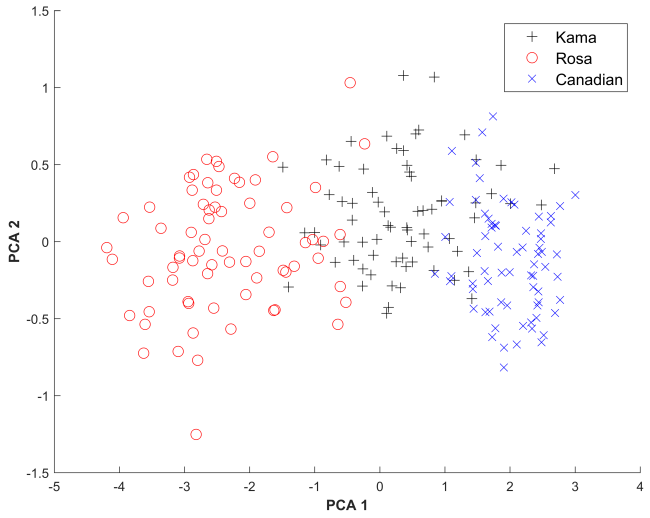

Figure 2: Data projected onto the first and second principal components.

Solution 2. The variation explained by each principal component is given by $\frac{\sigma_i^2}{\sum_{i'} \sigma_{i'}^2}$. As such we find:

$$VarExpPC1 = \frac{28.4^2}{28.4^2+5.5^2+1.2^2+0.5^2} = 0.9619 \quad (1)$$

$$VarExpPC2 = \frac{5.5^2}{28.4^2+5.5^2+1.2^2+0.5^2} = 0.0361 \quad (2)$$

$$VarExpPC3 = \frac{1.2^2}{28.4^2+5.5^2+1.2^2+0.5^2} = 0.0017 \quad (3)$$

$$VarExpPC4 = \frac{0.5^2}{28.4^2+5.5^2+1.2^2+0.5^2} = 0.0003 \quad (4)$$

As such the first PC accounts for more than 95% of the variance, the first two principal components accounts for $0.9619+0.0361 = 0.9980$ which is less than 99.9% of the variance. The fourth principal component accounts for 0.03% which is less than 0.05%. As the first principal component accounts for more than 95% of the variance the attributes are indeed very correlated and not the opposite.

Question 3. The data projected onto the two first principal components (as defined in Question 2) is given in Figure 2 where each class is indicated using different markers and colors. which one of the following statements pertaining to the PCA is *correct*?

A. A relatively long and narrow seed kernel will provide a large positive projection onto the second principal component.

B. The first principal component pertains to the general size of seeds.

C. A seed that has relatively small area and perimeter but large length and width of kernel will have a negative projection onto the third principal component.

D. As the third and fourth principal components account for a low amount of the variance in the data this is a difficult classification task.

E. Don't know.

Solution 3. As we for the second principal component have $\boldsymbol{v}_2^\top = [0.11 \ -0.13 \ -0.69 \ 0.71]$ a relatively long (large positive x_3) and relatively narrow (large negative x_4) will have a large negative projection onto this component. As the coefficients for the first principal component all are negative and generally have same magnitude for the four attributes, this appears to capture the general property of size of the seed, such that relatively large area, perimeter, kernel length and width will provide a negative projection and vice versa. The third principal component is defined by $\boldsymbol{v}_3^\top = [-0.39 \ -0.58 \ 0.53 \ 0.47]$, thus relatively small area and periphery (negative x_1 and x_2) but large kernel (i.e. positive x_3, and x_4) will have a positive projection onto this component. The PCA does not take information about the classes into account and therefore does not necessarily reflect features relevant for classification. In particular, the singular values are uninformed by the classes as this information is not available in the PCA analysis.

Question 4. A decision tree is fitted to the data projected onto the four principal components. At the root of the tree a split according to the projection of the standardized data onto the first principal component being larger than 0 is considered, i.e. $\tilde{\boldsymbol{x}}_n \boldsymbol{v}_1 \geq 0$. For impurity we will use the classification error given by $I(v) = 1 - \max_c p(c|v)$. Before the split we have 70 Kama, 70 Rosa, and 70 Canadian and after the split:

- 24 Kama, 70 Rosa, 0 Canadian below zero in the projection onto $\boldsymbol{v}_1$.
- 46 Kama, 0 Rosa, 70 Canadian above or equal to zero in the projection onto $\boldsymbol{v}_1$.

What is the purity gain of this split?

A. -1.0148

B. 0.0148

C. 0.3333

D. 0.6666

E. Don't know.

Solution 4. The purity gain is given by

$$\Delta = I(\text{parent}) - \sum_{j=1}^{2} \frac{N(v_j)}{N} I(v_j),$$

where

$$I(v) = 1 - \max_c p(c|v).$$

Inserting for the split defined by the projection onto the first PCA being greater or equal to zero we obtain

$$\begin{aligned}\Delta =& (1 - (\tfrac{70}{210})) \\ & -[\tfrac{94}{210}(1 - (\tfrac{70}{94})) \\ & +\tfrac{116}{210}(1 - (\tfrac{70}{116}))] \\ =& \frac{2}{3} - \frac{1}{3} = 1/3.\end{aligned}$$

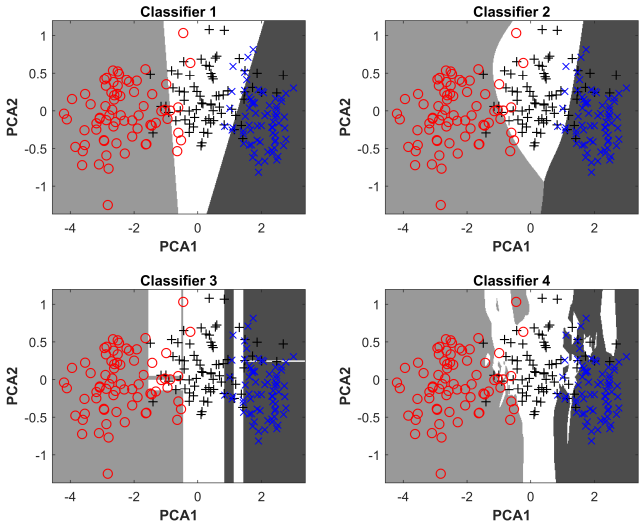

Figure 3: Decision boundaries for four different classifiers trained on the Seeds data projected onto the first two principal components.

Question 5. Four different classifiers are trained on the data projected onto the first two principal components (i.e., using the first and second principal components as features) and the decision boundary for each of the four classifiers is given in Figure 3. Which one of the following statements is *correct*?

A. Classifier 1 is a decision tree, Classifier 2 is an artificial neural network with three hidden units, Classifier 3 is a multinomial regression model, and Classifier 4 is a 3-nearest neighbor classifier.

B. Classifier 1 is an artificial neural network with three hidden units, Classifier 2 is a multinomial regression model, Classifier 3 is a 3-nearest neighbor classifier, and Classifier 4 is a decision tree.

C. Classifier 1 is an artificial neural network with three hidden units, Classifier 2 is a multinomial regression model, Classifier 3 is a decision tree, and Classifier 4 is a 3-nearest neighbor classifier.

D. Classifier 1 is a multinomial regression model, Classifier 2 is an artificial neural network with three hidden units, Classifier 3 is a decision tree, and Classifier 4 is a 3-nearest neighbor classifier.

E. Don't know.

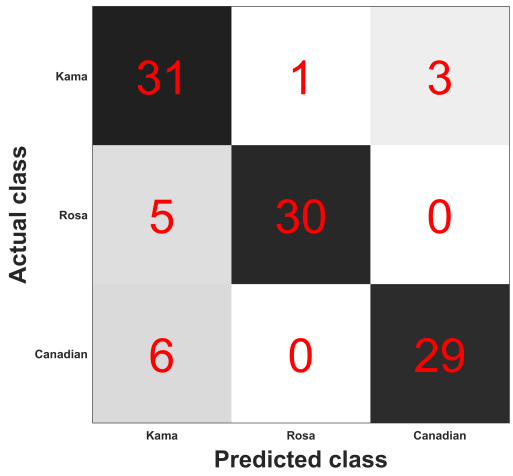

Figure 4: The confusion matrix of a 3-nearest neighbor classifier used to predict the Seeds data.

Solution 5. The decision boundary of classifier 1 is based on lines thus using multinomial regression. Classifier two has smooth boundaries that are non-linear thus based on ANN. Classifier three has axis aligned boundaries corresponding to the decision tree, leaving classifier four as the 3-nearest neighbour due to its very complex and non-smooth boundaries.

Question 6. The data is split in half and a KNN classifier used to predict the test-set based on the training set for K=3. The confusion matrix of the KNN classifier is given in Figure 4. What is the accuracy of the classifier?

A. 0.1429

B. 0.8571

C. 0.8911

D. 0.9574

E. Don't know.

Solution 6. The accuracy is given by the number of correctly classified observations out of the total classified observations which is $accurcy = \frac{31+30+29}{31+30+29+1+3+5+6} = 90/105 = 0.8571$.

	O1	O2	O3	O4	O5	O6	O7	O8	O9
O1	0	0.534	1.257	1.671	1.090	1.315	1.484	1.253	1.418
O2	0.534	0	0.727	2.119	1.526	1.689	1.214	0.997	1.056
O3	1.257	0.727	0	2.809	2.220	2.342	1.088	0.965	0.807
O4	1.671	2.119	2.809	0	0.601	0.540	3.135	2.908	3.087
O5	1.090	1.526	2.220	0.601	0	0.331	2.563	2.338	2.500
O6	1.315	1.689	2.342	0.540	0.331	0	2.797	2.567	2.708
O7	1.484	1.214	1.088	3.135	2.563	2.797	0	0.275	0.298
O8	1.253	0.997	0.965	2.908	2.338	2.567	0.275	0	0.343
O9	1.418	1.056	0.807	3.087	2.500	2.708	0.298	0.343	0

Table 2: Pairwise Euclidean distance between nine observations in the Seeds data. Black observations (i.e., O1, O2, O3) are observations corresponding to Kama seeds, red observations (i.e., O4, O5, O6) are observations corresponding to Rosa seeds, and blue observations (i.e., O7, O8, O9) are observations corresponding to Canadian seeds.

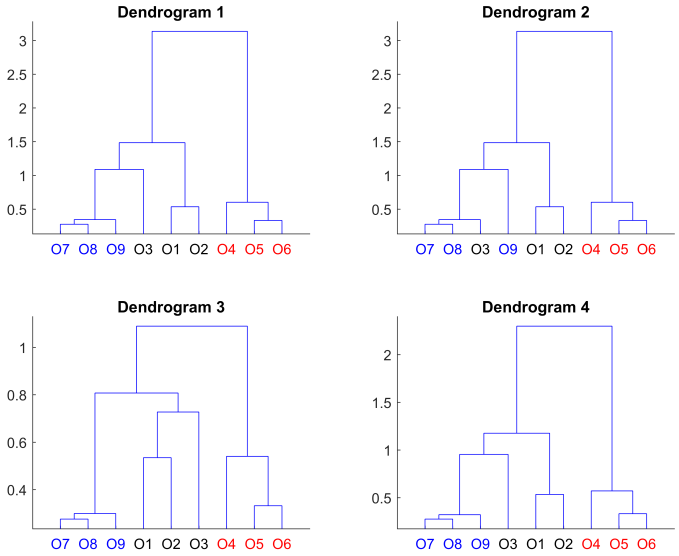

Figure 5: Four different dendrograms derived from the distances between the nine observation in Table 2.

Question 7. In Table 2 is given the pairwise Euclidean distances between nine observations of the Seeds data. A hierarchical clustering is used to cluster these nine observations using complete (i.e., maximum) linkage. Which one of the dendrograms given in Figure 5 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 7. In complete distance clusters are merged according to maximal distance between observations within each cluster. The dendrogram grows by first merging O7 and O8 at 0.275, then O5, O6 at level 0.331, then {O7,O8} with O9 at 0.343, then O1 and O2 at level 0.534, then O4 with {O5,O6} at 0.601, then O3, {O7,O8,O9} at level 1.0882, then {O1, O2} with {O3, O7,O8,O9} at 1.484, and finally {O1, O2, O3 O7,O8,O9} with {O4, O5,O6} at 3.135. Only Dendrogram 1 has these properties.

Question 8. We will consider thresholding Dendrogram 4 at the level of three clusters. We recall that the Rand index also denoted the simple matching coefficient (SMC) between the true labels and the extracted clusters is given by:

$$SMC = \frac{f_{00} + f_{11}}{K},$$

where f_{00} is the number of object pairs in different class assigned to different clusters and f_{11} is the number of object pairs in same class assigned to same cluster, whereas $K = N(N-1)/2$ is the total number of object pairs where N is the number of observations considered. What is the above SMC between the true labeling of the observations into the three classes Kama, Rosa, and Canadian, and the clustering defined by thresholding Dendrogam 4 at the level of three clusters?

A. 0.7500

B. 0.7778

C. 0.8611

D. 1.0000

E. Don't know.

Solution 8. When thresholding the clustering we obtain: the cluster indices: $[1\ 1\ 3\ 2\ 2\ 2\ 3\ 3\ 3]^\top$, whereas the true class labels are $[1\ 1\ 1\ 2\ 2\ 2\ 3\ 3\ 3]\top$. From this, we obtain: Total number of object pairs is:
$K = 9(9-1)/2 = 36$
$f_{00} = 2 \cdot (3+3) + 3 \cdot (3+1) = 24$
$f_{11} = 2{\cdot}(2{-}1)/2{+}3{\cdot}(3{-}1)/1{+}3{\cdot}(3{-}1)/2{+}1{\cdot}(1{-}1)/2 = 7$
$SMC = \frac{f_{00}+f_{11}}{K} = \frac{24+7}{36} = 0.8611$

Question 9. To determine the type of seed of an observation we will use a k-nearest neighbor (KNN) classifier to predict each of the nine observations based on the Euclidean distance between the observations given in Table 2. We will use leave-one-out cross-validation for the KNN in order to classify the nine considered observations using a two-nearest neighbor classifier, i.e. $K = 2$. For tied classes we will classify the observation according to its closest observation. The analysis will be based only on the data given in Table 2. Which one of the following statements is *correct*?

A. All the observations will be correctly classified.

B. One of the observations will be misclassified.

C. Two of the observations will be misclassified.

D. Three of the observations will be misclassified.

E. Don't know.

Solution 9. $N(O1, 2) = \{O2, O5\}$ as O2 is closest it will be correctly classified as Kama.
$N(O2, 2) = \{O1, O3\}$ and will be correctly classified as Kama.
$N(O3, 2) = \{O2, O9\}$ as O2 is closest it will be correctly classified as Kama.
$N(O4, 2) = \{O6, O5\}$ and will be correctly classified as Rosa.
$N(O5, 2) = \{O6, O4\}$ and will be correctly classified as Rosa.
$N(O6, 2) = \{O5, O4\}$ and will be correctly classified as Rosa.
$N(O7, 2) = \{O8, O9\}$ and will be correctly classified as Canadian.
$N(O8, 2) = \{O7, O9\}$ and will be correctly classified as Canadian.
$N(O9, 2) = \{O7, O8\}$ and will be correctly classified as Canadian.

Question 10. We suspect that observation O4 may be an outlier. In order to assess if this is the case we would like to calculate the average relative KNN density based on the observations given in Table 2 only. We recall that the KNN density and average relative density 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. Based on the data in Table 2, what is the average relative density for observation O4 for $K = 1$ nearest neighbors?

A. 0.54

B. 0.61

C. 1.63

D. 1.85

E. Don't know.

Solution 10.

$$\text{density}(\boldsymbol{x}_{O4}, 1) = (\frac{1}{1} \cdot 0.540)^{-1} = 1.8519$$

$$\text{density}(\boldsymbol{x}_{O6}, 1) = (\frac{1}{1} \cdot 0.331)^{-1} = 3.0211$$

$$\text{a.r.d.}(\boldsymbol{x}_{O4}, 1) = \frac{\text{density}(\boldsymbol{x}_{O4}, 1)}{\frac{1}{1}(\text{density}(\boldsymbol{x}_{O6}, 1))} = \frac{1.8519}{\frac{1}{1} \cdot 3.0211} = 0.61$$

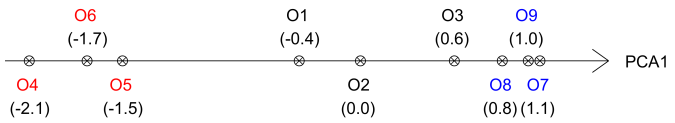

Figure 6: The nine observations considered in Table 2 projected onto the first principal component (the location of the projection is given in parenthesis).

Question 11. We will consider the nine observations projected onto the first principal component given in Figure 6. We will cluster this data using k-means with Euclidean distance into three clusters (i.e., k=3) and initialize the k-means algorithm with centroids located at observation O4, O6, and O5. Which one of the following statements is *correct*?

A. The converged solution will be {O4}, {O6},{O1, O2, O3, O5, O6, O7, O8, O9}.

B. The converged solution will be {O4, O5, O6},{O1, O2, O3}, {O7, O8, O9}.

C. The converged solution will be {O4, O5, O6},{O1, O2}, {O3, O7, O8, O9}.

D. The converged solution will be {O4},{O5, O6}, {O1, O2, O3, O7, O8, O9}.

E. Don't know.

Solution 11. With the described initialization, observation O4 will be assigned to the cluster located at O4, observation O6 will be assigned to the cluster located at O6, and the remaining observations {O1, O2, O3, O5, O7, O8, O9} assigned to the cluster located at O5. Thus, only cluster located at O5 will change location and the location updated to $\frac{-1.5+-0.4+0.0+0.6+0.8+1.0+1.1}{7} = 0.2286$. For this new location O5 is closer to cluster located at O6 than the cluster located at 0.2286, resulting in the updated clustering {O4},{O5, O6}, {O1, O2, O3, O7, O8, O9}. Thus the second cluster will change location to $\frac{-1.7+-1.5}{2} = -1.6$ whereas the third cluster will change location to $\frac{-0.4+0.0+0.6+0.8+1.0+1.1}{6} = 0.5167$. As O1 is still closest to cluster 3 there is no change of assignment and the k-means procedure has converged.

Feature(s)	Training Error Rate	Test Error Rate
No features	0.6667	0.6667
x_1	0.1143	0.1524
x_2	0.1143	0.1143
x_3	0.2190	0.1714
x_4	0.1524	0.1714
x_1 and x_2	0.0952	0.1619
x_1 and x_3	0.1143	0.1619
x_1 and x_4	0.1143	0.1619
x_2 and x_3	0.1238	0.1333
x_2 and x_4	0.1048	0.1429
x_3 and x_4	0.1143	0.1619
x_1 and x_2 and x_3	0.0571	0.1714
x_1 and x_2 and x_4	0.1048	0.1619
x_1 and x_3 and x_4	0.0857	0.1619
x_2 and x_3 and x_4	0.0762	0.1524
x_1 and x_2 and x_3 and x_4	0.0667	0.1810

Table 3: Error rate for the training and test set when using multinomial regression to predict the type of seed using different combinations of the four attributes (x_1–x_4) based on the hold-out method with 50 % of the observations hold-out for testing.

Question 12. A multinomial regression classifier is trained using different combinations of the four attributes x_1, x_2, x_3, and x_4. Table 3 gives the training and test performance of the multinomial regression classifier when trained using different combinations of the four attributes. Which one of the following statements is *correct*?

A. Forward selection will result in a better model being selected than backward selection.

B. Neither forward nor backward selection will identify the optimal feature combination for this problem.

C. Backward selection will use a model that includes three features.

D. Forward selection will select only one feature.

E. Don't know.

Solution 12. Using forward and backward selection we would like to minimize the test error rate. Thus, the forward selection would first select x_2 having lowest test error rate. As no combination of the feature

x_2 lead to improvement in the test error rate the forward selection method terminates. Backward selection starts with all features and an improvement is found removing feature x_1 providing a test error rate of 0.1524 for x_2 and x_3 and x_4. Subsequently, x_4 is removed providing a test error rate of 0.1333 for the features x_2 and x_3. Finally x_3 is removed as having only feature x_2 provides an error rate of 0.1143 and the method terminates.

Question 13. We would like to investigate if we can predict the width of a seed kernel (x_4) based on the area (x_1), perimeter (x_2), and length of kernel (x_3). For this purpose regularized least squares regression is applied based on minimizing with respect to $\boldsymbol{w}$ the cost function:

$$E(\boldsymbol{w}) = \sum_n (x_{n4} - [1 \quad x_{n1} \quad x_{n2} \quad x_{n3}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w},$$

where x_{nm} denotes the m'th feature of the n'th observation, and 1 is concatenated the data to account for the bias term. We will consider the following four different values of λ: $\lambda_1 = 1$, $\lambda_2 = 10$, $\lambda_3 = 100$, and $\lambda_4 = 1000$. We obtain the following four different solutions for $\boldsymbol{w}$ here given in random order of the values of λ considered:

$$\boldsymbol{w}_a = \begin{bmatrix} 0.0538 \\ 0.0558 \\ 0.1861 \\ -0.0596 \end{bmatrix}, \quad \boldsymbol{w}_b = \begin{bmatrix} 0.0089 \\ 0.0931 \\ 0.1093 \\ 0.0417 \end{bmatrix},$$

$$\boldsymbol{w}_c = \begin{bmatrix} 0.2811 \\ 0.0445 \\ 0.3379 \\ -0.4626 \end{bmatrix}, \quad \boldsymbol{w}_d = \begin{bmatrix} 0.0167 \\ 0.0698 \\ 0.1354 \\ 0.0403 \end{bmatrix}.$$

Which one of the following solutions to $\boldsymbol{w}$ corresponds to the correct value of λ?

A. $\boldsymbol{w}_a$ corresponds to λ_2.

B. $\boldsymbol{w}_b$ corresponds to λ_2.

C. $\boldsymbol{w}_c$ corresponds to λ_2.

D. $\boldsymbol{w}_d$ corresponds to λ_2.

E. Don't know.

Solution 13. As we increase λ we put more and more emphasis on the regularization penalization term $\|\boldsymbol{w}\|_2^2$. Thus, by evaluating this term for each solution vector we obtain: $\|\boldsymbol{w}_a\|_2^2 = 0.0442$, $\|\boldsymbol{w}_b\|_2^2 = 0.0224$, $\|\boldsymbol{w}_c\|_2^2 = 0.4092$, $\|\boldsymbol{w}_d\|_2^2 = 0.0251$. Thus sorting by these norm values we find that $\boldsymbol{w}_c$ corresponds to λ_1, $\boldsymbol{w}_a$ corresponds to λ_2, $\boldsymbol{w}_d$ corresponds to λ_3, $\boldsymbol{w}_b$ corresponds to λ_4.

No.	Attribute description
x_1	Occurrence of nausea
x_2	Lumbar pain
x_3	Urine pushing
x_4	Micturition pains
x_5	Burn/itch/swell urethra outlet
y	Inflammation of urinary bladder

Table 4: The attributes considered from the study on acute inflammation (taken from https://archive.ics.uci.edu/ml/datasets/Acute+ Inflammations). The attributes x_1–x_5 and y are binary where we use 1 for true and 0 for false.

Question 14. In a study of acute inflammation we would like to predict urinary bladder inflammation (the data is taken from https://archive.ics.uci.edu/ml/datasets/Acute+ Inflammations). We will consider a subset of the attributes, these attributes are given in Table 4. From the study we have

- 49.17 pct. of the persons have inflammation of urinary bladder.
- 32.20 pct. of the persons that have inflammation of urinary bladder have occurrence of nausea.
- 16.39 pct. of the persons that do not have inflammation of urinary bladder have occurrence of nausea.

What is the probability that a person that has occurrence of nausea, i.e. $x_1 = 1$, has inflammation of the urinary bladder, i.e. $y = 1$, according to this study?

A. 15.83 %

B. 32.20 %

C. 65.52 %

D. 98.82%

E. Don't know.

Solution 14. According to Bayes' theorem we have:

$$P(y=1|x_1=1) = \frac{P(x_1=1|y=1)P(y=1)}{P(x_1=1)}$$
$$= \frac{P(x_1=1|y=1)P(y_1=1)}{P(x_1=1|y=1)P(y_1=1)+P(x_1=1|y=0)P(y_1=0)}$$
$$= \frac{0.3220 \cdot 0.4917}{0.3220 \cdot 0.4917 + 0.1639 \cdot (1-0.4917)}$$
$$= 0.6552$$

	x_1	x_2	x_3	x_4	x_5	y
P1	1	1	1	1	0	1
P2	0	0	0	0	0	0
P3	1	1	0	1	0	0
P4	0	1	1	0	1	0
P5	1	1	1	1	1	1
P6	0	0	0	0	0	0
P7	1	1	0	1	0	0
P8	0	1	1	0	1	0
P9	1	1	1	1	0	1
P10	0	1	1	0	1	0
P11	0	0	0	0	0	0
P12	1	1	0	1	0	0
P13	0	1	1	0	1	0
P14	0	1	1	0	1	0

Table 5: Provided in the above table are the last 14 observations of the acute inflammation data.

Question 15. We will consider a subset of the acute inflammation data given by the last 14 observations provided in Table 5. We will consider this dataset a market basket with 14 persons (P1-P14) denoting the customers and six items denoted $x_1 - x_5$ and y corresponding to the five input attributes and output variable respectively of the features described in Table 4. What are all frequent itemsets with support greater than 40%?

A. $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$, $\{x_1, x_2\}$, $\{x_2, x_3\}$, $\{x_2, x_4\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$.

B. $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$, $\{x_1, x_2\}$, $\{x_1, x_4\}$, $\{x_2, x_3\}$, $\{x_2, x_4\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$.

C. $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$, $\{x_1, x_2\}$, $\{x_1, x_4\}$, $\{x_2, x_3\}$, $\{x_2, x_4\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$, $\{x_2, x_3, x_5\}$.

D. $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$, $\{x_1, x_2\}$, $\{x_1, x_4\}$, $\{x_2, x_3\}$, $\{x_2, x_4\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$, $\{x_1, x_2, x_4\}$,$\{x_2, x_3, x_5\}$.

E. Don't know.

Solution 15. For a set to have support more than 40% the set must occur at least $0.4 \cdot 14 = 5.6$, i.e. 6 out of the 14 times. All the itemsets that have this property are $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_4\}$, $\{x_5\}$, $\{x_1, x_2\}$, $\{x_1, x_4\}$, $\{x_2, x_3\}$, $\{x_2, x_4\}$, $\{x_2, x_5\}$, $\{x_3, x_5\}$, $\{x_1, x_2, x_4\}$, $\{x_2, x_3, x_5\}$.

Question 16. What is the confidence of the association rule $\{x_1, x_2, x_3, x_4, x_5\} \rightarrow \{y\}$?

A. 0.0%

B. 7.1 %

C. 21.4%

D. 100.0 %

E. Don't know.

Solution 16. The confidence is given as

$$P(y=1|x_1=1,x_2=1,x_3=1,x_4=1,x_5=1) = \frac{P(y=1,x_1=1,x_2=1,x_3=1,x_4=1,x_5=1)}{P(x_1=1,x_2=1,x_3=1,x_4=1,x_5=1)} = \frac{1/14}{1/14} = 1 = 100.0\%$$

Question 17. We would like to predict whether a subject has inflammation of urinary bladder ($y = 1$) or not ($y = 0$) using the data in Table 5 and the attributes x_1, and x_2 only. We will apply a Naïve Bayes classifier that assumes independence between the two attributes given y. Given that a person has $x_1 = 1$, and $x_2 = 1$ what is the probability that the person has an inflammation of urinary bladder ($y = 1$) according to the Naïve Bayes classifier?

A. 1/14

B. 3/14

C. 1/2

D. 11/19

E. Don't know.

Solution 17. According to the Naïve Bayes classifier we have

$$P(y=1|x_1=1,x_2=1) = \frac{\begin{pmatrix} P(x_1=1|y=1)\times \\ P(x_2=1|y=1)\times \\ P(y=1) \end{pmatrix}}{\begin{pmatrix} P(x_1=1|y=1)\times \\ P(x_2=1|y=1)\times \\ P(y=1) \end{pmatrix} + \begin{pmatrix} P(x_1=1|y=0)\times \\ P(x_2=1|y=0)\times \\ P(y=0) \end{pmatrix}} = \frac{3/3 \cdot 3/3 \cdot 3/14}{3/3 \cdot 3/3 \cdot 3/14 + 3/11 \cdot 8/11 \cdot 11/14} = .$$

Question 18. Considering the data in Table 5, we will use x_1 to classify whether a subject has inflammation of urinary bladder ($y = 1$) or not ($y = 0$). We will quantify how useful x_1 is for this purpose by calculating the area under curve (AUC) of the receiver operator characteristic (ROC). Which one of the ROC curves given in Figure 7 corresponds to using the feature x_1 to determine if a subject has inflammation of urinary bladder?

A. The curve with AUC=0.636.

B. The curve with AUC=0.864.

C. The curve with AUC=0.909.

D. The curve with AUC=1.000.

E. Don't know.

Solution 18. The ROC is defined by considering all conceivable thresholds and plotting the true positive

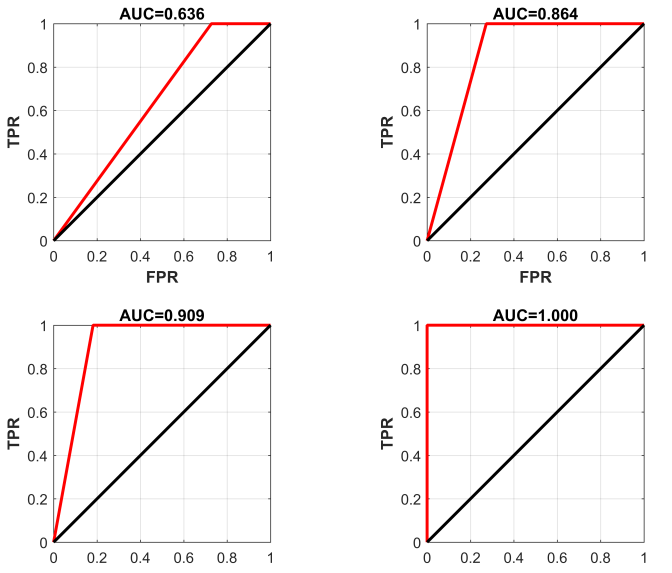

Figure 7: Four different receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC).

rate (TPR) against the false positive rate (FPR). For a threshold larger than 1 we have that TPR=FPR=0. For the threshold at 1 we have that 3 of the 11 observations having y=0 have $x_1 = 1$ thus FPR=3/11, whereas all three of the three observations with y=1 have $x_1 = 1$, thus TPR=1. lowering the threshold we get when thresholding above 0 that all 11 of the 11 observations or which y=0 are false positive, i.e. FPR=1, and all three of three observations where y=1 are true positive, thus TPR=1, giving an AUC =0.864.

Question 19. Considering the data in Table 5, we will calculate the similarity between $P1$ given as the vector $\boldsymbol{r} = [1\ 1\ 1\ 1\ 0\ 1]$ and $P3$ given by the vector $\boldsymbol{s} = [1\ 1\ 0\ 1\ 0\ 0]$ using Jaccard, Simple Matching Coefficient, and Cosine similarity given respectively by:

$$J(\boldsymbol{r},\boldsymbol{s}) = \frac{f_{11}}{M - f_{00}},$$
$$SMC(\boldsymbol{r},\boldsymbol{s}) = \frac{f_{11} + f_{00}}{M},$$
$$cos(\boldsymbol{r},\boldsymbol{s}) = \frac{f_{11}}{\|\boldsymbol{r}\|_2\|\boldsymbol{s}\|_2}.$$

Which one of the following statements regarding the similarity of $\boldsymbol{r}$ an $\boldsymbol{s}$ is *correct*?

A. $J(\boldsymbol{r},\boldsymbol{s}) < SMC(\boldsymbol{r},\boldsymbol{s})$

B. $J(\boldsymbol{r},\boldsymbol{s}) > cos(\boldsymbol{r},\boldsymbol{s})$

C. $SMC(\boldsymbol{r},\boldsymbol{s}) > cos(\boldsymbol{r},\boldsymbol{s})$

D. $cos(\boldsymbol{r},\boldsymbol{s}) = 3/15$

E. Don't know.

Solution 19. For $\boldsymbol{r}$ and ,$\boldsymbol{s}$ we have:

$$J(\boldsymbol{r},\boldsymbol{s}) = \frac{f_{11}}{M - f_{00}} = 3/5 = 0.6000,$$
$$SMC(\boldsymbol{r},\boldsymbol{s}) = \frac{f_{11} + f_{00}}{M} = 4/6 = 0.6667,$$
$$cos(\boldsymbol{r},\boldsymbol{s}) = \frac{f_{11}}{\|\boldsymbol{r}\|_2\|\boldsymbol{s}\|_2} = 3/(\sqrt{5}\sqrt{3}) = 0.7746.$$

Thus, J(P1,P3)<SMC(P1,P3) is the only correct statement.

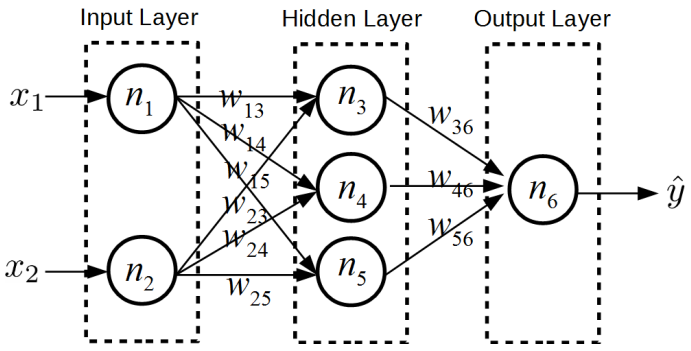

Figure 8: The architecture of the considered neural network having one hidden layer.

Question 20. A neural network is trained to separate persons with urinary inflammation ($y = 1$) from persons not having urinary inflammation based on the features x_1 and x_2. The structure of the neural network is outlined in Figure 8. The activation function used for all six neurons n_1, n_2, n_3, n_4, n_5, and n_6 is the rectified linear unit

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The neural network has no biases, i.e. all the biases of all units are zero. The weights of the network are:

$$\begin{array}{lll} w_{13} = 0.5, & w_{14} = 0.5, & w_{15} = -0.5, \\ w_{23} = 0.5, & w_{24} = -0.5, & w_{25} = 0.25, \\ w_{36} = 0.25, & w_{46} = -0.25, & w_{56} = 0.25. \end{array}$$

What will be the output ($\hat{y}$) of the neural network for an observation having $x_1 = 1$ and $x_2 = 1$?

A. 0

B. 0.25

C. 0.75

D. 1

E. Don't know.

Solution 20. The output of the neurons in the hidden layer will be:

$$\begin{aligned} n_3 &: f(0.5 \cdot 1 + 0.5 \cdot 1) = f(1) = 1, \\ n_4 &: f(0.5 \cdot 1 - 0.5 \cdot 1) = f(0) = 0, \\ n_5 &: f(-0.5 \cdot 1 + 0.25 \cdot 1) = f(-0.25) = 0. \end{aligned}$$

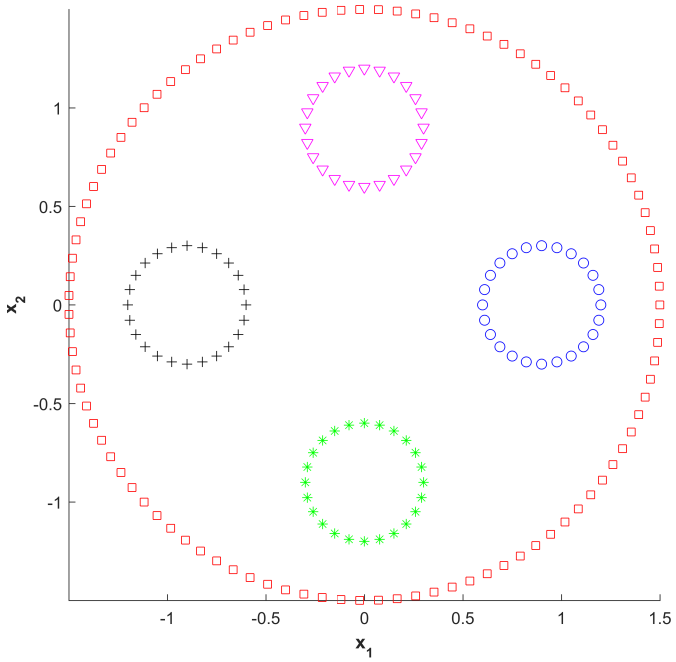

Figure 9: A dataset with five classes given respectively by a large circle and four smaller circles.

The output of the output neurons will therefore be:

$$n_6 : f(0.25 \cdot 1 - 0.25 \cdot 0 + 0.25 \cdot 0) = f(0.25) = 0.25.$$

Question 21. We will consider the dataset with five classes given in Figure 9 defined respectively by the four inner circles and the larger outer circle. We will cluster this dataset using hierarchical clustering. What would be a suitable measure of proximity and linkage in order to perfectly separate the five classes into five clusters?

A. Average linkage using the 2-norm (i.e. $\|\boldsymbol{x} - \boldsymbol{y}\|_2$) as proximity measure.

B. Single linkage using the 1-norm (i.e. $\|\boldsymbol{x} - \boldsymbol{y}\|_1$) as proximity measure.

C. Complete linkage using the 2-norm (i.e. $\|\boldsymbol{x} - \boldsymbol{y}\|_2$) as proximity measure.

D. Complete linkage using the 1-norm (i.e. $\|\boldsymbol{x} - \boldsymbol{y}\|_1$) as proximity measure.

E. Don't know.

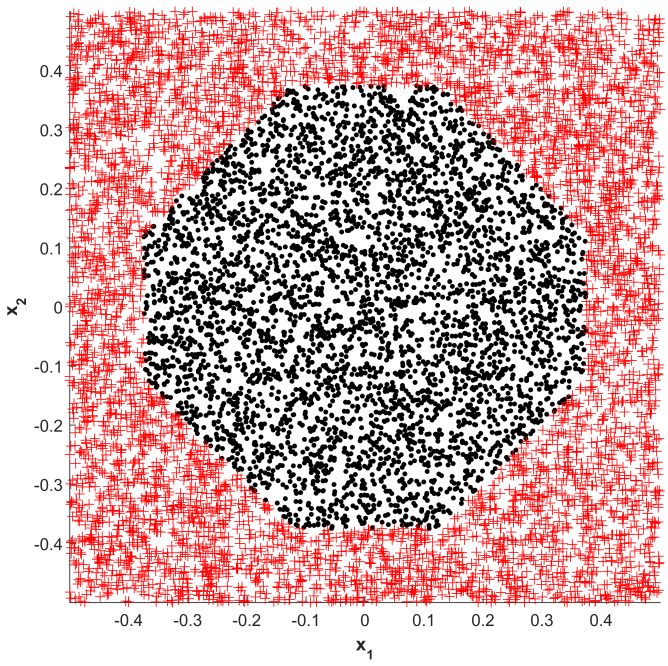

Figure 10: A two class classification problem.

Solution 21. The choice of norm will not generally influence the results much in this example as observations that are close in 1-norm will also be close in 2-norm. However, linkage function will heavily influence the results. As all clusters have the property that at least one observation is closer within the cluster than an observation in another cluster, thus, a contiguity based approach will be well-suited. Hence, single linkage clustering will perfectly separate the classes whereas the other approaches will fail.

Question 22. Consider the dataset with two classes given in Figure 10. Which one of the following decisions would lead to a perfect separation of the two classes?

