Technical University of Denmark

Written examination: December 18th 2018, 9 AM - 1 PM.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

Please hand in your answers using the electronic file. Only use this page in the case where digital handin is unavailable. In case you have to hand in the answers using the form on this sheet, please follow these instructions:

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

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

Answers:

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

Name: ____________________________________

Student number: __________________

PLEASE HAND IN YOUR ANSWERS DIGITALLY.

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

No.	Attribute description	Abbrev.
x_1	intercolumnar distance	interdist
x_2	upper margin	upperm
x_3	lower margin	lowerm
x_4	exploitation	exploit
x_5	row number	row nr.
x_6	modular ratio	modular
x_7	interlinear spacing	interlin
x_8	weight	weight
x_9	peak number	peak nr.
x_{10}	modular ratio/ interlinear spacing	mr/is
y	Who copied the text?	Copyist

Table 1: Description of the features of the Avila Bible dataset used in this exam. The dataset has been extracted from images of the 'Avila Bible', an XII century giant Latin copy of the Bible. The prediction task consists in associating each pattern to one of three copyist (copyist refers to the monk who copied the text in the bible), indicated by the y-value. Note that only a subset of the dataset is used. The dataset used here consist of $N = 525$ observations and the attribute y is discrete taking values $y = 1, 2, 3$ corresponding to the three different copyists.

Question 1.

The main dataset used in this exam is the Avila Bible dataset[1] shown in Table 1.

In Figure 1 and Figure 2 are shown respectively percentile plots and boxplots of the Avila Bible dataset based on the attributes x_2, x_3, x_9, x_{10} found in Table 1. Which percentile plots match which boxplots?

A. Boxplot 1 is mr/is, Boxplot 2 is lowerm, Boxplot 3 is upperm and Boxplot 4 is peak nr.

B. Boxplot 1 is upperm, Boxplot 2 is lowerm, Boxplot 3 is peak nr. and Boxplot 4 is mr/is

C. Boxplot 1 is upperm, Boxplot 2 is peak nr., Boxplot 3 is mr/is and Boxplot 4 is lowerm

D. Boxplot 1 is mr/is, Boxplot 2 is lowerm, Boxplot 3 is peak nr. and Boxplot 4 is upperm

E. Don't know.

[1]Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Avila`

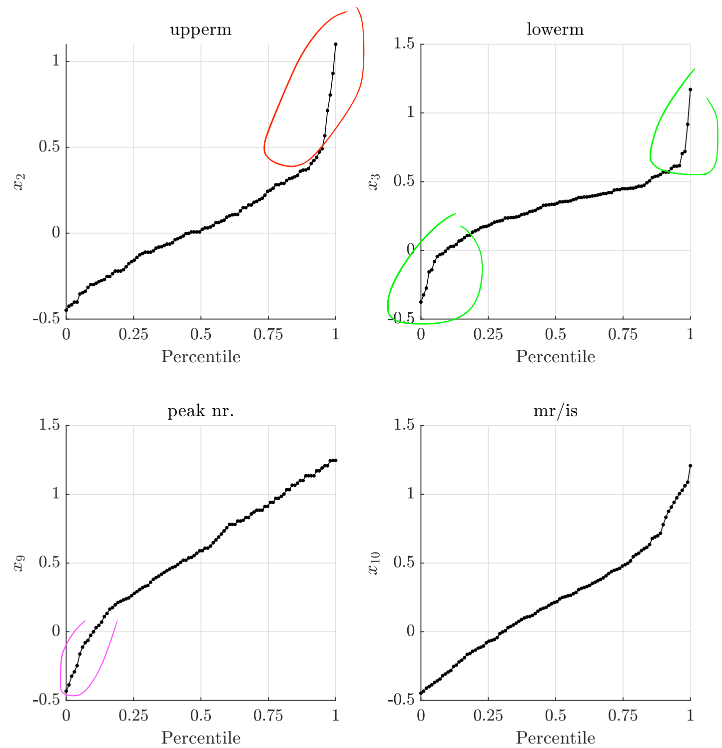

Figure 1: Plot of observations x_2, x_3, x_9, x_{10} of the Avila Bible dataset of Table 1 as percentile plots.

