Technical University of Denmark

Written examination: December 17th 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.

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

When you hand in your answers you have to upload two files:
1: Your answers to the multiple choice exam using the "answers.txt" file.
2: Your written full explanations of how you found the answer to each question not marked as "E" (Don't know) either as a .zip (with bmp, png, tiff, jpg as allowed file formats if you take photos of your handwritten explanations) or as a PDF file (if you write a document with your answers).
You have to hand in your answers to the exam in file 1 no later than 5 PM and file 2 no later than 5:15 PM.
Guessing on an answer is for the online exam not allowed, as each answer has to include an accompanying argumentation in writing for the answer.
Failing to timely upload both documents will count as not having handed in the exam!
Questions where we find answers in the "answers.txt" (file 1) that is different from the explanation or where explanations are insufficient in the accompanying file explaining the answers (file 2) will be treated as "Don't know". Systematic discrepancy between the answers in the two hand-in files will ultimately potentially count as attempt of cheating the exam.

Answers:

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

No.	Attribute description	Abbrev.
x_1	Month (1-12)	MONTH
x_2	$PM_{2.5}$ concentration ($\mu g/m^3$)	$PM_{2.5}$
x_3	PM_{10} concentration ($\mu g/m^3$)	PM_{10}
x_4	NO_2 concentration ($\mu g/m^3$)	NO_2
x_5	SO concentration ($\mu g/m^3$)	CO
x_6	O_3 concentration ($\mu g/m^3$)	O_3
x_7	Temperature (degree Celsius)	TEMP
x_8	Pressure (hPa)	PRES
x_9	Dew point temperature (degree Celsius)	DEWP
x_{10}	Precipitation/rainfall (mm)	RAIN
x_{11}	Wind speed (m/s)	WSPM
y	SO_2 concentration ($\mu g/m^3$)	pollution level

Table 1: Description of the features of the Beijing air pollution dataset used in this exam. It consists of measurements from 12 air-quality sites provided by the China Meteorological Administration. The measurements were taken hourly (March 1st, 2013 to February 28th, 2017), but we will only consider data from 2014, subsampled to every 8 hours, and with missing values removed. We consider the goal as predicting the SO_2 level both as regression and classification task. For regression tasks, y_r will refer to the continuous value in $\mu g/m^3$. For classification, the attribute y is discrete taking values $y = 1$ (corresponding to a light pollution level), $y = 2$ (corresponding to a medium pollution level), and $y = 3$ (corresponding to a high pollution level). There are $N = 981$ observations in total.

Question 1. The main dataset used in this exam is the Beijing air pollution dataset[1] described in Table 1. Table 2 contains summary statistics of four attributes from the Beijing air pollution dataset. Which boxplots

[1]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data`

	Mean	Std	$x_{p=25\%}$	$x_{p=50\%}$	$x_{p=75\%}$
$PM_{2.5}$	85.58	78.09	26	66	121.25
PM_{10}	113.2	85.18	48.75	97	156.25
NO_2	55.89	31.8	30	51	76.25
O_3	54.4	61.72	8	33	75

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

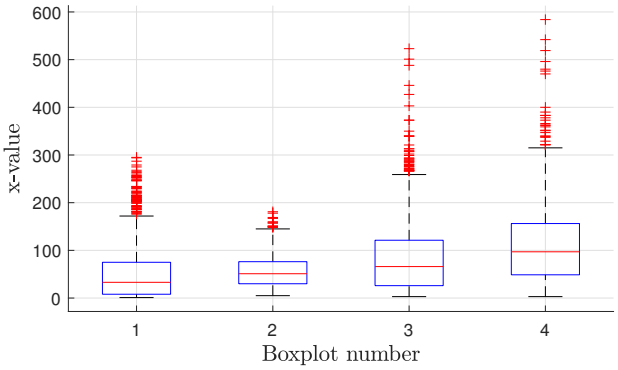

Figure 1: Boxplots corresponding to the variables with summary statistics indicated in Table 2 but not necessarily in that order.

in Figure 1 match which attributes?

A. Attribute $PM_{2.5}$ corresponds to boxplot 3 PM_{10} corresponds to boxplot 4 NO_2 corresponds to boxplot 1 and O_3 corresponds to boxplot 2

B. Attribute $PM_{2.5}$ corresponds to boxplot 4 PM_{10} corresponds to boxplot 3 NO_2 corresponds to boxplot 2 and O_3 corresponds to boxplot 1

C. Attribute $PM_{2.5}$ corresponds to boxplot 3 PM_{10} corresponds to boxplot 4 NO_2 corresponds to boxplot 2 and O_3 corresponds to boxplot 1

D. Attribute $PM_{2.5}$ corresponds to boxplot 1 PM_{10} corresponds to boxplot 3 NO_2 corresponds to boxplot 2 and O_3 corresponds to boxplot 4

E. Don't know.

Solution 1. To solve the problem, note that we can read of the median, 25'th, and 75'th percentiles from Table 2 as $q_{p=50\%}$, $q_{p=25\%}$, and $q_{p=75\%}$ respectively. These in turns can be matched to the boxplots in

Figure 1 by observing the median is the horizontal red line and the 25'th and 75'th percentiles corresponds to the top and bottom of the boxes. This easily rules out all options except C.

Question 2. A Principal Component Analysis (PCA) is carried out on the Beijing air pollution dataset in Table 1 based on the attributes x_1, x_3, x_5, x_8, x_{10}, x_{11}.

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 $\boldsymbol{USV}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.1 & -0.45 & -0.55 & 0.67 & -0.2 & 0.01 \\ -0.63 & -0.02 & -0.01 & -0.05 & -0.44 & -0.64 \\ -0.67 & 0.07 & 0.03 & 0.13 & -0.12 & 0.72 \\ -0.09 & 0.69 & 0.03 & 0.6 & 0.32 & -0.2 \\ 0.06 & -0.35 & 0.83 & 0.41 & -0.09 & -0.03 \\ 0.37 & 0.44 & 0.05 & 0.04 & -0.8 & 0.17 \end{bmatrix} \quad (1)$$

$$\boldsymbol{S} = \begin{bmatrix} 43.67 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 33.47 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 31.15 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 30.36 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 27.77 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 13.86 \end{bmatrix}.$$

Which one of the following statements is true?

A. The variance explained by the first five principal components is less than 0.9

B. The variance explained by the first three principal components is less than 0.715

C. The variance explained by the first principal component is less than 0.3

D. The variance explained by the last two principal components is less than 0.15

E. Don't know.

Solution 2. The correct answer is B. To see this, recall the variance explained by a given component k of the PCA is given by

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

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

$$\text{Var.Expl.} = \frac{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 + \sigma_5^2 + \sigma_6^2} = 0.6796.$$

Question 3. Consider again the PCA analysis for the Beijing air pollution 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 **PM**$_{10}$, a high value of **PRES**, and a low value of **WSPM** will typically have a negative value of the projection onto principal component number 5.

B. An observation with a high value of **PM**$_{10}$, a high value of **CO**, and a low value of **WSPM** will typically have a positive value of the projection onto principal component number 1.

C. An observation with a low value of **MONTH**, a low value of **PRES**, and a low value of **RAIN** will typically have a positive value of the projection onto principal component number 4.

D. An observation with a high value of MONTH, and a low value of RAIN will typically have a negative value of the projection onto principal component number 3.

E. Don't know.

Solution 3. The correct answer is D. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_3$ (i.e. column three of $\boldsymbol{V}$) is

$$b_3 = \boldsymbol{x}^\top \boldsymbol{v}_3 = \begin{bmatrix} x_1 & x_3 & x_5 & x_8 & x_{10} & x_{11} \end{bmatrix} \begin{bmatrix} -0.55 \\ -0.01 \\ 0.03 \\ 0.03 \\ 0.83 \\ 0.05 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and negative, this occurs if x_1, x_{10} has large magnitude and the sign convention given in option D.