A. If $\|\boldsymbol{x}\|_1 \leq \frac{1}{4}$ and $\|\boldsymbol{x}\|_2 \leq \frac{3}{8}$ then black dot, otherwise red plus.

B. If $\|\boldsymbol{x}\|_2 \leq \frac{3}{8}$ and $\|\boldsymbol{x}\|_\infty \leq \frac{1}{4}$ then black dot, otherwise red plus.

C. If $\|\boldsymbol{x}\|_1 \leq \frac{1}{2}$ and $\|\boldsymbol{x}\|_\infty \leq \frac{1}{2}$ then black dot, otherwise red plus.

D. If $\|\boldsymbol{x}\|_1 \leq \frac{1}{2}$ and $\|\boldsymbol{x}\|_\infty \leq \frac{3}{8}$ then black dot, otherwise red plus.

E. Don't know.

Solution 22. The decision boundary is formed by a hexagonal shape. Inspecting the position of the edges it is seen that the decision boundary traverses the coordinates $(0, 3/8)$ and $(0.25, 0.25)$ corresponding to a point where $\|\boldsymbol{x}\|_1 = \frac{1}{2}$ and $\|\boldsymbol{x}\|_\infty = \frac{3}{8}$ respectively.

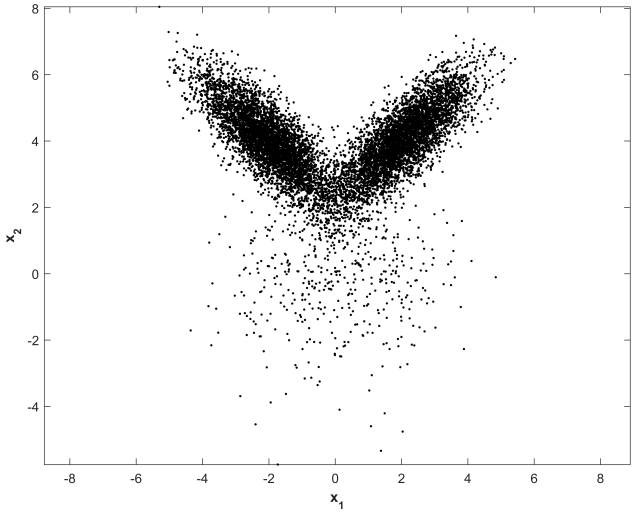

Figure 11: 10.000 data observations drawn from a Gaussian Mixture Model (GMM).

Question 23. Consider the 10.000 observations drawn from a Guassian Mixture Model (GMM) shown in Figure 11. We will in the following use: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$ to denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Which one of the following GMM densities best characterize the data?

A.

$$p(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix})$$

B.

$$p(\boldsymbol{x}) = 0.05 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}) + 0.475 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}) + 0.475 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix})$$

C.

$$p(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}) + 0.25 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}) + 0.25 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix})$$

D.

$$p(\boldsymbol{x}) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix})$$

E. Don't know.

Solution 23. There are three clusters and the centroids of these clusters are $\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} -2 \\ 4 \end{bmatrix}$. The cluster at $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is not very dense and should therefore have a very low weight, furthermore, the clusters at $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$ has positive covariance whereas the cluster at $\begin{bmatrix} -2 \\ 4 \end{bmatrix}$ has negative covariance. This property only holds for the answer option:

$$p(x) = 0.05 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}) + 0.475 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}) + 0.475 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix})$$

Question 24. We will consider a very large dataset with 100 mio. observations and ten features, i.e. $N = 100.000.000$ and $M = 10$. We would like to perform two-level cross-validation in order to select between 3 different settings of the parameters of a model (inner fold) and estimate the generalization error (outer fold). We are only allowed to train maximally 65 models in total. Which one of the following procedures satisfies this constraint?

A. Five fold cross-validation in both the outer and inner folds.

B. Leave-one-out cross-validation for the outer fold and hold-out 50 % for the inner fold.

C. Ten-fold cross-validation for the outer fold and two fold cross-validation for the inner fold.

D. Two-fold cross-validation for the outer fold and ten fold cross-validation for the inner fold.

E. Don't know.

Solution 24. In the inner fold we have to train as many models as we have folds to identify the optimal parameters. We then use the optimal parameters to train a model on the entire training set in order to evaluate this model on the test set defined in the outer fold. Thus we need for each outer fold to train the number of inner folds + one model (i.e. the model trained on the entire training data) times number of outer folds, i.e. $K_1(K_2 \cdot S + 1)$. This gives for:
Five fold cross-validation in both the outer and inner folds: $5(5 \cdot 3 + 1) = 80$ models
Leave-one-out cross-validation for the outer fold and hold-out: $100.000.000(1 \cdot 3 + 1) = 400.000.000$ models
Ten-fold cross-validation for the outer fold and two fold cross-validation for the inner fold: $10(2 \cdot 3 + 1) = 70$
Two-fold cross-validation for the outer fold and ten fold cross-validation for the inner fold: $2(10 \cdot 3 + 1) = 62$.

Question 25. We recall that the AdaBoost algorithm is given by updating the weight to the i'th data observation (w_i) based on the classifier f_t at round t according to:

$$w_i(t+1) = \frac{\tilde{w}_i(t+1)}{\sum_{j=1}^{N} \tilde{w}_j(t+1)}, \text{ where}$$

$$\tilde{w}_i(t+1) = \begin{cases} w_i(t)e^{-\alpha_t} & \text{if } f_t(\boldsymbol{x}_i) = y_i \\ w_i(t)e^{\alpha_t} & \text{if } f_t(\boldsymbol{x}_i) \neq y_i. \end{cases}$$

Here $\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$ (where log is the natural logarithm) and $\epsilon_t = \sum_{i=1}^{N} w_i\left(1 - \delta_{f_t(\boldsymbol{x}_i),y_i}\right)$, where $\delta_{f_t(\boldsymbol{x}_i),y_i} = 1$ if $f_t(\boldsymbol{x}_i) = y_i$ and zero otherwise. Initially the weights are uniform across samples, i.e. $w_1 = w_2 = \ldots = w_N = 1/N$ where N is the number of observations.

A dataset is sampled with replacement from this uniform distribution and the classifier is trained on this sampled data. Using this trained classifier 5 of the original 25 observations are misclassified. What will the updated weights be for these misclassified observations according to the AdaBoost algorithm?

A. 0.02

B. 0.025

C. 0.08

D. 0.1

E. Don't know.

Solution 25. As we have 5 misclassified observations the weighted error rate will be $\epsilon_1 = 5/25 = 1/5$, thus $\alpha_t = \frac{1}{2}\log(\frac{1-1/5}{1/5}) = \frac{1}{2}\log 4 = 0.6931$. Thus for correctly classified observations we have: $\tilde{w}_j(t+1) = w_j(t)e^{-\alpha_t} = 1/25e^{-0.6931}$ and for incorrectly classified observations $\tilde{w}_j(t+1) = w_j(t)e^{-\alpha_t} = 1/25e^{+0.6931}$. We thereby obtain for the updated weights for misclassified observations: $w_i(t+1) = \frac{1/25e^{+0.6931}}{20/25e^{-0.6931}+5/25e^{+0.6931}} = 0.1$

Question 26. For which of the following purposes is cross-validation *the least* well suited?

A. Select the number of hidden units in artificial neural networks (ANN).

B. Select the width of the Gaussian kernel in kernel density estimation (KDE).

C. Select the observations that minimize the training error.

D. Select the number of neighbors in KNN classification.

E. Don't know.

Solution 26. Cross-validation can trivially be used to quantify the number of hidden units in ANN and nearest neighbors in KNN classification by evaluating the performance predicting the output in these supervised

learning problems. As we have seen in the course cross-validation can also be used to quantify the width of the kernel density estimator. Cross-validation is used to quantify models generalization through the use of the test sets and not to minimize the training error.

Question 27. Which of the following statements regarding ensemble methods is correct?

A. In ensemble methods it is important that the different trained classifiers perform very similar.

B. In Random Forest features are randomly sampled at each node of the tree.

C. Random Forest is the same as fitting several decision trees and classifying according to the tree for which the leaf has highest purity.

D. In bagging the output class labels are randomly changed to introduce noise for robustness.

E. Don't know.

Solution 27. Using ensemble methods it is important that the different methods are not the same but as independent as possible. In Random Forest $m < M$ features are indeed randomly sampled at each node of the tree. Majority voting is used for the classification and not the tree with highest purity of the leaf. In bagging observations are uniformly sampled with replacement and there is no additional emphasis to misclassified observations.

Technical University of Denmark

Written examination: 23 May 2017, 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.

You must either use the electronic file or the form on this page to hand in your answers but not both. **We strongly encourage that you hand in your answers digitally using the electronic file.** If you hand in using the form on this page, please write your name and student number clearly.

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

No.	Attribute description	Abbrev.
x_1	Number of cylinders	cyl
x_2	Horsepower	hp
x_3	Weight	wt
x_4	Transmission (0=automatic, 1=manual)	am
x_5	Number of forward gears	gear
y	Miles pr. gallon	mpg

Table 1: The attributes of the Motor Trend Car Road Tests dataset taken from https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv. The output y is given by the miles pr. gallon the car drives. The dataset has 32 observations and we presently consider the five input features x_1–x_5.

Question 1. We will consider the data of Motor Trend Car Road Tests based on 32 automobiles (observations) taken from https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv for brevity denoted the Cars dataset in the following. The original data contains eleven attributes, however, we presently consider only five of these attributes given in Table 1 as well as the output attribute y given by how many miles the car drives pr. gallon of fuel.

Considering the attributes described in Table 1 which one of the following statements is *correct*?

A. The attribute x_5 is continuous.

B. The output variable y is ratio.

C. The attribute x_4 is ordinal.

D. The attribute x_1 is nominal.

E. Don't know.

Solution 1. As x_5 is given by the number of forward gears this attribute is discrete (i.e. defined by the integer numbers) and not continuous. As y has a meaningful zero value defining absence of miles pr. gallon y is ratio whereas we can talk about a car driving twice as far pr. gallon of fuel than another etc. Transmission is defined as either automatic or manual and thus categorical and not ordinal, (i.e., its is not meaningful to say that manual is better than automatic). x_1 is discrete ratio and thus not a nominal variable.

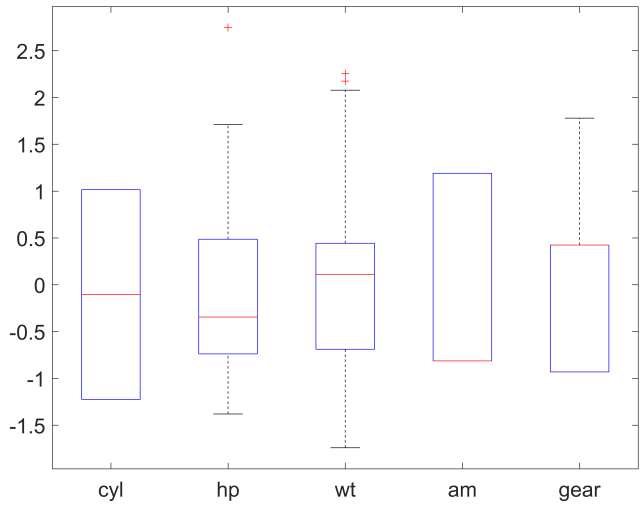

Figure 1: Boxplot of the five attributes x_1–x_5 after standardizing the data (i.e., subtracting the mean of each attribute and dividing the attribute by its standard deviation).

Question 2. In Figure 1 is given a boxplot of the five attributes x_1–x_5 after standardizing the data, i.e. subtracting the mean of each attribute and dividing each attribute by its standard deviation. Which one of the following statements is *correct*?

A. The majority of cars have automatic transmission.

B. The attribute x_5 (i.e., number of forward gears (gear)) appears to be normal distributed.

C. The attribute x_2 (i.e., horse power (hp)) has a clear outlier that should be removed.

D. From the boxplot it is clear that some of the attributes are highly correlated with each other.

E. Don't know.

Solution 2. As there are 32 cars and the median is placed at the lowest value the majority, i.e. at least 17 of cars indeed have to have automatic transmission. The attribute x_5 (gear) does not have a symmetric distribution and indeed seems far from normally distributed. In particular, its 50th and 75th percentile coincide. Even though the box plot indicates an outlier these should not be removed without very strong justification which is not provided here. Boxplots investigate each attribute separately and do not reveal

any aspects in regards to the relationship between attributes, i.e. if they are correlated.

Question 3. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_5, forming the standardized matrix $\tilde{\boldsymbol{X}}$, resulting in the following $\boldsymbol{S}$ and $\boldsymbol{V}$ matrices obtained from a singular value decomposition of $\tilde{\boldsymbol{X}}$:

$$\boldsymbol{S} = \begin{bmatrix} 10.2 & 0 & 0 & 0 & 0 \\ 0 & 6.1 & 0 & 0 & 0 \\ 0 & 0 & 2.8 & 0 & 0 \\ 0 & 0 & 0 & 2.2 & 0 \\ 0 & 0 & 0 & 0 & 1.6 \end{bmatrix},$$

$$\boldsymbol{V} = \begin{bmatrix} 0.49 & -0.31 & 0.42 & -0.14 & 0.69 \\ 0.39 & -0.62 & 0.05 & -0.24 & -0.63 \\ 0.51 & -0.06 & -0.55 & 0.66 & 0.08 \\ -0.44 & -0.46 & 0.42 & 0.65 & -0.02 \\ -0.40 & -0.55 & -0.59 & -0.27 & 0.35 \end{bmatrix}.$$

The data projected onto the first two principal components are given in Figure 2. Which one of the following statements is *correct*?

A. The first principal component accounts for less than 70 % of the variance.

B. The two first principal components account for less than 90 % of the variance.

C. The fifth principal component accounts for less than 1% of the variance.

D. As can be observed in Figure 2 there is a positive correlation between the projection of the data to the first and second principal component.

E. Don't know.

Solution 3. The variance explained by the th principal component is given by $\frac{\sigma_i^2}{\sum_{i'} \sigma_{i'}^2}$. As such we find:

$$VarExpPC1 = \frac{10.2^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.6648$$
$$VarExpPC2 = \frac{6.1^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.2378$$
$$VarExpPC3 = \frac{2.8^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.0501$$
$$VarExpPC4 = \frac{2.2^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.0309$$
$$VarExpPC5 = \frac{1.6^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.0164$$

As such, the first PC accounts for less than 70% of the variance, the first two principal components accounts for $0.6648 + 0.2378 = 0.9026$ which is not less than 90% of the variance. The fifth principal

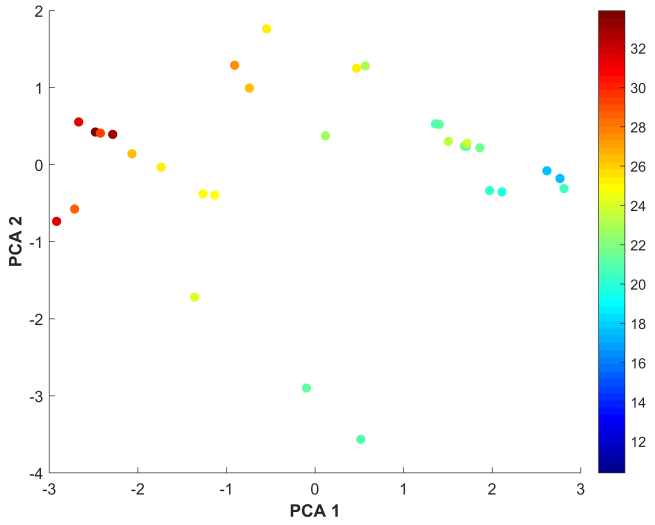

Figure 2: Data projected onto the first and second principal component. Each observation is color coded according to y, i.e. how many miles pr. gallon the car drives.

component accounts for 0.0164% which is not less than 1%. The data represented in the space of the PCA are uncorrelated, i.e. there is no correlation between the data projected onto the first and second principal components as $(\tilde{\boldsymbol{X}}\boldsymbol{v}_1)^\top(\tilde{\boldsymbol{X}}\boldsymbol{v}_2) = (\boldsymbol{U}_1 S_{11})^\top(\boldsymbol{U}_2 S_{22}) = S_{11}\boldsymbol{U}_1^\top \boldsymbol{U}_2 S_{22} = S_{11} 0 S_{22} = 0$.

Question 4. The data projected onto the two first principal components (as defined in Question 3) is given in Figure 2 where the output variable y is indicated by the color of each observation. which one of the following statements pertaining to the PCA is *correct*?

A. Cars with a relatively small number of cylinders, low horsepower, low weight, that have manual transmission, and many forward gears tend to drive longer pr. gallon of fuel.

B. The second principal component appears to provide a better description of how far a car drives pr. gallon of fuel than principal component direction one.

C. From the PCA plot it appears to be very difficult to predict fuel consumption based on the attributes x_1–x_5.

D. Cars with a relatively small number of cylinders, low horsepower, that are automatic, and with few forward gears will have a large negative projection onto the second principal component.

E. Don't know.

Solution 4. As the first three values of $\boldsymbol{v}_1$ are positive and the last two negative relatively small values of the first three attributes, i.e. cylinders, horsepower, and weight and relatively high value of the two last attribute, i.e. manual transmission and many forward gears will result in a negative projection onto the first principal component where most cars driving far pr. gallon of fuel are positioned, thus, this is correct. The second principal component direction does not appear to well characterize how far car drives pr. gallon when compared to the first principal component direction. From the plot it indeed seems feasible to predict fuel consumption. Finally, by inspecting $\boldsymbol{v}_2$ cars with a relatively small number of cylinders, low horsepower, that are automatic and with few forward gears will have a relatively large positive projection onto the second prinicipal component and not the reverse.

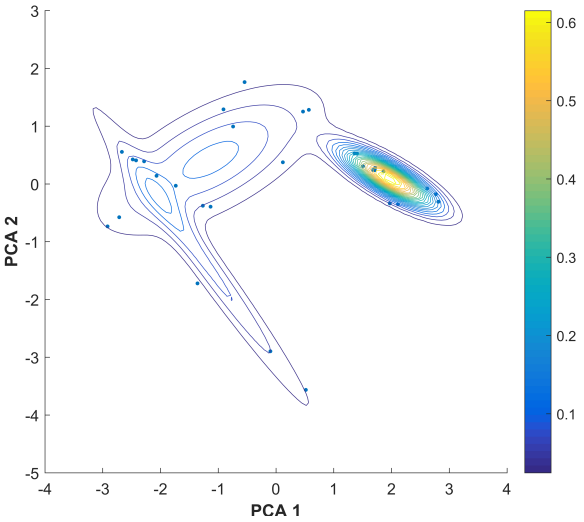

Figure 3: A Gaussian Mixture Model (GMM) with three clusters fitted to the Cars dataset projected onto the first two principal components.

Question 5. Consider the Gaussian Mixture Model (GMM) fitted to the Cars dataset projected onto the first two principal components for which the contours of the fitted GMM is given in Figure 3. We will in the following use:
$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$
to denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Which one of the following GMM densities corresponds to the fitted GMM?

A.

$$p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) + 0.176 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) + 0.375 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix})$$

B.

$$p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) + 0.176 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) + 0.375 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix})$$

C.

$$p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) + 0.176 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) + 0.375 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix})$$

D.

$$p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) + 0.176 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) + 0.375 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix})$$

E. Don't know.

Solution 5. Inspecting the contour plot we see that the cluster located at $\begin{bmatrix} -1.1287 \\ 0.4168 \end{bmatrix}$ has positive covariance, the cluster located at $\begin{bmatrix} -1.2978 \\ -1.2612 \end{bmatrix}$ has a negative covariance with high variance and the cluster located at $\begin{bmatrix} 1.9574 \\ 0.0913 \end{bmatrix}$ a negative covariance with lower variance as the other cluster with negative covariance - in particular this cluster has more spread in the PCA_1 direction than the PCA_2 direction whereas the other negative covariance cluster at $\begin{bmatrix} -1.2978 \\ -1.2612 \end{bmatrix}$ has more spread in the PCA_2 direction than the PCA_1 . This property only holds for the following answer option:

$$p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix})$$
$$+ 0.176 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix})$$
$$+ 0.375 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix})$$

Question 6. A least squares linear regression model is trained using different combinations of the five attributes x_1, x_2, x_3, x_4, and x_5. Table 2 gives the training and test root-mean-square error (RMSE=$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$) performance of the least squares linear regression model when trained using different combinations of the five attributes. Which one of the following statements is *correct*?

A. Forward selection will select the same optimal feature set as backward selection.

B. For this problem backward selection will identify the optimal feature combination.

C. Forward selection will end up selecting all the features, i.e., terminate at the feature set x_1, x_2, x_3, x_4, x_5.

D. Forward selection will as first feature select x_3.

E. Don't know.

Solution 6. Forward selection will first select x_1 with performance 3.0343, then improve most by including x_3 with 2.7692, subsequently include x_2 with performance 2.4869. Including additional features to the set x_1, x_2, x_3 provides no improvements on the test set and the forward selection will thus terminate. Backward selection will first remove x_5 with performance 2.5095, subsequently x_1 with performance 2.3423 thereby identifying the correct optimal feature set being x_2, x_3, x_4.

Feature(s)	Training RMSE	Test RMSE
x_1	3.2522	3.0343
x_2	3.1721	5.5072
x_3	2.907	3.4486
x_4	4.4608	5.8157
x_5	4.4637	9.0155
x_1 and x_2	3.043	3.8668
x_1 and x_3	2.4192	2.7692
x_1 and x_4	2.8918	3.3297
x_1 and x_5	3.2465	3.3177
x_2 and x_3	2.5325	2.5853
x_2 and x_4	2.8066	3.2509
x_2 and x_5	3.1646	4.6723
x_3 and x_4	2.8601	3.7105
x_3 and x_5	2.8230	4.3690
x_4 and x_5	3.9742	8.4346
x_1 and x_2 and x_3	2.4098	2.4869
x_1 and x_2 and x_4	2.7253	2.6258
x_1 and x_2 and x_5	3.0341	4.6252
x_1 and x_3 and x_4	2.3834	2.9486
x_1 and x_3 and x_5	2.3709	2.9676
x_1 and x_4 and x_5	2.7956	3.4894
x_2 and x_3 and x_4	2.4703	2.3423
x_2 and x_3 and x_5	2.5319	2.7017
x_2 and x_4 and x_5	2.7847	4.2531
x_3 and x_4 and x_5	2.8028	4.4864
x_1 and x_2 and x_3 and x_4	2.3679	2.5095
x_1 and x_2 and x_3 and x_5	2.3623	3.0234
x_1 and x_2 and x_4 and x_5	2.6249	4.3059
x_1 and x_3 and x_4 and x_5	2.2937	3.0251
x_2 and x_3 and x_4 and x_5	2.4561	2.8221
x_1 and x_2 and x_3 and x_4 and x_5	2.2759	3.1368

Table 2: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict fuel consumption of a car, i.e., mpg, using different combinations of the five attributes (x_1–x_5).

Question 7. Using the 32 observations of the Cars dataset we would like to predict the fuel consumption of cars (y) based on the five features ($x_1 - x_5$). For this purpose we consider regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function:

$$E(\boldsymbol{w}) = \sum_n (y_n - [1\ x_{n1}\ x_{n2}\ x_{n3}\ x_{n4}\ x_{n5}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w},$$

where x_{nm} denotes the m'th feature of the n'th observation, and 1 is concatenated the data to account for the bias term. We consider nine different values of λ and use leave-one-out cross-validation to quantify the performance of each of these different values of λ. The results of the leave-one-out cross-validation performance is given in Figure 4 where the optimal value of lambda is found to be $\lambda = 10^{-1.75}$ indicated with a black cross in the figure. Which one of the following statements is correct?

A. The identified test performance having RMSE=2.8 is an unbiased estimator of how the optimal model identified will generalize to new data.

B. To create the test error curve given in Figure 4 requires training 288 regularized least squares regression models.

C. Leave-one-out cross-validation is computationally more efficient than 10-fold cross-validation.

D. As λ increases the fitted models will have smaller and smaller bias.

E. Don't know.

Solution 7. The identified test performance is not an unbiased estimator of the optimal identified models generalization. To quantify the generalization performance of the optimally selected model would require two-layer cross-validation. In order to generate the test curve we need for each value of λ (9 values) to fit a model corresponding to each of the 32 observations left out which would require fitting $9 \cdot 32 = 288$ models, thus, this statement is correct. Leave-one-out cross-validation is not more computationally efficient as we have to fit more models than in 10-fold cross-validation and each of these models would also include more data for training. As we increase λ we also increase the models bias, i.e. eventually the model will predict everything as 0.

Question 8. We will build a model to determine if a car has a relatively high or low fuel consumption using the original data (i.e., the data is no longer standardized). For this purpose we will split the output y in two classes forming the new output variable z defined by thresholding at the median value of y, i.e. if $y_i > \text{median}(y)$ then $z_i = 1$, otherwise $z_i = 0$. We fit a logistic regression model using the features x_1–x_5 and use as output for the logistic regression model the binary variable z indicating if a car has relatively high ($z = 0$) or low ($z = 1$) fuel consumption. The predicted output probability is given by:

$$\hat{z} = \sigma(1257.6 - 46.8x_1 + 0.6x_2 - 271.1x_3 - 31.9x_4 - 44.7x_5),$$

where $\sigma(\cdot)$ is the logistic sigmoid function. Which one of the following statements regarding the logistic regression model is correct?

A. x_2 is the least important attribute for defining whether a car has high (i.e., $z = 0$) or low (i.e., $z = 1$) fuel consumption.

B. According to the model increasing a car's number of forward gears will make it more likely to have high fuel consumption (i.e., $z = 0$).

C. A new car with the following observation vector: $\boldsymbol{x}^* = [6\ 120\ 3.2\ 0\ 4]$ will be more likely to have high (i.e., $z = 0$) than low (i.e., $z = 1$) fuel consumption.

D. Logistic regression is not very suited for classification as it is a regression method.

E. Don't know.

Solution 8. We cannot interpret importance with respect to the amplitude of the coefficients as the scale of each attribute is very different and the attributes correlated. As x_5 has a negative coefficient increasing the number of forward gears will decrease the probability of low (i.e., $z = 1$) fuel consumption and thereby increase the probability of high (i.e., $z = 0$) fuel consumption. Thus, this statement is correct. We have for the probability according to the logistic regression model of the observation $\boldsymbol{x}^* = [6\ 120\ 3.2\ 0\ 4]$:

$$\hat{z}^* = \frac{1}{1+\exp(-(1257.6-46.8\cdot 6+0.6\cdot 120-271.1\cdot 3.2-31.9\cdot 0-44.7\cdot 4))} = 0.9227.$$

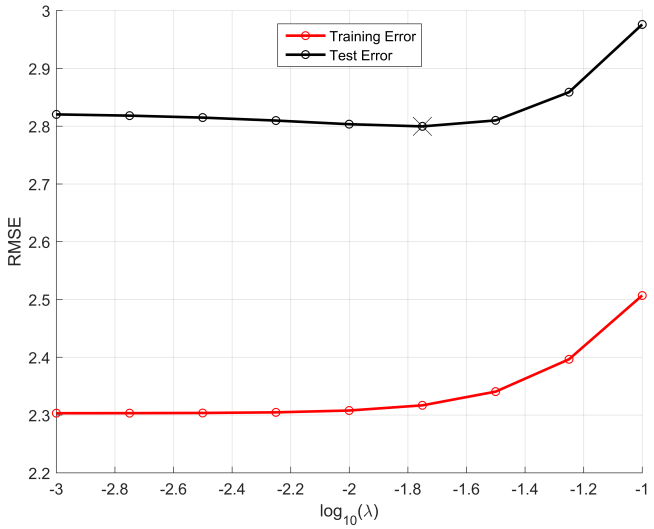

Figure 4: Regularized least squares regression applied to the Cars data in order to predict fuel consumption. Given is training and test performance as quantified by the root mean square error (RMSE) in red and black respectively as a function of the value of λ based on leave-one-out cross-validation.

Thus, this car is likely to have low (i.e, $z = 1$) fuel consumption. Logistic regression is a classification procedure and designed for this purpose.

Question 9. In order to improve the performance of the logistic regression model we will use ensembling based on the Adaboost algorithm using the 32 observations of the Cars dataset (note that for the Adaboost algorithm $\log(\cdot)$ is based on the natural logarithm). The first trained logistic regression classifier (i.e., for boosting round $t = 1$) has an error rate of 1/16. What will be the updated weight for each of the correctly classified observations?

A. 0.0081

B. 0.0167

C. 0.0332

D. 0.2500

E. Don't know.

Solution 9. In the first round (t=1) all samples are weighted equally thus $w_1 = w_2 = \ldots = w_{32} = 1/32$. As a result $\epsilon_1 = \sum_{i=1}^{N} w_i(1 - \delta_{f_t(x_i),y_i}) = \frac{1}{32}\sum_{i=1}^{N}(1 - \delta_{f_t(x_i),y_i}))$ which is the error rate, i.e. number of misclassified observations divided by the

	3 gears ($x_5 = 3$)	4 gears ($x_5 = 4$)	5 gears ($x_5 = 5$)
Low mpg ($z = 0$)	13	2	2
High mpg ($z = 1$)	2	10	3

Table 3: Number of low mpg and high mpg cars (i.e. $z = 0$ and $z = 1$) according to the number of gears, i.e. $x_5 = 3$, $x_5 = 4$, or $x_5 = 5$.

total number of observations (i.e., divided by 32) which is 1/16 as explained in the text. We then have $\alpha_1 = \frac{1}{2}\log(\frac{1-\epsilon_1}{\epsilon_1}) = 0.5\log(15)$. Thus $\tilde{w}_m = \frac{1}{32}exp(0.5\log(15))$ for a mis-classified observation and $\tilde{w}_c = \frac{1}{32}exp(-0.5\log(15))$ for a correctly classified observation. As two observations are misclassified and 30 correctly classified (i.e. the error rate is 1/16) we have $w_c = \frac{\frac{1}{32}exp(-0.5\log(15))}{2\cdot\frac{1}{32}exp(0.5\log(15))+30\cdot\frac{1}{32}exp(-0.5\log(15))} = 0.0167$

Question 10. A decision tree is fitted to the data considering as output whether the car has a relatively high (i.e., $z = 0$) or low (i.e., $z = 1$) fuel consumption. At the root of the tree three different splits according to the number of forward gears are considered based on the data given in Table 3. For impurity we will use the classification error given by $I(v) = 1 - \max_c p(c|v)$. For the three considered splits we have:

- Split A: 3 gear vs. 4 or 5 gears.
- Split B: 3 or 4 gears vs. 5 gears.
- Split C: 3 gears vs. 4 gears vs. 5 gears.

Thus, split A and B have two branches whereas split C has three branches. Which statement about the splits is correct?

A. Split B provides a higher purity gain than split A.

B. Split C provides a higher purity gain than split A.

C. The best obtainable purity gain is 9/32.

D. Split B provides a higher purity gain than split C.

E. Don't know.

Solution 10. The purity gain is given by

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

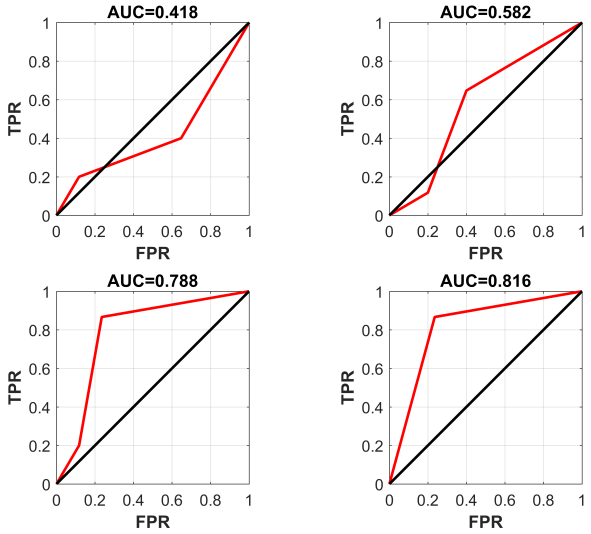

Figure 5: Four different receiver operator characteristic (ROC) curves and their area under curve (AUC) value.

where

$$I(v) = 1 - \max_c p(c|v).$$

Evaluating the purity gain for split A we have:

$$\begin{aligned}\Delta =&(1 - (\tfrac{17}{32}))\\ &-[\tfrac{15}{32}(1 - (\tfrac{13}{15}))\\ &+\tfrac{17}{32}(1 - (\tfrac{13}{17}))]\\ &= \frac{15}{32} - \frac{6}{32} = 9/32.\end{aligned}$$

Evaluating the purity gain for split B we have:

$$\begin{aligned}\Delta =&(1 - (\tfrac{17}{32}))\\ &-[\tfrac{27}{32}(1 - (\tfrac{15}{27}))\\ &+\tfrac{5}{32}(1 - (\tfrac{3}{5}))]\\ &= \frac{15}{32} - \frac{14}{32} = 1/32.\end{aligned}$$

Evaluating the purity gain for split C we have:

$$\begin{aligned}\Delta =&(1 - (\tfrac{17}{32}))\\ &-[\tfrac{15}{32}(1 - (\tfrac{13}{15}))\\ &+\tfrac{12}{32}(1 - (\tfrac{10}{12}))\\ &+\tfrac{5}{32}(1 - (\tfrac{3}{5}))]\\ &= \frac{15}{32} - \frac{6}{32} = 9/32.\end{aligned}$$

Question 11. We will evaluate the feature x_5 (gear) in its ability to discriminate low mpg, i.e., $z = 0$, (considered the negative class) from high mpg, i.e. $z = 1$, (considered the positive class) based on the data given in Table 3. For this purpose, we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature x_5 to discriminate cars with high (i.e., $z = 0$) from cars with low (i.e., $z = 1$) fuel consumption. Which one of the ROC curves given in Figure 5 corresponds to using x_5 to discriminate between high fuel consumption ($z = 0$) and low fuel consumption ($z = 1$)?

A. The curve having AUC=0.418

B. The curve having AUC=0.582

C. The curve having AUC=0.788

D. The curve having AUC=0.816

E. Don't know.

Solution 11. The ROC curve can be calculated by lowering the threshold, as no cars have more than 5 forward gears a threshold above 5 will result in the point (0,0). Lowering the threshold we find at the value 5 that 2/17 of the low mpg cars (FPR) are at 5 and 3/15 of the high mpg cars (TPR) are at 5 corresponding to the point (2/17,3/15). When lowering to a threshold of 4 gears or more we are at the point (4/17,13/15) and at a threshold at 3 gears or more we have (1,1). Thus, this curve corresponds to the curve having AUC=0.788.

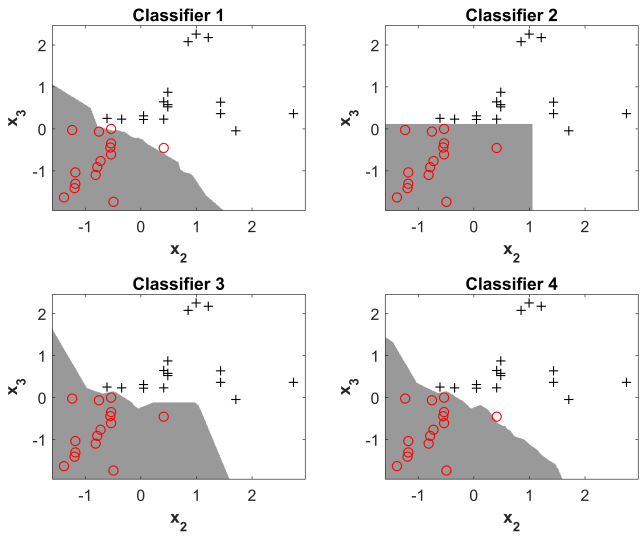

Figure 6: Decision boundaries for four different classifiers trained on the Cars dataset using only the two features x_2 and x_3.

Question 12. Four different classifiers are trained on the Cars dataset using only x_2 and x_3 as features (the features have been standardized) and the decision boundary for each of the four classifiers is given in Figure 6. Which one of the following statements is *correct*?

A. Classifier 1 is an artificial neural network with one hidden unit in the hidden layer.

B. Classifier 2 is a 3-nearest neighbor classifier.

C. Classifier 3 is a 1-nearest neighbor classifier.

D. Classifier 4 is a logistic regression classifier.

E. Don't know.

Solution 12. The decision boundary of classifier 1 cannot be a neural network with one hidden unit as this would correspond to linear decision boundary similar to logistic regression. Classifier 2 has straight vertical and horizontal lines resembling a decision tree and not a 3-nearest neighbor classifier. Classifier 3 is indeed a 1-nearest neighbor classifier as can be seen by the decision boundary following the most close-by observation. Classifier 4 cannot be a logistic regression classifier as this would require a decision boundary formed by a straight line (as for a neural network with one hidden unit).

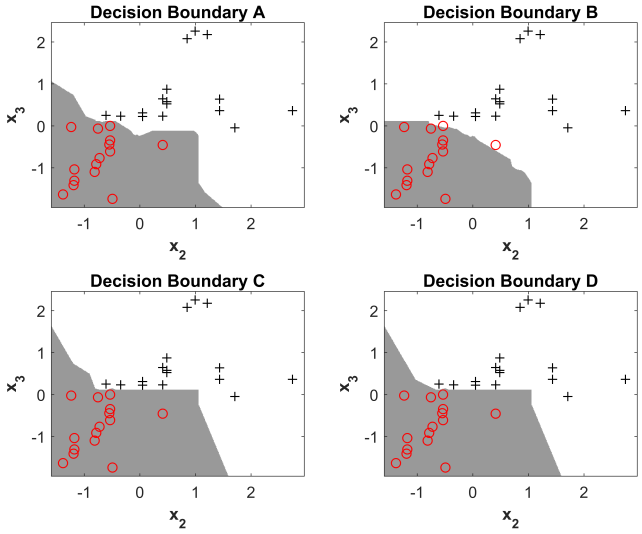

Figure 7: Decision boundaries for which one of the four decision boundaries corresponds to combining Classifier 1, Classifier 2, and Classifier 3 in Figure 6 using majority voting.

Question 13. In an attempt to make a stronger classifier the first three classifiers (i.e., Classifier 1, Classifier 2, and Classifier 3) are combined using majority voting. Which one of the decision boundaries given in Figure 7 corresponds to the combined classifier?

A. Decision boundary A.

B. Decision boundary B.

C. Decision boundary C.

D. Decision boundary D.

E. Don't know.

Solution 13. When combining the three classifiers the majority class is the class receiving two or more votes. Combining the three decision boundaries of the three first classifiers in Figure 6 to form a white region thus requires two of the classifiers being white in that region and vice versa for the gray region. This only holds for decision boundary A.

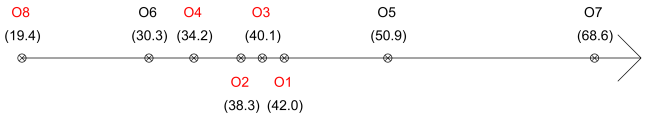

Figure 8: The eight first observations of the Cars dataset considered in regards to the feature $q = x_2/x_3$ (the value of q is given in parenthesis).

Question 14. We will consider the first eight observations of the Cars dataset and the new feature defined by the ratio of horse power to weight, i.e. defined by $q = x_2/x_3$. In Figure 8 is shown the value of q for the first eight observations that are colored according to low (red) and high (black) fuel consumption. We will cluster this data using k-means with Euclidean distance into two clusters (i.e., k=2) and initialize the k-means algorithm with centroids located at observation O8, and O6. Which one of the following statements is *correct*?