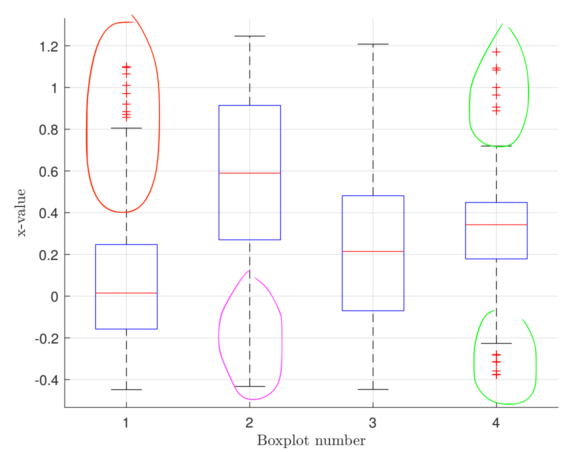

Figure 2: Boxplots corresponding to the variables plotted in Figure 1 but not necessarily in that order.

Question 2.

A Principal Component Analysis (PCA) is carried out on the Avila Bible dataset in Table 1 based on the attributes $x_1,\ x_3,\ x_5,\ x_6,\ x_7$.

The data is standardized by (i) substracting the mean and (ii) dividing each column by its standard deviation to obtain the standardized matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.04 & -0.12 & -0.14 & 0.35 & 0.92 \\ 0.06 & 0.13 & 0.05 & -0.92 & 0.37 \\ -0.03 & -0.98 & 0.08 & -0.16 & -0.05 \\ -0.99 & 0.03 & 0.06 & -0.02 & 0.07 \\ -0.07 & -0.05 & -0.98 & -0.11 & -0.11 \end{bmatrix} \quad (1)$$

$$\boldsymbol{S} = \begin{bmatrix} 14.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 8.19 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 7.83 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 6.91 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 6.01 \end{bmatrix}$$

Which one of the following statements is true?

A. The variance explained by the first principal component is greater than 0.45

B. The variance explained by the first four principal components is less than 0.85

C. The variance explained by the last four principal components is greater than 0.56

D. The variance explained by the first three principal components is less than 0.75

E. Don't know.

Question 3.

Consider again the PCA analysis fo the Avila Bible dataset. In Figure 3 the features x_5 and x_7 from Table 1 are plotted as black dots. We have indicated two special observations as colored markers (Point A and Point B).

We can imagine that the dataset, along with the two special observations, is projected onto the first two principal component directions given in $\boldsymbol{V}$ as computed earlier (see Equation (1)). Which one of the four plots

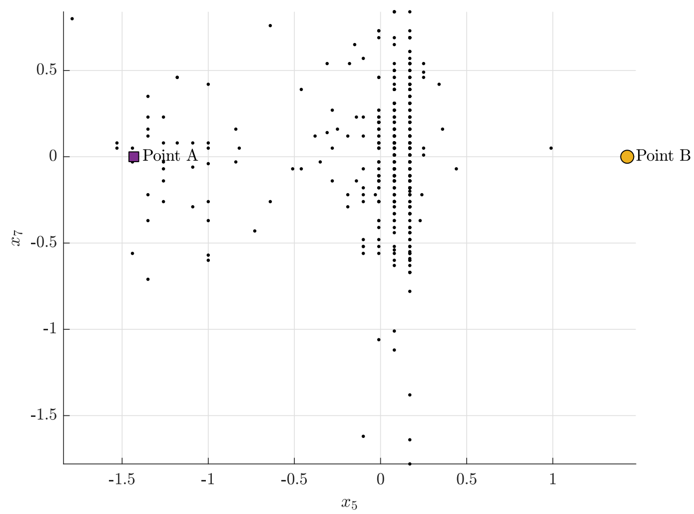

Figure 3: Black dots show attributes x_5 and x_7 of the Avila Bible dataset from Table 1. The two points corresponding to the colored markers indicate two specific observations A, B.