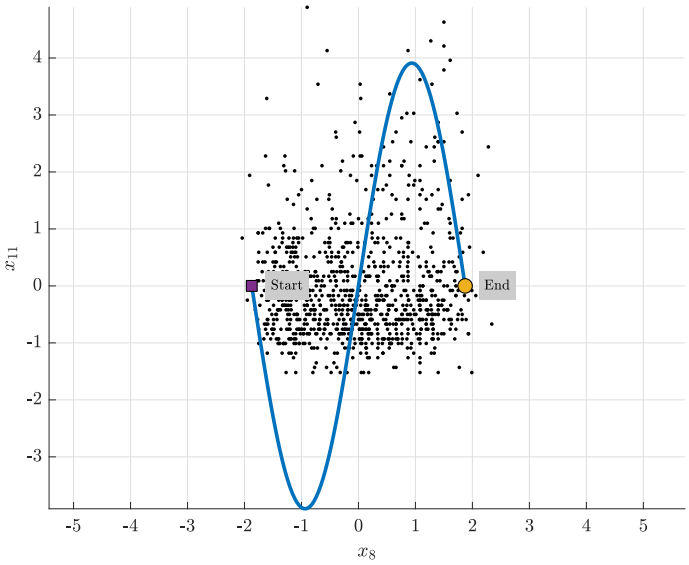

Figure 2: Black dots show attributes x_8 and x_{11} of the Beijing air pollution dataset from Table 1. The other attributes are kept fixed while x_8 and x_{11} are varied and thereby trace out the path indicated by the blue line, starting at the purple square and ending at the yellow circle.

Question 4. Consider again the Beijing air pollution dataset. In Figure 3 the features x_8 and x_{11} from Table 1 are plotted as black dots. Recall the data is temporally ordered, and suppose over a period of time the measurements undergoes an evolution indicated by the path here shown as a blue line which begins at the purple square, and ends at the yellow circle and where the other features can be considered fixed.

We can imagine the dataset, along with the path, is projected onto the first two principal components given in Equation (1). Which one of the four plots in Figure 2 shows the path?

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 4. Since we don't know the exact values of most of the x_i-coordinates, it is easier to work with the difference between the start and end-points in Figure 2 and translate them into the difference of the start

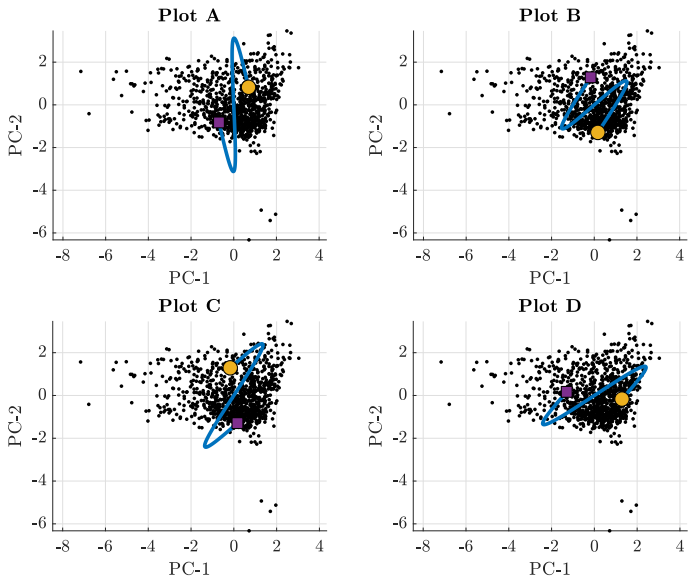

Figure 3: Candidate plots of the observations and path shown in Figure 2 projected onto the first two principal components considered in Equation (1). The start point is indicated by the purple square and the end point by the yellow circle.

and end points in the PCA projections. Notice from Figure 2 we can immediately compute:

$$\Delta \boldsymbol{x} = x_{\text{end}} - x_{\text{start}} = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 3.74 \\ 0.0 \\ 0.0 \end{bmatrix}$$

(this corresponds to the vector going from the start to end points). Then, all we need is to compute the PCA projection of this vector as:

$$\Delta \boldsymbol{b} = (\Delta \boldsymbol{x})^\top \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} -0.34 \\ 2.58 \end{bmatrix}$$

Which should be the vector beginning at the start-point and terminating at the end-point in the PCA projected plots. This rules out all plots except option C.

Question 5. Consider the Beijing air pollution dataset (but for this problem in the non-standardized version). The empirical covariance matrix of the first 5 attributes $x_1, \ldots, x_5$ is:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 12 & -29 & -21 & -12 & -317 \\ -29 & 6104 & 6026 & 1557 & 67964 \\ -21 & 6026 & 7263 & 1701 & 70892 \\ -12 & 1557 & 1701 & 1012 & 25415 \\ -317 & 67964 & 70892 & 25415 & 1212707 \end{bmatrix}.$$

What is the empirical correlation of MONTH and $\text{PM}_{2.5}$?

A. -5.38516

B. -0.0199

C. -0.10715

D. -0.0004

E. Don't know.

Solution 5. Recall the correlation is defined as

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

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

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

Therefore we get:

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

Then finally observe from Table 1 we get that MONTH corresponds to x_1 and $\text{PM}_{2.5}$ corresponds to x_2 so $i = 1$ and $j = 2$ and therefore by simple insertion option C is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	4.2	8.3	3.9	3.8	4.6	6.3	4.8	7.1	4.9
o_2	4.2	0.0	7.4	2.6	3.0	3.2	5.3	3.1	6.6	4.6
o_3	8.3	7.4	0.0	6.3	7.1	5.5	2.8	5.4	2.4	5.3
o_4	3.9	2.6	6.3	0.0	1.5	1.6	4.1	1.8	5.3	2.4
o_5	3.8	3.0	7.1	1.5	0.0	2.4	4.9	2.8	5.8	3.2
o_6	4.6	3.2	5.5	1.6	2.4	0.0	3.7	1.7	4.8	2.3
o_7	6.3	5.3	2.8	4.1	4.9	3.7	0.0	3.8	1.9	3.6
o_8	4.8	3.1	5.4	1.8	2.8	1.7	3.8	0.0	4.9	2.1
o_9	7.1	6.6	2.4	5.3	5.8	4.8	1.9	4.9	0.0	4.4
o_{10}	4.9	4.6	5.3	2.4	3.2	2.3	3.6	2.1	4.4	0.0

Table 3: The pairwise Euclidian distances, $d(o_i, o_j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 10 observations from the Beijing air pollution dataset (recall that $M = 11$). 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 light pollution level), the red observations $\{o_3,\ o_4,\ o_5,\ o_6\}$ belongs to class C_2 (corresponding to a medium pollution level), and the blue observations $\{o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to a high pollution level).

Question 6. To examine if observation o_5 may be an outlier, we will calculate the average relative density using the Euclidean distance based on 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 o_5 for $K = 2$ nearest neighbors?

A. 0.41

B. 0.82

C. 1.0

D. 0.51

E. Don't know.

Solution 6.

To solve the problem, first observe the $k = 2$ neighborhood of o_5 and density is:

$$N_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = \{o_4, o_6\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 5}}(\boldsymbol{x}_5) = 0.513$$

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

$$N_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = \{o_5, o_6\}, \quad N_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = \{o_4, o_8\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = 0.645, \text{density}_{\boldsymbol{X}_{\setminus 6}}(\boldsymbol{x}_6) = 0.606.$$

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

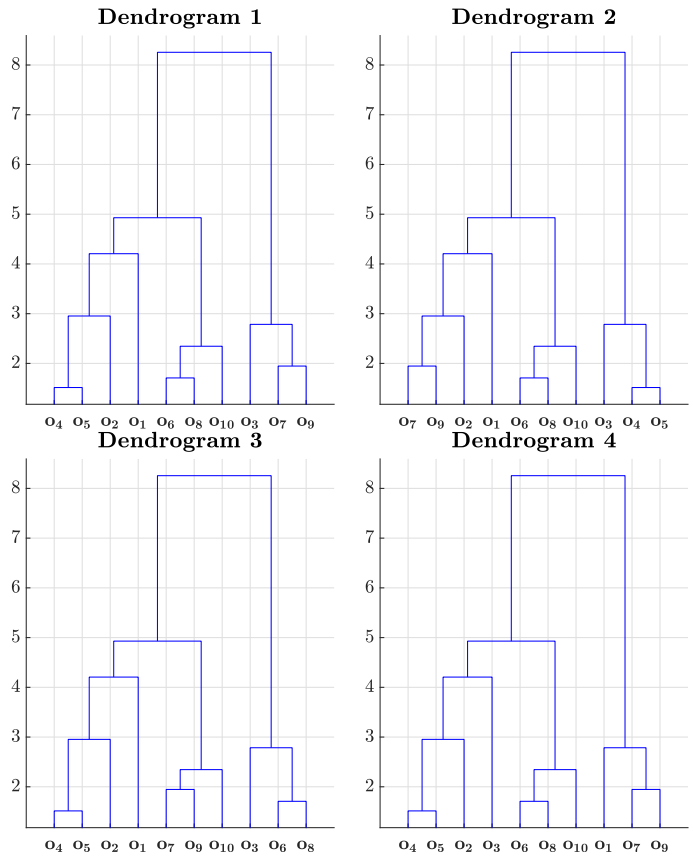

Figure 4: Proposed hierarchical clustering of the 10 observations in Table 3.