A. The converged solution will be {O8},{O1, O2, O3, O4, O5, O6, O7}.

B. The converged solution will be {O1, O2, O3, O4, O6, O8}, {O5, O7}.

C. The converged solution will be {O1, O2, O3, O4, O5, O6, O8}, {O7}.

D. The converged solution will be {O4, O6, O8}, {O1, O2, O3, O5, O7}.

E. Don't know.

Solution 14. With the described initialization, observation O8 will be assigned to the cluster located at O8, and the remaining observation will be assigned to the cluster located at O6, i.e. {O1, O2, O3, O4, O5, O6, O7}. Thus, only cluster located at O6 will change location and the location updated to $\frac{30.3+34.2+38.3+40.1+42.0+50.9+68.6}{7} = 43.5$. For this new location O6 is closer to cluster located at O8 than the cluster located at 43.5, resulting in the updated clustering {O6,O8}, {O1, O2, O3, O4, O5, O7}. Thus, the first cluster will change location to $\frac{19.4+30.3}{2} = 24.85$ and the second cluster to $\frac{34.2+38.3+40.1+42.0+50.9+68.6}{6} = 45.68$. Subsequently, the first cluster will be updated to contain {O4,O6,O8} with centroid at $\frac{19.4+30.3+34.2}{3} = 27.97$ and the second cluster with {O1, O2, O3, O5, O7} will have centroid

located at $\frac{38.3+40.1+42.0+50.9+68.6}{5} = 47.98$ which will form a converged solution as no observation change assignment.

Question 15. We will consider a clustering given by the two clusters {O2, O3, O4, O6, O8},{O1, O5, O7}. We will evaluate this clustering in terms of its correspondence with the class label information in which O1,O2,O3,O4, and O8 correspond to cars with low fuel consumption and O5, O6, and O7 correspond to cars with high fuel consumption. We recall that the Jaccard coefficient between the true labels and the extracted clusters is given by:

$$J = \frac{f_{11}}{K - f_{00}},$$

where f_{11} is the number of object pairs in same class assigned to same cluster, f_{00} is the number of object pairs in different class assigned to different clusters, and $K = N(N-1)/2$ is the total number of object pairs, where N is the number of observations considered. What is the above value of J between the true labeling of the observations in terms of high and low fuel consumption and the two clusters?

A. 0.1665

B. 0.3684

C. 0.5714

D. 0.7500

E. Don't know.

Solution 15. The cluster indices are given by the vector: $[2\ 1\ 1\ 1\ 2\ 1\ 2\ 1]^\top$, whereas the true class labels are given by the vector $[1\ 1\ 1\ 1\ 2\ 2\ 2\ 1]^\top$. From this, we obtain: Total number of object pairs is: $K = 8(8-1)/2 = 28$
$f_{00} = 4 \cdot 2 + 1 \cdot 1 = 9$
$f_{11} = 4{\cdot}(4{-}1)/2{+}1{\cdot}(1{-}1)/1{+}1{\cdot}(1{-}1)/2{+}2{\cdot}(2{-}1)/2 = 7$
$J = \frac{f_{11}}{K-f_{00}} = \frac{7}{28-9} = 0.3684$

Question 16. According to the Cars dataset we have that 59.38% of the cars have automatic transmission. Furthermore, 63.16% of the cars that have automatic transmission have eight cylinders, whereas 15.38% of the cars that have manual transmission have eight cylinders. According to the Cars dataset what is the probability that a car that has eight cylinders, i.e. $x_1 = 8$ will have automatic transmission, i.e. $x_4 = 0$?

A. 16.7 %

B. 37.5 %

C. 63.2%

D. 85.7 %

E. Don't know.

Solution 16. According to Bayes' theorem we have:

$$P(x_4 = 0|x_1 = 8) = \frac{P(x_1{=}8|x_4{=}0)P(x_4{=}0)}{P(x_1{=}8)}$$
$$= \frac{P(x_1{=}8|x_4{=}0)P(x_4{=}0)}{P(x_1{=}8|x_4{=}0)P(x_4{=}0)+P(x_1{=}8|x_4{=}1)P(x_4{=}1)}$$
$$= \frac{0.6316{\cdot}0.5938}{0.6316{\cdot}0.5938+0.1538{\cdot}(1-0.5938)} = 85.7\%$$

Question 17. We will consider an artificial neural network (ANN) trained to predict mpg (i.e., y) of a car. The ANN is based on the model:

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

where $h^{(1)}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the hyperbolic tangent function used as activation function in the hidden layer. We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -4 \\ 1 \\ 0.01 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} -10 \\ 1 \\ -0.02 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

and $w_0^{(2)} = 7$, $w_1^{(2)} = 8$, and $w_2^{(2)} = 9$.
What is the predicted fuel consumption of a car with observation vector $\boldsymbol{x}^* = [6\ 120\ 3.2\ 0\ 4]$?

A. 17.00

B. 20.85

C. 24.00

D. 33.40

E. Don't know.

Solution 17. The output is given by:

$$8 \cdot tanh(-4 + 1 \cdot 6 + 0.01 \cdot 120 + 1 \cdot 3.2 - 1 \cdot 0 - 1 \cdot 4)$$
$$+9 \cdot tanh(-10 + 1 \cdot 6 - 0.02 \cdot 120 + 1 \cdot 3.2 + 1 \cdot 0 + 1 \cdot 4)$$
$$+7 = 20.85$$

	hp_L	hp_H	wt_L	wt_H	am=0	am=1
O1	1	0	1	0	0	1
O2	1	0	1	0	0	1
O3	1	0	1	0	0	1
O4	1	0	1	0	1	0
O5	0	1	0	1	1	0
O6	1	0	0	1	1	0
O7	0	1	0	1	1	0
O8	1	0	1	0	1	0

Table 4: The first eight observations of the Cars dataset binarized considering the attribute x_2, x_3, and x_4 such that x_2 is split according to the median value in terms of low and high horse power (i.e. hp_L and hp_H), low and high weight (i.e. wt_L and wt_H), as well as whether the car has automatic (i.e., am=0) or manual (i.e., am=1) transmission. The eight observations are color coded in terms of low fuel consumption {O1, O2, O3, O4, O8} and high fuel consumption {O5, O6, O7}.

Question 18. Considering the dataset in Table 4 as a market basket problem with observation O1–O8 corresponding to customers and hp_L, hp_H ,wt_L, wt_H, am=0, am=1 corresponding to items, what is the confidence of the association rule $\{\text{wt}_H, \text{am=0}\} \rightarrow \{\text{hp}_H\}$?

A. 0.0%

B. 25.0

C. 66.7%

D. 100.0 %

E. Don't know.

Solution 18. The confidence is given as

$$P(\text{hp}_H = 1 | \text{wt}_H = 1, am = 0) = \frac{P(\text{hp}_H = 1, \text{wt}_H = 1, am = 0)}{P(\text{wt}_H = 1, am = 0)} = \frac{2/8}{3/8} = 2/3 = 66.7\%$$

Question 19. We will again consider the data in Table 4. What are all frequent itemsets with support greater than 30%?

A. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}.

B. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}.

C. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}, {hp$_L$, wt$_L$, am=1},

D. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}, {hp$_L$, wt$_L$, am=1}, {hp$_L$, wt$_L$, am=0},

E. Don't know.

Solution 19. For a set to have support more than 30% the set must occur at least $0.3 \cdot 8 = 2.4$, i.e. 3 out of the 8 times. All the itemsets that have this property are {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}, {hp$_L$, wt$_L$, am=1}.

Question 20. Considering the data in Table 4, we will calculate the similarity as well as distance between O1 given by the vector $\boldsymbol{a} = [1\ 0\ 1\ 0\ 0\ 1]$ and O4 given by the vector $\boldsymbol{b} = [1\ 0\ 1\ 0\ 1\ 0]$ using respectively the Jaccard (J), simple matching coefficient (SMC), cosine similarity (cos) and p-norm ($\|\cdot\|_p$) given by

$$J(\boldsymbol{a},\boldsymbol{b}) = \frac{f_{11}}{M - f_{00}},$$
$$SMC(\boldsymbol{a},\boldsymbol{b}) = \frac{f_{11} + f_{00}}{M},$$
$$cos(\boldsymbol{a},\boldsymbol{b}) = \frac{f_{11}}{\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2},$$
$$\|\boldsymbol{a} - \boldsymbol{b}\|_p = (\sum_{m=1}^{M} |a_m - b_m|^p)^{1/p}.$$

Which one of the following statements is correct?

A. $\|\boldsymbol{a} - \boldsymbol{b}\|_2 = 2$.

B. $\|\boldsymbol{a} - \boldsymbol{b}\|_1 < \|\boldsymbol{a} - \boldsymbol{b}\|_2$.

C. $J(\boldsymbol{a},\boldsymbol{b}) = 2/3$

D. $cos(\boldsymbol{a},\boldsymbol{b}) = SMC(\boldsymbol{a},\boldsymbol{b})$.

E. Don't know.

Solution 20. For $\boldsymbol{a}$ and ,$\boldsymbol{b}$ we have:

$$\|\boldsymbol{a} - \boldsymbol{b}\|_2 = \sqrt{0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2} = \sqrt{2},$$
$$\|\boldsymbol{a} - \boldsymbol{b}\|_1 = 0 + 0 + 0 + 0 + 1 + 1 = 2,$$
$$J(\boldsymbol{a},\boldsymbol{b}) = \frac{f_{11}}{M - f_{00}} = 2/(6-2) = 1/2,$$
$$SMC(\boldsymbol{a},\boldsymbol{b}) = \frac{f_{11} + f_{00}}{M} = 4/6 = 2/3,$$
$$cos(\boldsymbol{a},\boldsymbol{b}) = \frac{f_{11}}{\|\boldsymbol{r}\|_2 \|\boldsymbol{s}\|_2} = 2/(\sqrt{3}\sqrt{3}) = 2/3.$$

Hence, $cos(\boldsymbol{a},\boldsymbol{b}) = SMC(\boldsymbol{a},\boldsymbol{b})$ is correct.

	O1	O2	O3	O4	O5	O6	O7	O8
O1	0	0.2606	1.1873	2.4946	2.9510	2.5682	3.4535	2.4698
O2	0.2606	0	1.2796	2.4442	2.8878	2.4932	3.3895	2.4216
O3	1.1873	1.2796	0	2.8294	3.6892	2.9147	4.1733	2.2386
O4	2.4946	2.4442	2.8294	0	1.4852	0.2608	2.2941	1.8926
O5	2.9510	2.8878	3.6892	1.4852	0	1.5155	1.0296	3.1040
O6	2.5682	2.4932	2.9147	0.2608	1.5155	0	2.3316	1.8870
O7	3.4535	3.3895	4.1733	2.2941	1.0296	2.3316	0	3.7588
O8	2.4698	2.4216	2.2386	1.8926	3.1040	1.8870	3.7588	0

Table 5: Pairwise Euclidean distance between the first eight observations in the Cars dataset. Red observations (i.e., O1, O2, O3, O4, O8) are observations corresponding to low fuel consumption, whereas black observations (i.e., O5, O6, O7) are observations with high fuel consumption.

Question 21. We would like to predict whether a car has high fuel or low fuel consumption using the data in Table 4. We will apply a Naïve Bayes classifier that assumes independence between the attributes given the class label (i.e. high or low fuel consumption corresponding to observation indicated in black and red respectively in the table). Given that a car has $\text{hp}_L = 1$, and am=0 what is the probability that the car will have high fuel consumption according to the Naïve Bayes classifier derived from the data in Table 4?

A. 3/80

B. 1/8

C. 1/3

D. 3/5

E. Don't know.

Solution 21. We will let high fuel consumption be denoted by $z = 0$ and low fuel consumption by $z = 1$. According to the Naïve Bayes classifier we have

$$P(z=0|\text{hp}_L=1, am=0) = \frac{\begin{pmatrix} P(\text{hp}_L=1|z=0)\times \\ P(am=0|z=0)\times \\ P(z=0) \end{pmatrix}}{\begin{pmatrix} P(\text{hp}_L=1|z=0)\times \\ P(am=0|z=0)\times \\ P(z=0) \end{pmatrix} + \begin{pmatrix} P(\text{hp}_L=1|z=1)\times \\ P(am=0|z=1)\times \\ P(z=1) \end{pmatrix}}$$
$$= \frac{1/3 \cdot 3/3 \cdot 3/8}{1/3 \cdot 3/3 \cdot 3/8 + 5/5 \cdot 2/5 \cdot 5/8} = \frac{1/8}{1/8 + 2/8} = 1/3.$$

Question 22. To determine whether the fuel consumption of a car is high or low we will use a k-nearest neighbor (KNN) classifier to predict each of the eight observations based on the Euclidean distance between the observations given in Table 5. We will use leave-one-out cross-validation for the KNN in order to classify the eight considered observations using a one-nearest neighbor classifier, i.e. $K = 1$. The analysis will be based only on the data given in Table 5. Which one of the following statements is *correct*?

A. None of the observations will be misclassified.

B. One of the observations will be misclassified.

C. Two the observations will be misclassified.

D. Three of the observations will be misclassified.

E. Don't know.

Solution 22. $N(O1, 1) = \{O2\}$ as O2 is closest it will be correctly classified as having low fuel consumption.
$N(O2, 1) = \{O1\}$ as O1 is closest it will be correctly classified as having low fuel consumption.
$N(O3, 1) = \{O1\}$ as O1 is closest it will be correctly classified as having low fuel consumption.
$N(O4, 1) = \{O6\}$ as O6 is closest it will be incorrectly classified as having high fuel consumption.
$N(O5, 1) = \{O7\}$ as O7 is closest it will be correctly classified as having high fuel consumption.
$N(O6, 1) = \{O4\}$ as O4 is closest it will be incorrectly classified as having low fuel consumption.
$N(O7, 1) = \{O5\}$ as O5 is closest it will be correctly classified as having high fuel consumption.
$N(O8, 1) = \{O6\}$ as O6 is closest it will be incorrectly classified as having high fuel consumption.
Thus, three out of the eight observations will be misclassified.

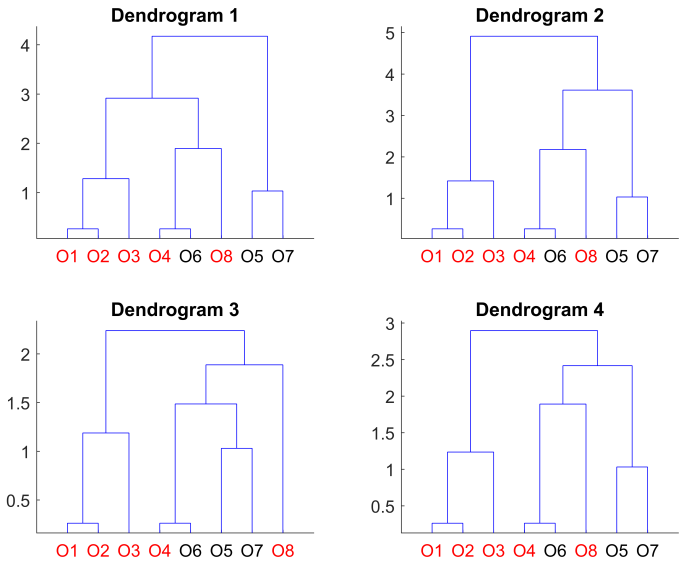

Figure 9: Four different dendrograms derived from the distances between the eight first cars given in Table 5.

Question 23. In Table 5 is given the pairwise Euclidean distances between the first eight observations of the Cars data. A hierarchical clustering is used to cluster these observations using average linkage. Which one of the dendrograms given in Figure 9 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 23. Using average linkage clusters are merged according to their average distance between the observations from each cluster. The dendrogram grows by first merging O1 and O2 at 0.2606, then O4 and O6 at 0.2608, then O5 and O7 at 1.0296, then {O1,O2} with O3 at (1.1873+1.2796)/2=1.2334, then {O4,O6} with O8 at (1.8926+1.8870)/2= 1.8898. Subsequently {O5, O7} merge with {O4,O6,O8} at (1.4852+2.2941+1.5155+ 2.3316+ 3.1040+ 3.7588)/6=2.4149, only dendrogram 4 has this property.

Question 24. We suspect that observation O8 may be an outlier. In order to assess if this is the case we would like to calculate the average relative KNN density based on the observations given in Table 5 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. Based on the data in Table 5, what is the average relative density for observation O8 for $K = 2$ nearest neighbors?

A. 0.4660

B. 0.4800

C. 0.5292

D. 1.8898

E. Don't know.

Solution 24.

$$\text{density}(\boldsymbol{x}_{O8}, 2) = (\frac{1}{2}(1.8870 + 1.8926))^{-1} = 0.5292$$

$$\text{density}(\boldsymbol{x}_{O6}, 2) = (\frac{1}{2}(0.2608 + 1.5155))^{-1} = 1.1259$$

$$\text{density}(\boldsymbol{x}_{O4}, 2) = (\frac{1}{2}(1.4852 + 0.2608))^{-1} = 1.1455$$

$$\text{a.r.d.}(\boldsymbol{x}_{O8}, 2) = \frac{\text{density}(\boldsymbol{x}_{O8}, 2)}{\frac{1}{2}(\text{density}(\boldsymbol{x}_{O6}, 2) + \text{density}(\boldsymbol{x}_{O4}, 2))} = \frac{0.5292}{\frac{1}{2}(1.1259 + 1.1455)} = 0.4660$$

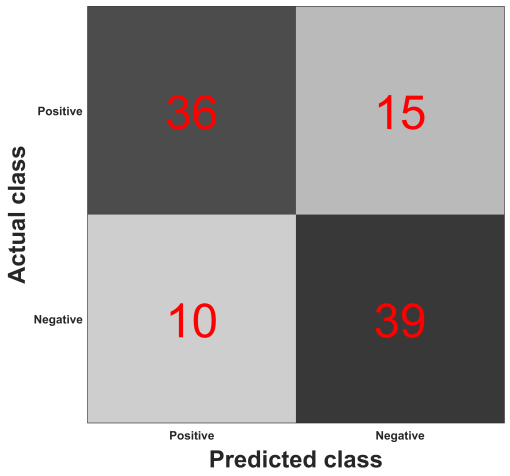

Figure 10: Confusion matrix of a classifier discriminating between 100 positive and negative test observations.

Question 25. We will consider a classifier classifying a dataset with 100 test observations into two classes (positive and negative) with confusion matrix given in Figure 10. Which statement regarding the classifier is correct?

A. The error rate of the classifier is 33.3 %.

B. The precision of the classifier is 75.0 %.

C. The recall of the classifier is 70.6 %.

D. There are more negative than positive examples in the test set.

E. Don't know.

Solution 25. The error rate of the classifier is (FP+FN)/(TP+FP+TN+FN)=(10+15)/100=25.0%. The Precision of the classifier is TP/(TP+FP)=36/(36+10)=78.3%. The Recall of the classifier is TP/(TP+FN)=36/(36+15)=70.6%. There are 51 positive examples and 49 negative examples in the test set.

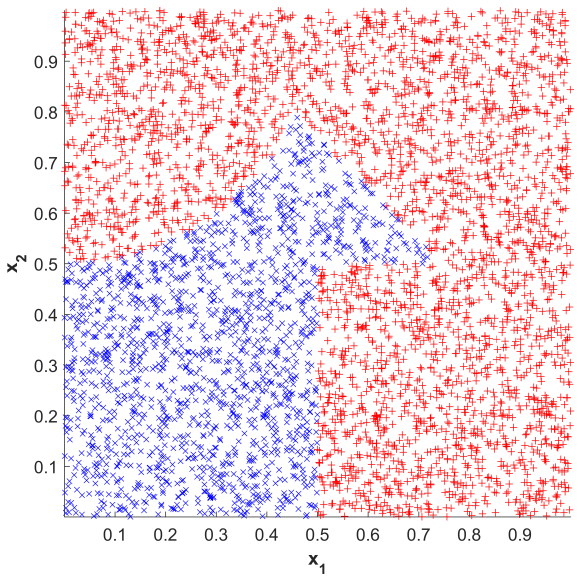

Figure 11: A two class classification problem with red plusses (i.e., +) and blue crosses (i.e., x) constituting the two classes.

Question 26. We will consider the two class classification problem given in Figure 11 in which the goal is to separate red plusses (i.e., +) from blue crosses (i.e., x). Which one of the following procedures will perfectly separate the two classes?

A. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_2 > 0.5$ **and** $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_\infty > 0.5$ **and** $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_1 > 0.75$ **then blue cross, otherwise red plus.**

B. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_\infty > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_2 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_1 > 0.75$ then blue cross, otherwise red plus.

C. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_2 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_1 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_\infty > 0.75$ then blue cross, otherwise red plus.

D. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_\infty > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_1 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_\infty > 0.75$ then blue cross, otherwise red plus.

E. Don't know.

Solution 26. The blue crosses are more than 0.5 in radius from the point $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Furthermore, they are more than 0.5 using the infinite norm (forming a box) from $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ (and more than 0.75 from $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ using the 1-norm). Thus, the solution is given by: $\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 1 \end{bmatrix}\|_2 > 0.5$ and $\|\boldsymbol{x} - \begin{bmatrix} 1 \\ 0 \end{bmatrix}\|_\infty > 0.5$ and $\|\boldsymbol{x} - \begin{bmatrix} 1 \\ 1 \end{bmatrix}\|_1 > 0.75$ then blue cross, otherwise red plus.

Question 27. Which one of the following statements is correct?

A. Unsupervised learning differs from supervised learning in that unsupervised learning both uses the input data and the outputs for training whereas supervised learning only uses the input data.

B. When using Gaussian Mixture Models (GMM) for outlier detection it is important that the observations evaluated for being outliers are included in the training of the GMM.

C. When training an artificial neural network for a dataset with very few observations it is important to include many hidden units in order to avoid overfitting.

D. Cross-validation can both be used for supervised and unsupervised learning.

E. Don't know.

Solution 27. Unsupervised learning differs from supervised learning in not having access to the output data y and not the reverse. It is important when evaluating outliers using a GMM to not include these data in the fitted density as the model may otherwise fit the density to the outliers and thereby not adequately identify these observations as outliers when they are included in defining the density. When having few observations an artificial neural network is very prone to overfitting if many hidden units are included and not the reverse. Indeed cross-validation can be used both for supervised and unsupervised learning. We used extensively cross-validation for supervised learning, i.e. classification and regression - however, we also used cross-validation in unsupervised learning in order to determine the number of clusters in a GMM and to quantify the kernel width in kernel density estimation.

Technical University of Denmark

Written examination: 19 December 2017, 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.

You must either use the electronic file or the form on this page to hand in your answers but not both. **We strongly encourage that you hand in your answers digitally using the electronic file.** If you hand in using the form on this page, please write your name and student number clearly.

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

No.	Attribute description	Abbrev.
x_1	Height (in feet)	Height
x_2	Weight (in pounds)	Weight
x_3	Percent of successful field goals (out of 100 attempted)	FG
x_4	Percent of successful free throws (out of 100 attempted)	FT
y	average points scored per game	PT

Table 1: The attributes of the Basketball dataset contains 54 observations of basketball players in terms of their height, weight and performance. The output y provides each player's average points scored per game.

Question 1. We will consider a basketball dataset containing 54 National Basketball Association (NBA) basketball players and their performance[1]. For brevity this dataset will be denoted the Basketball dataset in the following. In Table 1 the attributes of the data as well as the output attribute y defined by each player's average points scored per game are given. In Figure 1 is given a boxplot of the four attributes x_1–x_4 after standardizing the data, i.e. subtracting the mean of each attribute and dividing each attribute by its standard deviation. The worst performing player (i.e., the player with lowest value of y) is indicated by a black circle with an asteric inside. Considering the attributes described in Table 1 and the boxplot in Figure 1 which one of the following statements regarding the attributes x_1–x_4 and output variable y is correct?

A. The player with worst performance has Weight (i.e., x_2) that is between the 25th and 50th percentile.

B. The player with worst performance has FT (i.e., x_4) so low it is deemed an outlier.

C. The input attributes FT and FG (i.e., x_3 and x_4) are both ordinal variables.

D. The output y is ratio.

E. Don't know.

Solution 1. Inspecting the boxplot we observe that the value of x_2 is within the 50th and 75th percentiles.

[1]The dataset is taken from http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html

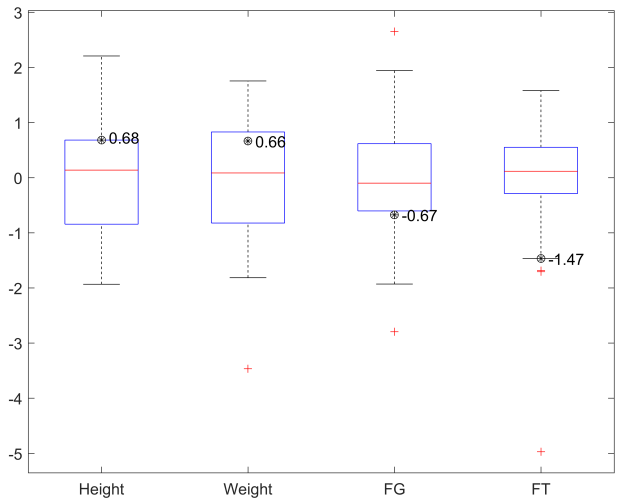

Figure 1: Boxplot of the four attributes x_1–x_4 after standardizing the data (i.e., subtracting the mean of each attribute and dividing each attribute by its standard deviation). The worst performing player (i.e., the player with lowest output value y) is indicated by a black circle with and asteric inside on top of each boxplot along with his standardized value of each attribute given as black digits.

The value of x_4 is positioned exactly at the smallest value for which the observation is still within the 25th percentile subtracted 1.5 times the interquartile range and therefore not an outlier as the lower whisker extends to this observation. In order to be ratio zero has to mean absence of what is being measured. As 0 percent of successful field goals and free throws implies absence of having scored these are ratio attributes and not just ordinal. We can here also talk about 25 % being half as frequent as 50 % etc. As zero average points scored per game implies absence of scoring this output variable is also ratio.

Question 2. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_4, forming the standardized matrix $\tilde{\boldsymbol{X}}$. The squared Frobenious norm of the standardized matrix is given by $\|\tilde{\boldsymbol{X}}\|_F^2 = 212$. A singular value decomposition is applied to the matrix $\tilde{\boldsymbol{X}}$ and we find that the first three singular values are $\sigma_1 = 11.1$, $\sigma_2 = 7.2$, $\sigma_3 = 5.2$. What is the value of the fourth singular value σ_4?

A. $\sigma_4 = 1.2$

B. $\sigma_4 = 2.3$

C. $\sigma_4 = 3.1$

D. $\sigma_4 = 9.9$

E. Don't know.

Solution 2. The variance explained by the i^{th} principal component is given by $\frac{\sigma_i^2}{\sum_{i'} \sigma_{i'}^2} = \frac{\sigma_i^2}{\|\tilde{\boldsymbol{X}}\|_F^2}$. Thus, $\sum_{i'} \sigma_{i'}^2 = \|\tilde{\boldsymbol{X}}\|_F^2$ and from this we know that $\sigma_4 = \sqrt{\|\tilde{\boldsymbol{X}}\|_F^2 - \sum_{i'=1}^3 \sigma_{i'}^2} = \sqrt{212 - 11.1^2 - 7.2^2 - 5.2^2} = 3.1$.

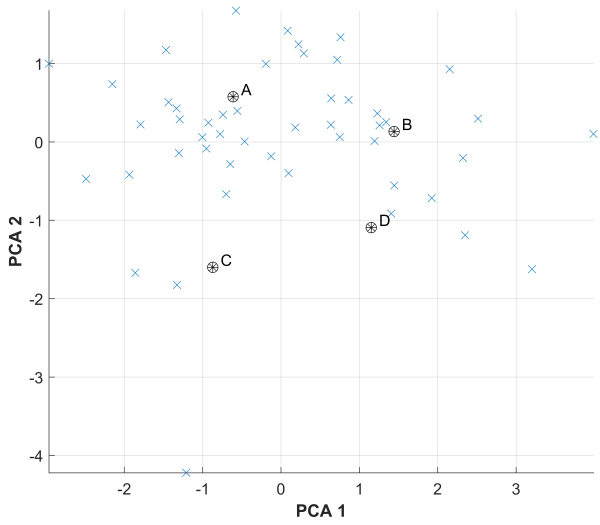

Figure 2: The Basketball data projected onto the first and second principal component. In the plot four observations are highlighted denoted A, B, C, and D of which one of these observation corresponds to the player with lowest output value y given by the circle with and asteric inside in Figure 1.

Question 3. From the singular value decomposition of $\tilde{\boldsymbol{X}}$ we further obtain the following $\boldsymbol{V}$ matrix:

$$\boldsymbol{V} = \begin{bmatrix} -0.60 & 0.02 & -0.41 & 0.69 \\ -0.61 & 0 & -0.33 & -0.72 \\ -0.46 & 0.46 & 0.76 & 0.04 \\ 0.25 & 0.89 & -0.39 & -0.04 \end{bmatrix}.$$

The data projected onto the first two principal components is given in Figure 2 including four observations denoted A, B, C, and D that are marked by a black circle with an asteric inside. One of these four observations corresponds to the worst performing player indicated by a similar black circle with an asteric inside in the boxplot of Figure 1. Which one of the four observations in Figure 2 corresponds to the worst performing player indicated in the boxplots of Figure 1?

A. Observation A.

B. Observation B.

C. Observation C.

D. Observation D.

E. Don't know.

Solution 3. From the boxplot we know that the worst performing player in terms of the output value y has

the standardized observation vector $\tilde{x}^* = [0.68\ 0.66\ -0.67\ -1.47]$ and will thus have the projection onto the two first principal components given by

$$[0.68\ 0.66\ -0.67\ -1.47] \begin{bmatrix} -0.60 & 0.02 \\ -0.61 & 0 \\ -0.46 & 0.46 \\ 0.25 & 0.89 \end{bmatrix} = [-0.8699 - 1.6029].$$

Thus, the observation will in the projection be located at (-0.8699 , -1.6029) which corresponds to the observation denoted C.

Question 4. A least squares linear regression model is trained using different combinations of the four attributes x_1, x_2, x_3,and x_4 in order to predict the average points scored per game y. Table 2 provides the training and test root-mean-square error (RMSE=$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$) performance of the least squares linear regression model when trained using different combinations of the four attributes. Which one of the following statements is correct?

A. Forward selection will terminate when all features x_1–x_4 are included in the feature set.

B. The solution identified by Forward selection will be worse than the solution identified by Backward selection.

C. Forward selection will terminate using two feature in the feature set.

D. Backward selection will terminate using two features in the feature set.

E. Don't know.

Solution 4. Forward selection will select x_4 with performance 5.6845, including additional features will not improve performance and the procedure will therefore terminate with the feature set x_4. Backward selection will remove feature x_2 terminating at the feature set x_1 and x_3 and x_4 with performance 5.5099 since removing additional features only increases the test error.

Feature(s)	Training RMSE	Test RMSE
No features	5.8977	5.8505
x_1	5.8760	6.0035
x_2	5.8841	5.9037
x_3	5.1832	5.9272
x_4	5.8727	5.6845
x_1 and x_2	5.6272	7.4558
x_1 and x_3	5.1482	5.6409
x_1 and x_4	5.8451	5.8269
x_2 and x_3	5.0483	5.6656
x_2 and x_4	5.8660	5.7461
x_3 and x_4	5.1125	5.7390
x_1 and x_2 and x_3	4.9836	6.2823
x_1 and x_2 and x_4	5.6261	7.3888
x_1 and x_3 and x_4	5.0839	5.5099
x_2 and x_3 and x_4	5.0113	5.5605
x_1 and x_2 and x_3 and x_4	4.9645	6.0892

Table 2: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict average points scored per game y using different combinations of the four attributes (x_1–x_4).

Question 5. We will encode the output attribute y in terms of three different classes, i.e. low performing players having performance below the 33.3 percentile, mid-performing players having performance in the range of the 33.3 percentile to 66.6 percentile and high-performing players having performance above the 66.6 percentile (i.e, each of the three classes will contain approximately one third of the data). We presently consider only the attributes FG and FT. Consider the three Gaussian distributions given in Figure 3 where each Gaussian is fitted to each of the three classes separately. We recall that the multivariate Gaussian distribution is given by:
$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The three fitted covariances (in arbitrary order) are given by $\boldsymbol{\Sigma}_a = \begin{bmatrix} 0.0035 & 0.0003 \\ 0.0003 & 0.0030 \end{bmatrix}$, $\boldsymbol{\Sigma}_b = \begin{bmatrix} 0.0028 & -0.0013 \\ -0.0013 & 0.0191 \end{bmatrix}$, and $\boldsymbol{\Sigma}_c = \begin{bmatrix} 0.0020 & 0.0001 \\ 0.0001 & 0.0061 \end{bmatrix}$. The Gaussians are plotted in Figure 3 in terms of lines indicating where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 5$, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 10$, and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 20$ thus defining ellipsoidal shapes at three different levels of the density functions. Which of the three classes correspond to which of the three covariance matrices $\boldsymbol{\Sigma}_a$, $\boldsymbol{\Sigma}_b$, and $\boldsymbol{\Sigma}_c$?

A. $\boldsymbol{\Sigma}_a$ corresponds to the low performing class, $\boldsymbol{\Sigma}_b$ corresponds to the mid performing class, $\boldsymbol{\Sigma}_c$ corresponds to the high performing class.

B. $\boldsymbol{\Sigma}_a$ corresponds to the low performing class, $\boldsymbol{\Sigma}_c$ corresponds to the mid performing class,

Solution 5.

Inspecting the covariance matrices indicated by the iso-line contours of each gaussian at $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 5$, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 10$, and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 20$ we observe that the high performing (i.e., blue class) has negative covariance which is only the case for $\boldsymbol{\Sigma}_b$ Of the low performing and mid-performing classes we observe that the low performing has a larger variance in the second dimension (i.e., the x_4 direction) than the mid-performing, thus, as $\boldsymbol{\Sigma}_c(2,2) = 0.0061 > 0.0030 = \boldsymbol{\Sigma}_a(2,2)$ we have that $\boldsymbol{\Sigma}_c$ corresponds to the low performing and $\boldsymbol{\Sigma}_a$ to the mid performing classes.

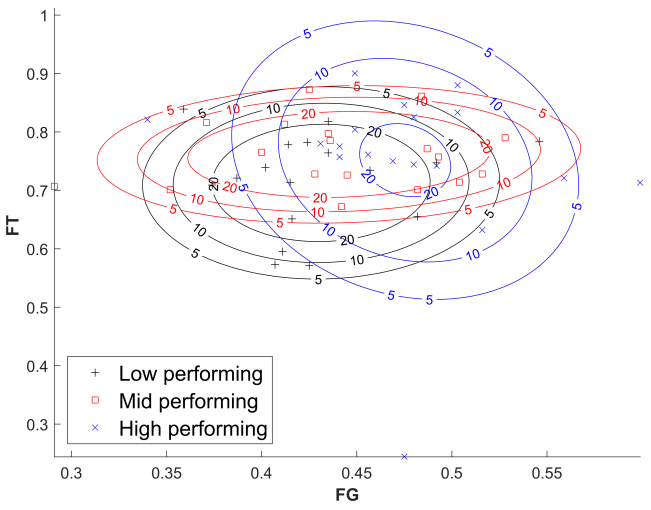

Figure 3: The 54 observations plotted in terms of percentage of successful field goals (FG) plotted against percentage of successful free throws (FT) for low, mid and high performing players respectively indicated by black plusses, red squares, and blue crosses. For each of these three classes a multivariate Gaussian distribution is fitted and the lines corresponding to density values of 5, 10, and 20 are plotted in black, red, and blue respectively.

Question 6. A decision tree is fitted to the data considering as output whether the basketball player was in the group performing low, mid, or high according to splitting the output value in terms of the 33.3 and 66.6 percentile as explained in the previous question. At the root of the tree it is considered to split according to Height (i.e., x_1), considering relatively short, medium, and tall players based on splitting x_1 also according to its 33.3 and 66.6 percentiles. For impurity we will use the Gini given by $I(v) = 1 - \sum_c p(c|v)^2$. Before the split, we have 18 low, 18 mid, and 18 high performing players and after the split we have

- Of the 18 short players we have that 6 have low, 9 have mid, and 3 have high performance.
- Of the 20 medium height players we have that 4 have low, 6 have mid, and 10 have high performance.
- Of the 16 tall players we have that 8 have low, 3 have mid, and 5 have high performance.

Which statement regarding the purity gain Δ of the split is correct?

A. $\Delta = 0.0505$

B. $\Delta = 0.1667$

C. $\Delta = 0.3333$

D. $\Delta = 0.6667$

E. Don't know.

Solution 6. The purity gain is given by

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

where

$$I(v) = 1 - \sum_c p(c|v)^2.$$

Evaluating the purity gain for the split we have:

$$\begin{aligned}\Delta =&(1 - ((\tfrac{18}{54})^2 + (\tfrac{18}{54})^2 + (\tfrac{18}{54})^2))\\ &-[\tfrac{18}{54}(1 - ((\tfrac{6}{18})^2 + (\tfrac{9}{18})^2 + (\tfrac{3}{18})^2))\\ &+\tfrac{20}{54}(1 - ((\tfrac{4}{20})^2 + (\tfrac{6}{20})^2 + (\tfrac{10}{20})^2))\\ &+\tfrac{16}{54}(1 - ((\tfrac{8}{16})^2 + (\tfrac{3}{16})^2 + (\tfrac{5}{16})^2))]\\ &= 0.0505\end{aligned}$$

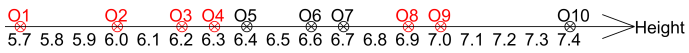

Figure 4: Considering the attribute Height for 10 observations in the Basketball data we inspect whether each of the 10 players here denoted O1,O2,...,O10 have a relatively high percentage of successful field goals (FG>45%) indicated in black and considered the positive class, i.e. observation O5,O6,O7, and O10 or a relatively low percentage of successful field goals (FG≤45%) indicated in red and considered the negative class, i.e. observations O1,O2,O3,O4,O8, and O9.

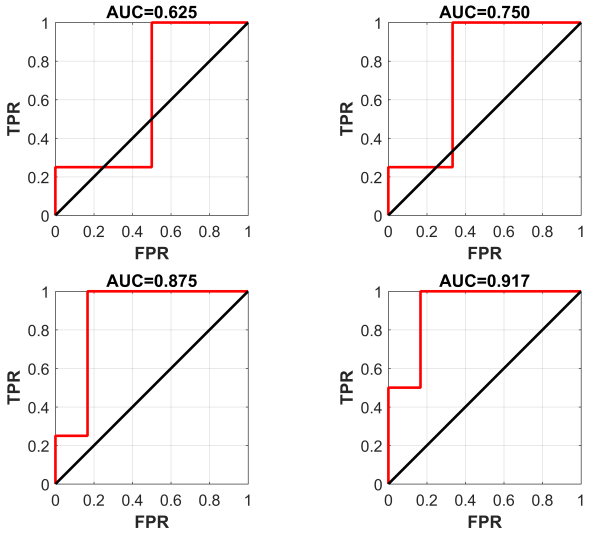

Figure 5: Four different receiver operator characteristic (ROC) curves and their corresponding area under curve (AUC) values.

Question 7. We suspect that a basketball player's Height (i.e., x_1) is predictive of whether the player is successful with his field goals FG (i.e., x_3). To quantify whether Height is predictive of FG we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature Height to discriminate between FG> 45% (positive class) or FG≤ 45% (negative class) considering the data given in Figure 4. Which one of the receiver operator characteristic (ROC) curves given in Figure 5 corresponds to the correct ROC curve?