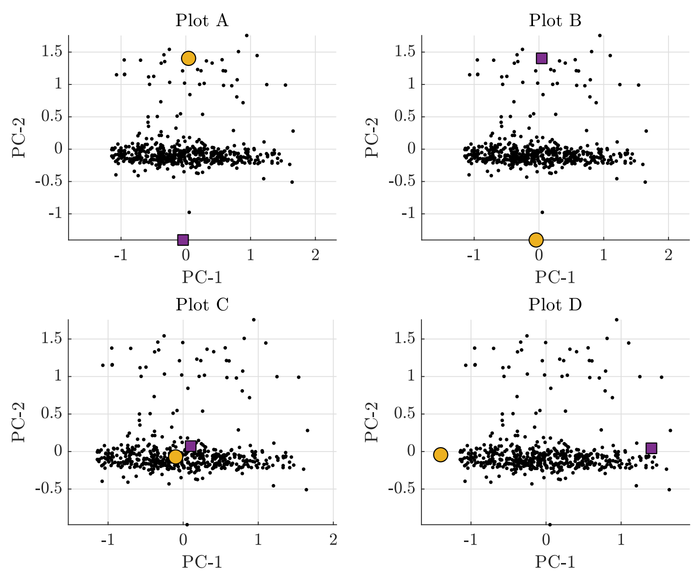

Figure 4: Candidate plots of the observations and path shown in Figure 3 projected onto the first two principal components considered in Equation (1). The colored markers still refer to points A and B, now in the coordinate system corresponding to the PCA projection.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	2.91	0.63	1.88	1.02	1.82	1.92	1.58	1.08	1.43
o_2	2.91	0.0	3.23	3.9	2.88	3.27	3.48	4.02	3.08	3.47
o_3	0.63	3.23	0.0	2.03	1.06	2.15	2.11	1.15	1.09	1.65
o_4	1.88	3.9	2.03	0.0	2.52	1.04	2.25	2.42	2.18	2.17
o_5	1.02	2.88	1.06	2.52	0.0	2.44	2.38	1.53	1.71	1.94
o_6	1.82	3.27	2.15	1.04	2.44	0.0	1.93	2.72	1.98	1.8
o_7	1.92	3.48	2.11	2.25	2.38	1.93	0.0	2.53	2.09	1.66
o_8	1.58	4.02	1.15	2.42	1.53	2.72	2.53	0.0	1.68	2.06
o_9	1.08	3.08	1.09	2.18	1.71	1.98	2.09	1.68	0.0	1.48
o_{10}	1.43	3.47	1.65	2.17	1.94	1.8	1.66	2.06	1.48	0.0

Table 2: The pairwise Euclidian distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 10 observations from the Avila Bible dataset (recall $M = 10$). Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1 (the data has been standardized). The colors indicate classes such that the black observations $\{o_1, o_2, o_3\}$ belongs to class C_1 (corresponding to copyist one), the red observations $\{o_4, o_5, o_6, o_7, o_8\}$ belongs to class C_2 (corresponding to copyist two), and the blue observations $\{o_9, o_{10}\}$ belongs to class C_3 (corresponding to copyist three).

in Figure 4 shows the correct PCA projection?

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Question 4. To examine if observation o_4 may be an outlier, we will calculate the average relative density based on euclidean distance and the observations given in Table 2 only. We recall that the KNN density and average relative density (ard) for the observation $\boldsymbol{x}_i$ are given by:

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

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

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors. What is the average relative density for observation o_4 for $K = 2$ nearest neighbors?

A. 1.0

B. 0.71

C. 0.68

D. 0.36

E. Don't know.

Question 5. Suppose a GMM model is applied to the Avila Bible dataset in the processed version shown in Table 2. The GMM is constructed as having $K = 3$ components, and each component k of the GMM is fitted by letting it's mean vectors $\boldsymbol{\mu}_k$ be equal to the location of the observations:

$$o_7, \ o_8, \ o_9$$

(i.e. each observation corresponds to exactly one mean vector) and setting the covariance matrix equal to $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix:

$$\mathcal{N}(\boldsymbol{o}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_k|}} e^{\frac{-d(\boldsymbol{o}_i, \boldsymbol{\mu}_k)^2}{2\sigma^2}}$$

where $|\cdot|$ is the determinant. The components of the GMM are weighted evenly.

If $\sigma = 0.5$, and denoting the density of the GMM as $p(\boldsymbol{x})$, what is the density as evaluated at observation o_3?