Question 7. A hierarchical clustering is applied to the 10 observations in Table 3 using *maximum* linkage. Which one of the dendrograms shown in Figure 4 corresponds to the distances given in Table 3?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

Solution 7. The correct solution is A. We can rule out the other solutions by observing the first merge operation at which they diverge from the correct solution.

- In dendrogram 2, merge operation number 1 should have been between the sets $\{f_4\}$ and $\{f_5\}$ at a height of 1.52, however in dendrogram 2 merge number 1 is between the sets $\{f_7\}$ and $\{f_9\}$.
- In dendrogram 3, merge operation number 2 should have been between the sets $\{f_6\}$ and $\{f_8\}$ at a height of 1.71, however in dendrogram 3 merge number 2 is between the sets $\{f_7\}$ and $\{f_9\}$.
- In dendrogram 4, merge operation number 5 should have been between the sets $\{f_3\}$ and $\{f_7, f_9\}$ at a height of 2.79, however in dendrogram 4 merge number 5 is between the sets $\{f_1\}$ and $\{f_7, f_9\}$.

Question 8. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two binary vectors of dimension $N = 1500$ such that $\boldsymbol{x}_1$ has one non-zero element and $\boldsymbol{x}_2$ has 1498 non-zero elements. What are the possible range of values of the Jaccard similarities of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$?

A. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00242]$

B. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00067]$

C. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00074]$

D. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) \in [0; 0.00206]$

E. Don't know.

Solution 8. To solve this problem, recall the Jaccard similarity is defined as

$$\frac{n_{11}}{N - n_{00}}$$

note it is possible for n_{11}, the number of coordinates where both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are non-zero, to be zero (in the case $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are sorted in ascending and descending order respectively) and hence the Jaccard similarity can be zero. On the other extreme, the maximal Jaccard similarity is obtained when n_{11} is as large as possible while n_{00} is as small as possible. This evidently occurs when the two arrays are sorted in the same order. In this case the overlap is

$$n_{11} = \min\{1, 1498\} = 1$$

and similarly

$$n_{00} = \min\{N - 1, N - 1498\} = 2.$$

Corresponding to a Jaccard similarity of 0.00067. We therefore see that B is correct.

Question 9. Consider again the Beijing air pollution dataset in Table 1. We would like to predict a pollution level 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_2, x_4, x_6, x_9, and x_{11} and in Table 4 we have pre-computed the estimated

Feature(s)	Training RMSE	Test RMSE
none	2.235	2.851
x_2	2.096	2.232
x_4	1.902	1.793
x_6	2.214	2.351
x_9	2.183	3.227
x_{11}	2.235	2.83
$x_2,\ x_4$	1.9	1.797
$x_2,\ x_6$	2.081	2.597
$x_4,\ x_6$	1.777	2.785
$x_2,\ x_9$	1.606	3.09
$x_4,\ x_9$	1.724	2.243
$x_6,\ x_9$	2.087	2.307
$x_2,\ x_{11}$	2.046	2.754
$x_4,\ x_{11}$	1.87	2.143
$x_6,\ x_{11}$	2.214	2.37
$x_9,\ x_{11}$	2.177	3.058
$x_2,\ x_4,\ x_6$	1.773	2.838
$x_2,\ x_4,\ x_9$	1.574	2.81
$x_2,\ x_6,\ x_9$	1.605	3.187
$x_4,\ x_6,\ x_9$	1.691	2.698
$x_2,\ x_4,\ x_{11}$	1.868	2.188
$x_2,\ x_6,\ x_{11}$	2.003	3.738
$x_4,\ x_6,\ x_{11}$	1.723	3.472
$x_2,\ x_9,\ x_{11}$	1.483	4.246
$x_4,\ x_9,\ x_{11}$	1.714	2.418
$x_6,\ x_9,\ x_{11}$	2.081	2.159
$x_2,\ x_4,\ x_6,\ x_9$	1.549	3.174
$x_2,\ x_4,\ x_6,\ x_{11}$	1.676	4.227
$x_2,\ x_4,\ x_9,\ x_{11}$	1.469	3.944
$x_2,\ x_6,\ x_9,\ x_{11}$	1.459	5.017
$x_4,\ x_6,\ x_9,\ x_{11}$	1.667	3.146
$x_2,\ x_4,\ x_6,\ x_9,\ x_{11}$	1.406	5.006

Table 4: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y_r in the Beijing air pollution dataset using different combinations of the features $x_2,\ x_4,\ x_6,\ x_9$, and x_{11}.

training and test error for the different variable combinations. Which of the following statements is correct?

A. Backward selection will select attributes $x_6,\ x_9,\ x_{11}$

B. Backward selection will select attributes $x_4,\ x_6,\ x_9,\ x_{11}$

C. Forward selection will select attributes $x_6,\ x_9,\ x_{11}$

D. Forward selection will select attributes $x_4,\ x_6,\ x_9,\ x_{11}$

E. Don't know.

Solution 9.

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 2.851.

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 $\{x_2\}$, $\{x_4\}$, $\{x_6\}$, $\{x_9\}$, $\{x_{11}\}$. Since the lowest error of the available sets is 1.793, which is lower than 2.851, we update the current selected variables to $\{x_4\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_4\}$ and adding each of the left-out variables thereby resulting in the sets $\{x_2, x_4\}$, $\{x_2, x_6\}$, $\{x_4, x_6\}$, $\{x_2, x_9\}$, $\{x_4, x_9\}$, $\{x_6, x_9\}$, $\{x_2, x_{11}\}$, $\{x_4, x_{11}\}$, $\{x_6, x_{11}\}$, $\{x_9, x_{11}\}$. 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 $\{x_2, x_4, x_6, x_9, x_{11}\}$ having an error of 5.006.

Step $i = 1$ The available variable sets to choose between is obtained by taking the current variable set $\{x_2, x_4, x_6, x_9, x_1 1\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_2, x_4, x_6, x_9\}$, $\{x_2, x_4, x_6, x_{11}\}$, $\{x_2, x_4, x_9, x_{11}\}$, $\{x_2, x_6, x_9, x_{11}\}$, $\{x_4, x_6, x_9, x_{11}\}$. Since the lowest error of the available sets is 3.146, which is lower than 5.006, we update the current selected variables to $\{x_4, x_6, x_9, x_{11}\}$

Step $i = 2$ The available variable sets to choose between is obtained by taking the current variable set $\{x_4, x_6, x_9, x_1 1\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_2, x_4, x_6\}$, $\{x_2, x_4, x_9\}$, $\{x_2, x_6, x_9\}$, $\{x_4, x_6, x_9\}$, $\{x_2, x_4, x_{11}\}$, $\{x_2, x_6, x_{11}\}$, $\{x_4, x_6, x_{11}\}$, $\{x_2, x_9, x_{11}\}$, $\{x_4, x_9, x_{11}\}$, $\{x_6, x_9, x_{11}\}$. Since the lowest error of the available sets is 2.159, which is lower than 3.146, we update the current selected variables to $\{x_6, x_9, x_{11}\}$

Step $i = 3$ The available variable sets to choose between is obtained by taking the current variable set $\{x_6, x_9, x_1 1\}$ and removing each of the left-out variables thereby resulting in the sets $\{x_2, x_4\}$, $\{x_2, x_6\}$, $\{x_4, x_6\}$, $\{x_2, x_9\}$, $\{x_4, x_9\}$, $\{x_6, x_9\}$, $\{x_2, x_{11}\}$, $\{x_4, x_{11}\}$, $\{x_6, x_{11}\}$, $\{x_9, x_{11}\}$. Since the lowest error of the newly constructed sets is not lower than the current error the algorithm terminates.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}
o_1	0	0	0	0	0	1	1	1	0	1	1
o_2	1	0	0	1	0	1	1	0	1	1	1
o_3	1	1	1	1	1	0	0	0	0	1	0
o_4	0	1	0	1	0	0	0	1	0	1	0
o_5	0	0	0	0	0	1	0	1	1	1	0
o_6	0	1	1	1	1	0	0	0	1	1	0
o_7	1	1	1	1	1	0	0	1	0	1	0
o_8	0	1	1	1	1	0	0	0	1	1	0
o_9	1	1	1	1	1	0	0	1	0	1	0
o_{10}	0	1	1	1	1	0	0	1	0	1	0

Table 5: Binarized version of the Beijing air pollution 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 light pollution level), the red observations $\{o_3,\ o_4,\ o_5,\ o_6\}$ belongs to class C_2 (corresponding to a medium pollution level), and the blue observations $\{o_7,\ o_8,\ o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to a high pollution level).