A. The curve having AUC=0.625

B. The curve having AUC=0.750

C. The curve having AUC=0.875

D. The curve having AUC=0.917

E. Don't know.

Solution 7. There are a total of 4 positive and 6 negative observations. When lowering the threshold for predicting high performance based on the value of Height we observe that the first observation to be above the threshold is O10 which belongs to the positive class, thus TPR=1/4, FPR=0/6. Subsequently we get two observations from the negative class thus TPR=1/4, FPR=2/6 and then three positive observations being above the threshold, i.e. TPR=4/4, FPR=2/6. Lowering the threshold further we obtain the remaining negative observations such that TPR=4/4, FPR=6/6. The only curve having this property is the curve with AUC=0.750.

Question 8. We will consider the ten observations of the Basketball dataset given in Figure 4. We will cluster this data using k-means with Euclidean distance into two clusters (i.e., k=2) . Which one of the following solutions constitutes a converged solution in the k-means clustering procedure?

A. {O1}, {O2, O3, O4, O5, O6, O7, O8, O9, O10}.

B. {O1, O2, O3, O4, O5}, {O6, O7, O8, O9, O10}.

C. {O1, O2, O3, O4, O5, O6, O7}, {O8, O9, O10}.

D. {O1, O2, O3, O4, O5, O6, O7, O8, O9}, {O10}.

E. Don't know.

Solution 8. The solution {O1}, {O2, O3, O4, O5, O6, O7, O8, O9, O10} has centroids at 5.7 and (6.0+6.2+6.3+6.4+6.6+6.7+6.9+7.0+74)/9=6.5889. As such, O2 is closer to the centroid at 5.7 than 6.6111 and will thus be reassigned to this centroid hence this is not a converged solution. The solution {O1, O2, O3, O4, O5}, {O6, O7, O8, O9, O10} has centroids at (5.7+6.0+6.2+6.3+6.4)/5=6.12 and (6.6+6.7+6.9+7.0+7.4)/5=6.92. As such, O5 is closer to the centroid at 6.12 and O6 is closer to the centroid at 6.92. This will thus form a converged solution The solution {O1, O2, O3, O4, O5, O6, O7}, {O8, O9, O10} has centroids at (5.7+6.0+6.2+6.3+6.4+6.6+6.7)/7= 6.2714 and (6.9+7.0+7.4)/3=7.1. As such, O7 is closer to the centroid at 7.1 than the one at 6.2714 and will thus be reassigned to this centroid, hence, this is also not a converged solution. The solution {O1, O2, O3, O4, O5, O6, O7, O8, O9}, {O10} has centroid at 6.4222 and 7.4. As such, O9 is closer to 7.4 than 6.4222 and will thus be reassigned to this centroid, hence, this is also not a converged solution.

Question 9. We suspect that observation O10 may be an outlier. In order to assess if this is the case we would like to calculate the average relative KNN density based on Euclidean distance and the observations given in Figure 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 (if observations are tied in terms of their distances to an observation, the observation with smallest observation number will be selected). Based on considering only the attribute Height and the ten observations in Figure 4, what is the average relative density for observation O10 for $K = 3$ nearest neighbors?

A. 0.409

B. 0.500

C. 0.533

D. 1.875

E. Don't know.

Solution 9.

$$\text{density}(\boldsymbol{x}_{O10}, 3) = (\frac{1}{3}(0.4 + 0.5 + 0.7))^{-1} = 1.8750$$

$$\text{density}(\boldsymbol{x}_{O9}, 3) = (\frac{1}{3}(0.1 + 0.3 + 0.4))^{-1} = 3.7500$$

$$\text{density}(\boldsymbol{x}_{O8}, 3) = (\frac{1}{3}(0.1 + 0.2 + 0.3))^{-1} = 5$$

$$\text{density}(\boldsymbol{x}_{O7}, 3) = (\frac{1}{3}(0.1 + 0.2 + 0.3))^{-1} = 5$$

$$\text{a.r.d.}(\boldsymbol{x}_{O10}, 3)) = \frac{\text{density}(\boldsymbol{x}_{O10}, 3)}{\frac{1}{3}(\text{density}(\boldsymbol{x}_{O9}, 3) + \text{density}(\boldsymbol{x}_{O8}, 3) + \text{density}(\boldsymbol{x}_{O7}, 3))} = \frac{1.8750}{\frac{1}{3}(3.75 + 5 + 5)} = 0.4091$$

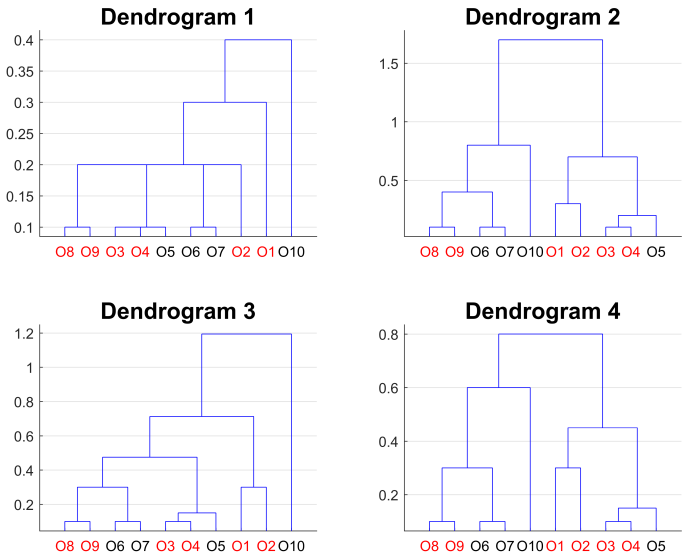

Figure 6: Four different dendrograms derived using the Euclidean distance between the 10 observations based on the attribute Height only. The value of Height for each of the 10 observations can be found in Figure 4. Red observations correspond to low values of FG and black observations to high values of FG.

Question 10. We will consider the Euclidean distance between the observations in Figure 4 based on the attribute Height only, (i.e., the Euclidean distance between observation O1 and O2 is $\sqrt{(5.7-6.0)^2} = 0.3$). A hierarchical clustering is used to cluster the observations based on their distances to each other using average linkage (when ties in the agglomerative procedure clusters containing the smallest observation numbers will merge first). Which one of the dendrograms given in Figure 6 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 10. Initially, O3, O4 will merge and O6, O7 will merge and O8, O9 will merge at the level of 0.1. Subsequently, O5 will merge onto {O3,O4} at the level of (0.1+0.2)/2=0.15. Next {O6,O7} will merge with {O8,O9} at the level of (0.3+0.3+0.2+0.3)/4=0.275. Next, O1,O2 will merge at 0.3 and subsequently {O1,O2} with {O3,O4,O5} at the level of (0.5+0.6+0.7+0.2+0.3+0.4)/6=0.45 and then O10 will merge with {O6,O7,O8,O9} at the level of (0.8+0.7+0.5+0.4)/4=0.6. The only dendrogram having these properties is dendrogram 4 and we can thus rule out the other dendrograms.

Question 11. We will cut dendrogram 2 at the level of two clusters and evaluate this clustering in terms of its correspondence with the class label information in which O1, O2, O3, O4, O8, and O9 correspond to low values of FG whereas O5, O6, O7, and O10 correspond to high values of FG. We recall that the Rand index also denoted the simple matching coefficient (SMC) between the true labels and the extracted clusters is given by $R = \frac{f_{11}+f_{00}}{K}$, where f_{11} is the number of object pairs in same class assigned to same cluster, f_{00} is the number of object pairs in different class assigned to different clusters, and $K = N(N-1)/2$ is the total number of object pairs, where N is the number of observations considered. What is the value of R between the true labeling of the observations in terms of high and low FG values and the two clusters?

A. 0.3226

B. 0.5333

C. 0.5778

D. 0.6222

E. Don't know.

Solution 11. The cluster indices are given by the vector: $[2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1]^\top$, whereas the true class labels are given by the vector $[1\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2]^\top$. From this, we obtain: Total number of object pairs is:
$K = 10(10-1)/2 = 45$
$f_{00} = 4 \cdot 3 + 1 \cdot 2 = 14$
$f_{11} = 4{\cdot}(4{-}1)/2{+}1{\cdot}(1{-}1)/1{+}3{\cdot}(3{-}1)/2{+}2{\cdot}(2{-}1)/2 = 10$
$R = \frac{f_{11}+f_{00}}{K} = \frac{10+14}{45} = 24/45.$

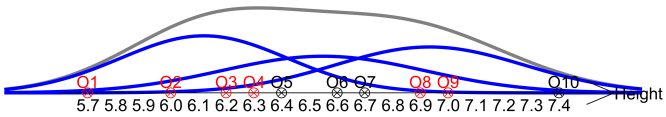

Figure 7: A Gaussian mixture model (GMM) with three clusters fitted to the 10 observations based only on the attribute Height. The overall probability density is given in gray and in blue the contribution from each of the three clusters to the density.

Question 12. We will fit a Gaussian mixture model (GMM) with three clusters to the 10 observations given in Figure 4. The fitted density is given in Figure 7 in which the overall density is given in gray and the contribution of each Gaussain given in blue. We recall that the Gaussian mixture model for 1-dimensional data is given by: $p(x) = \sum_k w_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$.
For the clustering we have:
$w_1 = 0.37, w_2 = 0.29, w_3 = 0.34,$
$\mu_1 = 6.12, \mu_2 = 6.55, \mu_3 = 6.93,$
$\sigma_1^2 = 0.09, \sigma_2^2 = 0.13, \sigma_3^2 = 0.12.$
What is the probability that observation O8 is assigned to cluster 2 according to the GMM?

A. 0.20

B. 0.29

C. 0.33

D. 0.37

E. Don't know.

Solution 12. The probability that the k'th cluster generated the observation O8 is given by $w_k N(6.9|\mu_k, \sigma_k^2)$ and we thus have:
$p(x_8 = 6.9, z_8 = 1) = 0.37 \frac{1}{\sqrt{2\pi 0.09}} \exp\left(-\frac{1}{2{\cdot}0.09}(6.9-6.12)^2\right) = 0.0168$
$p(x_8 = 6.9, z_8 = 2) = 0.29 \frac{1}{\sqrt{2\pi 0.13}} \exp\left(-\frac{1}{2{\cdot}0.13}(6.9-6.55)^2\right) = 0.2003$
$p(x_8 = 6.9, z_8 = 3) = 0.34 \frac{1}{\sqrt{2\pi 0.12}} \exp\left(-\frac{1}{2{\cdot}0.12}(6.9-6.93)^2\right) = 0.3901$
$p(z_8 = 2|x_8 = 6.9) = \frac{p(x_8=6.9, z_8=2)}{\sum_{k'} p(x_8=6.9, z_8=k')}$
$= \frac{0.2003}{0.0168+0.2003+0.3901} = 0.33.$

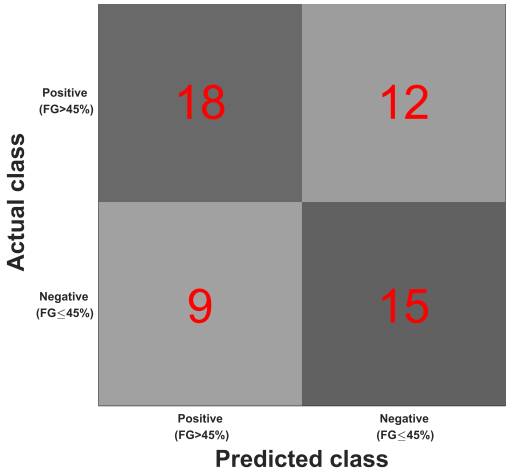

Figure 8: Confusion matrix based on a classifier's predictions of high or low success rate of field goals, (i.e., FG>45% considered the positive class or FG≤45% considered the negative class respectively).

Question 13. We will consider a simple classifier that predicts the 54 basketball players as having high success rate of field goals (FG>45%, considered the positive class) if they are taller than 6.65 foot and low otherwise ($FG \leq 45\%$, considered the negative class). The confusion matrix of the classifier is given in Figure 8. Which statement regarding the classifier is correct?

A. The recall of the classifier is 60.0 %.

B. The precision of the classifier is 61.1 %.

C. The accuracy of the classifier is 66.7 %.

D. The dataset is perfectly balanced.

E. Don't know.

Solution 13. The recall of the classifier is TP/(TP+FN)=18/(18+12)= 60.0%. The precision of the classifier is TP/(TP+FP)=18/(18+9)=66.7%. The accuracy rate of the classifier is (TP+TN)/(TP+FP+TN+FN)=(18+15)/54=61.1%. There are 30 positive examples and 24 negative examples in the test set.

Question 14. The National Basketball Association (NBA) is the top basketball league in USA and all males playing in the NBA earns more than several million dollars a year or more making them all have a very high salary. In USA we will assume approximately 0.2 % of the male population that are not playing in the NBA makes such similar very high salary. Furthermore, we will assume two out of a million American males are playing in the NBA. Assuming the above, what is the probability that a male in USA making such very high salary plays in the NBA?

A. 0.0002%

B. 0.0010%

C. 0.0999%

D. 0.2002%

E. Don't know.

Solution 14. What we are interested in is $P(\text{NBA}|\text{Very high salary}) = \frac{P(\text{Very high salary}|\text{NBA})P(NBA)}{P(\text{Very high salary}|\text{NBA})P(NBA)+P(\text{Very high salary}|\text{not NBA})P(\text{not NBA})} = \frac{\frac{1\cdot 2/1000000}{1\cdot 2/1000000+0.002\cdot(1-2/1000000)}}{} = \frac{2}{2+0.002*999998} = 0.0999\%$.

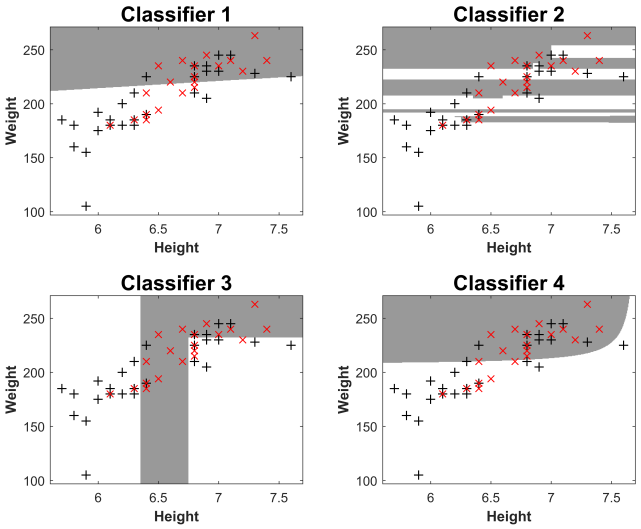

Figure 9: Decision boundaries for four different classifiers trained on the Basketball dataset considering the features Height and Weight. Gray regions classify into red crosses whereas white regions into black plusses.

Question 15. Four different classifiers are trained on the Basketball dataset considering the features Height and Weight in order to predict if the percentage of successful field goals is high (FG$>$ 45%) or low (FG$\leq$ 45%). The decision boundary for each of the four classifiers is given in Figure 9. Which one of the following statements is correct?

A. Classifier 1 corresponds to logistic regression considering as input x_1 and x_2, Classifier 2 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 3 is a decision tree including three decisions, Classifier 4 correponds to a logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$.

B. Classifier 1 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 2 is a decision tree including three decisions, Classifier 3 corresponds to logistic regression considering as input x_1 and x_2, Classifier 4 corresponds to a logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$.

C. Classifier 1 corresponds to a logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$, Classifier 2 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 3 is a decision tree including three decisions, Classifier 4 corresponds to logistic regression considering as input x_1 and x_2,

D. Classifier 1 is a decision tree including three decisions, Classifier 2 corresponds to logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$, Classifier 3 corresponds to a logistic regression considering as input x_1 and x_2, Classifier 4 is a 3-nearest neighbor classifier using Euclidean distance.

E. Don't know.

Solution 15. The decision boundary of classifier 1 is a straight line thus conforms to a logistic regression model using only the features x_1 and x_2 as inputs. Classifier 2 is a 3-nearest neighbor classifier and when using Euclidean distance the scale of Weight has much more variance than that of height thereby heavily influencing the distance measure. Classifier 3 is a decision tree with two vertical and one horizontal line corresponding to three decisions. Classifier 4 is non-linear and smooth corresponding to a logistic regression including a transformed variable, i.e. $x_1 \cdot x_2$.

Question 16. We will consider an artificial neural network (ANN) trained to predict the average score of a player (i.e., y). The ANN is based on the model:

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

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer (i.e., positive values are returned and negative values are set to zero). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix},$$

and $w_0^{(2)} = 2.84$, $w_1^{(2)} = 3.25$, and $w_2^{(2)} = 3.46$.
What is the predicted average score of a basketball player with observation vector $\boldsymbol{x}^* = [6.8\ 225\ 0.44\ 0.68]$?

A. 1.00

B. 3.74

C. 8.21

D. 11.54

E. Don't know.

Solution 16. The output is given by:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{w}) &= 2.84 \\ &+ 3.25 \cdot \max([1\ 6.8\ 225\ 0.44\ 0.68] \cdot \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, 0) \\ &+ 3.46 \max([1\ 6.8\ 225\ 0.44\ 0.68] \cdot \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix}, 0) \\ &= 2.84 + 3.25 \cdot \max(-1.027, 0) + 3.46 \max(2.516, 0) \\ &= 11.54 \end{aligned}$$

Question 17. Which statement regarding cross-validation is correct?

A. An advantage of five-fold cross-validation over three-fold cross-validation is that the datasets used for training are larger.

B. The more data used for training a model the more we can expect the model to overfit to the training data.

C. 10-fold cross-validation is more accurate but also more computationally expensive than leave-one-out cross-validation.

D. When upsampling data in order to avoid class imbalance issues the same observations should be included in the training and test set such that the training and test sets reflect the same properties.

E. Don't know.

Solution 17. When using k-fold cross-validation 1/k of the data is used for testing and (k-1)/k of the data for training during each fold. As such, five-fold cross-validation uses larger training sets than three-fold cross-validation. The more data used for training a model the less we can expect overfitting to occur as the model will be less prone to fit to specific aspects of the training set. 10-fold cross-validation should not be more accurate and in particular, it is not more expensive than leave-one-out cross-validation. Leave-one-out cross-validation is more expensive as we have to use as many folds as we have observations and each model is trained on a larger training set size. When upsampling the data it is important that the same observations do not occur in both the training and test set as we otherwise are training the model on parts of the test set and thereby fitting the model also to test data.

	H_L	H_H	W_L	W_H	$FG_{\leq 45\%}$	$FG_{>45\%}$	$FT_{<75\%}$	$FT_{>75\%}$
O1	1	0	1	0	1	0	1	0
O2	1	0	1	0	1	0	1	0
O3	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1
O5	1	0	1	0	0	1	0	1
O6	1	0	0	1	0	1	1	0
O7	0	1	1	0	0	1	0	1
O8	0	1	1	0	1	0	0	1
O9	0	1	0	1	1	0	1	0
O10	0	1	0	1	0	1	1	0

Table 3: The ten considered observations of the Basketball dataset binarized considering the attribute x_1–x_4. The attributes x_1 and x_2 are binarized according to whether they are below or above the median value of the attribute for the entire dataset of 54 observations. x_3 and x_4 are respectively threshold at a success rate of 45% for field goals (FG) and a success rate of 75% for free throw (FT). The ten observations are color coded in terms of average points scored per game (y) being in the low range {O2, O3, O6, O9} mid-range {O1, O4} and high range {O5, O7, O8, O10}.

Question 18. Considering the dataset in Table 3 as a market basket problem with observation O1–O10 corresponding to customers and H_L, H_H, W_L, W_H, $FG_{\leq 45\%}$, $FG_{>45\%}$, $FT_{\leq 75\%}$, and $FT_{>75\%}$ corresponding to items. What are all frequent itemsets with support greater than 35%?

A. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, and $\{FT_{>75\%}\}$.

B. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, and $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$.

C. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$, and $\{H_L,W_L,FG_{\leq 45\%}\}$.

D. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$, $\{H_L,W_L,FG_{\leq 45\%}\}$, and $\{H_L,W_L,FT_{\leq 75\%}\}$.

E. Don't know.

Solution 18. For a set to have support more than 35% the set must occur at least $0.35 \cdot 10 = 3.5$, i.e. 4 out of the 10 times. All the itemsets that have this property are:
$\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$, and $\{H_L,W_L,FG_{\leq 45\%}\}$.

Question 19. We consider again the data in Table 3 as a market basket problem. What is the confidence of the association rule $\{\mathrm{H}_L, \mathrm{W}_L\} \rightarrow \{\mathrm{FG}_{\leq 45\%}, \mathrm{FT}_{\leq 75\%}\}$?

A. 30 %

B. 40 %

C. 50 %

D. 60 %

E. Don't know.

Solution 19. The confidence is given as

$$P(\mathrm{FG}_{\leq 45\%}, \mathrm{FT}_{\leq 75\%} | \mathrm{H}_L, \mathrm{W}_L) = \frac{\mathrm{FG}_{\leq 45\%}, \mathrm{FT}_{\leq 75\%}, \mathrm{H}_L, \mathrm{W}_L)}{P(\mathrm{H}_L, \mathrm{W}_L)} = \frac{3/10}{5/10} = 3/5 = 60\%$$

Question 20. We would like to predict whether a basketball player has a high average score using the data in Table 3. We will apply a Naïve Bayes classifier that assumes independence between the attributes given the class label (i.e., the class label is given by the average points scored per game being low (black color), in the mid-range (red color) or high (blue color) respectively in the table). Given that a basketball player is relatively tall ($\mathrm{H}_H = 1$) relatively light weight ($\mathrm{W}_L = 1$) what is the probability that the basketball player will have a high average score according to the Naïve Bayes classifier derived from the data in Table 3?

A. 9/16

B. 9/11

C. 3/4

D. 1

E. Don't know.

Solution 20. Let HAS denote high average score. According to the Naïve Bayes classifier we have

$$P(HAS|\mathrm{H}_H = 1, \mathrm{W}_L = 1) = \frac{\begin{pmatrix} P(\mathrm{H}_H = 1||HAS)\times \\ P(\mathrm{W}_L = 1|HAS)\times \\ P(HAS) \end{pmatrix}}{\begin{pmatrix} P(\mathrm{H}_H = 1||LAS)\times \\ P(\mathrm{W}_L = 1|LAS)\times \\ P(LAS) + P(\mathrm{H}_H = 1|MAS)\times \\ P(\mathrm{W}_L = 1|MAS)\times \\ P(MAS) + P(\mathrm{H}_H = 1||HAS)\times \\ P(\mathrm{W}_L = 1|HAS)\times \\ P(HAS) \end{pmatrix}} = \frac{3/4 \cdot 3/4 \cdot 4/10}{1/4 \cdot 2/4 \cdot 4/10 + 0 \cdot 2/2 \cdot 2/10 + 3/4 \cdot 3/4 \cdot 4/10} = \frac{9/40}{2/40 + 0 + 9/40} = 9/11.$$

Question 21. Considering the data in Table 3, we will use a 3-nearest neighbor classifier to classify observation O10 (i.e., with binary observation vector [0 1 0 1 0 1 1 0]) based on observation O1–O9. We will classify according to the three neighboring observations with *highest* similarity according to the Jaccard (J) measure of similarity given by $J(\boldsymbol{a}, \boldsymbol{b}) = \frac{f_{11}}{M - f_{00}}$, where f_{11} and f_{00} are the number of one matches and zero matches respectively and M the total number of binary features. Which one of the following statements is correct?

A. O10 will be classified as black.

B. O10 will be classified as blue.

C. O10 will be classified as red.

D. The classifier will be tied between the classes black and blue.

E. Don't know.

Solution 21. For $O10$ we have:

$$J(O10, O1) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O2) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O3) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O4) = \frac{0}{8-0} = 0$$
$$J(O10, O5) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O6) = \frac{3}{8-3} = \frac{3}{5}$$
$$J(O10, O7) = \frac{3}{8-3} = \frac{3}{5}$$
$$J(O10, O8) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O9) = \frac{3}{8-3} = \frac{3}{5}$$

Hence the three nearest neighbors are O6,O7, and O9 with two black and one blue observation thus according to majority voting the observation will be classified as black.

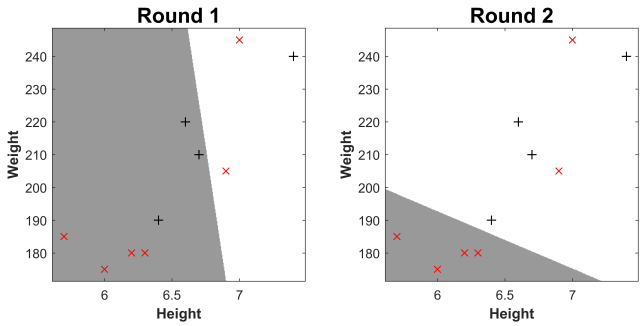

Figure 10: Decision boundaries for two rounds of boosting considering a logistic regression model using the features Height and Weight and the 10 observations also considered previously in Figure 4. Gray region indicates that the observation will be classified as red crosses, white regions that the observation will be classified as black plusses.

Question 22. We will consider classifying the 10 observations considered in Figure 4 using logistic regression and boosting by Adaboost (notice, the Adaboost algorithm uses the natural logarithm). For this purpose we include only two boosting rounds considering only the features Height (i.e., x_1) and Weight (i.e., x_2) as inputs. In the first round the data is sampled with equal probability $w_i = 1/10$ for $i = \{1, \ldots, 10\}$ and the logistic regression model with decision boundary given to the left of Figure 10 trained. A new dataset is subsequently sampled and a new logistic regression classifier given to the right of the Figure 10 trained. Based on these two rounds of the Adaboost algorithm what will an observation located at $x_1 = 6$ and $x_2 = 240$ be classified as?

A. The two classes will be tied for the Adaboost procedure.

B. The observation will be classified as red cross.

C. The observation will be classified as black plus.

D. The weights $w_1, \ldots, w_{10}$ are changed in round 1 of the Adaboost procedure.

E. Don't know.

Solution 22. We have for the first round that the weighted error rate $\epsilon_1 = 5/10$ with associated $\alpha_1 = \frac{1}{2}\log\frac{1-\epsilon_1}{\epsilon_1} = \frac{1}{2}\log 1 = 0$. The updated weights will thus be unchanged as $e^0 = 1$. In the next round

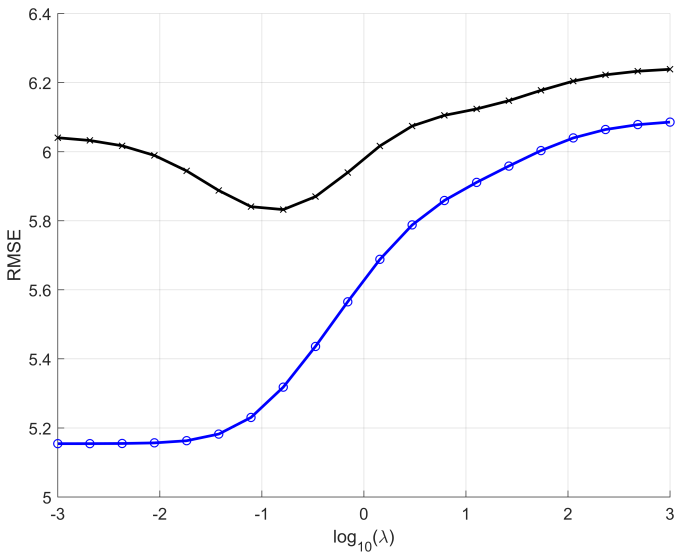

Figure 11: Root-mean square error (RMSE) curves as function of the regularization strenght λ for regularized least square regression predicting the average points scored per game y based on the attributes x_1–x_4.

$\epsilon_2 = 2/10$ and thus $\alpha_2 = \frac{1}{2}\log\frac{1-2/10}{2/10} = \frac{1}{2}\log 4 = 0.693$. When determining the class we weight each classifier by the importance α_t of the round t. However, as the first round has $\alpha_1 = 0$ this round has zero weight in the voting and thus the classifier will solely be based on the second round classifier for which $\alpha_2 = 0.693$. This classifier will deem the observation at $x_1 = 6$ and $x_2 = 240$ to belong to the class of black plusses as the decision region is white.

Question 23. Using the 54 observations of the Basketball dataset we would like to predict the average points scored per game (y) based on the four features (x_1–x_4). For this purpose we consider regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function:
$E(\boldsymbol{w}) = \sum_n (y_n - [1\ x_{n1}\ x_{n2}\ x_{n3}\ x_{n4}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$,
where x_{nm} denotes the m'th feature of the n'th observation, and 1 is concatenated the data to account for the bias term. We consider 20 different values of λ and use leave-one-out cross-validation to quantify the performance of each of these different values of λ. The results of the leave-one-out cross-validation performance is given in Figure 11. Inspecting the model for the value of $\lambda = 0.6952$ the following model is identified:
$f(\boldsymbol{x}) = 2.76 - 0.37x_1 + 0.01x_2 + 7.67x_3 + 7.67x_4$.
Which one of the following statements is correct?

A. In Figure 11 the blue curve with circles corresponds to the training error whereas the black curve with crosses corresponds to the test error.

B. According to the model defined for $\lambda = 0.6952$ increasing a players height will increase his average points scored per game.

C. There is no optimal way of choosing λ since increasing λ reduces the variance but increases the bias.

D. As we increase λ the 2-norm of the weight vector $\boldsymbol{w}$ will also increase.

E. Don't know.

Solution 23. The blue curve monotonically increases with λ reflecting a worse fit to the training set as we increase λ using regularization we can reduce the variance by introducing bias and the black curve indicates that an optimal tradeoff at around $10^{-0.8}$ as reflected by the test error indicated in the black curve being minimal. As we increase λ we will penalize the weights according to the squared 2-norm more and more and thus the 2-norm will be reduced. Finally, according to the fitted model we observe that the coefficient in front of x_1 (Height) is negative thus indicating that an increase in height will reduce the models prediction of average points scored per game.

Question 24. We will again consider the ridge regression described in the previous question. Which one of the following statements is correct?

A. Exhaustively evaluating all combinations of features would require the fitting of less models than the proposed ridge-regression procedure.

B. To generate the test curve we need to make predictions from a total of 1060 different models.

C. We can obtain an unbiased estimate of the generalization error of the best performing model from Figure 11.

D. The ridge regression model will be non-linear as the model includes regularization.

E. Don't know.

Solution 24. Exhaustively evaluating all feature combinations of four features would require evaluating $2^4 = 16$ models whereas we currently consider 20 different models. For each of the 20 values of λ we have to estimate 54 models according to the leave-one-out procedure, (i.e., 54 times we leave out an observation as the dataset contains 54 observations.) thus we need a total of $20 \cdot 54 = 1080$ different models for making the necessary predictions. In order to get an unbiased estimate of generalization of the best model we would need two-layer cross-validation (we currently only have one-level cross-validation). The ridge regression model will not be non-linear but still linear in the input data regardless of the regularization.

Question 25. Consider the clustering problem given in Figure 12. Which clustering approach is *most* suited for correctly separating the data into the four groups indicated by black crosses, red circles, magenta plusses, and blue asterics?

A. A well-separated clustering approach.

B. A contiguity-based clustering approach.

C. A center-based clustering approach.

D. A conceptual clustering approach.

E. Don't know.

Solution 25. As the observation in each cluster is at least closest to one other observation in its cluster than to an observation in another cluster a contiguity based approach is most suited.

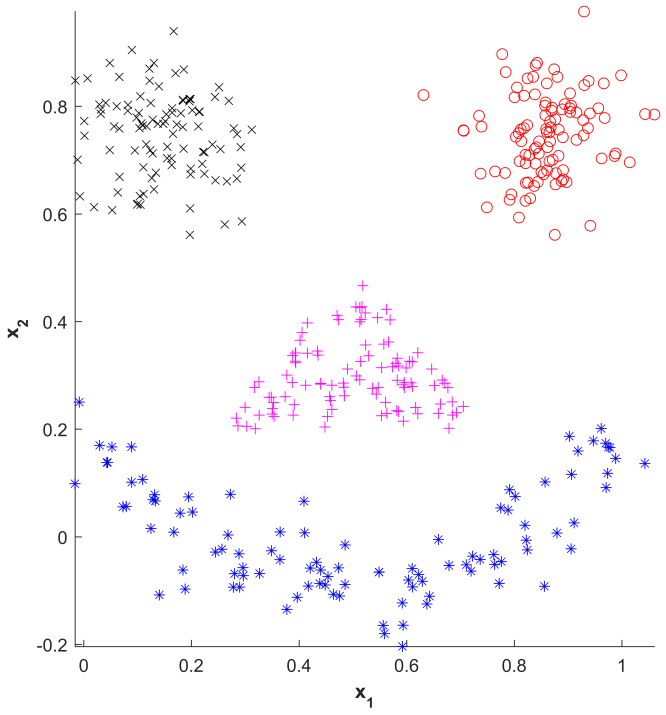

Figure 12: A clustering problem containing four clusters indicated by black crosses, red circles, magenta plusses and blue stars.

Question 26. Which one of the following statements is correct?

A. Multinomial regression can only handle classification problems where the problem is to classify between two classes.

B. Decision trees return the probability that an observation is in a given class.

C. k-means, Gaussian Mixture Models (GMM) and Artificial Neural Networks (ANN) are all prone to local minima issues and thus it is recommended to run the procedures using multiple initializations.

D. The accuracy is a good performance measure when facing severe class imbalance issues in a two class classification problem.

E. Don't know.

Solution 26. Multinomial regression is a generalization of two class logistic regression to handle multiple

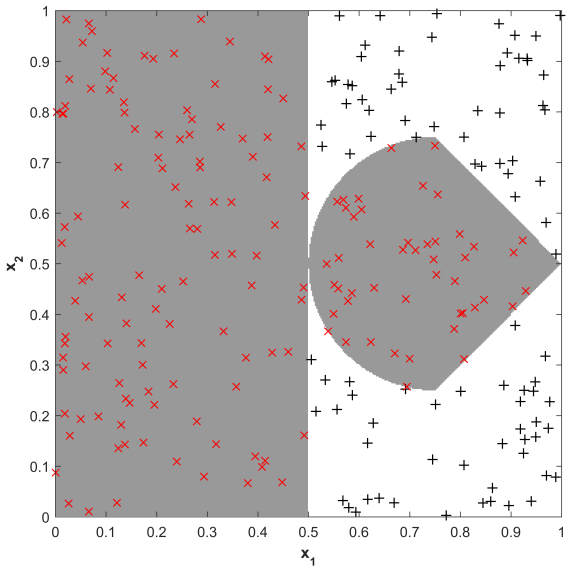

Figure 13: A two class classification problem with red crosses (i.e., x) and black plusses (i.e., +) constituting the two classes as well as the associated decision boundaries of the two classes indicated respectively by gray and white regions.

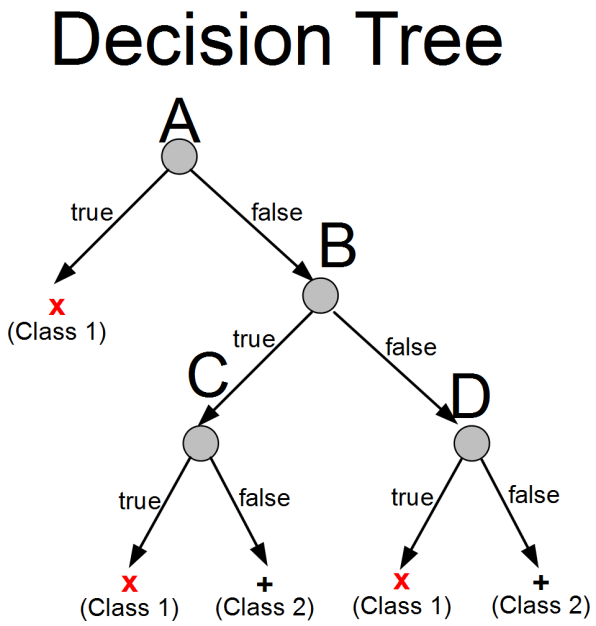

Figure 14: A decision tree with four decisions (A,B,C, and D) forming the decision boundaries given in Figure 13 if adequately defined.

Question 27. We will consider the two class classification problem given in Figure 13 in which the goal is to separate red crosses (i.e., x) from black plusses (i.e., +) based on the decision boundaries in gray and white indicated in the top panel of the figure. Which one of the following procedures based on the decision tree given in Figure 14 will perfectly separate the two classes?

A. **A=**$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$**,**
B=$x_1 < 0.75$**,**
C= $\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_2 < 0.25$**,**
D=$\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_1 < 0.25$**.**

B. A=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_1 < 0.5$,
B=$x_1 < 0.75$,
C= $\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_2 < 0.25$,
D=$\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_\infty < 0.25$.

C. A=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$,
B=$x_1 < 0.75$,
C= $\|\boldsymbol{x} - \begin{bmatrix} 0.5 \\ 0.75 \end{bmatrix}\|_2 < 0.25$,
D=$\|\boldsymbol{x} - \begin{bmatrix} 0.5 \\ 0.75 \end{bmatrix}\|_1 < 0.25$.

D. A=$x_1 < 0.75$,
B=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$,

classes. Decision trees do not return probabilities of being in each class but hard assigns observations to the classes based on majority voting in each terminal leaf. K-means, Gaussian Mixture Models and Artificial Neural Networks (ANN) are indeed all prone to local minima and it is therefore advised to use multiple restarts selecting the initialization with best solution. Accuracy is not a good performance measure when facing severe class-imbalance issues as we may trivially obtain a very high accuracy simply by classifying by chance. The AUC of the receiver operator characteristic would here be more appropriate as it is not influenced by the relative sizes of the two classes.

Solution 27. All observations for which $x_1 < 0.5$ are red crosses which can be captured by the initial decision A=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$. For the remaining observations it appears two different norms are at play depending on whether $x_1 < 0.75$ or not, thus B=$x_1 < 0.75$. If $x_1 < 0.75$ we observe that decision C should have a circular shape defined by C= $\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_2 < 0.25$ whereas if $x_1 \geq 0.75$ we have a diamond shape defined by D=$\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_1 < 0.25$. The other solutions will not similarly correctly define the decision boundaries.

Technical University of Denmark

Written examination: 24 May 2018, 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.

Your answers to the questions are to be handed in using the electronic file. Use only this page for hand in if you are unable to hand in digitally. In case you have to hand in the answers using the form on this sheet, please print your name and student number clearly.

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

No.	Attribute description	Abbrev.
x_1	Number of seats times kilometers pr. week	S*KM/Week
x_2	Incidences 1985-1999	Inc. 85-99
x_3	Fatal accidents 1985-1999	FA 85-99
x_4	Fatalities 1985-1999	Fat. 85-99
x_5	Incidences 2000-2014	Inc. 00-14
x_6	Fatal accidents 2000-2014	FA 00-14
y	Fatalities 2000-2014	Fat. 00-14

Table 1: The attributes of the airline safety dataset that contains 56 observations of different airline companies and their properties in terms of number of seats times number of kilometers per week, incidences, fatal accidents, and fatalities accumulated over the period of 1985-1999 and 2000-2014 respectively. We presently consider as output y the number of fatalities from 2000-2014.

Question 1. We will consider the airline safety dataset consisting of 56 airline companies and their number of flights as quantified by number of seats times kilometers per week as well as incidences, fatal accidents, and fatalities quantified for the period of 1985-1999 and 2000-2014 respectively[1]. For brevity this dataset will be denoted the airline safety dataset. In Table 1 is given the attributes of the data as well as the output attribute y defined by the number of fatalities from 2000-2014. In Figure 1 is shown a matrix plot of the six attributes x_1–x_6.