A. $p(o_3) = 0.048402$

B. $p(o_3) = 0.076$

C. $p(o_3) = 0.005718$

D. $p(o_3) = 0.114084$

E. Don't know.

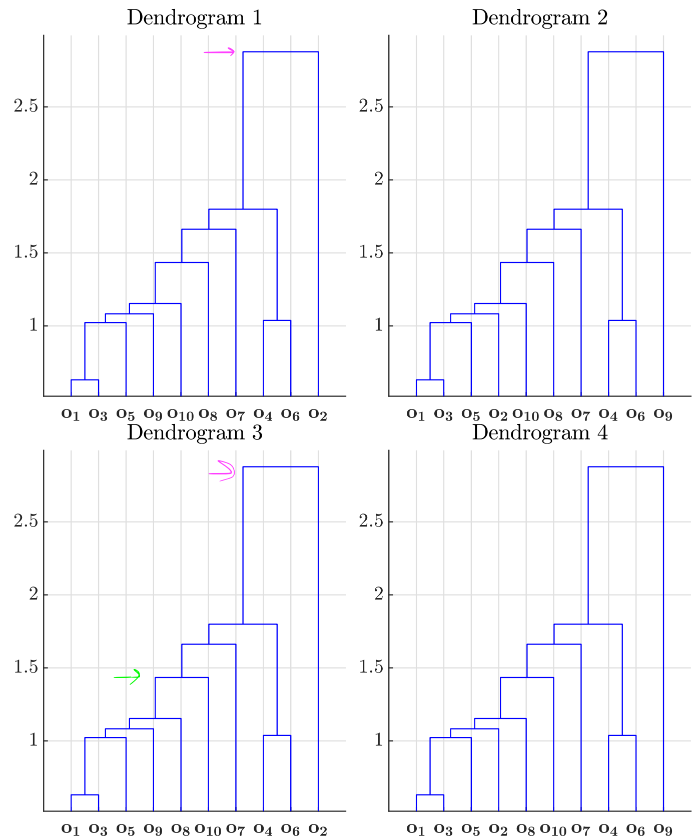

Figure 5: Proposed hierarchical clustering of the 10 observations in Table 2.

Question 6. A hierarchical clustering is applied to the 10 observations in Table 2 using *minimum* linkage. Which of the dendrograms shown in Figure 5 corresponds to the clustering?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

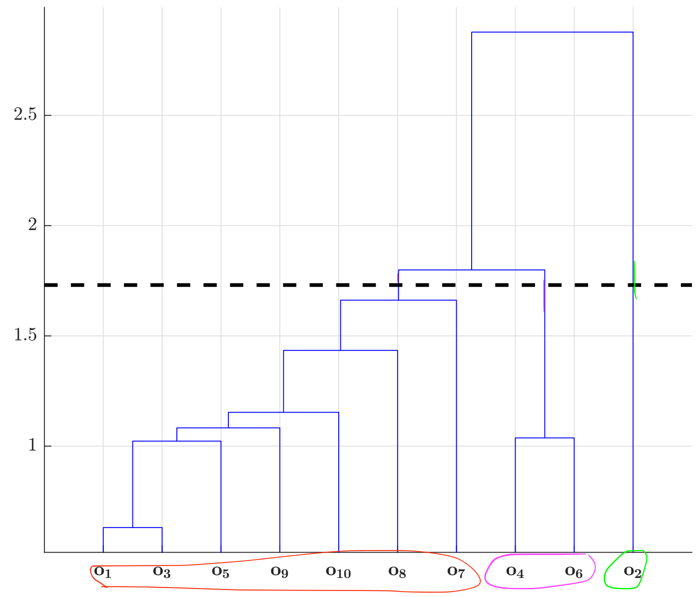

Figure 6: Dendrogram 1 from Figure 5 with a cutoff indicated by the dotted line, thereby generating 3 clusters.

Question 7.
Consider dendrogram 1 from Figure 5. Suppose we apply a cutoff (indicated by the black line) thereby generating three clusters. We wish to compare the quality of this clustering, Q, to the ground-truth clustering, Z, indicated by the colors in Table 2. Recall the *normalized mutual information* of the two clusterings Z and Q is defined as

$$\text{NMI}[Z, Q] = \frac{\text{MI}[Z, Q]}{\sqrt{H[Z]}\sqrt{H[Q]}}$$

where MI is the *mutual information* and H is the entropy. Assuming we always use an entropy based on the natural logarithm,

$$H = -\sum_{i=1}^{n} p_i \log p_i, \quad \log(e) = 1,$$

what is the normalized mutual information of the two clusterings?