Question 10. We again consider the Beijing air pollution dataset from Table 1 and the $N = 10$ observations we already encountered in Table 3. The data is processed to produce 11 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 $N \times M = 10 \times 11$ binary matrix in Table 5. We wish to apply a Bayesian classifier to the dataset and as part of this task we have to estimate the probability

$$p(f_2 = 0,\ f_{11} = 0|y = 2).$$

For better numerical stability, we will use robust estimation to obtain the probability by introducing a regularization factor of α, viz.:

$$p(A|B) = \frac{\{\text{Occurences matching } A \text{ and } B\} + \alpha}{\{\text{Occurences matching } B\} + 2\alpha}.$$

[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.

What is the probability estimate if $\alpha = 1$?

A. $p(f_2 = 0,\ f_{11} = 0|y = 2) = \frac{1}{3}$

B. $p(f_2 = 0,\ f_{11} = 0|y = 2) = \frac{3}{5}$

C. $p(f_2 = 0,\ f_{11} = 0|y = 2) = 0$

D. $p(f_2 = 0,\ f_{11} = 0|y = 2) = 1$

E. Don't know.

Solution 10. Of the observations in class $y = 2$ only 1 have simultaneously $f_2 = 0,\ f_{11} = 0$. As this class contains *four* observations, we see the answer is

$$\frac{1+\alpha}{4+2\alpha} = \frac{2}{6}$$

Therefore, answer A is correct.

Question 11. Consider the binarized version of the Beijing air pollution dataset shown in Table 5.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 11$ items $f_1, f_2, \ldots, f_{11}$. Which of the following options represents all (non-empty) itemsets with support greater than 0.65 (and only itemsets with support greater than 0.65)?

A. $\{f_4\}$, $\{f_{10}\}$, $\{f_4, f_{10}\}$

B. $\{f_2\}$, $\{f_4\}$, $\{f_{10}\}$, $\{f_2, f_4\}$, $\{f_2, f_{10}\}$, $\{f_4, f_{10}\}$, $\{f_2, f_4, f_{10}\}$

C. $\{f_2\}$, $\{f_4\}$, $\{f_{10}\}$, $\{f_2, f_4\}$, $\{f_2, f_{10}\}$, $\{f_4, f_{10}\}$

D. $\{f_{10}\}$

E. Don't know.

Solution 11. 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 0.65, an itemset needs to be contained in 7 rows. It is easy to see this rules out all options except B.

Question 12. We again consider the binary matrix from Table 5 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 11$ items $f_1, \ldots, f_{11}$.

What is the *confidence* of the rule $\{f_1, f_3, f_4, f_5, f_8\} \rightarrow \{f_2, f_{10}\}$?

A. The confidence is 1

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

C. The confidence is $\frac{2}{7}$

D. The confidence is $\frac{9}{20}$

E. Don't know.

Solution 12. The confidence of the rule is easily computed as

$$\frac{\text{support}(\{f_1, f_3, f_4, f_5, f_8\} \cup \{f_2, f_{10}\})}{\text{support}(\{f_1, f_3, f_4, f_5, f_8\})} = \frac{\frac{1}{5}}{\frac{1}{5}} = 1.$$

Therefore, answer A is correct.

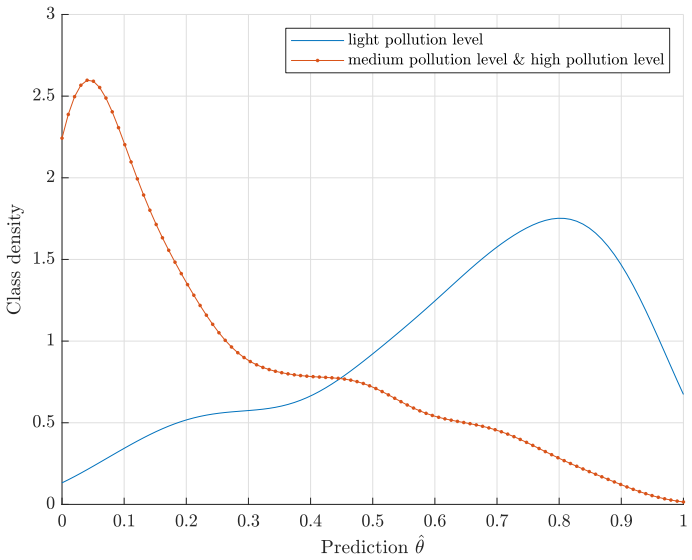

Figure 5: Class density (as function of the predictions of a logistic regression classifier $\hat{\theta}$) of the two-class problem of predicting *light pollution level* vs. *medium pollution level & high pollution level*.

Question 13. A logistic regression classifier is applied to the Beijing air pollution dataset described in Table 1 to solve the binary classification problem of *light pollution level* (positive class) vs. *medium pollution level & high pollution level* (negative class). The output of the classifier is the class-assignment probability $\hat{\theta}$, and for each threshold value θ_0 we assign observations with $\hat{\theta} > \theta_0$ to the positive class *light pollution level* (and otherwise to the negative class *medium pollution level & high pollution level*).

Suppose the class-density for each class is as indicated in Figure 5, which of the reciever operator characteristic (ROC) curves in Figure 5 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. Recall we compute the ROC curve from the false positive rate (FPR) and true positive rate (TPR) for particular choices of threshold value $\hat{\theta}_0$

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

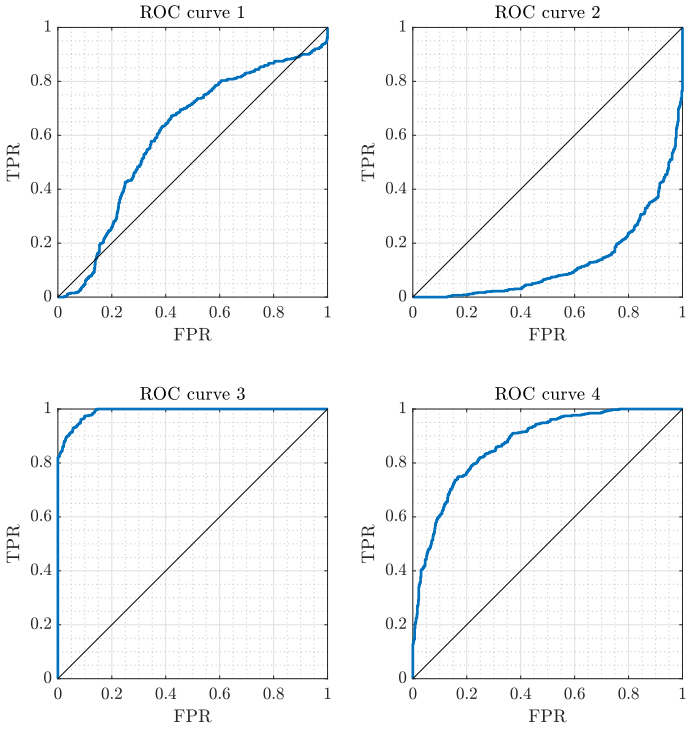

Figure 6: Proposed ROC curves for the two-class classifier described in Figure 5.

Furthermore, compute e.g. the TPR, one assumes every observation predicted to belong to *medium pollution level & high pollution level* (positive class) if $\hat{\theta} > \theta_0$ and otherwise to *light pollution level* (negative class)

We then divide the total number of observations belonging to positive class *and which are predicted to belong to the positive class* 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 the positive class *but in fact belongs to the negative class*, divided by the total number of observations in the *negative* class. For concreteness, we have inserted the true values of the TPR and FPR as function of θ_0 in Figure 7.

