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.
**Answers:**

---

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.

| 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.

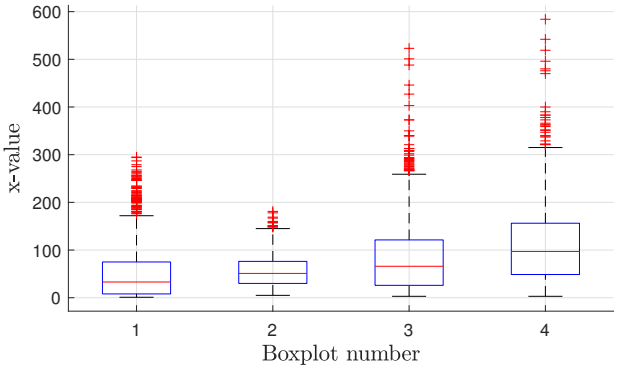


Figure 1: Boxplots corresponding to the variables with summary statistics indicated in Table 2 but not necessarily in that order.

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.

**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.

**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 $\mathbf{PM}_{10}$, 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 $\mathbf{PM}_{10}$, 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.

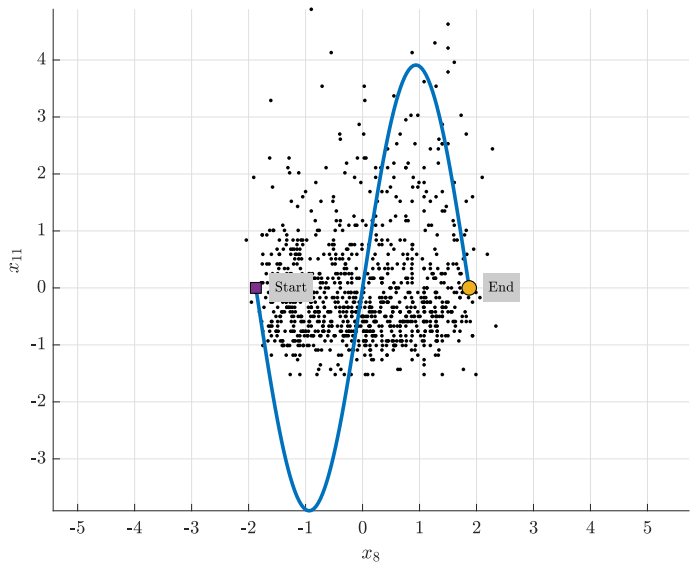


Figure 2: Black dots show attributes $x_8$ and $x_{11}$ of the Beijing air pollution dataset from Table 1. The other attributes are kept fixed while $x_8$ and $x_{11}$ are varied and thereby trace out the path indicated by the blue line, starting at the purple square and ending at the yellow circle.

**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.

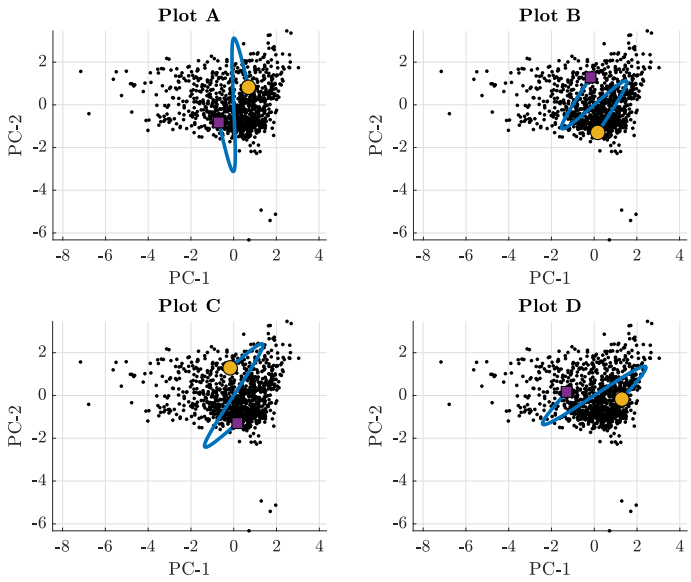


Figure 3: Candidate plots of the observations and path shown in Figure 2 projected onto the first two principal components considered in Equation (1). The start point is indicated by the purple square and the end point by the yellow circle.

**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:

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

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

A. $-5.38516$

B. $-0.0199$

C. $-0.10715$

D. $-0.0004$

E. Don't know.

| | $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ | $o_6$ | $o_7$ | $o_8$ | $o_9$ | $o_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $o_1$ | 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.