A. $\text{NMI}[Z, Q] \approx 0.313$

B. $\text{NMI}[Z, Q] \approx 0.302$

C. $\text{NMI}[Z, Q] \approx 0.32$

D. $\text{NMI}[Z, Q] \approx 0.274$

E. Don't know.

x_9-interval	$y = 1$	$y = 2$	$y = 3$	
$x_9 \leq 0.13$	108	112	56	← b1
$0.13 < x_9$	58	75	116	← b2

Table 3: Proposed split of the Avila Bible dataset based on the attribute x_9. We consider a 2-way split where for each interval we count how many observations belonging to that interval has the given class label.

Question 8. Consider the distances in Table 2 based on 10 observations from the Avila Bible dataset. The class labels C_1, C_2, C_3 (see table caption for details) will be predicted using a k-nearest neighbour classifier based on the distances given in Table 2. Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 1-nearest neighbour classifier (i.e. $k = 1$). What is the error rate computed for all $N = 10$ observations?

A. error rate $= \frac{4}{10}$

B. error rate $= \frac{9}{10}$

C. error rate $= \frac{2}{10}$

D. error rate $= \frac{6}{10}$

E. Don't know.

Question 9.

Suppose we wish to build a classification tree based on Hunt's algorithm where the goal is to predict Copyist which can belong to three classes, $y = 1$, $y = 2$, $y = 3$. The first split we consider is a two-way split based on the value of x_9 into the intervals indicated in Table 3. For each interval, we count how many observations belong to each of the three classes and the result is indicated in Table 3. Suppose we use the *classification error* impurity measure, what is then the purity gain Δ?

A. $\Delta \approx 0.485$

B. $\Delta \approx 0.078$

C. $\Delta \approx 0.566$

D. $\Delta \approx 1.128$

E. Don't know.

Question 10. Consider the split in Table 3. Suppose we build a classification tree with *only* this split and evaluate it on the same data it was trained on. What is the accuracy?

A. Accuracy is: 0.64

B. Accuracy is: 0.29

C. Accuracy is: 0.35

D. Accuracy is: 0.43

E. Don't know.

Question 11. Suppose s_1 and s_2 are two text documents containing the text:

$$s_1 = \left\{\begin{matrix}\text{the bag of words representation} \\ \text{should not give you a hard time}\end{matrix}\right\}$$

$$s_2 = \left\{\begin{matrix}\text{remember the representation should} \\ \text{be a vector}\end{matrix}\right\}$$

The documents are encoded using a bag-of-words encoding assuming a total vocabulary size of $M = 10000$. No stopwords lists or stemming is applied to the dataset. What is the cosine similarity between documents s_1 and s_2?

A. cosine similarity of s_1 and s_2 is 0.047619

B. cosine similarity of s_1 and s_2 is 0.000044

C. cosine similarity of s_1 and s_2 is 0.000400

D. cosine similarity of s_1 and s_2 is 0.436436

E. Don't know.

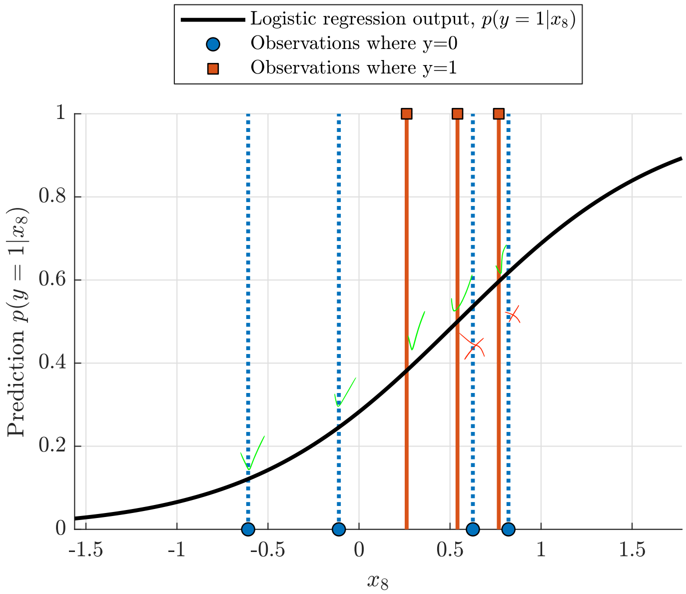

Figure 7: Output of a logistic regression classifier trained on 7 observations from the dataset.