To reason about the options, first, observe that the location of the two humps in Figure 5 implies an AUC well over 0.5: Consider for instance a value of $_0$ between them in which the TPR is much larger than the FPR.

Next, consider the case where θ_0 is very small and increases, and recall this corresponds to the point $(1, 1)$ on the ROC curve. Since for low values of θ_0 the negative class has a large density (and the positive class has a low but non-zero density), this implies the FPR has to decrease much less rapidly than the TPR, because the number negative predictions falsely

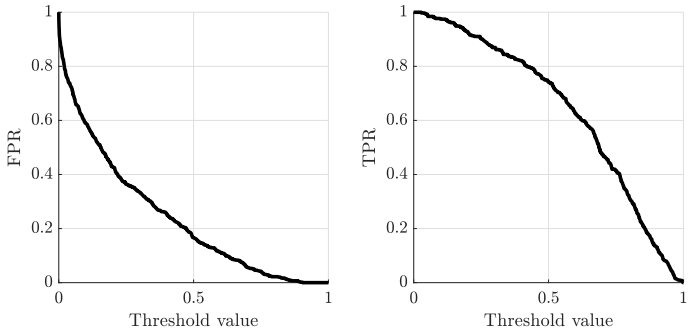

Figure 7: TPR, FPR curves for the classifier.

classified as positive decreases much more rapidly than the number of positive class members predicted to be negative. In other words, the ROC curve must initially stay above the line of identity.

At the same time, this rules out the case that e.g. the FPR decreases while the TPR remains at 1, as the two classes overlap. In other words, if the FPR ¡ 1 then so will the TPR. These observations together allows us to rule out all options except D.

	$i=1$	$i=2$	$i=3$	$i=4$
x_i	2	5	6	7
y_i	6	7	7	9

Table 6: Simple 1d regression dataset

Question 14. Consider the small 1d dataset shown in Table 6 comprised of $N = 4$ observations and where the goal is to predict y_i given x_i. Suppose we apply ridge regression to the problem in the form described in the lecture notes, Section 14.1.

If $\lambda = 2$, what is the ridge regression cost function assuming the weight-vector is

$$\boldsymbol{w} = \begin{bmatrix} 0.6 \end{bmatrix}$$

i.e. $E_\lambda(\boldsymbol{w}, w_0)$?

A. $E_\lambda(\boldsymbol{w}, w_0) = 1.205$

B. $E_\lambda(\boldsymbol{w}, w_0) = 1.97$

C. $E_\lambda(\boldsymbol{w}, w_0) = 1.033$

D. $E_\lambda(\boldsymbol{w}, w_0) = 2.662$

E. Don't know.

Solution 14. The cost function is defined as

$$E_\lambda(\boldsymbol{w}, w_0) = \sum_{i=1}^{4} (y_i - \hat{y}_i)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w}.$$

Where $\hat{y}_i$ are the predictions. According to Section 14.1 these are computed from the *standardized* feature matrix as:

$$\hat{y} = \frac{x - \mu}{\sigma} w + \mathbb{E}[y].$$

Where μ and σ is the mean and standard deviations of x as computed on the training set in Table 6, i.e. $\mu = 5.0$ and $\sigma = 2.16$. Since $\mathbb{E}[y] = \frac{1}{N}\sum_{i=1}^{N} y_i = 7.25$ we find that the predicted values of y are:

$$\hat{\boldsymbol{y}} = \begin{bmatrix} 6.417 & 7.25 & 7.528 & 7.805 \end{bmatrix}.$$

Inserting these in the cost function we get:

$$E_\lambda(\boldsymbol{w}, w_0) = 0.42^2 + 0.25^2 + 0.53^2 + 1.19^2 + \lambda 0.36 = 2.662$$

hence D is correct.

	1	2	3	4	5	6
x_7	-1.76	-0	0.06	0.08	0.65	1.3
y_r	12	6	8	10	4	2

Table 7: Values of x_7 and the corresponding value of y_r.

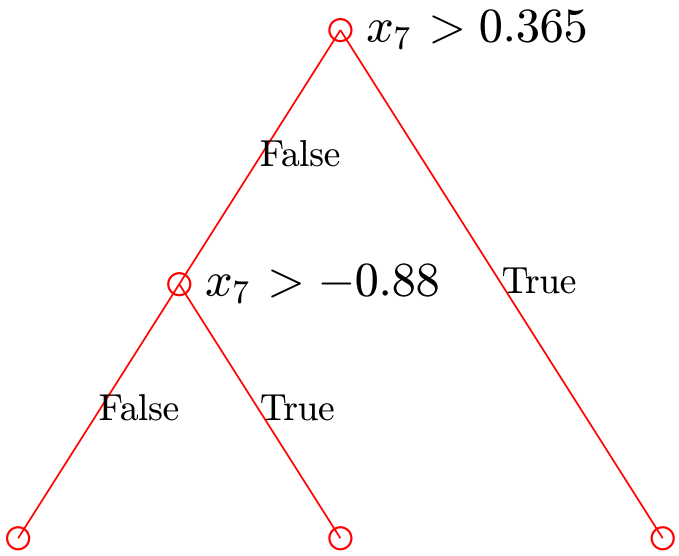

Figure 8: Structure of decision tree. The goal is to determine the splitting rules.

Question 15. We will consider the first 6 observations of the Beijing air pollution dataset shown in Table 3. Table 7 shows their corresponding value of x_7 and y_r. We fit a small regression tree to this dataset, the structure (and binary splitting rules) is depicted in Figure 8. What is the predicted value $\hat{y}_r$ as evaluated at $x_7 = 0.5$?

A. $\hat{y}_r = 2.85$

B. $\hat{y}_r = 3.0$

C. $\hat{y}_r = 2.52$

D. $\hat{y}_r = 0.98$

E. Don't know.

Solution 15.

The predicted value for a given input is computed as the average y-value of those observations in the training set which is assigned to the same leaf node v as the input, i.e.

$$y(v) = \frac{1}{N(v)} \sum_{i \in v} y_i$$

(see the section on regression trees in lecture notes). Therefore, we first need to find out which leaf node the observation is assigned to. To do this, start at the root and compare $x_7 = 0.5$ to the rule in the split

$$x_7 > 0.365$$

and we continue down the right branch. Continuing in this manner, we see $x_7 = 0.5$ is classified to leaf number one from the left. Then, proceeding in the same manner with the x_7 observations in Table 7, we see the observations o_5, and o_6 are also assigned to leaf one (counted from the left). According to the above the prediction is then simply the average of their y-value

$$\hat{y} = \frac{1}{2}(4 + 2)$$

or $\hat{y} = 3.0$, hence B is correct.

Question 16. In this problem, we will again consider the 6 observations from the Beijing air pollution dataset shown in Table 7. Recall Figure 8 shows the structure of the small regression tree fitted to this dataset using Hunt's algorithm along with the thereby obtained binary splitting rules. What was the purity gain Δ of the first split Hunt's algorithm accepted?

A. $\Delta = 11.67$

B. $\Delta = 3.67$

C. $\Delta = 8.0$

D. $\Delta = 56.0$

E. Don't know.

Solution 16.

The first split Hunt's algorithm accepted must be the split at the root, i.e.

$$x > 0.365$$

Partitioning the observations in Table 7 according to this split results in the two sets

$$v_1 = \{1, 2, 3, 4\}, \quad v_2 = \{5, 6\}$$

at the two legs. The impurity of these two sets, and the impurity of all y-values, is computed using the impurity measure appropriate for regression trees

$$I(v) = \frac{1}{N(v)} \sum_{i \in v_1} (y_i - y(v))^2$$

where $y(v)$ is the average of the y-values in v_i. Specifically

$$y(v_1) = 9.0, \quad y(v_2) = 3.0$$

And therefore, with a similar calculation for the set at the root node v_0 corresponding to all 6 observations,

$$y(v_0) = 7$$

Therefore:

$$I(v_0) = 11.67, \quad I(v_1) = 5.0, \quad I(v_2) = 1.0$$

these are finally combined to the impurity gain as

$$\Delta = I(v_0) - \sum_{k=1}^{2} \frac{N(v_k)}{N} I(v_k)$$

where for instance $N(v_1) = 4$ are the number of observations in branch 1. We find by insertion that $\Delta = 8.0$ and hence C is correct.