Considering the attributes described in Table 1 and the matrix plot in Figure 1 which one of the following statements regarding the attributes x_1–x_6 is correct?

A. At least one of the attributes appears to be normal distributed.

B. x_2 corresponding to Inc. 85-99 and x_3 corresponding to FA 85-99 are negatively correlated.

C. All the six attributes are ratio.

D. All the attributes are continuous.

E. Don't know.

Solution 1. Inspecting the histograms along the diagonal of the matrix plot it is clearly seen that they all have a mode around zero/low values and decrease with no negative observations on the left side of the mode. This is not corresponding to the bell shape of a normal distribution and thus none of the attributes apear normally distributed. Inspecting, the plot of x_2 vs. x_3 we observe that airline with more incidences in 85-99 also have more fatal accidents in 85-99, there is thus a positive correlation between these two attributes. As for all attributes zero means absence of what is being measured and it makes sense to talk about a company having twice as many seats times kilometers per week, or incidents, or accidents, or fatalities, thus, all these attributes are ratio. As incidents, accidents and fatalities are counts they are discrete integer variables and not continuous.

[1]The dataset is taken from `https://github.com/fivethirtyeight/data/tree/master/airline-safety`

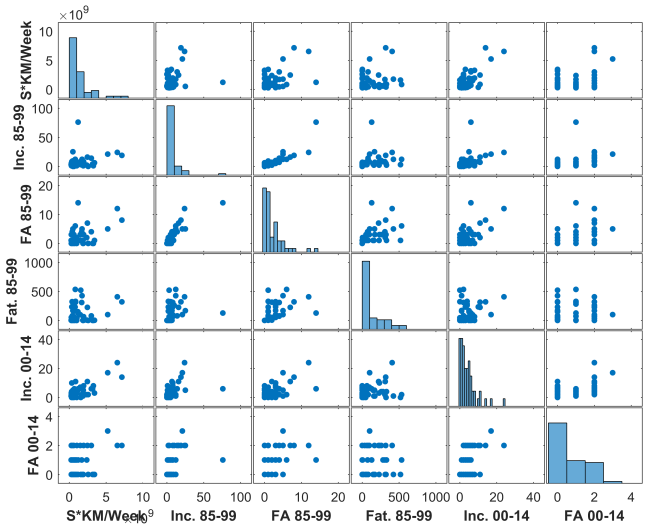

Figure 1: Matrix plot of the six attributes x_1–x_6. Along the diagonal of the matrix plot is given the histogram of each attribute.

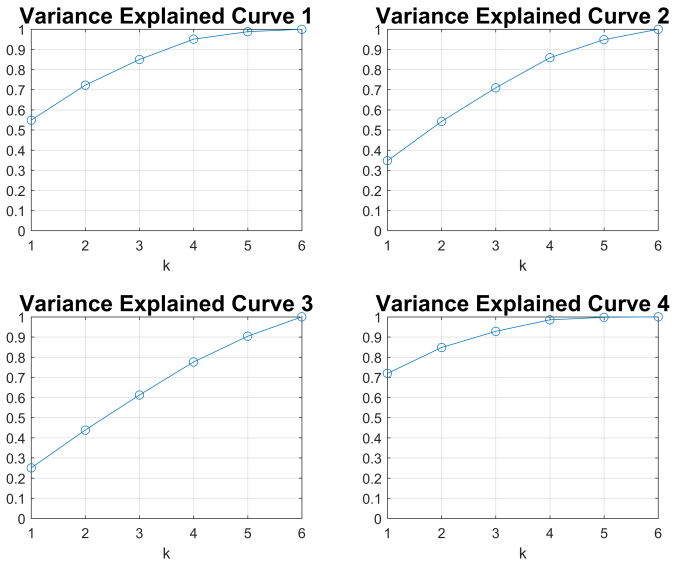

Figure 2: Four different curves of the variance explained as function of keeping the first k principal components when performing a PCA analysis. One of the four curves correctly corresponds to the PCA of the standardized airline safety data.

Question 2. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_6, forming the standardized matrix $\tilde{\boldsymbol{X}}$ (i.e., each attribute has been subtracted its mean and divided by its standard deviation). A singular value decomposition is applied to the standardized data matrix, i.e. $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ and we find the following solution in terms of the $\boldsymbol{S}$ and $\boldsymbol{V}$ matrices:

$$\boldsymbol{S} = \begin{bmatrix} 13.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 7.6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5.8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2.0 \end{bmatrix}.$$

$$\boldsymbol{V} = \begin{bmatrix} 0.38 & -0.51 & 0.23 & 0.47 & -0.55 & 0.11 \\ 0.41 & 0.41 & -0.53 & 0.24 & 0.00 & 0.58 \\ 0.50 & 0.34 & -0.13 & 0.15 & -0.05 & -0.77 \\ 0.29 & 0.48 & 0.78 & -0.17 & 0.00 & 0.23 \\ 0.45 & -0.42 & 0.09 & 0.03 & 0.78 & 0.04 \\ 0.39 & -0.23 & -0.20 & -0.82 & -0.30 & 0.04 \end{bmatrix}.$$

In Figure 2 is given the pct. of variance explained by retaining the first k principal components as a function of k. Which one of the four curves corresponds to the correct curve of variance explained as function of the number of principal components retained?

A. Variance Explained Curve 1.

B. Variance Explained Curve 2.

C. Variance Explained Curve 3.

D. Variance Explained Curve 4.

E. Don't know.

Solution 2. The variance explained by the first k principal components is given by $\frac{\sum_1^k \sigma_k^2}{\sum_{i'=1}^6 \sigma_{i'}^2}$. We thereby get that the curve should have the following values:

k=1: $\frac{13.5^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.5487$

k=2: $\frac{13.5^2+7.6^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.7226$

k=3: $\frac{13.5^2+7.6^2+6.5^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.8498$

k=4: $\frac{13.5^2+7.6^2+6.5^2+5.8^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.9511$

k=5: $\frac{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 0.9880$

k=6: $\frac{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+.02^2}{13.5^2+7.6^2+6.5^2+5.8^2+3.5^2+2.0^2} = 1$

Only Variance Explained Curve 1 has this property.

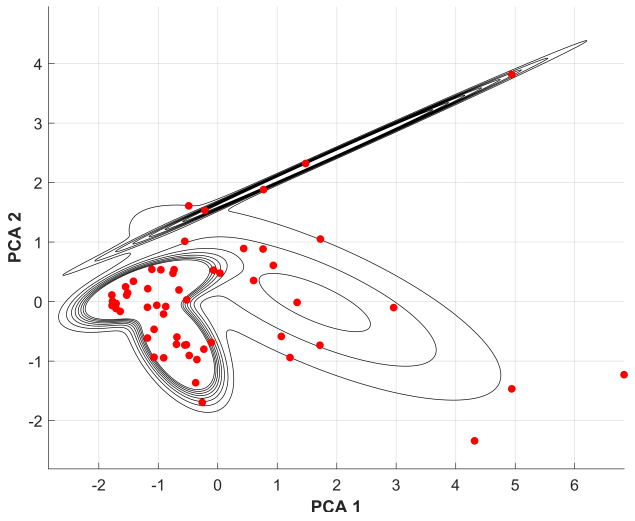

Figure 3: A Gaussian Mixture Model (GMM) using four clusters fitted to the standardized airline safety data projected onto the first two principal components.

Question 3. According to the extracted PCA directions given by the matrix $\boldsymbol{V}$ in the above what will be the coordinate of the standardized observation $\tilde{\boldsymbol{x}}^* = [-0.1 \;\; 0.2 \;\; 0.1 \;\; -0.3 \;\; 1 \;\; 0.5]$ when projected onto the first two principal components?

A. (-0.753, 0.206)

B. (0.652, -0.512)

C. (0.680, 0.019)

D. (0.671, -0.139)

E. Don't know.

Solution 3. The observation $\tilde{\boldsymbol{x}}^* = [-0.1 \;\; 0.2 \;\; 0.1 \;\; -0.3 \;\; 1 \;\; 0.5]$ will have the projection onto the two first principal components given by

$$[-0.1 \;\; 0.2 \;\; 0.1 \;\; -0.3 \;\; 1 \;\; 0.5] \begin{bmatrix} 0.38 & -0.51 \\ 0.41 & 0.41 \\ 0.50 & 0.34 \\ 0.29 & 0.48 \\ 0.45 & -0.42 \\ 0.39 & -0.23 \end{bmatrix} = [0.652 \;\; -0.512].$$

Thus, the observation will in the projection be located at (0.652, -0.512).

Question 4. Which one of the following statements regarding Gaussian Mixture Modeling (GMM) is correct?

A. The number of clusters used in the GMM can be determined by selecting the number of clusters that provides the best likelihood of the training data used for training the density.

B. For high-dimensional data, i.e. where the number of features M is large, it can be beneficial to constrain the covariance of each cluster to be diagonal, i.e. enforcing off-diagonal terms of the covariance matrices to be zero, in order to reduce the number of parameters in the GMM model.

C. The GMM is guaranteed to find the optimal clustering for a given dataset.

D. Similar to the k-means algorithm that assigns observations to the cluster in closest proximity, the EM-algorithm used to estimate the parameters of the GMM considers only the cluster each observation is the most likely to belong to when estimating the parameters in the M-step.

E. Don't know.

Solution 4. For the GMM we can use cross-validation to determine the number of clusters, however, this selection of the number of clusters must be based on the likelihood of the test data and not the training data, as this otherwise would result in overfitting of the density to the data. For high-dimensional data the covariance matrix can be ill-determined and it can therefore be beneficial to reduce the covariance matrix to a diagonal matrix for which only the variance of each features need to be estimated, which substantially reduces the number of free parameters in the GMM. The GMM is prone to local minima solutions and therefore it is recommended to fit the model using many restarts taking the best of these randomly initialized models. In the E-step of the GMM it is quantified how likely it is for the observations to belong to each cluster and thereby the observations are "soft" assigned to each cluster. Thus, in the subsequent M-step this soft-assignment is used taking into account how likely it is for the observations to belong to each of the clusters and contributing in the update of the cluster according to this, such that each clusters mean and covariance is a weighted average of the observations contribution weighted according to this probability.

Question 5. We fit a Gaussian Mixture Model (GMM) to the standardized data projected onto the first two principal component directions using four mixture components (i.e., 4 clusters). We recall that the multivariate Gaussian distribution is given by: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Which one of the following GMM densities corresponds to the fitted density given in Figure 3?

A.

$$p(\boldsymbol{x}) = 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix}) + 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix})$$

B.

$$p(\boldsymbol{x}) = 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix}) + 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix})$$

C.

$$p(\boldsymbol{x}) = 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix}) + 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix})$$

D.

$$p(\boldsymbol{x}) = 0.0673 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}, \begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}) + 0.3360 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -1.2222 \\ 0.1830 \end{bmatrix}, \begin{bmatrix} 0.2639 & 0.0803 \\ 0.0803 & 0.0615 \end{bmatrix}) + 0.2992 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -0.6687 \\ -0.7343 \end{bmatrix}, \begin{bmatrix} 0.1166 & -0.0771 \\ -0.0771 & 0.1729 \end{bmatrix}) + 0.2975 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}, \begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix})$$

E. Don't know.

Solution 5. Inspecting the GMM density we observe that the cluster located at $\begin{bmatrix} 1.8422 \\ 2.4306 \end{bmatrix}$ will have the lowest mixing proportion as only few observations belong to this cluster. Furthermore, it must have a large positive correlation and variance as given by $\begin{bmatrix} 3.8237 & 1.7104 \\ 1.7104 & 0.7672 \end{bmatrix}$. Only aswer option 2 and 4 have this property. The cluster located at $\begin{bmatrix} 1.6359 \\ -0.0183 \end{bmatrix}$ has negative covariance but the variance of the cluster is also much larger than the variance of the other cluster having negative covariance. Thus, this cluster must have the covariance $\begin{bmatrix} 4.0475 & -1.5818 \\ -1.5818 & 1.1146 \end{bmatrix}$. Only answer option 4 has this property.

Question 6. We would like to predict the safety of a given airline company. However, in order to take into account the volume of flights of the company when evaluating its safety we define a new output variable given by $\tilde{y} = y/x_1$. By defining $\tilde{y}$ as the number of fatalities divided by the number of seats times kilometers per week of the company, fatalities are quantified relative to the volume of flights that has been catered by the company. A least squares linear regression model is trained using different combinations of the five attributes x_2, x_3, x_4, x_5, and x_6 in order to predict $\tilde{y}$. Table 2 provides the training and test root-mean-square error (RMSE=$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{y}_i)^2}$) performance of the least squares linear regression model when trained using different combinations of the five attributes. Which one of the following statements is correct?

A. Forward and backward selection will result in the same features being selected.

B. Forward selection will terminate with four features in the feature set.

C. Backward selection will not remove any features.

D. x_4 will be among the features selected by forward selection.

E. Don't know.

Solution 6. Forward selection will result in x_6 being selected with performance 0.18749 and subsequently x_3 with performance 0.17624 and then x_5 with performance 0.17082 as no improvement can be achieved by adding additional features it will terminate at the set x_3, x_5, and x_6. Backward selection will result in removing x_4 to have x_2, x_3, x_5, x_6 with performance 0.17299 and then remove x_2 to attain the performance 0.17082 by se the feature set x_3, x_5, and x_6 upon which no further improvements can be achieved by removing features.

Feature(s)	Training RMSE	Test RMSE
none	0.11279	0.20677
x_2	0.10930	0.22301
x_3	0.10974	0.21773
x_4	0.10911	0.21362
x_5	0.11254	0.20729
x_6	0.09301	0.18749
x_2, x_3	0.10914	0.22247
x_2, x_4	0.10756	0.22145
x_2, x_5	0.10909	0.22513
x_2, x_6	0.09108	0.18555
x_3, x_4	0.10837	0.21768
x_3, x_5	0.10961	0.21800
x_3, x_6	0.09108	0.17624
x_4, x_5	0.10910	0.21368
x_4, x_6	0.09234	0.19121
x_5, x_6	0.08993	0.17657
x_2, x_3, x_4	0.10753	0.22138
x_2, x_3, x_5	0.10887	0.22435
x_2, x_3, x_6	0.09071	0.18029
x_2, x_4, x_5	0.10731	0.22315
x_2, x_4, x_6	0.08947	0.19339
x_2, x_5, x_6	0.08900	0.17610
x_3, x_4, x_5	0.10828	0.21795
x_3, x_4, x_6	0.08805	0.17900
x_3, x_5, x_6	0.08896	0.17082
x_4, x_5, x_6	0.08891	0.18062
x_2, x_3, x_4, x_5	0.10730	0.22314
x_2, x_3, x_4, x_6	0.08782	0.18371
x_2, x_3, x_5, x_6	0.08878	0.17299
x_2, x_4, x_5, x_6	0.08727	0.18336
x_3, x_4, x_5, x_6	0.08603	0.17440
x_2, x_3, x_4, x_5, x_6	0.08595	0.17685

Table 2: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict $\tilde{y}$ using different combinations of the five attributes (x_2–x_6).

Question 7. We would again like to predict $\tilde{y}$ based on x_2, x_3, x_4, x_5, and x_6. For this purpose, we will use the regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function: $E(\boldsymbol{w}) = \sum_n (\tilde{y}_n - [1\ x_{n2}\ x_{n3}\ x_{n4}\ x_{n5}\ x_{n6}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$, We will consider 20 different values of λ and use 10-fold cross-validation to select for the optimal value of λ. Which one of the following statement regarding the described regularized least squares regression procedure is correct?

A. Increasing λ will result in an increase in the 2-norm of the trained $\boldsymbol{w}$, i.e. in an increase of the quantity $\|\boldsymbol{w}\|_2$.

B. 10-fold cross-validation will require the fitting of 10 models in total to quantify the best value of λ.

C. The test error obtained for the optimal value of λ is a biased estimate of the generalization error.

D. When using regularization in least squares regression the model becomes more prone to overfitting.

E. Don't know.

Solution 7. When increasing λ the quantity $\boldsymbol{w}^\top \boldsymbol{w} = \|\boldsymbol{w}\|_2^2$ will be penalized more and therefore reduced (and not increased in terms of the quantity $\|\boldsymbol{w}\|_2$). To quantify the best value of λ using 10-fold cross-validation we need to carry out the cross-validation for each of the 20 considered values of λ resulting in $10 \times 20 = 200$ models to be fitted. The test error obtained for the optimal value of λ is biased as it is the best among 20 selected test-performances. To get an unbiased estimate thus requires two-level cross-validation. Regularization reduce overfitting as the model can less well fit the training data due to the regularization.

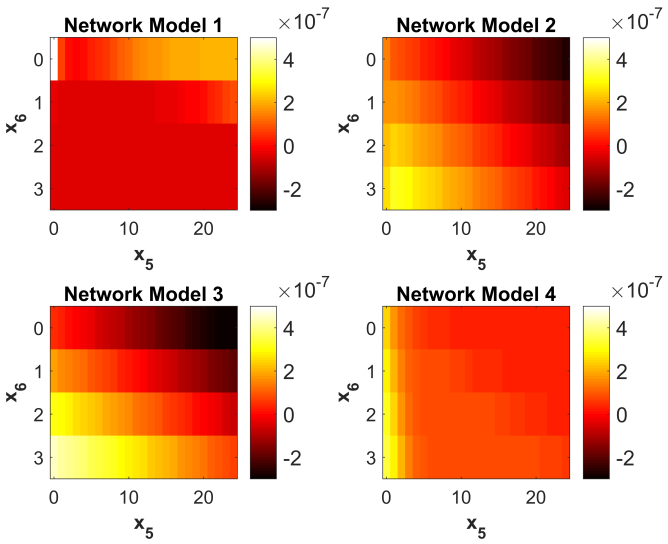

Figure 4: Four different artifical neural network models trained on the airline data to predict $\tilde{y}$ based on the features x_5 and x_6. In each image plot is given the output of one of the four networks using different combinations of x_5 and x_6. As such, $x_5 = 0$ and $x_6 = 0$ is given at the upper left corner whereas $x_5 = 24$ and $x_6 = 3$ is given at the lower right corner of each image.

Question 8. We will consider an artificial neural network (ANN) trained to predict $\tilde{y}$ based only on using x_5 and x_6 as inputs, i.e., incidences and fatal accidents during 2000-2014. The trained model is given by $f(\boldsymbol{x}, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x_5\ x_6]\boldsymbol{w}_j^{(1)})$, where $h^{(1)}(z) = 1/(1+exp(-z))$ is the logistic function used as activation function in the hidden layer (i.e., values are mapped to be between 0 and 1). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} 0.0189 \\ 0.9159 \\ -0.4256 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} 3.7336 \\ -0.8003 \\ 5.0741 \end{bmatrix},$$

and $w_0^{(2)} = 0.3799 \cdot 10^{-6}$, $w_1^{(2)} = -0.3440 \cdot 10^{-6}$, and $w_2^{(2)} = 0.0429 \cdot 10^{-6}$. Which one of the resulting outputs as function of x_5 and x_6 given in Figure 4 corresponds to the trained network?

A. Network Model 1.

B. Network Model 2.

C. Network Model 3.

D. Network Model 4.

E. Don't know.

Solution 8. Consider for instance the two corners located at $x_5 = 0$, $x_6 = 3$ and $x_5 = 24$, $x_6 = 0$ at these two locations we have that the outputs are respectively given by:

$$\begin{aligned} f([03], \boldsymbol{w}) &= 0.3799 \cdot 10^{-6} \\ &\quad - 0.3440 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 0\ 3] \cdot \begin{bmatrix} 0.0189 \\ 0.9159 \\ -0.4256 \end{bmatrix}))} \\ &\quad + 0.0429 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 0\ 3] \cdot \begin{bmatrix} 3.7336 \\ -0.8003 \\ 5.0741 \end{bmatrix}))} \\ &= 0.3799 \cdot 10^{-6} - 0.3440 \cdot 10^{-6} \cdot 0.2213 \\ &\quad + 0.0429 \cdot 10^{-6} \cdot 1.0000 \\ &= 3.4667 \cdot 10^{-7} \end{aligned}$$

and

$$\begin{aligned} f([240], \boldsymbol{w}) &= 0.3799 \cdot 10^{-6} \\ &\quad - 0.3440 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 24\ 0] \cdot \begin{bmatrix} 0.0189 \\ 0.9159 \\ -0.4256 \end{bmatrix}))} \\ &\quad + 0.0429 \cdot 10^{-6} \cdot \frac{1}{1+exp(-([1\ 24\ 0] \cdot \begin{bmatrix} 3.7336 \\ -0.8003 \\ 5.0741 \end{bmatrix}))} \\ &= 0.3799 \cdot 10^{-6} - 0.3440 \cdot 10^{-6} \cdot 1 \\ &\quad + 0.0429 \cdot 10^{-6} \cdot 4.2410e \cdot 10^{-7} \\ &= 3.5900 \cdot 10^{-8}. \end{aligned}$$

The only network that has this property is Network Model 4.

Question 9. We would like to use two level cross-validation to select for the optimal number of hidden units in an artificial neural network (ANN) with one hidden layer as well as quantify the generalization of the selected model. For this purpose, we will use two-level cross-validation in which we in the outer fold use 5-fold cross-validation and in the inner fold (i.e., the fold in which we quantify the optimal number of hidden units in the hidden layer) use 10-fold cross-validation. As ANNs are prone to local minima issues we will train three models for each specification of the number of hidden unit based on three different random initializations and use the model out of these three with best training error. We have a computational budget of training 1000 models and would like to evaluate in steps of 1 from 1 to H hidden units (i.e., if H=3 we will evaluate ANNs with 1, 2, and 3 hidden units). What is the largest value of H for which no more than 1000 models will be trained?

A. 6

B. 19

C. 20

D. 66

E. Don't know.

Solution 9. In two level cross-validation we have for each inner fold $10 \cdot H$ different models. For the optimal of these selected models, we will have to train an additional model on the full dataset used for the inner fold to predict the test data of the outer fold. This thus requires the training of $5 \cdot (10 \cdot H + 1)$ models. As we for each trained model use three random initializations we obtain a total of $3 \cdot 5 \cdot (10 \cdot H + 1)$ models to be trained. We thereby obtain $3 \cdot 5 \cdot (10 \cdot H + 1) = 1000 \Rightarrow (10 \cdot H + 1) = 1000/15 \Rightarrow H = (1000/15 - 1)/10 = 6.5667$. We can thus maximally evaluate for H=6.

Question 10. Some people are afraid of flying and thus prefer to take for instance the car or bus. According to the economist Ian Savage[2] the probability of dying travelling 600 km is approximately:
- Chance dying travelling by car is 0.000271 %.
- Chance dying travelling by bus is 0.000004 %.
- Chance dying travelling by plane is 0.000003 %.

The distance travelling from Copenhagen to Oslo is 600 km regardless of the trip being based on car, bus, or plane. We will assume when travelling from Copenhagen to Oslo 30 % of people take the car, 10 % of people take the bus, and 60 % of people take the plane. Given a person died travelling between Copenhagen and Oslo what is the probability it was from travelling by plane?

A. $1.80 \cdot 10^{-4}$ %

B. 1.08 %

C. 2.16 %

D. 10.0 %

E. Don't know.

Solution 10. Let D denote the event dying travelling between Copenhagen and Oslo. Let C denote the event car, B the event Buss and F the event plane (i.e., flight). According to the numbers given we have P(D—C)=0.000271 %, P(D—B)=0.000004 %, and P(D—F)=0.000003 %. Using Bayes theorem we have:

$$\begin{aligned} P(F|D) &= \frac{P(D|F)P(F)}{P(D|F)P(F) + P(D|B)P(B) + P(D|C)P(C)} \\ &= \frac{0.000003\% \cdot 60\%}{0.000003\% \cdot 60\% + 0.000004\% \cdot 10\% + 0.000271\% \cdot 30\%} \\ &= 2.16\% \end{aligned}$$

[2] http://faculty.wcas.northwestern.edu/~ipsavage/MosesLecture.pdf

Question 11. We will predict whether an airline company is relatively safe (considered the positive class) or unsafe (considered the negative class) based on thresholding $\tilde{y}$ at its median value, i.e. if $\tilde{y} < \text{median}(\tilde{y})$ it is considered safe otherwise it is considered unsafe.

A decision tree is subsequently fitted to the data. At the root of the tree it is considered to split according to the median value of the number of incidences in 2000-2014 (i.e., x_5). For impurity we will use the classification error given by $I(v) = 1 - \max_c p(c|v)$. Before the split, we have 32 safe and 24 unsafe airline companies, and after the split we have

- 23 safe and 8 unsafe airline companies with relatively few incidences.
- 9 safe and 16 unsafe airline companies with relatively many incidences.

Which statement regarding the purity gain Δ of the split is correct?

A. $\Delta = -0.2679$

B. $\Delta = 0.1250$

C. $\Delta = 0.2500$

D. $\Delta = 0.4286$

E. Don't know.

Solution 11. The purity gain is given by

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

where

$$I(v) = 1 - \max_c p(c|v).$$

Evaluating the purity gain for the split we have:

$$\begin{aligned}\Delta =& (1 - 32/56) \\ & -[\tfrac{31}{56}(1 - (\tfrac{23}{31})) \\ & +\tfrac{25}{56}(1 - (\tfrac{16}{25}))] \\ =& \, 7/56\end{aligned}$$

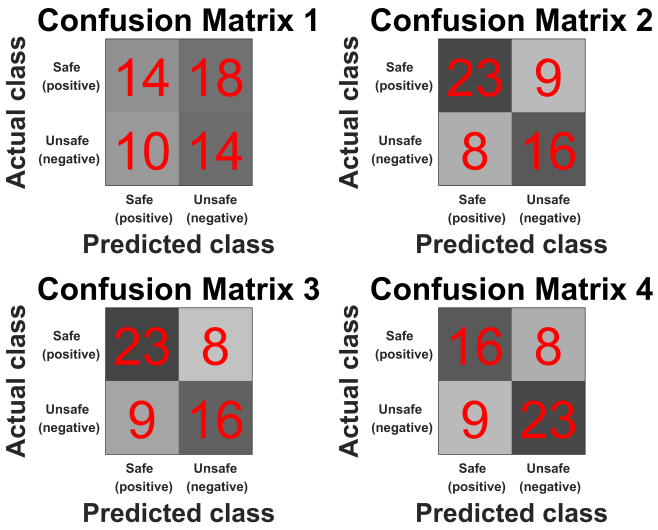

Figure 5: Four different confusion matrices where one corresponds to the confusion matrix of the decision tree with one split according to the median value of the number of incidences in 2000-2014 (i.e., x_5).

Question 12. We will consider the decision tree given by having only the above split (defined in question 11) as a decision and classifying according to this split using the largest class (i.e., using majority voting). In Figure 5 is given four different confusion matrices. Which one of the four confusion matrices corresponds to the decision tree's classification of the 56 observations?

A. Confusion Matrix 1.

B. Confusion Matrix 2.

C. Confusion Matrix 3.

D. Confusion Matrix 4.

E. Don't know.

Solution 12. The decision tree will use majority voting at each leaf in order to classify the 56 observations. For the left branch we have that the majority is safe and thus the 23 safe observations will be correctly classified whereas the 8 unsafe classification will be misclassified as safe. Likewise for the right branch, the majority is unsafe and thus the 16 unsafe observations will be classified as unsafe whereas the 8 safe observations will be misclassified as unsafe. This corresponds to confusion matrix 2.

Question 13. Which statement regarding classification is correct?

A. In classification the output value is continuous.

B. Logistic regression is not a classification approach but a regression method.

C. The k-means algorithm is a supervised classification method.

D. The softmax function is used to provide the probability that an observation is assigned to each class.

E. Don't know.

Solution 13. We make the distinction between a classification and a regression problem based on the property of the output variable being either categorical or continuous. Thus, for continuous outputs we use regression methods and not classification methods. Logistic regression is designed for binary outputs and thus a classification method. The k-means algorithm is used for unsupervised learning and thus not a supervised classification method as it only relies on the input data $\boldsymbol{X}$ and not on output values $\boldsymbol{y}$. The softmax function is used in multinomial regression and artificial neural networks for multi-class classification problems in order to provide outputs interpreted as the probability an observation is assigned to each class similar to the role of the logistic function in logistic regression.

Question 14. We will consider Confusion Matrix 1 given in Figure 5 and we regard the class safe as the *positive* class and unsafe as the *negative* class. Which statement regarding a classifier having performance given by the performance indicated by Confusion Matrix 1 is correct?

A. The classifier's precision is 7/12.

B. The classifier's recall is 1/2.

C. The false positive rate (FPR) of the classifier is 5/14.

D. The accuracy of the classifier is better than guessing everything to be the largest class.

E. Don't know.

Solution 14. The precision is 14/(14+10)=7/12. The recall is 14/(14+18)=7/16. The FPR is 10/(10+14)=5/12. Guessing everything to be the largest class would correspond to guessing everything as unsafe with accuracy of 32/56 whereas the accuracy of the classifier is 28/56.

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0	8.55	0.43	1.25	1.14	3.73	2.72	1.63	1.68	1.28
O2	8.55	0	8.23	8.13	8.49	6.84	8.23	8.28	8.13	7.66
O3	0.43	8.23	0	1.09	1.10	3.55	2.68	1.50	1.52	1.05
O4	1.25	8.13	1.09	0	1.23	3.21	2.17	1.29	1.33	0.56
O5	1.14	8.49	1.10	1.23	0	3.20	2.68	1.56	1.50	1.28
O6	3.73	6.84	3.55	3.21	3.20	0	2.98	2.66	2.50	3.00
O7	2.72	8.23	2.68	2.17	2.68	2.98	0	2.28	2.30	2.31
O8	1.63	8.28	1.50	1.29	1.56	2.66	2.28	0	0.25	1.46
O9	1.68	8.13	1.52	1.33	1.50	2.50	2.30	0.25	0	1.44
O10	1.28	7.66	1.05	0.56	1.28	3.00	2.31	1.46	1.44	0

Table 3: Pairwise Euclidean distances between the first ten observations of the standardized airline safety data. Black observations (i.e., O1, O3, O4, O5, O10) are observations corresponding to relatively safe airline companies, red observations (i.e., O2, O6, O7, O8, O9) are observations corresponding to relatively unsafe airline companies.

Question 15. To determine whether an airline company is relatively safe or unsafe we will use a k-nearest neighbor (KNN) classifier to predict each of the ten observations based on the Euclidean distances between the observations given in Table 3. We will use leave-one-out cross-validation for the KNN in order to classify the ten considered observations and use $K = 1$, i.e., a one nearest neighbor classifier. The analysis will be based only on the data given in Table 3. What will be the error rate of the classifier?

A. 0 %

B. 10 %

C. 20 %

D. 30 %

E. Don't know.

Solution 15. $N(O1, 1) = \{O3\}$ as O3 is closest it will be correctly classified as safe.
$N(O2, 1) = \{O6\}$ as O6 is closest it will be correctly classified as unsafe.
$N(O3, 1) = \{O1\}$ as O1 is closest it will be correctly classified as safe.
$N(O4, 1) = \{O10\}$ as O10 is closest it will be correctly classified as safe.
$N(O5, 1) = \{O3\}$ as O3 is closest it will be correctly classified as safe.
$N(O6, 1) = \{O9\}$ as O9 is closest it will be correctly classified as unsafe.
$N(O7, 1) = \{O4\}$ as O4 is closest it will be incorrectly classified as safe.
$N(O8, 1) = \{O9\}$ as O9 is closest it will be correctly classified as unsafe.
$N(O9, 1) = \{O8\}$ as O8 is closest it will be correctly classified as unsafe.
$N(O10, 1) = \{O4\}$ as O4 is closest it will be correctly classified as safe.
Thus, one out of the ten observations will be misclassified.

Question 16. We will again consider the Euclidean distances between the first ten observations given in Table 3. Agglomerative hierarchical clustering is used to cluster these ten observations based on their distances to each other using average linkage. Which one of the dendrograms given in Figure 6 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 16. As O2 merges last to the cluster containing all the remaining observations we can simply evaluate at what level O2 will merge which is given by O2's average distance to the observations {O1,O3,O4,O5,O6,O7,O8,O9,O10} which is given by $(8.55 + 8.23 + 8.13 + 8.49 + 6.84 + 8.24 + 8.28 + 8.13 + 7.66)/9 = 8.0611$. Only dendrogram 4 has this property.

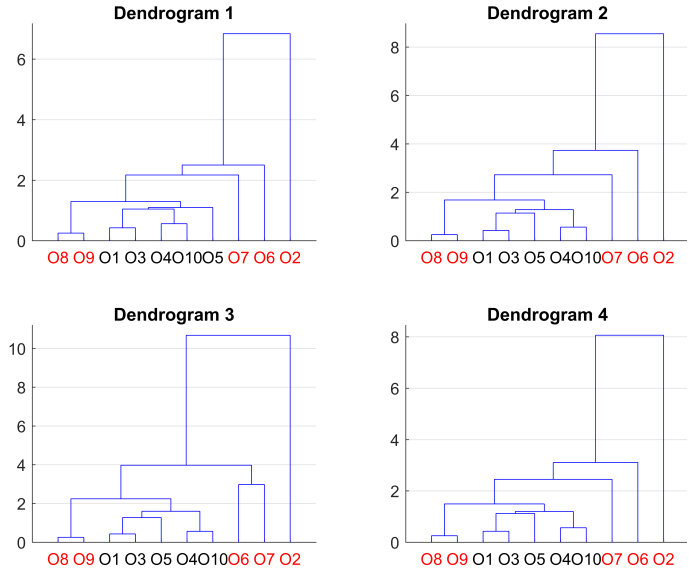

Figure 6: Four different dendrograms derived using the Euclidean distances between the first 10 observations in the airline safety data. Black observations correspond to relatively safe companies whereas red to relatively unsafe companies.

Question 17. We will cut dendrogram 1 at the level of three clusters and evaluate this clustering in terms of its correspondence with the class label information in which black observations, i.e., O1, O3, O4, O5, O10, are observations corresponding to relatively safe airline companies, and red observations, i.e., O2, O6, O7, O8, O9, are observations corresponding to relatively unsafe airline companies. We recall that the Rand index also denoted the simple matching coefficient (SMC) between the true labels and the extracted clusters is given by $R = \frac{f_{11}+f_{00}}{K}$, where f_{11} is the number of object pairs in same class assigned to same cluster, f_{00} is the number of object pairs in different class assigned to different clusters, and $K = N(N-1)/2$ is the total number of object pairs, where N is the number of observations considered. What is the value of R between the true labeling of the observations and the three extracted clusters?

A. 0.40

B. 0.47

C. 0.51

D. 0.60

E. Don't know.

Solution 17. The cluster indices are given by the vector: $[1211131111]^\top$, whereas the true class labels are given by the vector $[1211122221]^\top$. From this, we obtain:
$K = 10(10-1)/2 = 45$
$f_{00} = 5 \cdot 1 + 5 \cdot 1 + 1 \cdot 0 = 10$
$f_{11} = 5{\cdot}(5{-}1)/2{+}3{\cdot}(3{-}1)/2{+}1{\cdot}(1{-}1)/2{+}1{\cdot}(1{-}1)/2 = 13$
$R = \frac{f_{11}+f_{00}}{K} = \frac{13+10}{45} = 23/45.$

Question 18. We suspect that observation O2 may be an outlier. In order to quantify if this could be the case we will calculate the average relative KNN density based on Euclidean distance and the observations given in Table 3 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 O2 for $K = 2$ nearest neighbors?

A. 0.085

B. 0.138

C. 0.169

D. 0.356

E. Don't know.

Solution 18.

$$\text{density}(\boldsymbol{x}_{O2}, 2) = (\frac{1}{2}(6.84 + 7.66))^{-1} = 0.1379$$
$$\text{density}(\boldsymbol{x}_{O6}, 2) = (\frac{1}{2}(2.50 + 2.66))^{-1} = 0.3876$$
$$\text{density}(\boldsymbol{x}_{O10}, 2) = (\frac{1}{2}(0.56 + 1.05))^{-1} = 1.2422$$
$$\text{a.r.d.}(\boldsymbol{x}_{O2}, 2)) = \frac{\text{density}(\boldsymbol{x}_{O2}, 2)}{\frac{1}{2}(\text{density}(\boldsymbol{x}_{O6}, 2) + \text{density}(\boldsymbol{x}_{O10}, 2))} = \frac{0.1379}{\frac{1}{2}(0.3876 + 1.2422)} = 0.169$$

	x_1^L	x_1^H	x_2^L	x_2^H	x_3^L	x_3^H	x_4^L	x_4^H	x_5^L	x_5^H	x_6^L	x_6^H
O1	1	0	1	0	1	0	1	0	1	0	1	0
O2	0	1	0	1	0	1	0	1	0	1	0	1
O3	1	0	0	1	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1	0	1	1	0
O5	0	1	1	0	1	0	1	0	1	0	1	0
O6	0	1	0	1	0	1	0	1	0	1	0	1
O7	0	1	1	0	1	0	0	1	0	1	0	1
O8	1	0	1	0	1	0	1	0	0	1	0	1
O9	0	1	0	1	1	0	1	0	0	1	0	1
O10	1	0	0	1	0	1	0	1	0	1	1	0

Table 4: The ten first observations of the airline safety dataset binarized considering the attribute x_1–x_6. The attributes are all binarized according to being below or equal (denoted L) or above the median value (denoted H). The ten observations are color coded in terms of relatively safe {O1, O3, O4, O5, O10} or unsafe {O2, O6, O7, O8, O9} airline companies.

Question 19. We will binarize each feature in the airline safety data according to the median value of the feature denoting below or equal the median value using the superscript L and above the median value using the superscript H. In Table 4 is given the first 10 observations after this binarization and we will consider these 10 observations as a dataset used for market basket analysis with observation O1–O10 corresponding to customers. What is the support for the association rule $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

Solution 19. The support of $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$ is given by the number of times out of the total number of customers that customers have relatively high values of both x_2, x_3, x_4, x_5, and x_6, i.e., given by the support of the itemset $\{x_2^H, x_3^H, x_4^H, x_5^H, x_6^H\}$. Only customer O2 and O6 have this property out of the 10 customers, thus the support is 2/10.

Question 20. We consider again the data in Table 4 as a market basket problem. What is the confidence of the association rule $\{x_2^H, x_3^H, x_4^H, x_5^H\} \rightarrow \{x_6^H\}$?

A. 0.0 %

B. 20.0 %

C. 66.7 %

D. 100.0 %

E. Don't know.

Solution 20. The confidence is given as

$$P(x_6^H = 1 | x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1) = \frac{P(x_6^H = 1, x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1)}{P(x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1)} = \frac{2/10}{3/10} = 2/3 = 66.7\%$$

Question 21. We would like to predict whether an airline company is relatively safe or unsafe considering only the data given in Table 4. We will apply a Naïve Bayes classifier that assumes independence between the attributes given the class label (i.e., the class label is given by safe airlines in black, i.e., O1, O3, O4, O5, and O10, and unsafe airlines in red, i.e., O2, 06, O7, O8, and O9). Given that an airline company has $x_2^H = 1$, $x_3^H = 1$, $x_4^H = 1$, $x_5^H = 1$ what is the probability that the airline company is considered safe according to the Naïve Bayes classifier derived from the data in Table 4?

A. 0

B. 4/625

C. 4/49

D. 1/3

E. Don't know.

Solution 21. Let $\tilde{y} = 1$ denote that the airline is safe.