Question 12. Consider again the Avila Bible dataset. We are particularly interested in predicting whether a bible copy was written by copyist 1, and we therefore wish to train a logistic regression classifier to distinguish between copyist one vs. copyist two and three.

To simplify the setup further, we select just 7 observations and train a logistic regression classifier using only the feature x_8 as input (as usual, we apply a simple feature transformation to the inputs to add a constant feature in the first coordinate to handle the intercept term). To be consistent with the lecture notes, we label the output as $y = 0$ (corresponding to copyist one) and $y = 1$ (corresponding to copyist two and three).

In Figure 7 is shown the predicted output probability an observation belongs to the positive class, $p(y = 1|x_8)$. What are the weights?

A. $\begin{bmatrix} -0.93 \\ 1.72 \end{bmatrix}$

B. $\begin{bmatrix} -2.82 \\ 0.0 \end{bmatrix}$

C. $\begin{bmatrix} 1.36 \\ 0.4 \end{bmatrix}$

D. $\begin{bmatrix} -0.65 \\ 0.0 \end{bmatrix}$

E. Don't know.

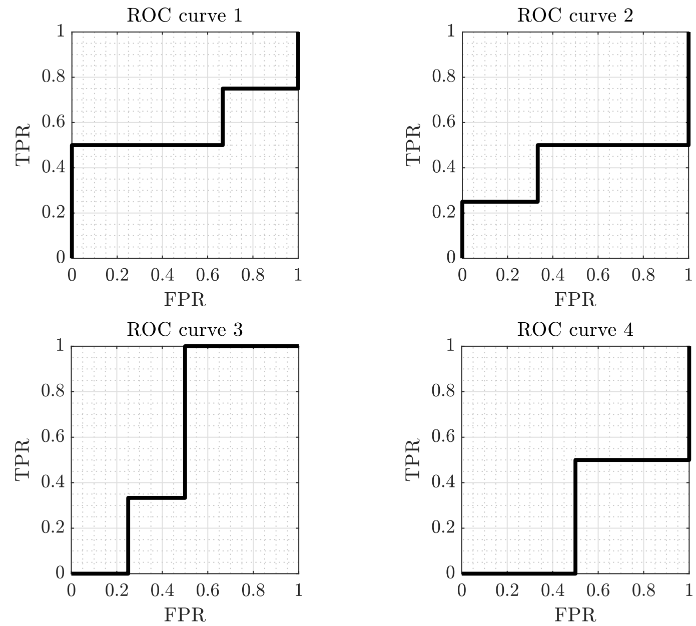

Figure 8: Proposed ROC curves for the logistic regression classifier in Figure 7.

Question 13.

To evaluate the classifier Figure 7, we will use the *area under curve* (AUC) of the *reciever operator characteristic* (ROC) curve as computed on the 7 observations in Figure 7. In Figure 8 is given four proposed ROC curves, which one of the curves corresponds to the classifier?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

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

Table 4: Binarized version of the Avila Bible dataset. Each of the features f_i are obtained by taking a feature x_i and letting $f_i = 1$ correspond to a value x_i greater than the median (otherwise $f_i = 0$).The colors indicate classes such that the black observations $\{o_1,\ o_2,\ o_3\}$ belongs to class C_1 (corresponding to copyist one), the red observations $\{o_4,\ o_5,\ o_6,\ o_7,\ o_8\}$ belongs to class C_2 (corresponding to copyist two), and the blue observations $\{o_9,\ o_{10}\}$ belongs to class C_3 (corresponding to copyist three).

Question 14. We again consider the Avila Bible dataset from Table 1 and the $N = 10$ observations we already encountered in Table 2. The data is processed to produce 10 new, binary features such that $f_i = 1$ corresponds to a value x_i greater than the median[2], and we thereby arrive at the $N \times M = 10 \times 10$ binary matrix in Table 4. Suppose we train a naïve-Bayes classifier to predict the class label y from only the features $f_1,\ f_2,\ f_6$. If for an observations we observe

$$f_1 = 1,\ f_2 = 1,\ f_6 = 0$$

what is then the probability that $y = 1$ according to the Naïve-Bayes classifier?

A. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{50}{77}$

B. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{25}{43}$

C. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{5}{11}$

D. $p_{\text{NB}}(y = 1 | f_1 = 1,\ f_2 = 1,\ f_6 = 0) = \frac{10}{19}$

E. Don't know.

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

Question 15.