Question 17. Consider once more the Beijing air pollution dataset treated as a regression problem where the goal is to predict y_r. We wish to do this using KNN regression using $K = 3$ neighbors. We will simplify the problem by only considering the first $N = 6$ observations whose pairwise distances are given in Table 3, and their corresponding y_r-value can be found in Table 7.

Suppose we evaluate the leave-one-out estimate of the generalization error defined as

$$E = \frac{1}{N}\sum_{i=1}^{N} L(y_{r,i}, \hat{y}_{r,i})$$

where $y_{r,i}$ is the y_r-value of observation i, $\hat{y}_{r,i}$ is the predicted value and L is the standard squared (Euclidian) loss.

It is too time-consuming to compute the full LOO estimate of the generalization error, but what is the contribution from observation $i = 1$?

A. $L(y_{r,1}, \hat{y}_{r,1}) = 6.667$

B. $L(y_{r,1}, \hat{y}_{r,1}) = 28.444$

C. $L(y_{r,1}, \hat{y}_{r,1}) = 6.0$

D. $L(y_{r,1}, \hat{y}_{r,1}) = 7.111$

E. Don't know.

Solution 17.

First, notice that by a simple lookup in Table 7 that $y_{r,1} = 12$. To compute the predicted value, note that the K-nearest neighbors to observation $i = 1$ (but not including 1 itself) are observations

$$\{o_2, o_4, o_5\}$$

according to Table 3. The predicted value is then the mean of the corresponding y-values according to table Table 7 or

$$\hat{y}_{r,1} \approx 6.667$$

We can then simply compute the squared loss as:

$$L(y_{r,1}, \hat{y}_{r,1}) = (y_{r,1} - \hat{y}_{r,1})^2 = 28.444$$

and therefore option B is correct.

Fold	M_1/M_2	$M_1/\overline{M}_2$	$\overline{M}_1/M_2$	$\overline{M}_1/\overline{M}_2$
1	134	40	24	47
2	141	31	26	48
3	131	23	25	66
4	132	30	25	58

Table 8: Outcome of cross-validation. Rows are combination of outcomes of the two models.

Question 18. We will consider the Beijing air pollution dataset, and compare two models for predicting the class label y. Specifically, let M_1 be a $K = 1$ nearest neighbor classification model and M_2 a $K = 5$ nearest neighbor classification model. To compare them statistically, we perform $K = 4$ fold cross-validation, and for each fold we record the number of observations where both models are correct (as M_1/M_2), M_1 is correct and M_2 wrong (as $M_1/\overline{M_2}$), and so on. The outcome can be found in Table 8.

These results are sufficient to perform the McNemar test to compare the performance difference, i.e. the difference in accuracy, of model M_1 and M_2. According to the McNemar test, what is the estimated difference in accuracy

$$\hat{\theta} = \text{acc}(M_1) - \text{acc}(M_2)$$

of the two models?

A. $\hat{\theta} = 0.75$

B. $\hat{\theta} = 0.07$

C. $\hat{\theta} = 0.11$

D. $\hat{\theta} = 0.02$

E. Don't know.

Solution 18.

While the accuracies can easily be computed explicitly from the information in Table 8, a simpler solution (which is more true to the lecture notes) is to observe the differences in accuracies is

$$\hat{\theta} = \frac{n_{12} - n_{21}}{N}$$

Where n_{12} are the number of times model M_1 is correct and M_2 is false (i.e. the sum of column 2) and similarly n_{21} is the sum of column 3 and finally $N = 981$ is the number of observations. We find

$$\hat{\theta} = \frac{124 - 100}{N} = 0.02$$

and therefore D is correct.

Question 19. We will again consider the result of the two KNN models in Table 8 as evaluated over the $K = 4$ folds. What is the Jeffreys $\alpha = 0.05$ confidence interval $[\theta_L, \theta_U]$ of the model M_2?

A.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 538.5, b = 443.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 538.5, b = 443.5)$$

B.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 638.5, b = 343.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 638.5, b = 343.5)$$

C.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 538.5, b = 219.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 538.5, b = 219.5)$$

D.

$$\theta_L = \text{cdf}_B^{-1}(0.025|a = 662.5, b = 319.5),$$
$$\theta_U = \text{cdf}_B^{-1}(0.975|a = 662.5, b = 319.5)$$

E. Don't know.

Solution 19.

Since the cross-validation folds are non-overlapping, we can easily find the number of times model M_2 makes a correct prediction as the sum of columns 1 and 3 in Table 8 or $n^+ = 638$. Similarly, the sum of all entries in column 2 and 4 are the number of wrong guesses or $n^- = 343$.

The lower limit of the Jeffreys interval is now defined as

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

and the upper limit can be found from the same expression by replacing $\frac{\alpha}{2}$ with $1 - \frac{\alpha}{2}$. Therefore, B is correct.

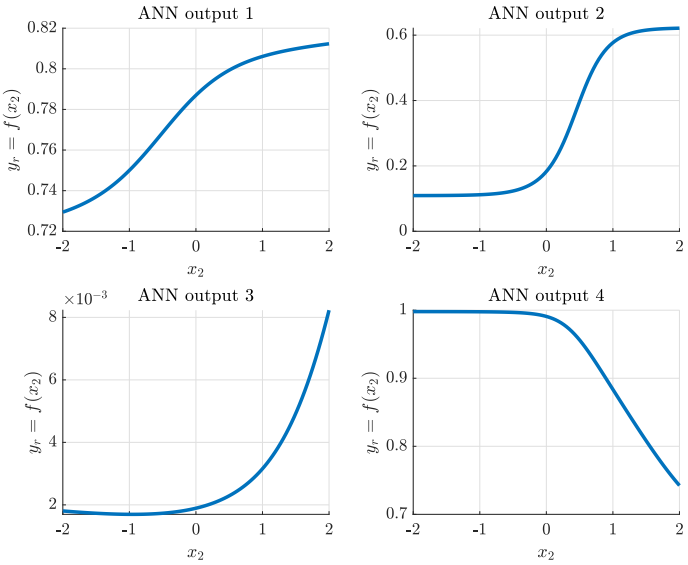

Figure 9: Suggested outputs of an ANN trained on the attribute x_2 from the Beijing air pollution dataset to predict y_r.

Question 20. ***Notice:*** *The version of question 20 in the main exam set contains a minor misprint in the axis on Figure 8 and the text to Figure 8. The misprint has been corrected in this version. Use this version when answering the question.*

We will consider an artificial neural network (ANN) trained on the Beijing air pollution dataset described in Table 1 to predict y_r from the attribute x_2. Suppose the neural network takes the form:

$$f(x, \boldsymbol{w}) = h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}([1\ x]\boldsymbol{w}_j^{(1)})\right).$$

where the activation functions are selected as $h^{(1)}(x) = \sigma(x)$ (the logistic sigmoid activation function) and $h^{(2)}(x) = \sigma(x)$ (the logistic sigmoid activation function) and the weights are given as:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -0.5 \\ -0.1 \end{bmatrix}, \quad \boldsymbol{w}_2^{(1)} = \begin{bmatrix} 0.9 \\ 2.0 \end{bmatrix},$$
$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -1.0 \\ 0.4 \end{bmatrix}, \quad w_0^{(2)} = 1.4.$$

Which one of the curves in Figure 9 will then corre-

spond 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 20.

It suffices to compute the activation of the neural network at $x_2 = -2$. The activation of each of the two hidden neurons is:

$$n_1 = h^{(1)}(\begin{bmatrix}1 & -2\end{bmatrix} \boldsymbol{w}_1^{(1)}) = 0.426$$
$$n_2 = h^{(1)}(\begin{bmatrix}1 & -2\end{bmatrix} \boldsymbol{w}_2^{(1)}) = 0.043.$$

The final output is then computed as:

$$f(x, \boldsymbol{w}) = h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} h^{(1)}(\begin{bmatrix}1 & x_1 & x_2\end{bmatrix} \boldsymbol{w}_j^{(1)})\right)$$
$$= h^{(2)}\left(w_0^{(2)} + \sum_{j=1}^{2} w_j^{(2)} n_j\right) = 0.729.$$

This rules out all options except A.