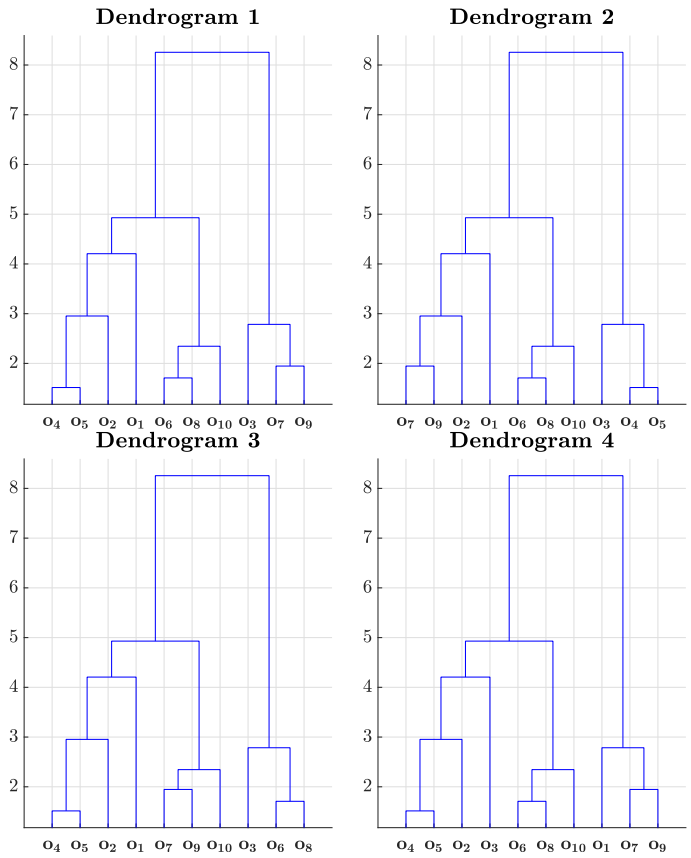


Figure 4: Proposed hierarchical clustering of the 10 observations in Table 3.

**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.

**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.

**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 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.

| 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 $x_{11}$.

| | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ | $f_9$ | $f_{10}$ | $f_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $o_1$ | 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 $\alpha$, 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.

**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.

**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 $\{f_1, f_3, f_4, f_5, f_8\} \to \{f_2, f_{10}\}$?

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.

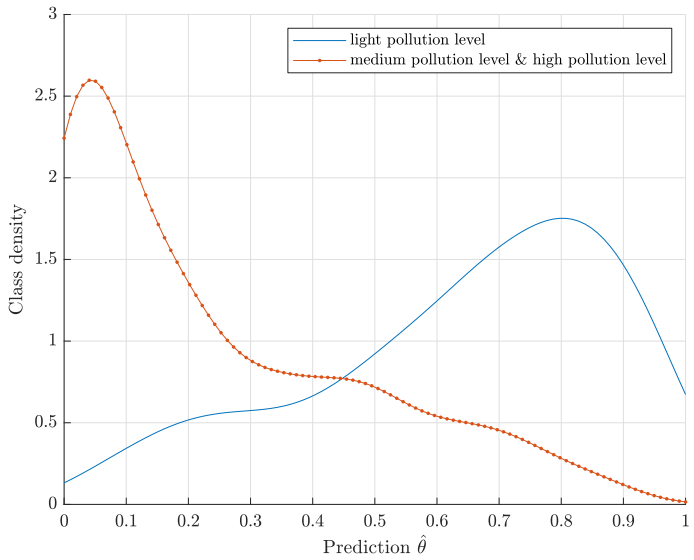


Figure 5: Class density (as function of the predictions of a logistic regression classifier $\hat{\theta}$) of the two-class problem of predicting *light pollution level* vs. *medium pollution level & high pollution level*.

**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 $\theta_0$ 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.

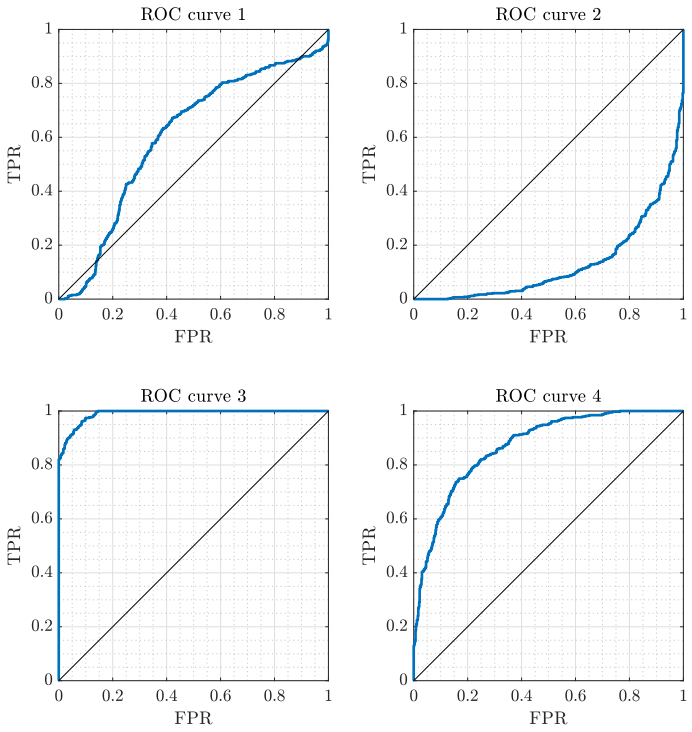


Figure 6: Proposed ROC curves for the two-class classifier described in Figure 5.