Consider the binarized version of the Avila Bible dataset shown in Table 4.

The matrix can be considered as representing $N = 10$ transactions $o_1, o_2, \ldots, o_{10}$ and $M = 10$ items $f_1, f_2, \ldots, f_{10}$. Which of the following options represents all (non-empty) itemsets with support greater than 0.55 (and only itemsets with support greater than 0.55)?

A. $\{f_1\}$, $\{f_2\}$, $\{f_6\}$, $\{f_7\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_6, f_{10}\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_6\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_4\}$, $\{f_6\}$, $\{f_7\}$, $\{f_8\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_2\}$, $\{f_2, f_3\}$, $\{f_2, f_4\}$, $\{f_3, f_4\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_2, f_7\}$, $\{f_3, f_7\}$, $\{f_6, f_7\}$, $\{f_2, f_8\}$, $\{f_3, f_8\}$, $\{f_7, f_8\}$, $\{f_2, f_9\}$, $\{f_3, f_9\}$, $\{f_6, f_9\}$, $\{f_7, f_9\}$, $\{f_8, f_9\}$, $\{f_1, f_{10}\}$, $\{f_2, f_{10}\}$, $\{f_6, f_{10}\}$, $\{f_2, f_3, f_4\}$, $\{f_1, f_2, f_6\}$, $\{f_2, f_3, f_7\}$, $\{f_2, f_3, f_8\}$, $\{f_2, f_3, f_9\}$, $\{f_6, f_7, f_9\}$, $\{f_2, f_8, f_9\}$, $\{f_3, f_8, f_9\}$, $\{f_7, f_8, f_9\}$, $\{f_1, f_2, f_{10}\}$, $\{f_1, f_6, f_{10}\}$, $\{f_2, f_6, f_{10}\}$, $\{f_2, f_3, f_8, f_9\}$, $\{f_1, f_2, f_6, f_{10}\}$

D. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_6\}$, $\{f_7\}$, $\{f_8\}$, $\{f_9\}$, $\{f_{10}\}$, $\{f_1, f_2\}$, $\{f_2, f_3\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_6, f_7\}$, $\{f_7, f_9\}$, $\{f_8, f_9\}$, $\{f_2, f_{10}\}$, $\{f_6, f_{10}\}$, $\{f_1, f_2, f_6\}$, $\{f_2, f_6, f_{10}\}$

E. Don't know.

Question 16. We again consider the binary matrix from Table 4 as a market basket problem consisting of $N = 10$ transactions $o_1, \ldots, o_{10}$ and $M = 10$ items $f_1, \ldots, f_{10}$.

What is the *confidence* of the rule $\{f_1, f_3, f_8, f_9\} \rightarrow \{f_2, f_6, f_7\}$

A. Confidence is $\frac{1}{10}$

B. Confidence is 1

C. Confidence is $\frac{1}{2}$

D. Confidence is $\frac{3}{20}$

E. Don't know.

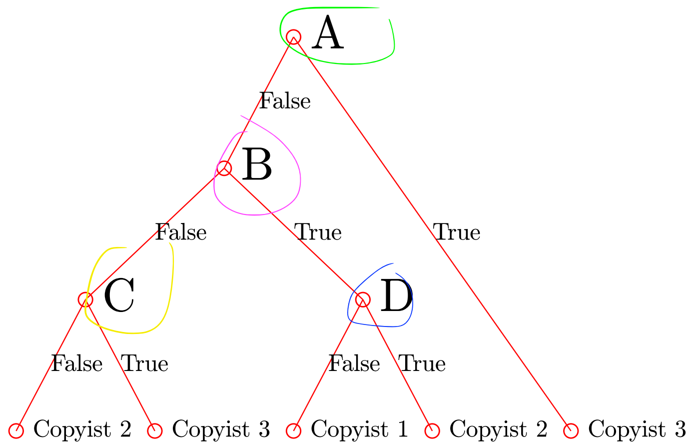

Figure 9: Example classification tree.