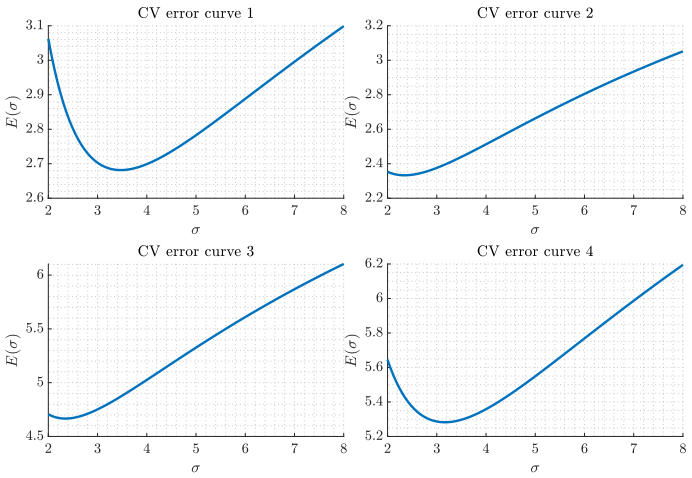

Figure 10: Estimated negative log-likelihood as obtained using hold-out cross validation on a small, $N = 3$ one-dimensional dataset as a function of kernel width σ.

Question 21. Consider the following $N = 3$ observations of the attribute CO from the Beijing air pollution dataset described in Table 1.

$$x_5 : \quad \begin{bmatrix}4.5 & -0.5 & 1.2\end{bmatrix}.$$

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 hold-out cross validation (CV) using the average (per observation) negative log-likelihood

$$E(\sigma) = \frac{-1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \log p_\sigma(x_i).$$

We construct the hold out split by considering the first 2 observations a training set and the last observations as a test set.

Which of the cross validation curves in Figure 10 shows the cross-validation estimate of the generalization error $E(\sigma)$?

A. CV error curve 1

B. CV error curve 2

C. CV error curve 3

D. CV error curve 4

E. Don't know.

Solution 21. To solve the problem, we will compute the hold-out 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:

$$p_\sigma(x_i) = \frac{1}{N-1}\sum_{j\neq i}\mathcal{N}(x_i|x_j, \sigma = 2)$$

Therefore, training on the two first observations and testing on the last simply corresponds to evaluating this expression for $i = 3$ i.e.:

$$p_\sigma(x_3) = 0.095$$

The CV hold-out error is the average of the test set, but since the test set only contains a single observation it is equal to minus the log of the above expression. In other words

$$\begin{aligned} E(\sigma = 2) &= \frac{1}{N^{\text{test}}}\sum_{i=1}^{N^{\text{test}}} -\log p_\sigma(x_i) \\ &= -\log p_\sigma(x_3) = 2.353. \end{aligned}$$

Therefore, the correct answer is B.

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

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) for the first round of boosting $t = 1$.

Question 22. Consider again the Beijing air pollution 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_6 and x_9. We use a KNN classification model ($K = 1$) to this dataset and apply AdaBoost to improve the performance. After the first round of boosting, we obtain predictions and the true class labels as tabulated in Table 9. Given this information, how will the AdaBoost update the weights $\boldsymbol{w}$?

A. $[0.173 \quad 0.103 \quad 0.173 \quad 0.103 \quad 0.103 \quad 0.173 \quad 0.173]$

B. $[0.146 \quad 0.138 \quad 0.146 \quad 0.138 \quad 0.138 \quad 0.146 \quad 0.146]$

C. $[0.125 \quad 0.167 \quad 0.125 \quad 0.167 \quad 0.167 \quad 0.125 \quad 0.125]$

D. $[0.102 \quad 0.198 \quad 0.102 \quad 0.198 \quad 0.198 \quad 0.102 \quad 0.102]$

E. Don't know.

Solution 22.

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

$$\{y_2, y_4, y_5\}$$

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

$$\epsilon_{t=1} = \sum_i w_i(t)(1 - \delta_{f_t(x_i),y_i}) = 0.429$$

From this, we compute α_t as

$$\alpha_t = \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} = 0.144$$

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

Question 23. Consider the Beijing air pollution dataset from Table 1 consisting of $N = 981$ observations, and suppose the attribute O_3 concentration ($\mu g/m^3$) has been binarized into low and high values. We still consider the goal to predict the pollution level. Given the following information

- Of the 391 observations with light pollution level, 64 had a high value of O_3 concentration ($\mu g/m^3$)
- Of the 241 observations with medium pollution level, 66 had a high value of O_3 concentration ($\mu g/m^3$)
- Of the 349 observations with high pollution level, 206 had a high value of O_3 concentration ($\mu g/m^3$)

and supposing a particular observation has a low value of O_3 concentration ($\mu g/m^3$), what is the probability of observing medium pollution level?

A. 0.271

B. 0.192

C. 0.044

D. 0.141

E. Don't know.

Solution 23. The problem is solved by applying Bayes rule. Introducing the binary variable x such that $x = 1$ if an observation has a high value of O_3 concentration ($\mu g/m^3$) (and otherwise $x = 0$) the question asked is equivalent to computing $p(y = 2|x = 0)$. Applying Bayes' theorem we get:

$$p(y = 2|x = 0) = \frac{p(x = 0|y = 2)p(y = 2)}{\sum_{k=1}^{3} p(x = 0|y = k)p(y = k)}$$

Recall that $p(x = 0|y) = p(x = 1|y)$, we can obtain the required probabilities from each of the three bullet points above. We obtain:

- $p(y = 1) = \frac{391}{N}$ and $p(x = 1|y = 1) = \frac{64}{391}$.
- $p(y = 2) = \frac{241}{N}$ and $p(x = 1|y = 2) = \frac{66}{241}$.
- $p(y = 3) = \frac{349}{N}$ and $p(x = 1|y = 3) = \frac{206}{349}$.

Plugging these into Bayes theorem, and using that $p(x = 0|y) = 1 - p(x = 1|y)$ because x is binary, we see $p(y = 2|x = 0) = 0.271$ and hence that option A is correct.

Question 24. Consider again the Beijing air pollution dataset. We consider a multinomial regression model applied to the dataset projected onto the first two principal directions, i.e. applied to observations of the form $\boldsymbol{x} = \begin{bmatrix} b_1 & b_2 \end{bmatrix}^\top$ where b_1 and b_2 are the coordinates of the PCA projections.

In the notation of the lecture notes, suppose the weight-vectors in the multinomial regression model are

$$w_1 = \begin{bmatrix} 0.04 \\ 1.32 \\ -1.48 \end{bmatrix}, \quad w_2 = \begin{bmatrix} -0.03 \\ 0.7 \\ -0.85 \end{bmatrix}.$$

What is the class-assignment probability vector $\tilde{\boldsymbol{y}}$ for the input observation with coordinates $b_1 = -5.52$, $b_2 = -4.69$?

A. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.77 & 0.23 & 0.0 \end{bmatrix}^\top$

B. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.26 & 0.39 & 0.35 \end{bmatrix}^\top$

C. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.16 & 0.24 & 0.6 \end{bmatrix}^\top$

D. $\tilde{\boldsymbol{y}} = \begin{bmatrix} 0.22 & 0.07 & 0.72 \end{bmatrix}^\top$

E. Don't know.

Solution 24. Let $\boldsymbol{b}$ be the input vector. Then:

$$\tilde{\boldsymbol{b}} = \begin{bmatrix} 1.0 \\ -5.52 \\ -4.69 \end{bmatrix}.$$

Recall the class-assignment probability vector is computed as

$$P(y = k|\boldsymbol{x}) = \begin{cases} \frac{e^{\hat{y}_k}}{1+\sum_{k'=1}^{2} e^{\hat{y}_{k'}}}. & \text{if } k \leq 2 \\ \frac{1}{1+\sum_{k'=1}^{2} e^{\hat{y}_{k'}}} & \text{if } k = 3. \end{cases}$$

in the case of multinomial regression we have

$$\hat{y}_1 = \tilde{\boldsymbol{b}}^T \boldsymbol{w}_1 \approx -0.305 \quad \hat{y}_2 = \tilde{\boldsymbol{b}}^T \boldsymbol{w}_2 \approx 0.093$$

Simply inserting these number we get that the first coordinate of the class-assignment probability vector is:

$$p(y = 1|\boldsymbol{x}) = \frac{e^{\hat{y}_1}}{1 + e^{\hat{y}_1} + e^{\hat{y}_2}} = 0.26$$

(and similar for the other values of y). From this B is evidently correct.