| | $i=1$ | $i=2$ | $i=3$ | $i=4$ |
|---|---|---|---|---|
| $x_i$ | 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.

| | 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 $y_r$.

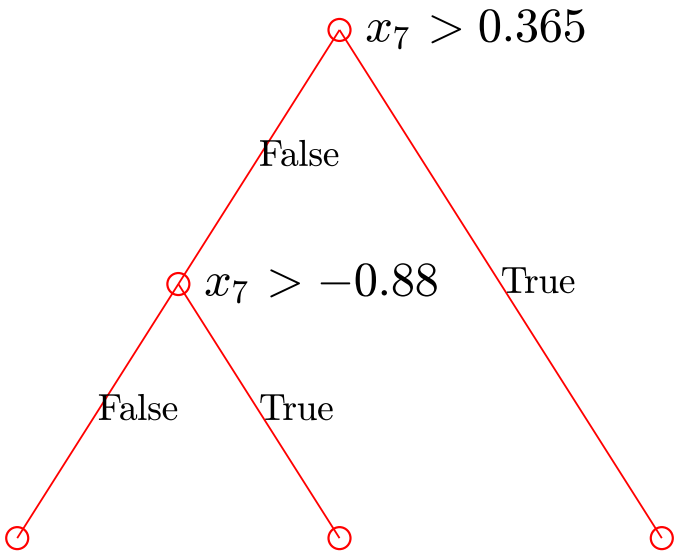


Figure 7: Structure of decision tree. The goal is to determine the splitting rules.

**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 7. 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.

**Question 16.** In this problem, we will again consider the 6 observations from the Beijing air pollution dataset shown in Table 7. Recall Figure 7 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 $\Delta$ 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.

**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.

| 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.

**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.

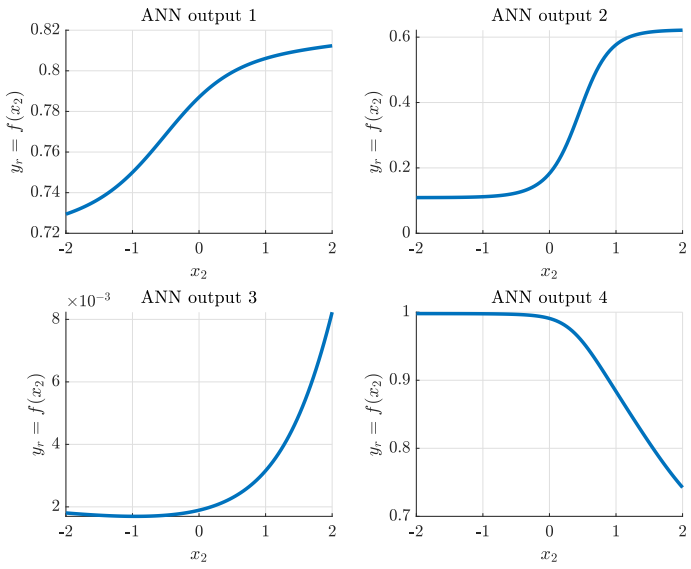


Figure 8: Suggested outputs of an ANN trained on the attribute $x_2$ from the Beijing air pollution dataset to predict $y_r$.

**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 8 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.

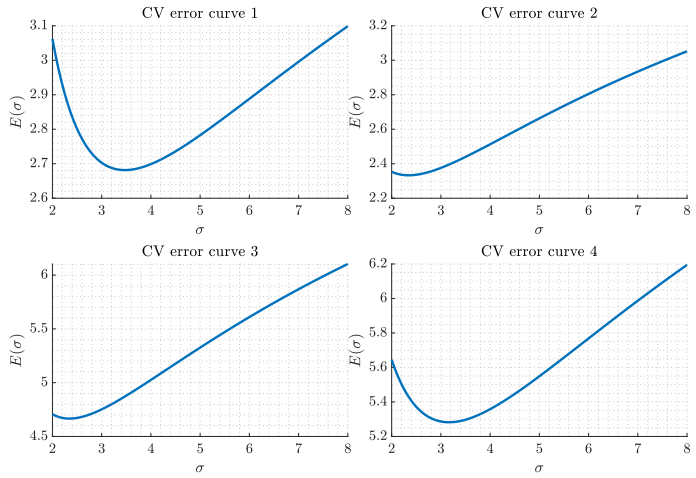


Figure 9: Estimated negative log-likelihood as obtained using hold-out cross validation on a small, $N = 3$ one-dimensional dataset as a function of kernel width $\sigma$.