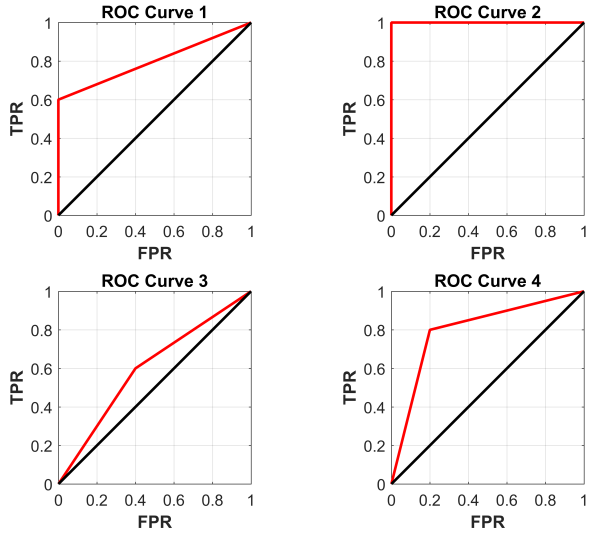

Figure 7: Four different receiver operator characteristic (ROC) curves.

According to the Naïve Bayes classifier we have

$$P(\tilde{y} = 1 | x_2^H = 1, x_3^H = 1, x_4^H = 1, x_5^H = 1) = \frac{\begin{pmatrix} P(x_2^H = 1|\tilde{y} = 1)\times \\ P(x_3^H = 1|\tilde{y} = 1)\times \\ P(x_4^H = 1|\tilde{y} = 1)\times \\ P(x_5^H = 1|\tilde{y} = 1)\times \\ P(\tilde{y} = 1) \end{pmatrix}}{\begin{pmatrix} P(x_2^H = 1|\tilde{y} = 1)\times \\ P(x_3^H = 1|\tilde{y} = 1)\times \\ P(x_4^H = 1|\tilde{y} = 1)\times \\ P(x_5^H = 1|\tilde{y} = 1)\times \\ P(\tilde{y} = 1) \\ +P(x_2^H = 1|\tilde{y} = 0)\times \\ P(x_3^H = 1|\tilde{y} = 0)\times \\ P(x_4^H = 1|\tilde{y} = 0)\times \\ P(x_5^H = 1|\tilde{y} = 0)\times \\ P(\tilde{y} = 0) \end{pmatrix}}$$

$$= \frac{2/5 \cdot 1/5 \cdot 2/5 \cdot 2/5 \cdot 5/10}{2/5 \cdot 1/5 \cdot 2/5 \cdot 2/5 \cdot 5/10 + 3/5 \cdot 2/5 \cdot 3/5 \cdot 5/5 \cdot 5/10}$$

$$= \frac{40/(5^4 \cdot 10)}{40/(5^4 \cdot 10) + 450/(5^4 \cdot 10)} = 40/490 = 4/49$$

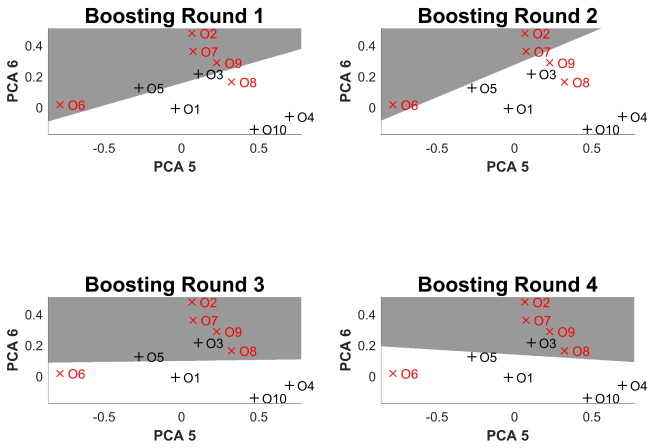

Figure 8: Decision boundaries for four rounds of boosting considering a logistic regression model using the fifth and sixth principal components as features and the first 10 observations of the airline safety data. Gray region indicates that the observation will be classified as unsafe (red crosses), white regions that the observation will be classified as safe (black plusses).

	Round 1	Round 2	Round 3	Round 4
O1	0.1000	0.0714	0.0469	0.0319
O2	0.1000	0.0714	0.0469	0.0319
O3	0.1000	0.1667	0.1094	0.2059
O4	0.1000	0.0714	0.0469	0.0319
O5	0.1000	0.1667	0.1094	0.2059
O6	0.1000	0.0714	0.0469	0.0882
O7	0.1000	0.0714	0.0469	0.0319
O8	0.1000	0.1667	0.3500	0.2383
O9	0.1000	0.0714	0.1500	0.1021
O10	0.1000	0.0714	0.0469	0.0319

Table 5: The weights for the first four rounds of AdaBoost.

Question 22. We will use x_5^L to determine if an airline is safe *considered the positive class* or unsafe *considered the negative class*. In Figure 7 are given four different receiver operator characteristic curves (ROC). Which one of the four ROC curves corresponds to using x_5^L to determine if an airline is safe (positive class) or unsafe (negative class)?

A. ROC curve 1.

B. ROC curve 2.

C. ROC curve 3.

D. ROC curve 4.

E. Don't know.

Solution 22. The ROC curve starts at (0,0) for which we threshold above 1. When tresholding at one we obtain that 3 out of 5 safe airlines (positive class) have $x_5^L = 1$ and 0 out of 5 unsafe (negative class) have $x_5^L = 1$ thus the ROC curve will be at the point (0,3/5). When we subsequently further lower the threshold to be at 0 all airline companies that are safe and unsafe will be at this threshold value or above, thus, the ROC curve will end here at (1,1). Only ROC curve 1 has this property.

Question 23. We would like to build a model for classifying whether and airline is safe or not based on the first 10 observations of the airline safety dataset. In order to do so, we will use the resulting classifier obtained by using the AdaBoost algorithm and a logistic regression classifier. The weights of the first four rounds of the AdaBoost procedure is given in Figure 8 and the associated sampling weights used for each round is given in Table 5. How will observation O5 and O6 be classified according to the ensemble classifier obtained by combining the four boosting rounds using the voting procedure defined by the AdaBoost algorithm?

A. Observation O5 and O6 will be tied between safe and unsafe by the AdaBoost classifier.

B. Both observation O5 and O6 will be correctly classified by the AdaBoost classifier.

C. Only one of the two observations O5 and O6 will be correctly classified by the AdaBoost classifier.

D. Neither of the two observations O5 and O6 will be correctly classified by the AdaBoost classifier.

E. Don't know.

Solution 23. The resulting boosting procedure weight each classifier according to their importance α_t for the t^{th} round, where $\alpha_t = 0.5 \log \frac{1-e_t}{e_t}$ such that $e_t = \sum_n w_n(t)(1 - \delta_{f_t(x_n),y_n})$ is the error rate weighted according to the weights of the t^{th} round. During the first round O3, O5 and O8 are misclassified and thus $e_1 = 0.1 + 0.1 + 0.1 = 0.3$ and consequently $\alpha_1 = 0.5 \log \frac{1-0.3}{0.3} = 0.4236$. In the second round O8 and O9 are mis-classified and thus $e_2 = 0.1667 + 0.0714 = 0.2381$ and therefore $\alpha_2 = 0.5 \log \frac{1-0.2381}{0.2381} = 0.5816$. In the third round O3, O5, and O6 are misclassified thus $e_3 = 0.1094 + 0.1094 + 0.0469 = 0.2657$ and therefore $\alpha_3 = 0.5 \log \frac{1-0.2657}{0.2657} = 0.5083$. Finally, in the fourth round O3 and O6 are misclassified and thus $e_4 = 0.2059 + 0.0882 = 0.2941$ and therefore $\alpha_4 = 0.5 \log \frac{1-0.2657}{0.2657} = 0.4378$. For observation O5 both classifier of round 2 and 4 vote for it being safe with the strength of vote given by $\alpha_2 + \alpha_4 = 0.5816 + 0.4378 = 1.0194$ whereas the strength of vote for unsafe is lower, i.e. $\alpha_1 + \alpha_3 = 0.4236 + 0.5083 = 0.9319$. Observation O6 will be classified as unsafe as classifier 1 and 2 classifies it as unsafe with strength $\alpha_1 + \alpha_2 = 0.4236 + 0.5816 = 1.0052$ and safe according to classifier 3 and 4 with the lower voting strength of $\alpha_3 + \alpha_4 = 0.5083 + 0.4378 = 0.9461$.

Question 24. Four different classifiers are trained on the airline safety data considering only the first 10 observations projected onto the fifth and sixth principal component to determine if an airline is relatively safe or unsafe. The decision boundary for each of the four classifiers is given in Figure 9 when only using as input the data projected onto the fifth (PCA 5) and sixth (PCA 6) principal component, i.e. each method has only these two inputs. Which one of the following statements is correct?

A. Classifier 1 corresponds to a logistic regression classifier, Classifier 2 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 3 is a decision tree classifier, and Classifier 4 corresponds to an artificial neural network (ANN).

B. Classifier 1 is a Naive Bayes classifier based on the use of univariate normal distributions, Classifier 2 is a 3-nearest neighbor classifier, Classifier 3 is a Decision Tree classifier, and Classifier 4 is a 1-nearest neighbor classifier using Euclidean distance.

C. Classifier 1 is a Naive Bayes classifier based on the use of univariate normal distributions, Classifier 2 is a 3-nearest neighbor classifier, Classifier 3 is a 1-nearest neighbor classifier, and Classifier 4 is a Decision Tree classifier.

D. Classifier 1 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 2 is a Naive Bayes classifier based on the use of univariate normal distributions, Classifier 3 is a Decision Tree classifier, and Classifier 4 is a 1-nearest neighbor classifier using Euclidean distance.

E. Don't know.

Solution 24. Classifier 1 is a 3-nearest neighbor classifier as such one red cross and one black plus are within the wrong decision boundary due to the majority voting of the three nearest neighbors. Classifier 2 has smooth decision boundaries which would correspond to the Naive Bayes classifier based on the use of univariate normal distributions. Classifier 3 is a Decision Tree classifier due to its vertical and horizontal decision boundaries. Classifier 4 is a 1-nearest neighbor classifier using Euclidean distance as the decision boundary clearly follows the most nearby observation.

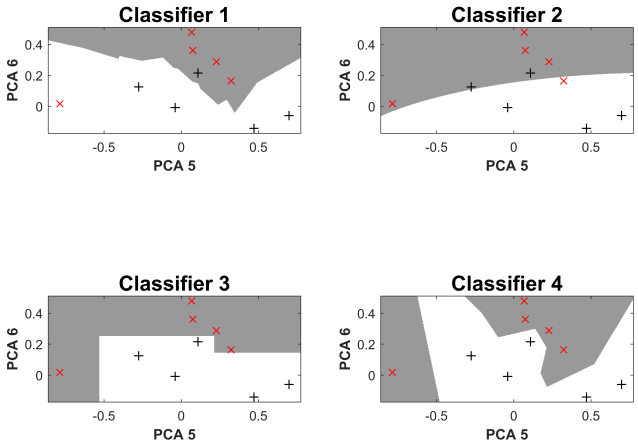

Figure 9: Decision boundaries for four different classifiers trained on the airline safety data using principal component 5 and 6 as input to the classifiers. Gray regions classify into red crosses whereas white regions into black plusses.

Question 25. Consider a dataset with eight observations located at {1.0, 1.2, 1.5, 2.0, 2.2, 2.5, 3.0, 3.2}. We will cluster the dataset using the k-means algorithm using $k = 3$ clusters and initialize the clusters at the locations of the first three observations, i.e. cluster 1 will be initially located at 1.0, cluster 2 at 1.2 and cluster 3 at 1.5. What will be the converged clustering of the eight observations using the k-means procedure based on Euclidean distance as dissimilarity?

A. {1.0}, {1.2, 1.5}, {2.0, 2.2, 2.5, 3.0, 3.2}

B. {1.0, 1.2}, {1.5}, {2.0, 2.2, 2.5, 3.0, 3.2}

C. {1.0, 1.2, 1.5}, {2.0, 2.2, 2.5}, { 3.0, 3.2}

D. {1.0, 1.2, 1.5}, {2.0, 2.2}, {2.5, 3.0, 3.2}

E. Don't know.

Solution 25. The cluster located at 1.5 will be closest to the observations located at 2.0, 2.2, 2.5, 3.0, and 3.2 and will therefore be assigned these whereas cluster located at 1.0 and 1.2 will only be assigned respectively the observation located at 1.0 and 1.2. The location of the centroid for cluster 3 will thereby be changed such that cluster 3 is updated to be located at: $(1.5+2.0+2.2+2.5+3.0+3.2)/6 = 2.4$. Subsequently, observation 1.5 will be closer to cluster located at 1.2 than cluster located at 2.4 and will therefore as only observation change assignment, such that cluster 2 will

be updated to be located at $(1.2 + 1.5)/2 = 1.35$ whereas cluster 3 will be updated to be located at $(2.0+2.2+2.5+3.0+3.2)/5 = 2.58$. As no observation will change assignment based on the location of the updated clusters the algorithm will converge to the clustering given by:
{1.0}, {1.2, 1.5}, {2.0, 2.2, 2.5, 3.0, 3.2}.

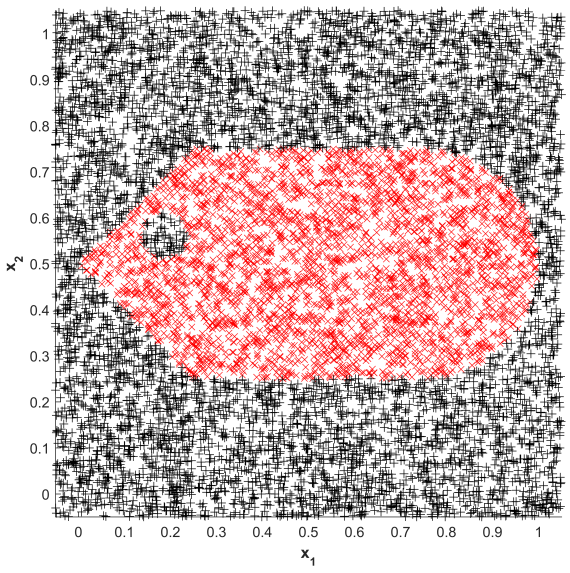

Figure 10: A two class classification problem with red crosses (i.e., x) and black plusses (i.e., +) constituting the two classes.

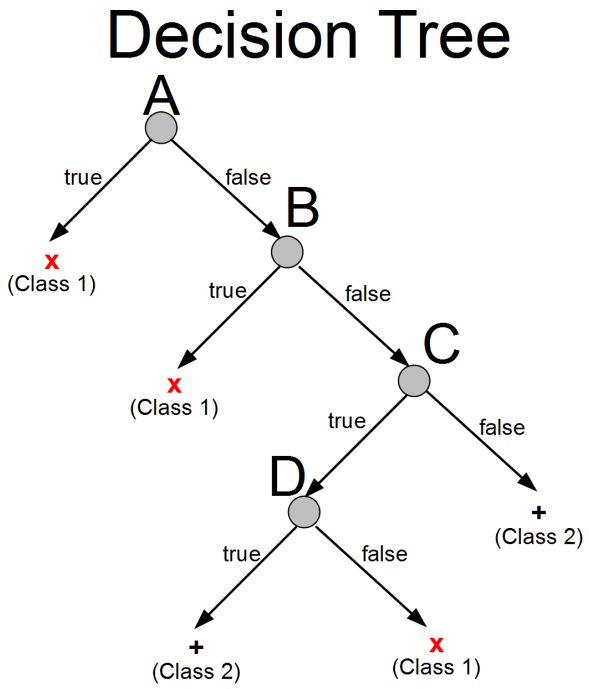

Figure 11: A decision tree with four decisions (A, B, C, and D) perfectly separating the black plusses from red crosses in Figure 10 if adequately defined.

Question 26. We will consider the two class classification problem given in Figure 10 in which the goal is to separate red crosses (i.e., x) from black plusses (i.e., +). Which one of the following procedures based on the decision tree given in Figure 11 will perfectly separate the two classes?

A. $\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathrm{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_2 \le 1/20,$
$\mathrm{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathrm{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/4.$

B. $\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathrm{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_2 \le 1/4,$
$\mathrm{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathrm{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/20.$

C. $\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathrm{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathrm{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_2 \le 1/4,$
$\mathrm{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/20.$

D. $\mathbf{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_\infty \le 1/4,$
$\mathbf{B}=\|\boldsymbol{x}-\begin{bmatrix}3/4\\1/2\end{bmatrix}\|_2 \le 1/4,$
$\mathbf{C}=\|\boldsymbol{x}-\begin{bmatrix}1/4\\1/2\end{bmatrix}\|_1 \le 1/4,$
$\mathbf{D}=\|\boldsymbol{x}-\begin{bmatrix}3/16\\9/16\end{bmatrix}\|_2 \le 1/20.$

E. Don't know.

Solution 26. The red crosses contains a circle located at (0.75,0.5) and a square located at (0.5, 0.5) and a diamond located at (0.25, 0.5) all with radius 0.25 corresponding to:
$\mathrm{A}=\|\boldsymbol{x}-\begin{bmatrix}1/2\\1/2\end{bmatrix}\|_\infty \le 1/4,$

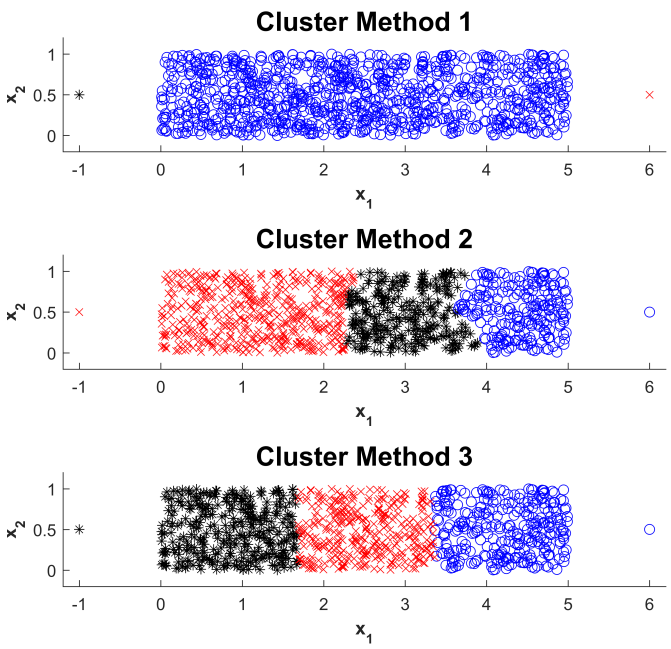

Figure 12: A dataset with two features x_1 and x_2 and 100 observations clustered using three different clutering approaches.

B=$\|\boldsymbol{x} - \begin{bmatrix} 3/4 \\ 1/2 \end{bmatrix}\|_2 \leq 1/4$,

C=$\|\boldsymbol{x} - \begin{bmatrix} 1/4 \\ 1/2 \end{bmatrix}\|_1 \leq 1/4$.

Finally, there are black plusses located at (0.1875, 0.5625) with a small radius of 0.05 corresponding to:

D=$\|\boldsymbol{x} - \begin{bmatrix} 3/16 \\ 9/16 \end{bmatrix}\|_2 \leq 1/20$.

This only holds for the last answer option.

Question 27. We will consider a dataset with two features x_1 and x_2 and 1000 observations that is clustered using three different clustering approaches all based on Euclidean distance as measure of distance. The clustering extracted by each of the three considered approaches are given in Figure 12. Which one of the following statements is correct?

A. Cluster Method 1 corresponds to k-means,
Cluster Method 2 corresponds to hierarchical clustering using single linkage,
Cluster method 3 corresponds to hierarchical clustering using complete linkage.

B. Cluster Method 1 corresponds to hierarchical clustering using single linkage,
Cluster Method 2 corresponds to k-means,
Cluster method 3 corresponds to hierarchical clustering using complete linkage.

**C. Cluster Method 1 corresponds to hierarchical clustering using single linkage,
Cluster method 2 corresponds to hierarchical clustering using complete linkage,
Cluster Method 3 corresponds to k-means.**

D. Cluster Method 1 corresponds to hierarchical clustering using complete linkage,
Cluster method 2 corresponds to hierarchical clustering using single linkage,
Cluster Method 3 corresponds to k-means.

E. Don't know.

Solution 27. Single linkage clusters consecutively by merging according to the two observations of each cluster that is closest to each other. As such, the clustering will be influenced by the gaps between the observations and therefore the two outlying observations at (-1,0.5) and (6,0.5) will be merged the latest in the dendrogram resulting in the three clusters given by Cluster Method 1. Complete linkage clusters consecutively by merging according to the two observations of each cluster that is the furthest apart. As such, the clustering will be influenced by how far the cluster extends as well as influenced by earlier merge decisions corresponding to the clustering given by Cluster Method 2. K-means will cluster observations according to their proximity to the center of the cluster which corresponds to a clustering given by Cluster Method 3. Neither Cluster Method 1

and Cluster Method 2 can be k-means as the distance to the centroid of each cluster would result in a different clustering configuration than the ones obtained. Furthermore, cluster method 2 cannot be single linkage as the major gaps to (-1,0.5) and (6,0.5) will make these merge latest in the dendrogram. Thus the only correct answer option is:
Cluster Method 1 corresponds to hierarchical clustering using single linkage,
Cluster method 2 corresponds to hierarchical clustering using complete linkage,
Cluster Method 3 corresponds to k-means.

Technical University of Denmark

Written examination: December 18th 2018, 9 AM - 1 PM.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

No.	Attribute description	Abbrev.
x_1	intercolumnar distance	interdist
x_2	upper margin	upperm
x_3	lower margin	lowerm
x_4	exploitation	exploit
x_5	row number	row nr.
x_6	modular ratio	modular
x_7	interlinear spacing	interlin
x_8	weight	weight
x_9	peak number	peak nr.
x_{10}	modular ratio/ interlinear spacing	mr/is
y	Who copied the text?	Copyist

Table 1: Description of the features of the Avila Bible dataset used in this exam. The dataset has been extracted from images of the 'Avila Bible', an XII century giant Latin copy of the Bible. The prediction task consists in associating each pattern to one of three copyist (copyist refers to the monk who copied the text in the bible), indicated by the y-value. Note that only a subset of the dataset is used. The dataset used here consist of $N = 525$ observations and the attribute y is discrete taking values $y = 1,\ 2,\ 3$ corresponding to the three different copyists.

Question 1.

The main dataset used in this exam is the Avila Bible dataset[1] shown in Table 1.

In Figure 1 and Figure 2 are shown respectively percentile plots and boxplots of the Avila Bible dataset based on the attributes $x_2,\ x_3,\ x_9,\ x_{10}$ found in Table 1. Which percentile plots match which boxplots?

A. Boxplot 1 is mr/is, Boxplot 2 is lowerm, Boxplot 3 is upperm and Boxplot 4 is peak nr.

B. Boxplot 1 is upperm, Boxplot 2 is lowerm, Boxplot 3 is peak nr. and Boxplot 4 is mr/is

C. Boxplot 1 is upperm, Boxplot 2 is peak nr., Boxplot 3 is mr/is and Boxplot 4 is lowerm

D. Boxplot 1 is mr/is, Boxplot 2 is lowerm, Boxplot 3 is peak nr. and Boxplot 4 is upperm

E. Don't know.

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

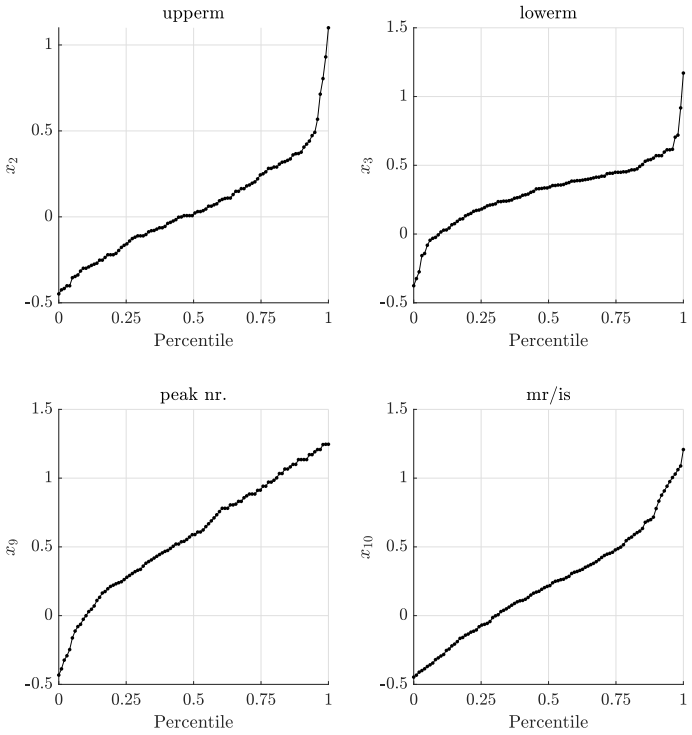

Figure 1: Plot of observations $x_2,\ x_3,\ x_9,\ x_{10}$ of the Avila Bible dataset of Table 1 as percentile plots.

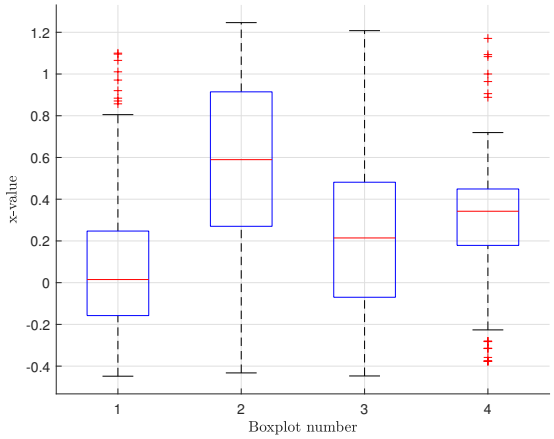

Figure 2: Boxplots corresponding to the variables plotted in Figure 1 but not necessarily in that order.

Solution 1. The correct answer is C. To see this, recall that by the definition of a boxplot the horizontal red line indicates the 50th percentile. We can read these of from the percentile plots by observing the values corresponding to 0.5. These are:

$$x_2 = 0.0, \quad x_3 = 0.3, \quad x_9 = 0.6, \quad x_{10} = 0.2.$$

In a similar manner, we know the upper-part of the box must correspond to the 75th percentile. These can also be read of from the percentile plots (the value corresponding to 0.75) and are:

$$x_2 = 0.2, \quad x_3 = 0.4, \quad x_9 = 0.9, \quad x_{10} = 0.5.$$

Taken together these rule out all but option C.

Question 2.

A Principal Component Analysis (PCA) is carried out on the Avila Bible dataset in Table 1 based on the attributes $x_1,\ x_3,\ x_5,\ x_6,\ x_7$.

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

$$\boldsymbol{V} = \begin{bmatrix} 0.04 & -0.12 & -0.14 & 0.35 & 0.92 \\ 0.06 & 0.13 & 0.05 & -0.92 & 0.37 \\ -0.03 & -0.98 & 0.08 & -0.16 & -0.05 \\ -0.99 & 0.03 & 0.06 & -0.02 & 0.07 \\ -0.07 & -0.05 & -0.98 & -0.11 & -0.11 \end{bmatrix} \tag{1}$$

$$\boldsymbol{S} = \begin{bmatrix} 14.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 8.19 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 7.83 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 6.91 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 6.01 \end{bmatrix}$$

Which one of the following statements is true?

A. The variance explained by the first principal component is greater than 0.45

B. The variance explained by the first four principal components is less than 0.85

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

D. The variance explained by the first three principal components is less than 0.75

E. Don't know.

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

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

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

$$\text{Var.Expl.} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_3^2 + \sigma_5^2 + \sigma_6^2 + \sigma_7^2} = 0.4942.$$

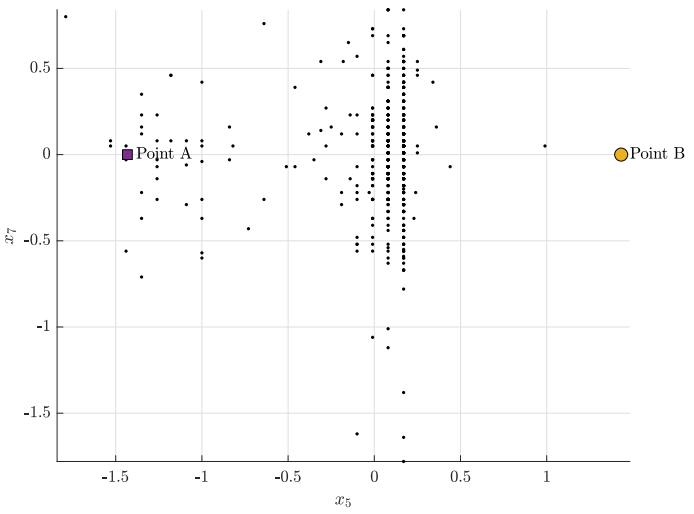

Figure 3: Black dots show attributes x_5 and x_7 of the Avila Bible dataset from Table 1. The two points corresponding to the colored markers indicate two specific observations A, B.

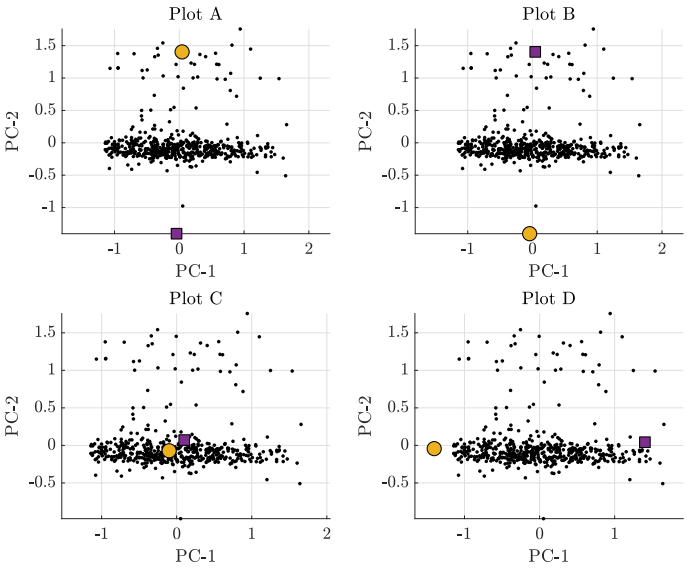

Figure 4: Candidate plots of the observations and path shown in Figure 3 projected onto the first two principal components considered in Equation (1). The colored markers still refer to points A and B, now in the coordinate system corresponding to the PCA projection.

Question 3.

Consider again the PCA analysis fo the Avila Bible dataset. In Figure 3 the features x_5 and x_7 from Table 1 are plotted as black dots. We have indicated two special observations as colored markers (Point A and Point B).

We can imagine that the dataset, along with the two special observations, is projected onto the first two principal component directions given in $\boldsymbol{V}$ as computed earlier (see Equation (1)). Which one of the four plots in Figure 4 shows the correct PCA projection?

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 3. Since we don't know the exact values of most of the x_i-coordinates, it is easier to work with the difference between observation A and B in Figure 3 and translate them into the difference in the PCA projections. Notice from Figure 3 we can immediately compute:

$$\Delta \boldsymbol{x} = \boldsymbol{x}_{\text{end}} - \boldsymbol{x}_{\text{start}} = \begin{bmatrix} 0.0 \\ 0.0 \\ 2.86 \\ 0.0 \\ 0.0 \end{bmatrix}$$

(this corresponds to the vector going from Point A to Point B). Then, all we need is to compute the PCA projection of this vector as:

$$\Delta \boldsymbol{b} = \left((\Delta \boldsymbol{x})^\top \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \right)^\top = \begin{bmatrix} -0.09 \\ -2.8 \end{bmatrix}$$

Which should be the vector beginning at Point A and terminating at B in the PCA projected plots. This rules out all plots except option B.

Question 4. To examine if observation o_4 may be an outlier, we will calculate the average relative density based on euclidean distance and the observations given in Table 2 only. We recall that the KNN density and average relative density (ard) for the observation $\boldsymbol{x}_i$ are given by:

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

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

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors. What is the average relative density for observation o_4 for $K = 2$ nearest neighbors?

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	2.91	0.63	1.88	1.02	1.82	1.92	1.58	1.08	1.43
o_2	2.91	0.0	3.23	3.9	2.88	3.27	3.48	4.02	3.08	3.47
o_3	0.63	3.23	0.0	2.03	1.06	2.15	2.11	1.15	1.09	1.65
o_4	1.88	3.9	2.03	0.0	2.52	1.04	2.25	2.42	2.18	2.17
o_5	1.02	2.88	1.06	2.52	0.0	2.44	2.38	1.53	1.71	1.94
o_6	1.82	3.27	2.15	1.04	2.44	0.0	1.93	2.72	1.98	1.8
o_7	1.92	3.48	2.11	2.25	2.38	1.93	0.0	2.53	2.09	1.66
o_8	1.58	4.02	1.15	2.42	1.53	2.72	2.53	0.0	1.68	2.06
o_9	1.08	3.08	1.09	2.18	1.71	1.98	2.09	1.68	0.0	1.48
o_{10}	1.43	3.47	1.65	2.17	1.94	1.8	1.66	2.06	1.48	0.0

Table 2: The pairwise Euclidian distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 10 observations from the Avila Bible dataset (recall $M = 10$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1 (the data has been standardized). The colors indicate classes such that the black observations $\{o_1,\ o_2,\ o_3\}$ belongs to class C_1 (corresponding to copyist one), the red observations $\{o_4,\ o_5,\ o_6,\ o_7,\ o_8\}$ belongs to class C_2 (corresponding to copyist two), and the blue observations $\{o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to copyist three).

A. 1.0

B. 0.71

C. 0.68

D. 0.36

E. Don't know.

Solution 4.

To solve the problem, first observe the $k = 2$ neighborhood of o_4 and density is:

$$N_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = \{o_6, o_1\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = 0.685$$

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

$$N_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = \{o_4, o_{10}\}, \quad N_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = \{o_3, o_5\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = 0.704, \text{density}_{\boldsymbol{X}_{\setminus 1}}(\boldsymbol{x}_1) = 1.212.$$

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

Question 5.

Suppose a GMM model is applied to the Avila Bible dataset in the processed version shown in Table 2. The GMM is constructed as having $K = 3$ components, and each component k of the GMM is fitted by letting it's mean vectors $\boldsymbol{\mu}_k$ be equal to the location of the observations:

$$o_7, \; o_8, \; o_9$$

(i.e. each observation corresponds to exactly one mean vector) and setting the covariance matrix equal to $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix:

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

where $|\cdot|$ is the determinant. The components of the GMM are weighted evenly.

If $\sigma = 0.5$, and denoting the density of the GMM as $p(\boldsymbol{x})$, what is the density as evaluated at observation o_3?

A. $p(o_3) = 0.048402$

B. $p(o_3) = 0.076$

C. $p(o_3) = 0.005718$

D. $p(o_3) = 0.114084$

E. Don't know.

Solution 5.

Since the mixture components are weighted equally, the density of the test observation becomes:

$$p(\boldsymbol{o}_i) = \sum_{k=1}^{3} \frac{1}{3} \mathcal{N}(\boldsymbol{o}_i | \boldsymbol{\mu}_k, \sigma^2 \boldsymbol{I}).$$

To correctly evaluate this density, we also need to know the dimensionality of the multivariate normal distributions. This can be found in Table 2 to be $M = 10$. The density of a single mixture component is therefore:

$$\mathcal{N}(\boldsymbol{o}_i | \boldsymbol{o}_j, \sigma^2 \boldsymbol{I}) = \frac{1}{(2\pi\sigma^2)^{\frac{10}{2}}} e^{-\frac{d(\boldsymbol{o}_i, \boldsymbol{o}_j)^2}{2\sigma^2}}.$$

where the distances can be found in Table 2. Plugging in the values we see option C is correct.

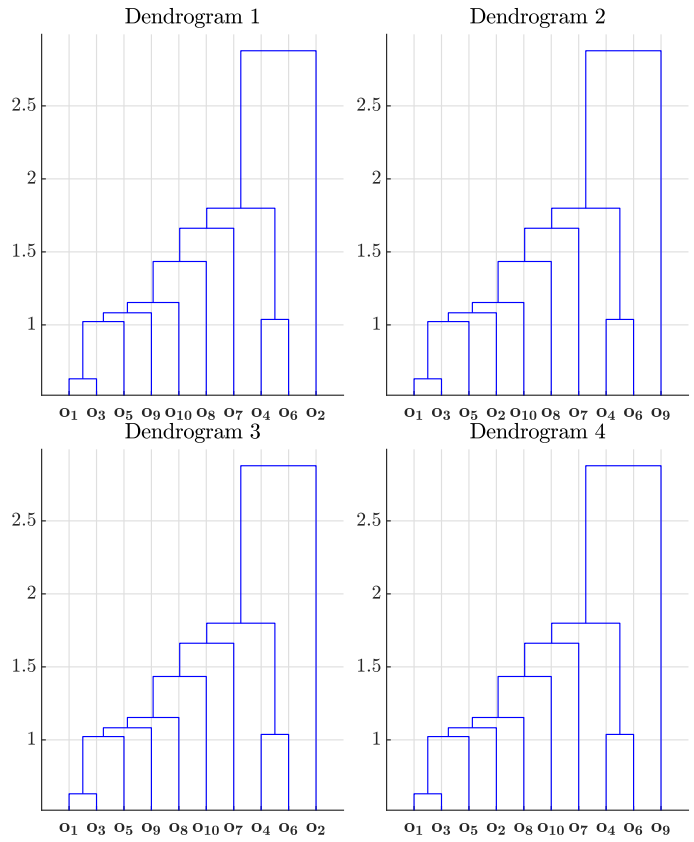

Figure 5: Proposed hierarchical clustering of the 10 observations in Table 2.

Question 6. A hierarchical clustering is applied to the 10 observations in Table 2 using *minimum* linkage. Which of the dendrograms shown in Figure 5 corresponds to the clustering?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

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

- In dendrogram 1, merge operation number 5 should have been between the sets $\{f_8\}$ and $\{f_9, f_5, f_1, f_3\}$ at a height of 1.15, however in dendrogram 1 merge number 5 is between the sets $\{f_{10}\}$ and $\{f_9, f_5, f_1, f_3\}$.
- In dendrogram 2, merge operation number 4 should have been between the sets $\{f_9\}$ and $\{f_5, f_1, f_3\}$ at a height of 1.08, however in dendrogram 2 merge number 4 is between the sets $\{f_2\}$ and $\{f_5, f_1, f_3\}$.
- In dendrogram 4, merge operation number 4 should have been between the sets $\{f_9\}$ and $\{f_5, f_1, f_3\}$ at a height of 1.08, however in dendrogram 4 merge number 4 is between the sets $\{f_2\}$ and $\{f_5, f_1, f_3\}$.

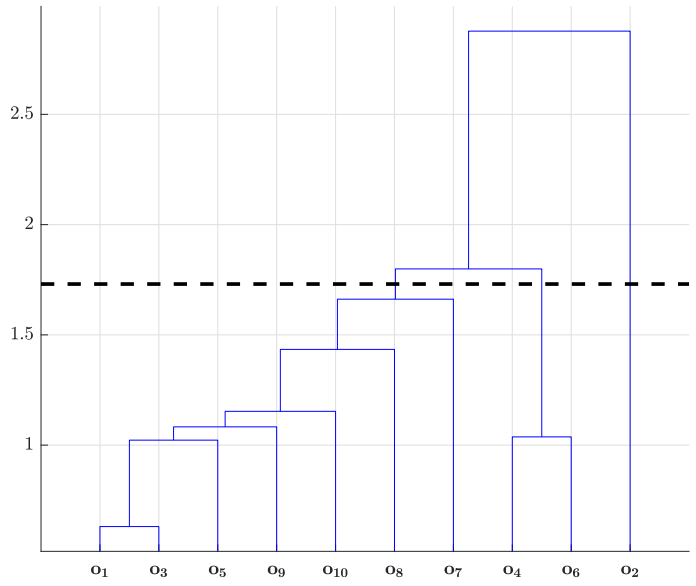

Figure 6: Dendrogram 1 from Figure 5 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

Question 7.