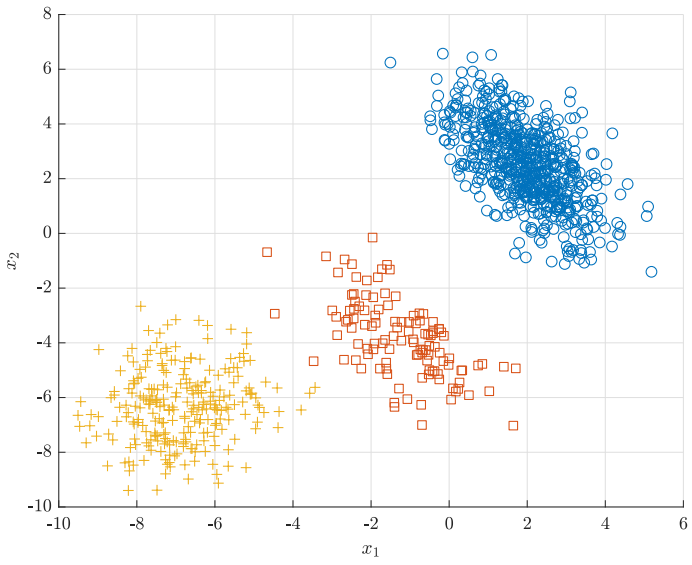

Figure 11: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

Question 25. Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In Figure 11 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 model.
Which one of the following GMM densities was used to

generate the data?

A.

$$p(\boldsymbol{x}) = \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right) + \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right)$$

B.

$$p(\boldsymbol{x}) = \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right) + \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right) + \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right)$$

C.

$$p(\boldsymbol{x}) = \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right) + \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right)$$

D.

$$p(\boldsymbol{x}) = \frac{1}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}2.0\\2.4\end{bmatrix}, \begin{bmatrix}1.3 & 0.3\\0.3 & 2.0\end{bmatrix}\right) + \frac{5}{8}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.2\\-3.8\end{bmatrix}, \begin{bmatrix}1.0 & -0.8\\-0.8 & 1.9\end{bmatrix}\right) + \frac{1}{4}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-6.9\\-6.3\end{bmatrix}, \begin{bmatrix}1.1 & -0.9\\-0.9 & 2.2\end{bmatrix}\right)$$

E. Don't know.

Solution 25.

B 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 12 is shown the densities for densities corresponding to option B (upper left), C (upper right) and D (bottom center).

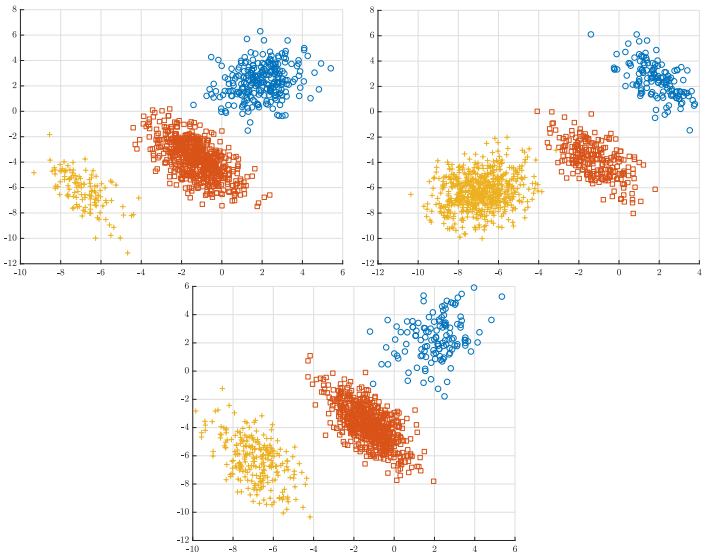

Figure 12: GMM mixtures corresponding to alternative options.

Question 26. Consider the following four classifiers:

MREG: Multinomial regression

ANN: Artificial neural network with 5 hidden units

CT: Classification tree with regular axis-aligned splits ($b_i < c$)

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

Suppose the classifiers are trained on a subset of the Beijing air pollution dataset described in Table 1 after it has been projected onto the first two principal components b_1 and b_2 from Equation (1). The decision boundary for each of the four classifiers is given in Figure 13. Which one of the following statements is correct?

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

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

C. Classifier 1 corresponds to **CT**,
Classifier 2 corresponds to **MREG**,

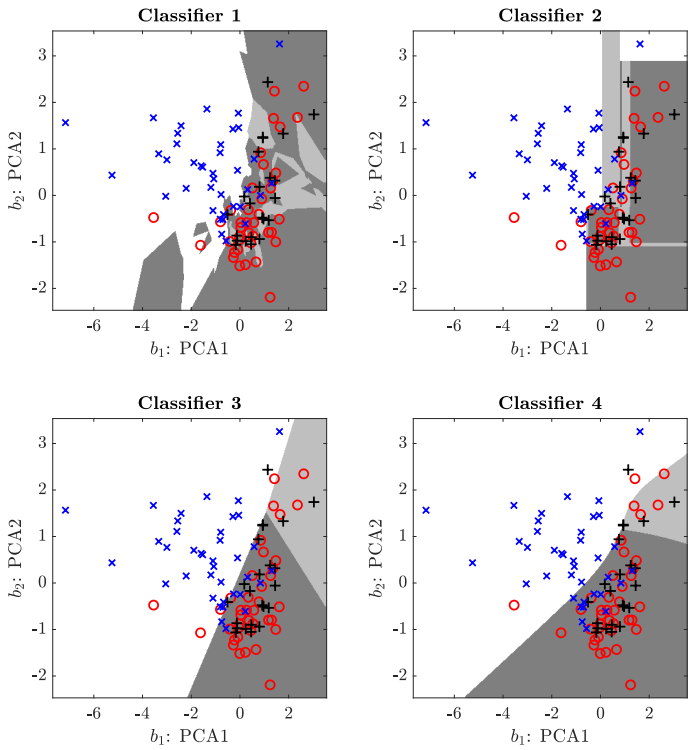

Figure 13: Decision boundaries for four different classifiers trained on the Beijing air pollution dataset when projected onto the first two principal components.

Classifier 3 corresponds to **ANN**,

Classifier 4 corresponds to **KNN**.

D. Classifier 1 corresponds to **ANN**,

Classifier 2 corresponds to **KNN**,

Classifier 3 corresponds to **MREG**,

Classifier 4 corresponds to **CT**.

E. Don't know.

Solution 26. To solve this problem, we have to use our intuition about what the typical decision boundaries for the different methods look like:

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

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

Question 27. Consider a small dataset comprised of $N = 10$ observations

$$x = \begin{bmatrix} 0.4 & 0.5 & 1.1 & 2.2 & 2.6 & 3.0 & 3.6 & 3.7 & 4.9 & 5.0 \end{bmatrix}.$$

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

$$\mu_1 = 2.4, \ \mu_2 = 3.3, \ \mu_3 = 3.5$$

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

A. $\{0.4, 0.5, 1.1, 2.2\}$, $\{2.6, 3, 3.6\}$, $\{3.7, 4.9, 5\}$

B. $\{0.4, 0.5, 1.1\}$, $\{2.2, 2.6, 3, 3.6\}$, $\{3.7, 4.9, 5\}$

C. $\{0.4, 0.5, 1.1, 2.2, 2.6\}$, $\{3, 3.6, 3.7, 4.9\}$, $\{5\}$

D. $\{0.4, 0.5, 1.1, 2.2, 2.6\}$, $\{3, 3.6, 3.7\}$, $\{4.9, 5\}$

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. Given the initial centroids, the K-means algorithm assign observations to the nearest centroid resulting in the partition:

$$\{0.4, 0.5, 1.1, 2.2, 2.6\}, \{3\}, \{3.6, 3.7, 4.9, 5\}.$$

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

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

$$\mu_1 = 1.36, \ \mu_2 = 3, \ \mu_3 = 4.3.$$

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

$$\{0.4, 0.5, 1.1\}, \{2.2, 2.6, 3, 3.6\}, \{3.7, 4.9, 5\}.$$

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

$$\mu_1 = 0.666667,\ \mu_2 = 2.85,\ \mu_3 = 4.53333.$$

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

$$\{0.4, 0.5, 1.1\},\ \{2.2, 2.6, 3, 3.6\},\ \{3.7, 4.9, 5\}.$$

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