**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 $\sigma$ (i.e., $\sigma$ is the standard deviation of the Gaussian kernels), and we wish to find $\sigma$ 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 9 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.

| Variable | $y^{\text{true}}$ | $t = 1$ |
|---|---|---|
| $y_1$ | 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 $y^{\text{true}}$ (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. $\begin{bmatrix} 0.173 & 0.103 & 0.173 & 0.103 & 0.103 & 0.173 & 0.173 \end{bmatrix}$

B. $\begin{bmatrix} 0.146 & 0.138 & 0.146 & 0.138 & 0.138 & 0.146 & 0.146 \end{bmatrix}$

C. $\begin{bmatrix} 0.125 & 0.167 & 0.125 & 0.167 & 0.167 & 0.125 & 0.125 \end{bmatrix}$

D. $\begin{bmatrix} 0.102 & 0.198 & 0.102 & 0.198 & 0.198 & 0.102 & 0.102 \end{bmatrix}$

E. Don't know.

**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.

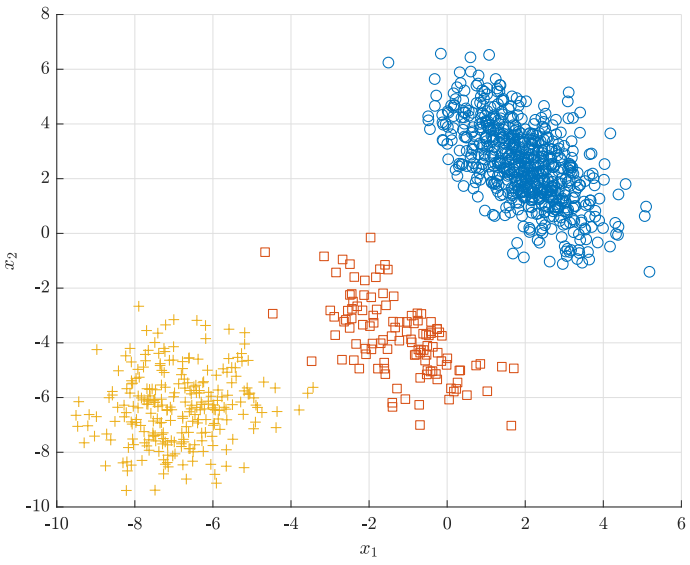


Figure 10: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

**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.

**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 10 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.

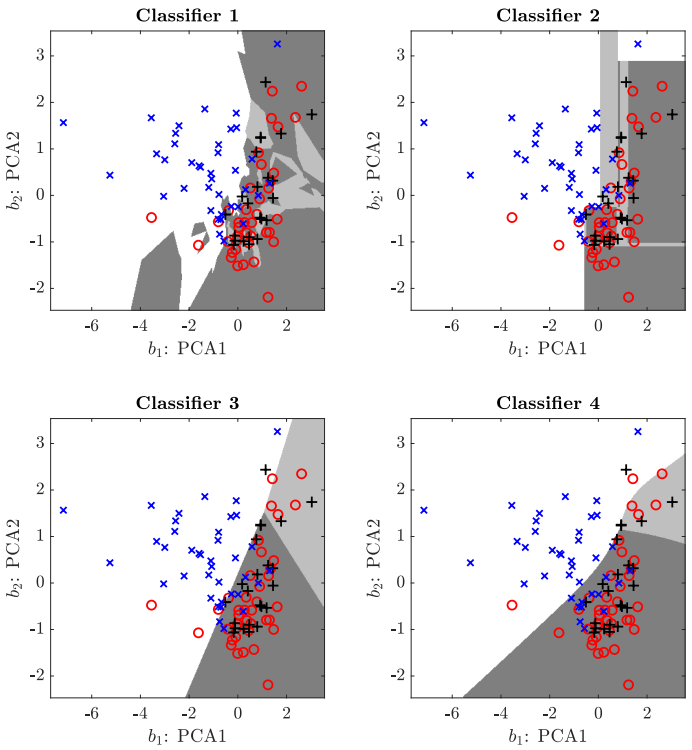


Figure 11: Decision boundaries for four different classifiers trained on the Beijing air pollution dataset when projected onto the first two principal components.

**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 11. 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**,
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.

**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.