Consider dendrogram 1 from Figure 5. Suppose we apply a cutoff (indicated by the black line) thereby generating three clusters. We wish to compare the quality of this clustering, Q, to the ground-truth clustering, Z, indicated by the colors in Table 2. Recall the *normalized mutual information* of the two clusterings Z and Q is defined as

$$\text{NMI}[Z, Q] = \frac{\text{MI}[Z, Q]}{\sqrt{H[Z]}\sqrt{H[Q]}}$$

where MI is the *mutual information* and H is the entropy. Assuming we always use an entropy based on the natural logarithm,

$$H = -\sum_{i=1}^{n} p_i \log p_i, \quad \log(e) = 1,$$

what is the normalized mutual information of the two clusterings?

A. $\text{NMI}[Z, Q] \approx 0.313$

B. $\text{NMI}[Z, Q] \approx 0.302$

C. $\text{NMI}[Z, Q] \approx 0.32$

D. $\text{NMI}[Z, Q] \approx 0.274$

E. Don't know.

Solution 7. To compute the MI, we will use the relation

$$\text{MI}[p, q] = H[P] + H[Q] - H[P, Q]$$

Where P is the clustering corresponding to the colors in Table 2 and Q the clustering obtained by cutting the dendrogram in Figure 6:

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

From this information we can define the matrix of probabilities $p(i, j)$ such that

$$p(i, j) = \frac{\text{Observations in cluster } i \text{ in } P \text{ and } j \text{ in } Q}{N}$$
$$= \frac{1}{N}\begin{bmatrix} 0 & 2 & 1 \\ 2 & 3 & 0 \\ 0 & 2 & 0 \end{bmatrix}$$

From these we can define the probabilities corresponding to the clustering of P and Q as: $p_P(i) = \sum_j p(i, j)$, $p_Q(j) = \sum_i p(i, j)$. The mutual information is then

$$\begin{aligned} \text{MI} &= H[P] + H[Q] - H[P, Q] \\ &= 0.802 + 1.03 - 1.557 \\ &= 0.274. \end{aligned}$$

We can then simply use the equation for the NMI given in the problem to see answer B is correct.

Question 8. Consider the distances in Table 2 based on 10 observations from the Avila Bible dataset. The class labels C_1, C_2, C_3 (see table caption for details) will be predicted using a k-nearest neighbour classifier based on the distances given in Table 2. Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 1-nearest neighbour classifier (i.e. $k = 1$). What is the error rate computed for all $N = 10$ observations?

A. error rate $= \frac{4}{10}$

B. error rate $= \frac{9}{10}$

C. error rate $= \frac{2}{10}$

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

E. Don't know.

Solution 8.

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

The model considered is KNN classifier with $k = 1$. To figure out the error for a particular observation i (i.e. the test set for this fold), we train a model on the other observations and predict on observation i. To do that, simply find the observation different than i closest to i according to Table 2 and predict i as belonging to it's class. Concretely, we find: $N(o_i, k) = \{o_3\}$, $N(o_i, k) = \{o_5\}$, $N(o_i, k) = \{o_1\}$, $N(o_i, k) = \{o_6\}$, $N(o_i, k) = \{o_1\}$, $N(o_i, k) = \{o_4\}$, $N(o_i, k) = \{o_{10}\}$, $N(o_i, k) = \{o_3\}$, $N(o_i, k) = \{o_1\}$, and $N(o_i, k) = \{o_1\}$.

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

$$\{o_1, o_3, o_4, o_6\}$$

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

Question 9.

Suppose we wish to build a classification tree based on Hunt's algorithm where the goal is to predict Copyist which can belong to three classes, $y = 1$, $y = 2$, $y = 3$. The first split we consider is a two-way split based on the value of x_9 into the intervals indicated in Table 3. For each interval, we count how many observations belong to each of the three classes and the result is indicated in Table 3. Suppose we use the *classification error* impurity measure, what is then the purity gain Δ?

x_9-interval	$y = 1$	$y = 2$	$y = 3$
$x_9 \leq 0.13$	108	112	56
$0.13 < x_9$	58	75	116

Table 3: Proposed split of the Avila Bible dataset based on the attribute x_9. We consider a 2-way split where for each interval we count how many observations belonging to that interval has the given class label.

A. $\Delta \approx 0.485$

B. $\Delta \approx 0.078$

C. $\Delta \approx 0.566$

D. $\Delta \approx 1.128$

E. Don't know.

Solution 9. Recall the information gain Δ is given as:

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

These quantities are easiest computed by forming the matrix R_{ki}, defined as the number of observations in split k belonging to class i:

$$R = \begin{bmatrix} 108 & 112 & 56 \\ 58 & 75 & 116 \end{bmatrix}.$$

We obtain $N(r) = \sum_{ki} R_{ki} = 525$ as the total number of observations and the number of observations in each branch is simply:

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

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

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

and the impurity I_0 from

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

In particular we obtain:

$$I_0 = 0.644, I(v_1) = 0.594,\ I(v_2) = 0.534.$$

Combining these we see that $\Delta = 0.078$ and therefore option B is correct.

Question 10. Consider the split in Table 3. Suppose we build a classification tree with *only* this split and evaluate it on the same data it was trained on. What is the accuracy?

A. Accuracy is: 0.64

B. Accuracy is: 0.29

C. Accuracy is: 0.35

D. Accuracy is: 0.43

E. Don't know.

Solution 10. We will first form the matrix R_{ki}, defined as the number of observations in split k belonging to class i:

$$R = \begin{bmatrix} 108 & 112 & 56 \\ 58 & 75 & 116 \end{bmatrix}.$$

From this we obtain $N = \sum_{ki} R_{ki} = 525$ as the total number of observations. For each split, the number of observations in the largest classes, n_k, is:

$$n_1 = \max_i R_{ik} = 112,\ n_2 = \max_i R_{ik} = 116.$$

Therefore, the accuracy is:

$$\text{Accuracy:} \quad \frac{112 + 116}{525}$$

and answer D is correct.

Question 11. Suppose s_1 and s_2 are two text documents containing the text:

$$s_1 = \begin{Bmatrix} \text{the bag of words representation} \\ \text{should not give you a hard time} \end{Bmatrix}$$
$$s_2 = \begin{Bmatrix} \text{remember the representation should} \\ \text{be a vector} \end{Bmatrix}$$

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

A. cosine similarity of s_1 and s_2 is 0.047619

B. cosine similarity of s_1 and s_2 is 0.000044

C. cosine similarity of s_1 and s_2 is 0.000400

D. cosine similarity of s_1 and s_2 is 0.436436

E. Don't know.

Solution 11. The correct answer is D. Since we are computing the cosine similarity, the length of the vocabulary is irrelevant. We then observe that document s_1 contains $n_1 = 12$ unique words and document s_2 contains $n_2 = 7$ unique words, and the two documents have $f_{11} = 4$ words in common. The cosine similarity is therefore:

$$\cos(s_1, s_2) = \frac{f_{11}}{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|} = \frac{f_{11}}{\sqrt{n_1}\sqrt{n_2}} \approx 0.44.$$

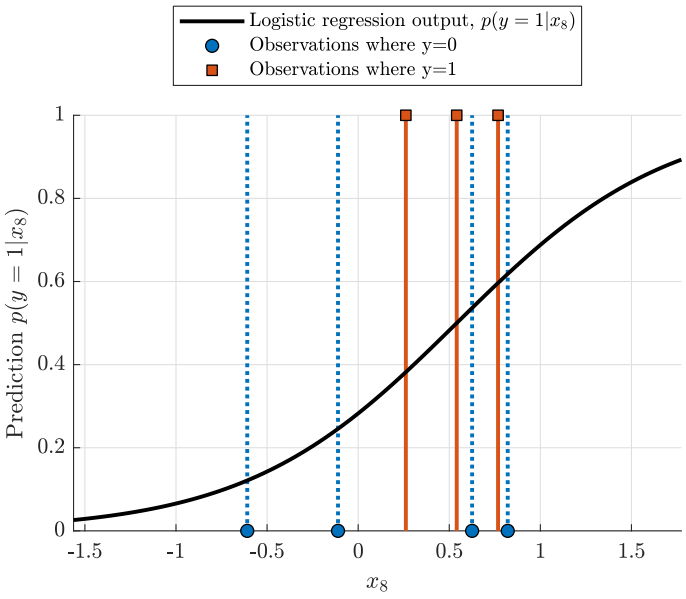

Figure 7: Output of a logistic regression classifier trained on 7 observations from the dataset.

Question 12. Consider again the Avila Bible dataset. We are particularly interested in predicting whether a bible copy was written by copyist 1, and we therefore wish to train a logistic regression classifier to distinguish between copyist one vs. copyist two and three.

To simplify the setup further, we select just 7 observations and train a logistic regression classifier using only the feature x_8 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). To be consistent with the lecture notes, we label the output as $y = 0$ (corresponding to copyist one) and $y = 1$ (corresponding to copyist two and three).

In Figure 7 is shown the predicted output probability an observation belongs to the positive class, $p(y = 1|x_8)$. What are the weights?

A. $\begin{bmatrix} -0.93 \\ 1.72 \end{bmatrix}$

B. $\begin{bmatrix} -2.82 \\ 0.0 \end{bmatrix}$

C. $\begin{bmatrix} 1.36 \\ 0.4 \end{bmatrix}$

D. $\begin{bmatrix} -0.65 \\ 0.0 \end{bmatrix}$

E. Don't know.

Solution 12. The solution is easily found by simply computing the predicted $\hat{y} = p(y = 1|x_8)$-value for an appropriate choice of x_8. Notice that

$$p(y = 1|x_8) = \sigma(\tilde{\boldsymbol{x}}_8^T \boldsymbol{w})$$

If we select $x_8 = 1$ and select the weights as in option A we find $p(y = 1|x_8) = 0.69$, in good agreement with the figure. On the other hand, for the weights in option C we obtain $\hat{y} = 0.85$, for D that $\hat{y} = 0.34$ and finally for B that $\hat{y} = 0.06$. We can therefore conclude that A is correct.

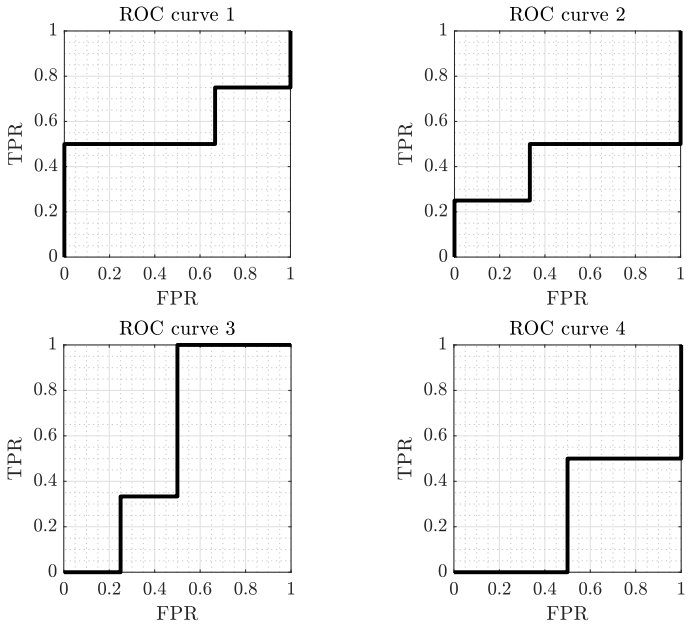

Figure 8: Proposed ROC curves for the logistic regression classifier in Figure 7.

Question 13.

To evaluate the classifier Figure 7, we will use the *area under curve* (AUC) of the *reciever operator characteristic* (ROC) curve as computed on the 7 observations in Figure 7. In Figure 8 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.

Solution 13. 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

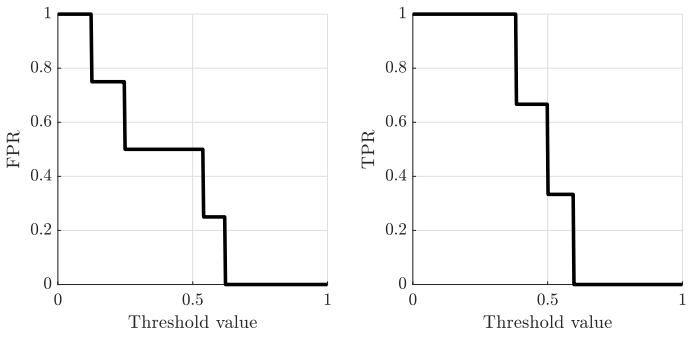

Figure 9: TPR, FPR curves for the logistic regression classifier in Figure 7.

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 9. 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 C.

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

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

Question 14. We again consider the Avila Bible dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce 10 new, binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[2], and we thereby arrive at the $N \times M = 10 \times 10$ binary matrix in Table 4. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_1,\ f_2,\ f_6$. If for an observations we observe

$$f_1 = 1,\ f_2 = 1,\ f_6 = 0$$

what is then the probability that $y = 1$ according to the Naïve-Bayes classifier?

A. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{50}{77}$

B. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{25}{43}$

C. $p_{\mathbf{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{5}{11}$

D. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{10}{19}$

E. Don't know.

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

Solution 14. To solve this problem, we simply use the general form of the naïve-Bayes approximation and plug in the relevant numbers. We get:

$$p_{\text{NB}}(y=1|f_1=1,\ f_2=1,\ f_6=0) =$$
$$\frac{p(f_1=1|y=1)p(f_2=1|y=1)p(f_6=0|y=1)p(y=1)}{\sum_{j=1}^{3} p(f_1=1|y=j)p(f_2=1|y=j)p(f_6=0|y=j)p(y=j)}$$
$$= \frac{\frac{1}{1}\frac{2}{3}\frac{1}{3}\frac{3}{10}}{\frac{1}{1}\frac{2}{3}\frac{1}{3}\frac{3}{10} + \frac{2}{5}\frac{1}{1}\frac{2}{5}\frac{1}{2} + \frac{1}{2}\frac{0}{1}\frac{0}{1}\frac{1}{5}}$$
$$= \frac{5}{11}.$$

Therefore, answer C is correct.

Question 15.

Consider the binarized version of the Avila Bible dataset shown in Table 4.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 10$ items $f_1, f_2, \ldots, f_{10}$. Which of the following options represents all (non-empty) itemsets with support greater than 0.55 (and only itemsets with support greater than 0.55)?

A. $\{f_1\}$, $\{f_2\}$, $\{f_6\}$, $\{f_7\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_6, f_{10}\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_6\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_6\}$, $\{f_7\}$, $\{f_8\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_2\}$, $\{f_2, f_3\}$, $\{f_2, f_4\}$, $\{f_3, f_4\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_2, f_7\}$, $\{f_3, f_7\}$, $\{f_6, f_7\}$, $\{f_2, f_8\}$, $\{f_3, f_8\}$, $\{f_7, f_8\}$, $\{f_2, f_9\}$, $\{f_3, f_9\}$, $\{f_6, f_9\}$, $\{f_7, f_9\}$, $\{f_8, f_9\}$, $\{f_1, f_{10}\}$, $\{f_2, f_{10}\}$, $\{f_6, f_{10}\}$, $\{f_2, f_3, f_4\}$, $\{f_1, f_2, f_6\}$, $\{f_2, f_3, f_7\}$, $\{f_2, f_3, f_8\}$, $\{f_2, f_3, f_9\}$, $\{f_6, f_7, f_9\}$, $\{f_2, f_8, f_9\}$, $\{f_3, f_8, f_9\}$, $\{f_7, f_8, f_9\}$, $\{f_1, f_2, f_{10}\}$, $\{f_1, f_6, f_{10}\}$, $\{f_2, f_6, f_{10}\}$, $\{f_2, f_3, f_8, f_9\}$, $\{f_1, f_2, f_6, f_{10}\}$

D. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_6\}$, $\{f_7\}$, $\{f_8\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_2\}$, $\{f_2, f_3\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_6, f_7\}$, $\{f_7, f_9\}$, $\{f_8, f_9\}$, $\{f_2, f_{10}\}$, $\{f_6, f_{10}\}$, $\{f_1, f_2, f_6\}$, $\{f_2, f_6, f_{10}\}$

E. Don't know.

Solution 15. 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 0.55, an itemset needs to be contained in 6 rows. It is easy to see this rules out all options except *B*.

Question 16. We again consider the binary matrix from Table 4 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 10$ items $f_1, \ldots, f_{10}$.

What is the *confidence* of the rule $\{f_1, f_3, f_8, f_9\} \to \{f_2, f_6, f_7\}$

A. Confidence is $\frac{1}{10}$

B. Confidence is 1

C. Confidence is $\frac{1}{2}$

D. Confidence is $\frac{3}{20}$

E. Don't know.

Solution 16. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_1, f_3, f_8, f_9\} \cup \{f_2, f_6, f_7\})}{\text{support}(\{f_1, f_3, f_8, f_9\})} = \frac{\frac{1}{10}}{\frac{1}{10}} = 1.$$

Therefore, answer B is correct.

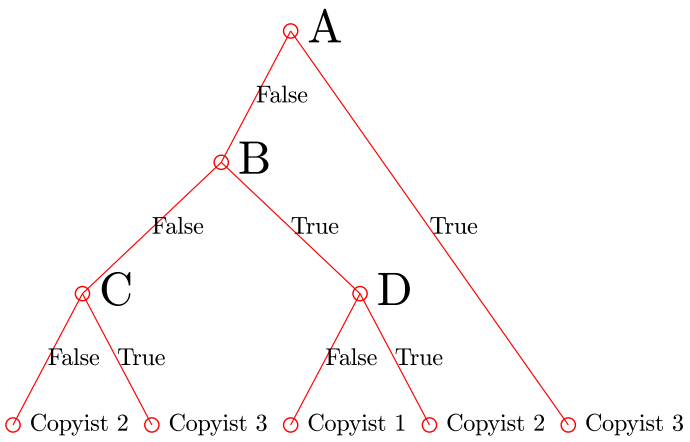

Figure 10: Example classification tree.

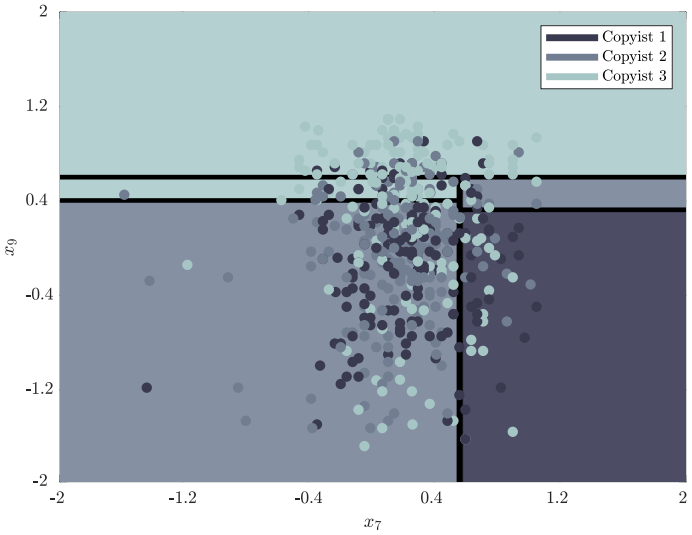

Figure 11: classification boundary.

Question 17.

Consider again the Avila Bible dataset. Suppose we train a decision tree to classify which of the 3 classes, Copyist 1, Copyist 2, Copyist 3, an observation belongs to. Since the attributes of the dataset are continuous, we will consider binary splits of the form $x_i \geq z$ for different values of i and z, and for simplicity we limit ourselves to the attributes x_7 and x_9. Suppose the trained decision tree has the form shown in Figure 10, and that according to the tree the predicted label assignment for the $N = 525$ observations are as given in Figure 11, what is then the correct rule assignment

to the nodes in the decision tree?

A. ***A***: $x_7 \geq 0.5$, ***B***: $x_9 \geq 0.54$, ***C***: $x_9 \geq 0.35$, ***D***: $x_9 \geq 0.26$

B. ***A***: $x_7 \geq 0.5$, ***B***: $x_9 \geq 0.26$, ***C***: $x_9 \geq 0.54$, ***D***: $x_9 \geq 0.35$

C. *A*: $x_9 \geq 0.54$, *B*: $x_7 \geq 0.5$, *C*: $x_9 \geq 0.35$, *D*: $x_9 \geq 0.26$

D. ***A***: $x_9 \geq 0.26$, ***B***: $x_7 \geq 0.5$, ***C***: $x_9 \geq 0.35$, ***D***: $x_9 \geq 0.54$

E. Don't know.

Solution 17.

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 12 and it follows answer C is correct.

Question 18. We will again consider the binarized version of the Avila Bible dataset already encountered in Table 4, however we will now only consider the first $M = 6$ features f_1, f_2, f_3, f_4, f_5, f_6.

We wish to apply the Apriori algorithm (the specific variant encountered in chapter 19 of the lecture notes) to find all itemsets with support greater than $\varepsilon = 0.15$. Suppose at iteration $k = 3$ we know that:

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

Recall the key step in the Apriori algorithm is to construct L_3 by first considering a large number of candidate itemsets C'_3, and then rule out some of them using the downwards-closure principle thereby saving many (potentially costly) evaluations of support. Suppose L_2 is given as above, which of the following itemsets does the Apriori algorithm *not* have to evaluate the support of?

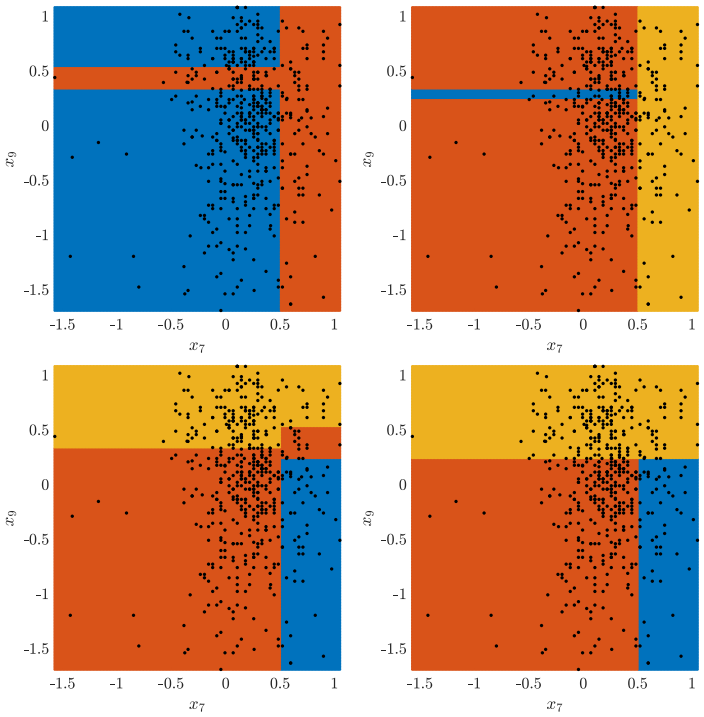

Figure 12: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

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

B. $\{f_1, f_2, f_6\}$

C. $\{f_2, f_3, f_6\}$

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

E. Don't know.

Solution 18. Recall the Apriori algorithm obtain L_3 from L_2 in three steps. First, the Apriori algorithm construct C'_3 by, for each itemset I in L_2, loop over all items not already in I and consider all such combinations where I is enlarged by a single item as a candidate

itemset in C_3'. Specifically we get:

$$C_3' = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The downwards closure principle is then applied by removing and itemset I in C_3' if I contains a subset of 2 items not found in L_2. We thereby get:

$$C_3 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Finally, L_3 is constructed from C_3 by removing those itemsets with a support lower than ε. Thus, the itemsets we don't have to compute support from are those itemsets found in C_3' but not in C_3, or as an even simpler criteria, those which have a subset of size 2 not found in L_2. This rules out all options except D.

Question 19.

Consider again the Avila Bible dataset in Table 1. We would like to predict the copyist using a linear regression, and since we would like the model to be as interpretable as possible we will use variable selection to obtain a parsimonious model. We limit ourselves to the 5 features $x_1,\ x_5,\ x_6,\ x_8,\ x_9$ and in Table 5 we have pre-computed the estimated training and test error for different variable combinations of the dataset. Which of the following statements is correct?

A. Backward selection will select attributes x_1

B. Backward selection will select attributes $x_1,\ x_5,\ x_6,\ x_8$

C. Forward selection will select attributes $x_1,\ x_8$

D. Forward selection will select attributes $x_1,\ x_5,\ x_6,\ x_8$

E. Don't know.

Solution 19.

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

Forward selection: The method is initialized with the set $\{\}$ having an error of 4.163.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1\}$, $\{x_5\}$, $\{x_6\}$, $\{x_8\}$, $\{x_9\}$. Since the lowest error of the available sets is 3.252, which is lower than 4.163, we update the current selected variables to $\{x_1\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_1, x_5\}$, $\{x_1, x_6\}$, $\{x_5, x_6\}$, $\{x_1, x_8\}$, $\{x_5, x_8\}$, $\{x_6, x_8\}$, $\{x_1, x_9\}$, $\{x_5, x_9\}$, $\{x_6, x_9\}$, $\{x_8, x_9\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

Backward selection: The method is initialized with the set $\{x_1, x_5, x_6, x_8, x_9\}$ having an error of 5.766.

Feature(s)	Training RMSE	Test RMSE
none	3.429	4.163
x_1	3.043	3.252
x_5	3.303	4.52
x_6	3.424	4.274
x_8	3.399	4.429
x_9	2.866	5.016
$x_1,\ x_5$	3.001	3.44
$x_1,\ x_6$	3.031	3.423
$x_5,\ x_6$	3.297	4.641
$x_1,\ x_8$	3.017	3.42
$x_5,\ x_8$	3.299	4.485
$x_6,\ x_8$	3.396	4.519
$x_1,\ x_9$	2.644	4.267
$x_5,\ x_9$	2.645	5.495
$x_6,\ x_9$	2.787	5.956
$x_8,\ x_9$	2.71	5.536
$x_1,\ x_5,\ x_6$	2.988	3.607
$x_1,\ x_5,\ x_8$	3.0	3.453
$x_1,\ x_6,\ x_8$	3.007	3.574
$x_5,\ x_6,\ x_8$	3.292	4.61
$x_1,\ x_5,\ x_9$	2.523	4.704
$x_1,\ x_6,\ x_9$	2.562	5.184
$x_5,\ x_6,\ x_9$	2.544	6.552
$x_1,\ x_8,\ x_9$	2.517	4.686
$x_5,\ x_8,\ x_9$	2.628	5.532
$x_6,\ x_8,\ x_9$	2.629	6.569
$x_1,\ x_5,\ x_6,\ x_8$	2.988	3.614
$x_1,\ x_5,\ x_6,\ x_9$	2.425	5.725
$x_1,\ x_5,\ x_8,\ x_9$	2.491	4.734
$x_1,\ x_6,\ x_8,\ x_9$	2.433	5.687
$x_5,\ x_6,\ x_8,\ x_9$	2.53	6.597
$x_1,\ x_5,\ x_6,\ x_8,\ x_9$	2.398	5.766

Table 5: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y in the avila dataset using different combinations of the features $x_1,\ x_5,\ x_6,\ x_8,\ x_9$.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_5, x_6, x_8, x_9\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_5, x_6, x_8\}$, $\{x_1, x_5, x_6, x_9\}$, $\{x_1, x_5, x_8, x_9\}$, $\{x_1, x_6, x_8, x_9\}$, $\{x_5, x_6, x_8, x_9\}$. Since the lowest error of the available sets is 3.614, which is lower than 5.766, we update the current selected variables to $\{x_1, x_5, x_6, x_8\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_5, x_6, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_5, x_6\}$, $\{x_1, x_5, x_8\}$, $\{x_1, x_6, x_8\}$, $\{x_5, x_6, x_8\}$, $\{x_1, x_5, x_9\}$, $\{x_1, x_6, x_9\}$, $\{x_5, x_6, x_9\}$, $\{x_1, x_8, x_9\}$, $\{x_5, x_8, x_9\}$, $\{x_6, x_8, x_9\}$. Since the lowest error of the available sets is 3.453, which is lower than 3.614, we update the current selected variables to $\{x_1, x_5, x_8\}$

Step $i = 3$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_5, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1, x_5\}$, $\{x_1, x_6\}$, $\{x_5, x_6\}$, $\{x_1, x_8\}$, $\{x_5, x_8\}$, $\{x_6, x_8\}$, $\{x_1, x_9\}$, $\{x_5, x_9\}$, $\{x_6, x_9\}$, $\{x_8, x_9\}$. Since the lowest error of the available sets is 3.42, which is lower than 3.453, we update the current selected variables to $\{x_1, x_8\}$

Step $i = 4$ The available variable sets to choose between is obtained by taking the current variable set $\{x_1, x_8\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_1\}$, $\{x_5\}$, $\{x_6\}$, $\{x_8\}$, $\{x_9\}$. Since the lowest error of the available sets is 3.252, which is lower than 3.42, we update the current selected variables to $\{x_1\}$

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

Question 20.

Consider the Avila Bible dataset from Table 1. We wish to predict the copyist based on the attributes *upperm* and *mr/is*.

$p(\tilde{x}_2, \tilde{x}_{10}\|y)$	$y=1$	$y=2$	$y=3$
$\tilde{x}_2=0,\ \tilde{x}_{10}=0$	0.19	0.3	0.19
$\tilde{x}_2=0,\ \tilde{x}_{10}=1$	0.22	0.3	0.26
$\tilde{x}_2=1,\ \tilde{x}_{10}=0$	0.25	0.2	0.35
$\tilde{x}_2=1,\ \tilde{x}_{10}=1$	0.34	0.2	0.2

Table 6: Probability of observing particular values of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on y.

Therefore, suppose the attributes have been binarized such that $\tilde{x}_2 = 0$ corresponds $x_2 \leq -0.056$ (and otherwise $\tilde{x}_2 = 1$) and $\tilde{x}_{10} = 0$ corresponds $x_{10} \leq -0.002$ (and otherwise $\tilde{x}_{10} = 1$). Suppose the probability for each of the configurations of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on the copyist y are as given in Table 6. and the prior probability of the copyists is

$$p(y=1) = 0.316,\ p(y=2) = 0.356,\ p(y=3) = 0.328.$$

Using this, what is then the probability an observation was authored by copyist 1 given that $\tilde{x}_2 = 1$ and $\tilde{x}_{10} = 0$?

A. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.25$

B. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.313$

C. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.262$

D. $p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = 0.298$

E. Don't know.

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

$$p(y=1|\tilde{x}_2=1,\tilde{x}_{10}=0) = \frac{p(\tilde{x}_2=1,\tilde{x}_{10}=0|y=1)p(y=1)}{\sum_{k=1}^{3} p(\tilde{x}_2=1,\tilde{x}_{10}=0|y=k)p(y=k)}$$

The values of $p(y)$ are given in the problem text and the values of $p(\tilde{x}_2=1,\tilde{x}_{10}=0|y)$ in Table 6. Inserting the values we see option D is correct.

Variable	$t=1$	$t=2$	$t=3$	$t=4$
y_1	1	2	2	2
y_2	1	2	2	1
y_3	2	2	2	1
y_4	1	1	1	2
y_5	1	1	1	1
y_6	2	2	2	1
y_7	1	2	2	1
y_8	2	1	1	2
y_9	2	2	2	2
y_{10}	1	1	2	2
y_{11}	2	2	1	2
y_{12}	2	1	1	2
y_1^{test}	2	1	1	2
y_2^{test}	2	2	1	2
ϵ_t	0.583	0.657	0.591	0.398
α_t	-0.168	-0.325	-0.185	0.207

Table 7: Tabulation of each of the predicted outputs of the AdaBoost classifiers, as well as the intermediate values α_t and ϵ_t, when the AdaBoost algorithm when evaluated for $T = 4$ steps. Note the table includes the prediction of the two test points in Figure 13.

Question 21.

Consider again the Avila Bible dataset of Table 1. Suppose we limit ourselves to $N = 12$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_6 and x_8. We wish to apply a KNN classification model ($K = 2$) to this dataset and apply AdaBoost to improve the performance. During the first $T = 4$ rounds of boosting, we obtain the decision boundaries shown in Figure 13. The figure also contains two test observations (marked by a cross and a square).

The prediction of the intermediate AdaBoost classifiers, as well as the values of α_t and ϵ_t, are given in Table 7. Given this information, how will the AdaBoost classifier, as obtained by combining the $T = 4$ weak classifiers, classify the two test observations?

A. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}1 & 1\end{bmatrix}$

B. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}2 & 1\end{bmatrix}$

C. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}1 & 2\end{bmatrix}$

D. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}2 & 2\end{bmatrix}$

E. Don't know.

Solution 21.

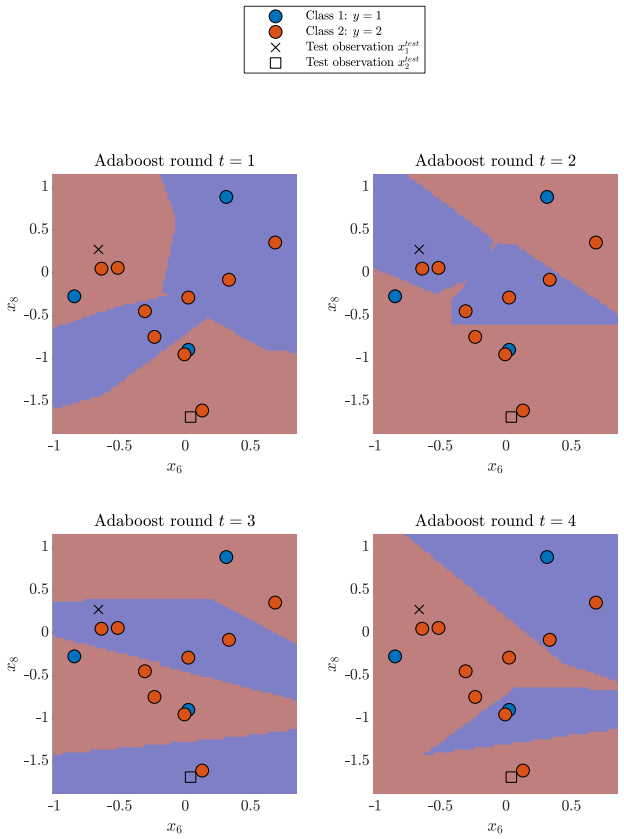

Figure 13: Decision boundaries for a KNN classifier for the first $T = 4$ rounds of boosting. Notice that in addition to the training data, the plot also indicate the location of two test points.

According to the AdaBoost algorithm, the classification rule when combining T AdaBoost algorithms is:

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

In other words, the classification rule is obtained by summing the α_t where $f_t(\boldsymbol{x}) = 1$ (as F_1) and those where $f_t(\boldsymbol{x}) = 2$ (as F_2) and then selecting the y corresponding to the largest value. We get for the two test points:

$$\begin{aligned}
F_1(\boldsymbol{x}_1^{\text{test}}) &= \alpha_2 + \alpha_3 = -0.51 \\
F_2(\boldsymbol{x}_1^{\text{test}}) &= \alpha_1 + \alpha_4 = 0.039 \\
F_1(\boldsymbol{x}_2^{\text{test}}) &= \alpha_3 = -0.185 \\
F_2(\boldsymbol{x}_2^{\text{test}}) &= \alpha_1 + \alpha_2 + \alpha_4 = -0.286.
\end{aligned}$$

Therefore, we get

$$\begin{bmatrix} \tilde{y}_1^{\text{test}} \\ \tilde{y}_2^{\text{test}} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and option B is correct.

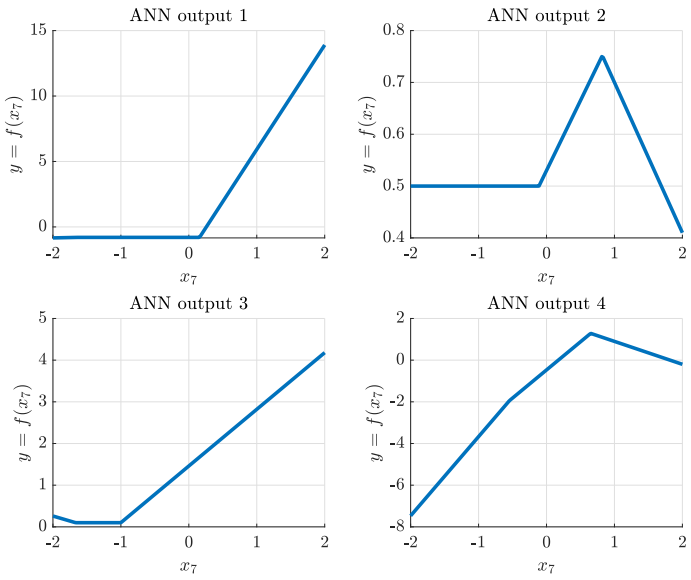

Figure 14: Suggested activation curves for an ANN applied to the feature x_7 from Avila Bible dataset.

Question 22.

We will consider an artificial neural network (ANN) applied to the Avila Bible dataset described in Table 1 and trained to predict based on just the feature x_7; that is, the neural network is a function that maps from a single real number to a single real number: $f(x_7) = y$

Suppose the neural network takes the form:

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

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer and the weights are given as:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -1.8 \\ -1.1 \end{bmatrix}$$
$$\boldsymbol{w}_2^{(1)} = \begin{bmatrix} -0.6 \\ 3.8 \end{bmatrix}$$
$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -0.1 \\ 2.1 \end{bmatrix},$$
$$w_0^{(2)} = -0.8.$$

Which of the curves in Figure 14 will then correspond to the function f?

A. ANN output 4

B. ANN output 1

C. ANN output 3

D. ANN output 2

E. Don't know.

Solution 22.

It suffices to compute the activation of the neural network at $x_7 = 2$. The activation of each of the two hidden neurons is:

$$n_1 = h^{(1)}(\begin{bmatrix} 1 & 2 \end{bmatrix} \boldsymbol{w}_1^{(1)}) = 0$$
$$n_2 = h^{(1)}(\begin{bmatrix} 1 & 2 \end{bmatrix} \boldsymbol{w}_2^{(1)}) = 7.$$

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

$$f(x, \boldsymbol{w}) = w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x]\boldsymbol{w}_j^{(1)})$$
$$= w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} n_j = 13.9.$$

This rules out all options except B.

Question 23. Suppose a neural network is trained to translate documents. As part of training the network, we wish to select between four different ways to encode the documents (i.e., $S = 4$ 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 $K_2 = 4$-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. 1020 minutes

B. 2040 minutes

C. 300 minutes

D. 960 minutes

E. Don't know.

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

$$K_1(K_2S + 1) = 51$$

Multiplying by the time taken to train a single model we obtain a total training time of 1020 minutes and therefore answer A is correct.

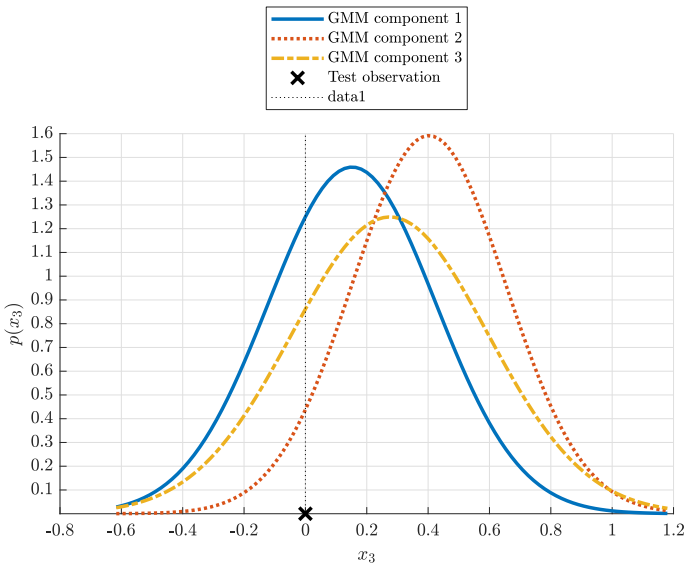

Figure 15: Mixture components in a GMM mixture model with $K = 3$.

Question 24.

We wish to apply the EM algorithm to fit a 1D GMM mixture model to the single feature x_3 from the Avila Bible dataset. At the first step of the EM algorithm, the $K = 3$ mixture components has densities as indicated by each of the curves in Figure 15 (i.e. each curve is a normalized, Gaussian density $\mathcal{N}(x; \mu_k, \sigma_k)$). In the figure, we have indicated the x_3-value of a single observation i from the dataset as a black cross.

Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step. We assume the weights of the components are

$$\boldsymbol{\pi} = \begin{bmatrix} 0.15 & 0.53 & 0.32 \end{bmatrix}$$

and the mean/variances of the components are those indicated in the figure.

According to the EM algorithm, what is the (approximate) probability the black cross is assigned to mixture component 3 (γ_{ik})?

A. 0.4

B. 0.86

C. 0.28

D. 0.58

E. Don't know.

Solution 24.

Recall γ_{ik} is the posterior probability that observation i is assigned to mixture component 3 which can easily be obtained using Bayes' theorem. We see that:

$$\gamma_{i,3} = \frac{p(x_i|z_{i,3}=1)\pi_3}{\sum_{k=1}^{3} p(x_i|z_{ik}=1)\pi_k}.$$

To use Bayes' theorem, we need to read off the probabilities from Figure 15. These are (approximately):

$$p(x_i|z_{i1}=1) = 1.25$$
$$p(x_i|z_{i2}=1) = 0.45$$
$$p(x_i|z_{i3}=1) = 0.85$$

Combining these with the class-assignment probabilities we obtain:

$$\gamma_{i,3} = 0.39.$$

Note this answer is not *exactly* the answer given in the question because we lost precision when we read off the probabilities from the figure. However, the answer is close enough to the answer 0.4 (and far enough away from the other) we can conclude the solution is A.

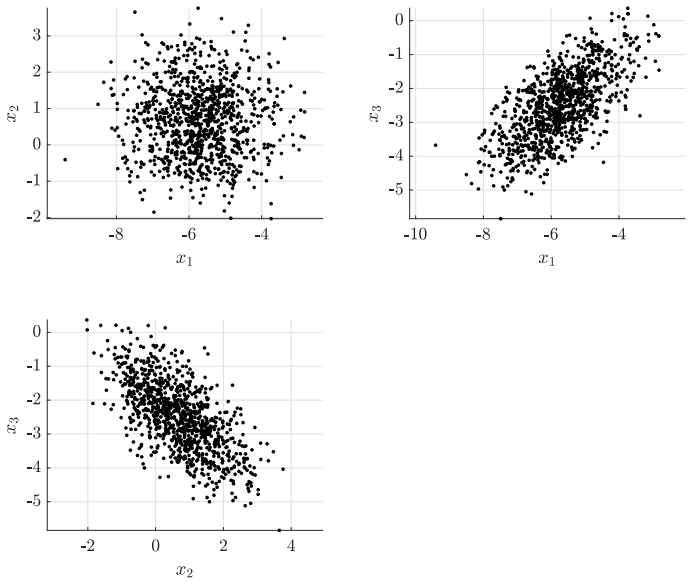

Figure 16: Scatter plot of each pairs of attributes of a vectors $\boldsymbol{x}$ drawn from a multivariate normal distribution of 3 dimensions.

Question 25. Consider a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ and mean μ and suppose we generate 1000 random samples from it:

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

Plots of each pair of coordinates of the draws $\boldsymbol{x}$ is shown in Figure 16. What is the most plausible covariance matrix?

A. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.65 & -0.65 \\ 0.65 & 1.0 & 0.0 \\ -0.65 & 0.0 & 1.0 \end{bmatrix}$

B. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & 0.65 \\ 0.0 & 1.0 & -0.65 \\ 0.65 & -0.65 & 1.0 \end{bmatrix}$

C. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & -0.65 & 0.0 \\ -0.65 & 1.0 & 0.65 \\ 0.0 & 0.65 & 1.0 \end{bmatrix}$

D. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & -0.65 \\ 0.0 & 1.0 & 0.65 \\ -0.65 & 0.65 & 1.0 \end{bmatrix}$

E. Don't know.

Solution 25. To solve this problem, recall that the correlation between coordinates x_i, x_j of an observation drawn from a multivariate normal distribution is

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

We can therefore use the scatter plots of variables x_i, x_j to read of the sign off Σ_{ij} (or whether it is zero). This rules out all but option B.

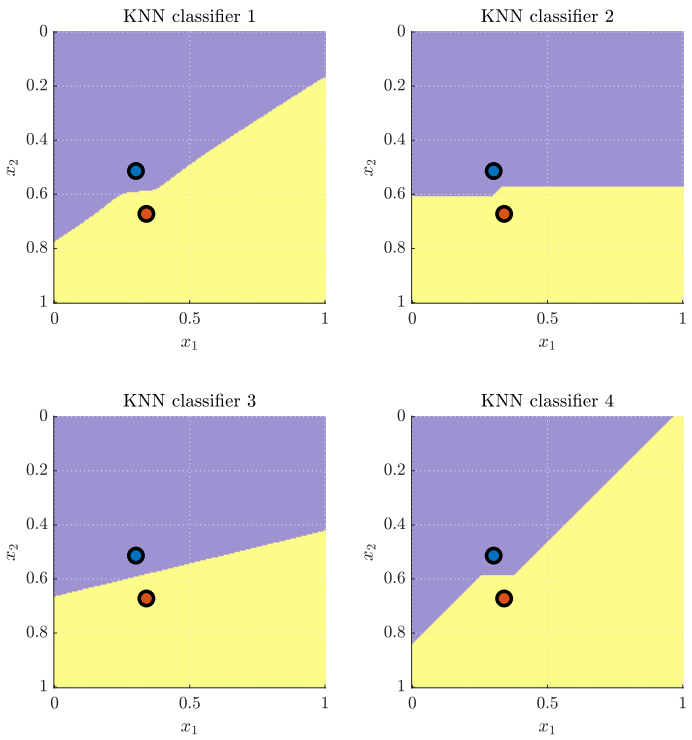

Figure 17: Decision boundaries for a KNN classifier, $K = 1$, computed for the two observations marked by circles (the colors indicate class labels), but using four different p-distances $d_p(\cdot, \cdot)$ to compute k-neighbors.

Question 26.

We consider a K-nearest neighbor (KNN) classifier with $K = 1$. Recall in a KNN classifier, we find the nearest neighbors by computing the distances using a distance measure $d(\boldsymbol{x}, \boldsymbol{y})$. For this problem, we will consider KNN classifiers based on different distance measures based on p-norms

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}}, p \geq 1$$

and what decision surfaces they induce.

In Figure 17 are shown four different decision boundaries obtained by training the KNN ($K = 1$) classifiers using the training observations (marked by the two circles in the figure):

$$\boldsymbol{x}_1 = \begin{bmatrix} 0.301 \\ 0.514 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 0.34 \\ 0.672 \end{bmatrix}$$

and with corresponding class labels $y_1 = 0$ and $y_2 = 1$, but with distance measures based on $p = 1,\ 2,\ 4,\ \infty$ (not necessarily plotted in that order).

Which norms were used in the four KNN classifiers?

A. KNN classifier 1 corresponds to $p = \infty$, KNN classifier 2 corresponds to $p = 2$, KNN classifier 3 corresponds to $p = 4$, KNN classifier 4 corresponds to $p = 1$

B. KNN classifier 1 corresponds to $p = 4$, KNN classifier 2 corresponds to $p = 2$, KNN classifier 3 corresponds to $p = 1$, KNN classifier 4 corresponds to $p = \infty$

C. KNN classifier 1 corresponds to $p = 4$, KNN classifier 2 corresponds to $p = 1$, KNN classifier 3 corresponds to $p = 2$, KNN classifier 4 corresponds to $p = \infty$

D. KNN classifier 1 corresponds to $p = \infty$, KNN classifier 2 corresponds to $p = 1$, KNN classifier 3 corresponds to $p = 2$, KNN classifier 4 corresponds to $p = 4$

E. Don't know.

Solution 26.

To solve this problem, one could simply consider points on the decision boundary and verify under which norm they had the same distance to the two test-observations. As this may feel a bit ad-hoc, we will here present a more general solution:

For simplicity, notice that (i) translating (moving) the coordinate system does not affect the distance calculation (ii) rotating the coordinate system by $\frac{\pi}{2}$ only corresponds to interchanging the role of x and y (iii) reflecting the coordinate system around an axis will not change the distance computation.

We can therefore consider the case where the location of the coordinates of the two points are $(-x, -y)$ and (x, y) where $y > x \geq 0$. Consider a point (d, h) *on* the decision boundary. The criteria for being so is:

$$(|d + x|^p + |h + y|^p)^{\frac{1}{p}} = (|d - x|^p + |h - y|^p)^{\frac{1}{p}}$$

Suppose now that $p = 2$. This is the standard Euclidean distance, and it is clear the decision boundary is a straight line passing through $(0, 0)$ and perpendicular to the vector (x, y).

For the other choices of p, suppose we limit ourselves to the case where $d < x$. For $p = 1$ we obtain:

$$|-d + x|^p + |-h + y|^p = |d + x|^p + |h + y|^p$$

Since the quantities within the absolute value operators are positive this becomes:

$$-d + x + -h + y = d + x + h + y$$

and therefore $d = -h$. That is, the decision boundary (for small d) must be a straight line at an 45-degree angle to the coordinate system.

For the case $p = \infty$, assume that $d + x < y$ (which is always possible since $y > x$). We then get:

$$\max\{|-d + x|, |-h + y|\} = \max\{|d + x|, |h + y|\}$$

One can either approach this expression with some algebra, but notice if $h = 0$ we get:

$$\max\{|-d + x|, |-0 + y|\} = \max\{|d + x|, |0 + y|\}$$

Therefore, if d is so small that $d + x < y$ this is trivially satisfied. In other words, when $d + x < y$ the horizontal line $h = 0$ is a solution.

Finally, the case $p = 4$ can be obtained by the process of illumination, or by noting the decicion bounday must look somewhat like a crossover between the $p = 2$ and $p = \infty$ case. We can therefore rule out all possibilities except C.

Question 27. Consider a small dataset comprised of $N = 9$ observations

$$x = \begin{bmatrix} 0.1 & 0.3 & 0.5 & 1.0 & 2.2 & 3.0 & 4.1 & 4.4 & 4.7 \end{bmatrix}.$$

Suppose a k-means algorithm is applied to the dataset with $K = 4$ and using Euclidian distances. At a given stage of the algorithm the data is partitioned into the blocks:

$$\{0.1, 0.3\},\ \{0.5, 1\},\ \{2.2, 3, 4.1\},\ \{4.4, 4.7\}$$

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

A. $\{0.1, 0.3, 0.5, 1\}$, $\{2.2\}$, $\{\}$, $\{3, 4.1, 4.4, 4.7\}$

B. $\{0.1, 0.3\}$, $\{0.5, 1\}$, $\{2.2, 3\}$, $\{4.1, 4.4, 4.7\}$

C. $\{0.1, 0.3\}$, $\{0.5\}$, $\{1, 2.2\}$, $\{3, 4.1, 4.4, 4.7\}$

D. $\{0.1, 0.3\}$, $\{0.5, 1, 2.2, 3\}$, $\{4.1, 4.4\}$, $\{4.7\}$

E. Don't know.

Solution 27. Recall the K-means algorithm iterates between assigning the observations to their nearest centroids, and then updating the centroids to be equal to the average of the observations assigned to them. Therefore, the subsequent steps in the K-means algorithm are:

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

$$\mu_1 = 0.2,\ \mu_2 = 0.75,\ \mu_3 = 3.1,\ \mu_4 = 4.55.$$

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

$$\{0.1, 0.3\},\ \{0.5, 1\},\ \{2.2, 3\},\ \{4.1, 4.4, 4.7\}.$$

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

$$\mu_1 = 0.2,\ \mu_2 = 0.75,\ \mu_3 = 2.6,\ \mu_4 = 4.4.$$

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

$$\{0.1, 0.3\},\ \{0.5, 1\},\ \{2.2, 3\},\ \{4.1, 4.4, 4.7\}.$$

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

Technical University of Denmark

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

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

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

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

Answers:

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

Name: _____

Student number: _____

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

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

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

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

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

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

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

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

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

E. Don't know.

Solution 1.

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

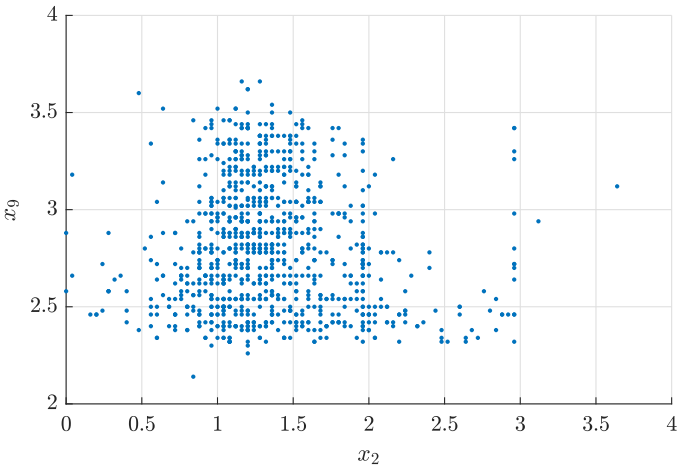

Figure 1: Scatter plot of observations x_2 and x_9 of the travel review dataset described in Table 1.

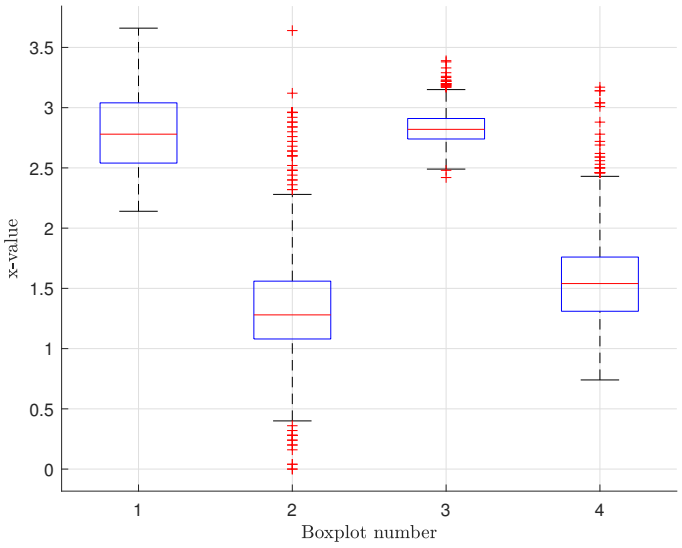

Figure 2: Four boxplots in which two of the boxplots correspond to the two variables plotted in Figure 1.

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

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

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

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

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

Which one of the following statements is true?

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

A. 0.41

B. 1.0

C. 0.51

D. 0.83

E. Don't know.

Solution 4.

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

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

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

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

and

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

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

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

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

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

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

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

E. Don't know.

Solution 5.

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

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

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

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

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

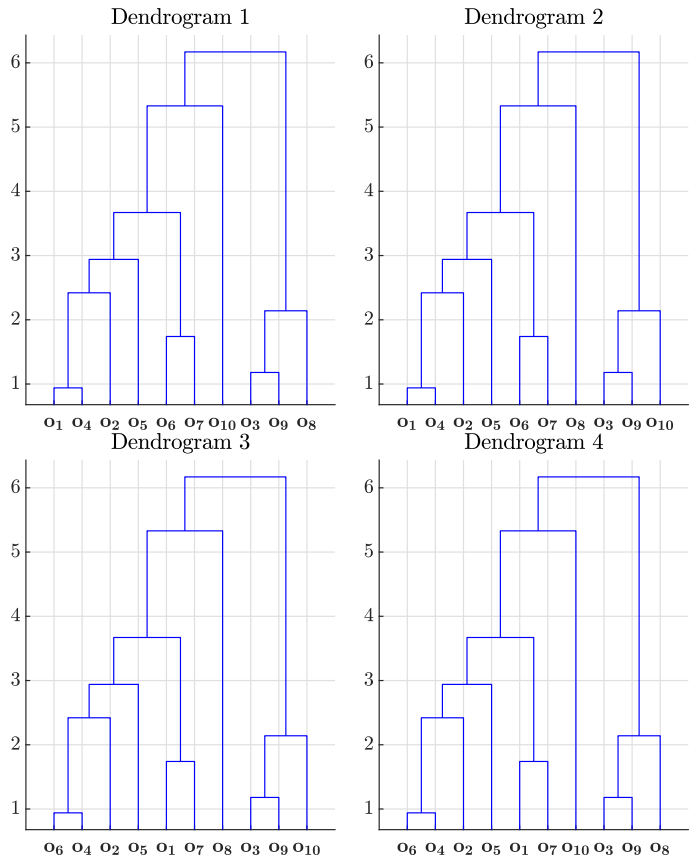

Figure 3: Proposed hierarchical clustering of the 10 observations in Table 2.

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

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

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

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

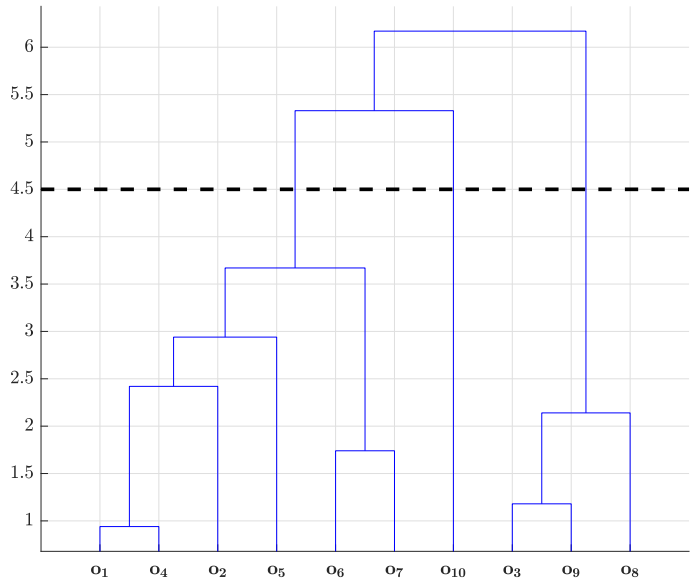

Figure 4: Dendrogram 1 from Figure 3 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

$$S = 4, D = 17$$

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

and the impurity I_0 from

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

In particular we obtain:

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

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

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

A. The accuracy is: 0.42

B. The accuracy is: 0.685

C. The accuracy is: 0.338

D. The accuracy is: 0.097

E. Don't know.

Solution 9.

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

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

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

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

Therefore, the accuracy is:

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

and answer A is correct.

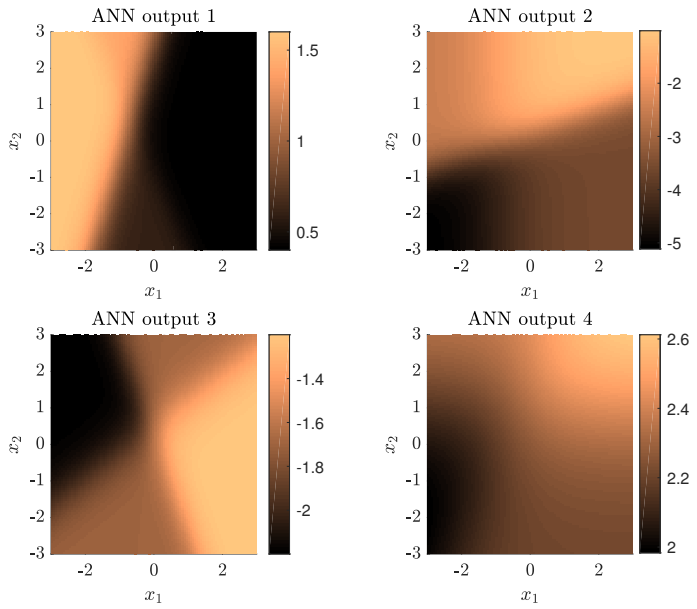

Figure 5: Suggested outputs of an ANN trained on the two attributes x_1 and x_2 from the travel review dataset to predict y.

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

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

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

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

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

A. ANN output 1

B. ANN output 2

C. ANN output 3

D. ANN output 4

E. Don't know.

Solution 10.

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

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

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

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

This rules out all options except D.

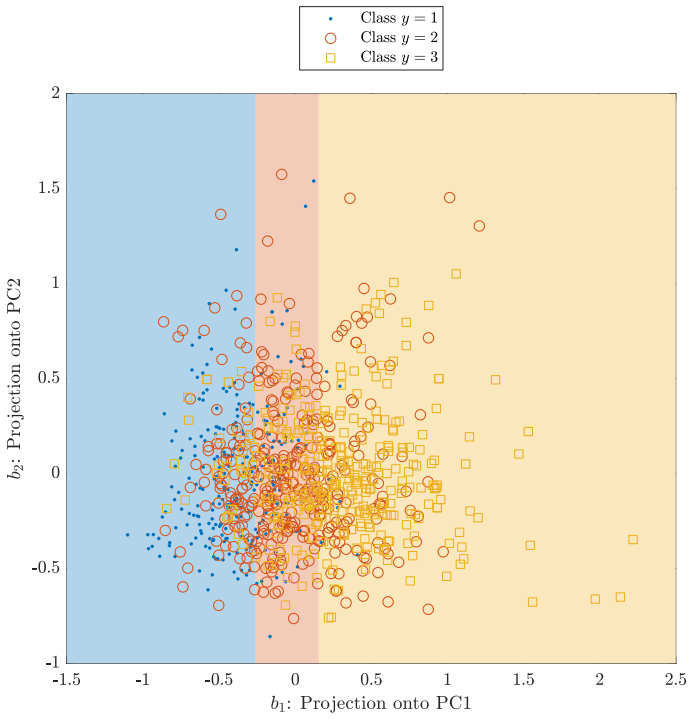

Figure 6: Output of a logistic regression classifier trained on observations from the travel review dataset.

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

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

and then use the softmax transformation in the form:

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

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

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

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

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

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

E. Don't know.

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

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

The projections onto the four options are, in order,

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Question 13. We again consider the travel review dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce 9 new, binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[2], and we thereby arrive at the $N \times M = 10 \times 9$ binary matrix in Table 4. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_2,\ f_5,\ f_8$. If for an observations we observe

$$f_2 = 0,\ f_5 = 0,\ f_8 = 1$$

what is then the probability it has high rating ($y = 3$) according to the Naïve-Bayes classifier?

A. $p_{\mathbf{NB}}(y = 3 | f_2 = 0,\ f_5 = 0,\ f_8 = 1) = \frac{934}{2527}$

B. $p_{\mathrm{NB}}(y = 3 | f_2 = 0,\ f_5 = 0,\ f_8 = 1) = \frac{6477}{9304}$

C. $p_{\mathrm{NB}}(y = 3 | f_2 = 0,\ f_5 = 0,\ f_8 = 1) = \frac{1307}{6073}$

D. $p_{\mathrm{NB}}(y = 3 | f_2 = 0,\ f_5 = 0,\ f_8 = 1) = \frac{2123}{3857}$

E. Don't know.

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

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

$$p_{\text{NB}}(y=3|f_2=0,\ f_5=0,\ f_8=1) = \frac{p(f_2=0|y=3)p(f_5=0|y=3)p(f_8=1|y=3)p(y=3)}{\sum_{j=1}^{3} p(f_2=0|y=j)p(f_5=0|y=j)p(f_8=1|y=j)p(y=j)}$$
$$= \frac{\frac{5}{7}\frac{2}{7}\frac{2}{7}\frac{1}{2}}{\frac{3}{4}\frac{3}{4}\frac{1}{4}\frac{1}{5} + \frac{3}{5}\frac{3}{5}\frac{1}{5}\frac{3}{10} + \frac{5}{7}\frac{2}{7}\frac{2}{7}\frac{1}{2}}$$
$$= \frac{934}{2527}.$$

Therefore, answer A is correct.

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

C. The confidence is 1

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

E. Don't know.

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

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

Therefore, answer B is correct.

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

Therefore, answer B is correct.

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

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

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

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

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

E. Don't know.

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

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

and therefore option B is correct.

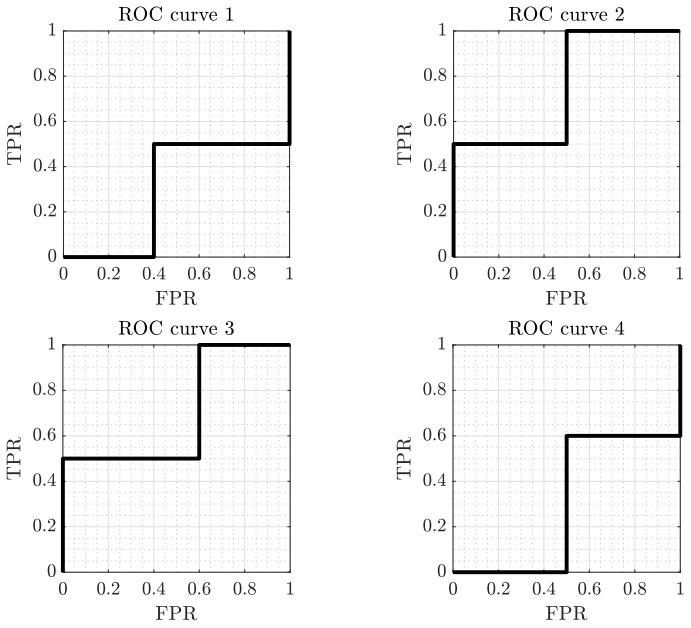

Figure 7: Proposed ROC curves for the neural network classifier with predictions/true class labels given in Table 5

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

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

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

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

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

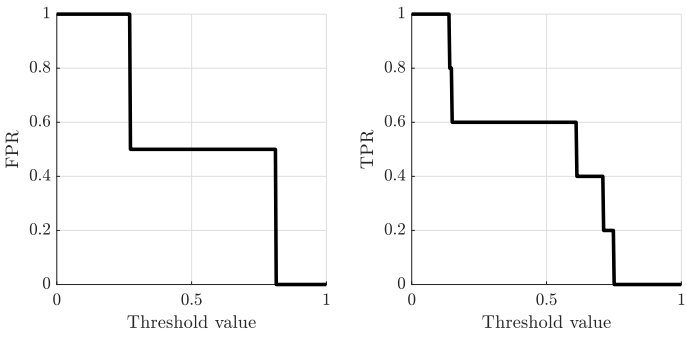

Figure 8: TPR, FPR curves for the classifier.

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

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

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

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

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

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

A. Forward selection will select attributes x_6

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

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

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

E. Don't know.

Solution 19.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

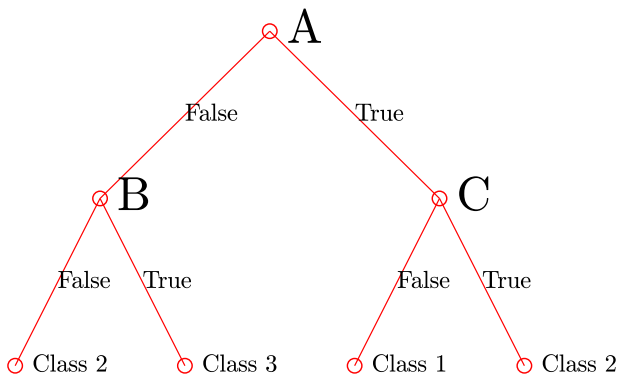

Figure 9: Example classification tree.

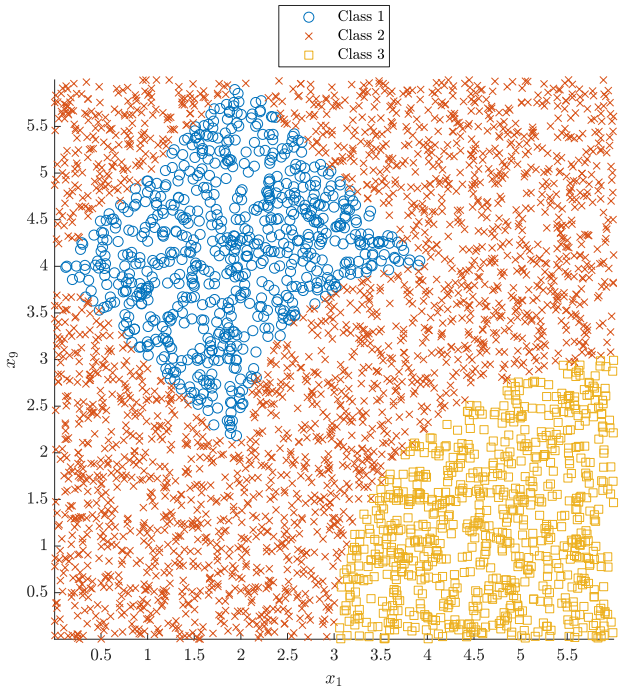

Figure 10: classification boundary.

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

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

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

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

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

E. Don't know.

Solution 21.

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

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

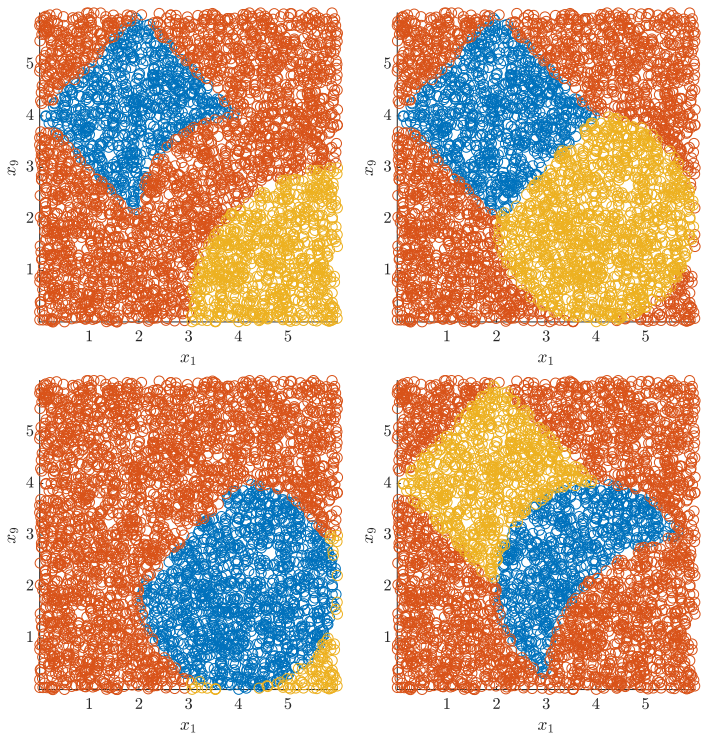

Figure 11: Classification trees induced by each of the options. (Top row: option A and B, bottom row: C and D)

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

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

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

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

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

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

A. 208 models

B. 100 models

C. 200 models

D. 104 models

E. Don't know.

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

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

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

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

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

and suppose that the identified values of the mixture weights are

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

and the parameters of the multivariate normal densities:

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

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

A. 0.49

B. 0.31

C. 0.08

D. 0.68

E. Don't know.

Solution 23.

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

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

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

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

Combining these with the class-assignment probabilities we obtain:

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

and conclude the solution is B.

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

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

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

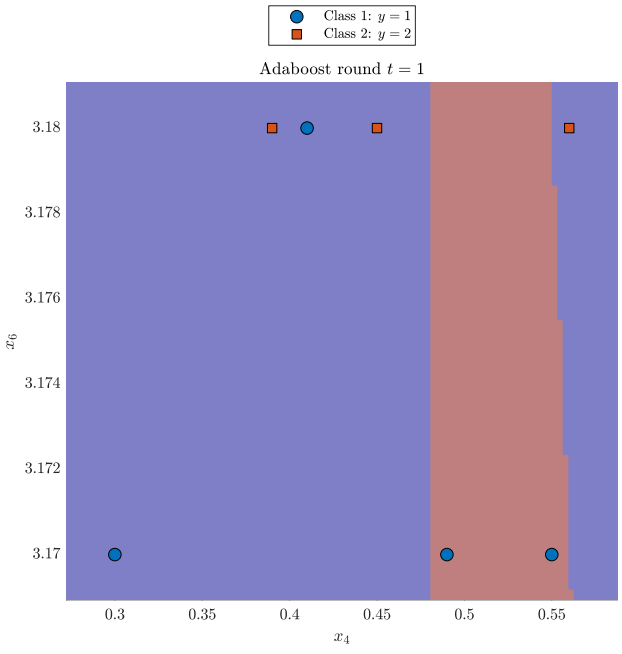

Figure 12: Decision boundaries for a KNN classifier for the first $T = 1$ rounds of boosting.

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

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

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

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

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

E. Don't know.

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

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

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

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

From this, we compute α_t as

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

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

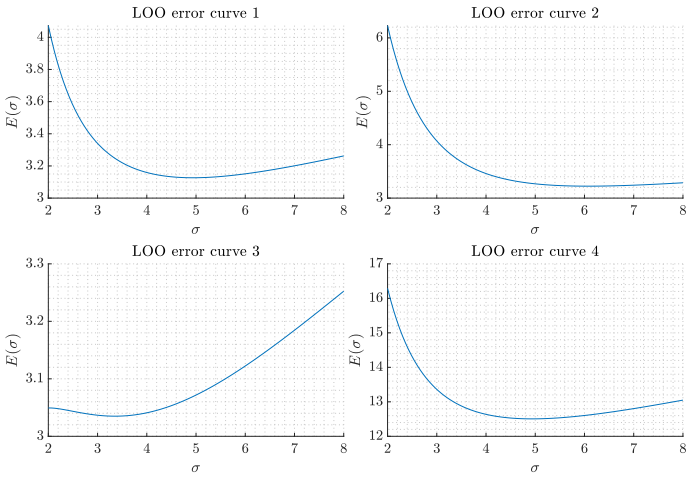

Figure 13: Estimated negative log-likelihood as obtained using LOO cross-validation on a small, $N = 4$ one-dimensional dataset as a function of kernel width σ.

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

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

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

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

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

A. LOO curve 1

B. LOO curve 2

C. LOO curve 3

D. LOO curve 4

E. Don't know.

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

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

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

These values are approximately:

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

The LOO error is then:

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

Therefore, the correct answer is A.

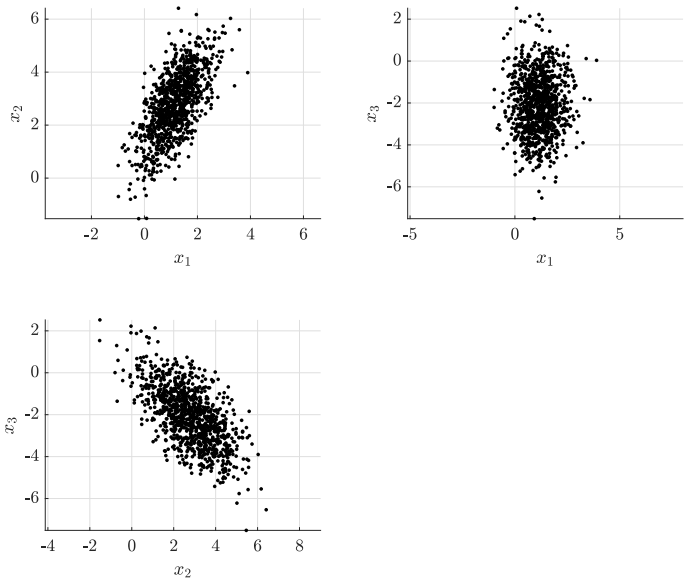

Figure 14: Scatter plots of all pairs of attributes of a vector $\boldsymbol{x}$ when $\boldsymbol{x}$ is a random vector distributed as a multivariate normal distribution of 3 dimensions.

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

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

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

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

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

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

A. The correlation between x_1 and x_2 is 0.647

B. The correlation between x_1 and x_2 is -0.611

C. The correlation between x_1 and x_2 is 0.747

D. The correlation between x_1 and x_2 is 0.56

E. Don't know.

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

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

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

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

The correlation is then given as:

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

and therefore answer A is correct.

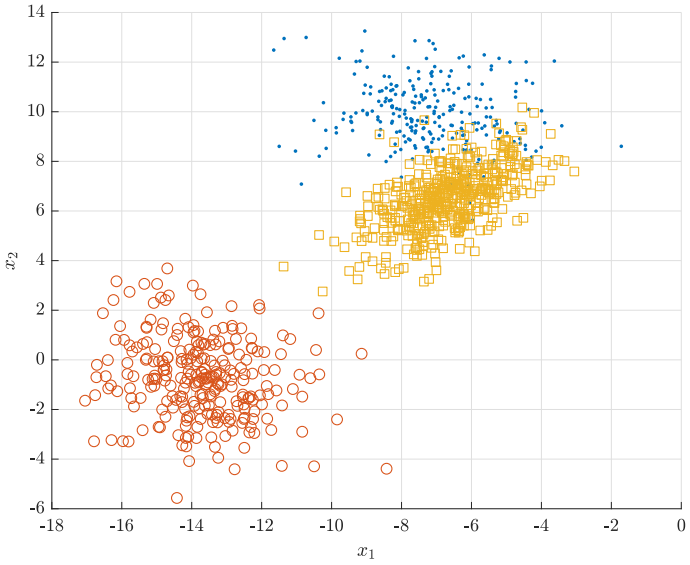

Figure 15: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

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

generate the data?

A.

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

B.

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

C.

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

D.

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

E. Don't know.

Solution 27.

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

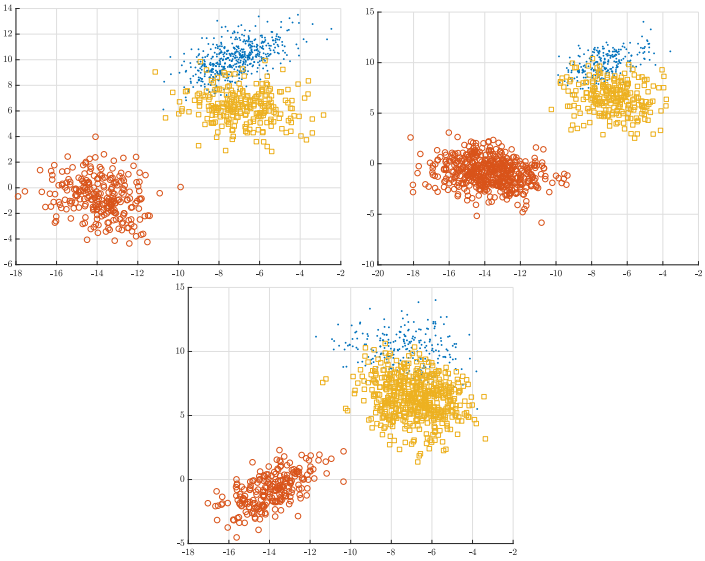

Figure 16: GMM mixtures corresponding to alternative options.