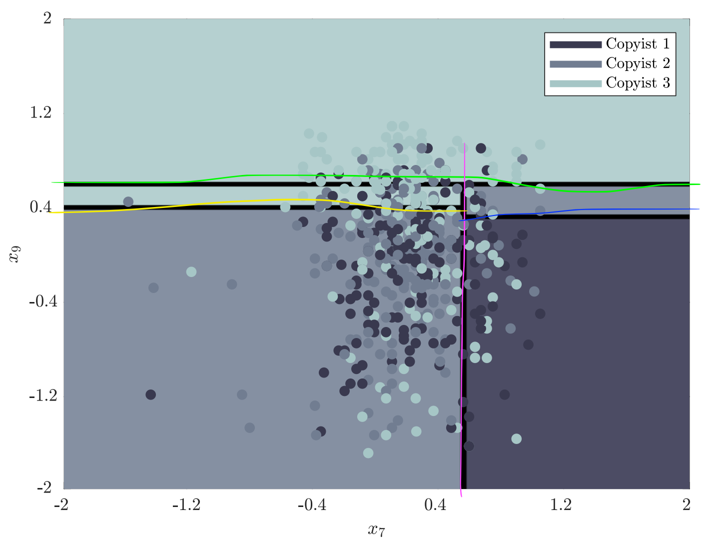

Figure 10: classification boundary.

Question 17.

Consider again the Avila Bible dataset. Suppose we train a decision tree to classify which of the 3 classes, Copyist 1, Copyist 2, Copyist 3, an observation belongs to. Since the attributes of the dataset are continuous, we will consider binary splits of the form $x_i \geq z$ for different values of i and z, and for simplicity we limit ourselves to the attributes x_7 and x_9. Suppose the trained decision tree has the form shown in Figure 9, and that according to the tree the predicted label assignment for the $N = 525$ observations are as given in Figure 10, what is then the correct rule assignment to the nodes in the decision tree?

A. $\cancel{\boldsymbol{A}\text{: } x_7 \geq 0.5}$, $\boldsymbol{B}$: $x_9 \geq 0.54$, $\boldsymbol{C}$: $x_9 \geq 0.35$, $\boldsymbol{D}$: $x_9 \geq 0.26$

B. $\cancel{\boldsymbol{A}\text{: } x_7 \geq 0.5}$, $\boldsymbol{B}$: $x_9 \geq 0.26$, $\boldsymbol{C}$: $x_9 \geq 0.54$, $\boldsymbol{D}$: $x_9 \geq 0.35$

C. $\boldsymbol{A}$: $x_9 \geq 0.54$, $\boldsymbol{B}$: $x_7 \geq 0.5$, $\boldsymbol{C}$: $x_9 \geq 0.35$, $\boldsymbol{D}$: $x_9 \geq 0.26$

D. $\boldsymbol{A}$: $x_9 \geq 0.\cancel{26}$, $\boldsymbol{B}$: $x_7 \geq 0.5$, $\boldsymbol{C}$: $x_9 \geq 0.35$, $\boldsymbol{D}$: $x_9 \geq 0.54$

E. Don't know.

Question 18. We will again consider the binarized version of the Avila Bible dataset already encountered in Table 4, however we will now only consider the first $M = 6$ features f_1, f_2, f_3, f_4, f_5, f_6.

We wish to apply the Apriori algorithm (the specific variant encountered in chapter 19 of the lecture notes) to find all itemsets with support greater than $\varepsilon = 0.15$. Suppose at iteration $k = 3$ we know that:

$$L_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Recall the key step in the Apriori algorithm is to construct L_3 by first considering a large number of candidate itemsets C_3', and then rule out some of them using the downwards-closure principle thereby saving many (potentially costly) evaluations of support. Suppose L_2 is given as above, which of the following itemsets does the Apriori algorithm *not* have to evaluate the support of?

A. $\{f_2, f_3, f_4\}$

B. $\{f_1, f_2, f_6\}$

C. $\{f_2, f_3, f_6\}$

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

E. Don't know.

Question 19.

Consider again the Avila Bible dataset in Table 1. We would like to predict the copyist using a linear regression, and since we would like the model to be as interpretable as possible we will use variable selection to obtain a parsimonious model. We limit ourselves to the 5 features x_1, x_5, x_6, x_8, x_9 and in Table 5 we have pre-computed the estimated training and test error for different variable combinations of the dataset. Which of the following statements is correct?

A. Backward selection will select attributes x_1

B. Backward selection will select attributes x_1, x_5, x_6, x_8

C. Forward selection will select attributes x_1, x_8

D. Forward selection will select attributes x_1, x_5, x_6, x_8

E. Don't know.

Question 20.

Consider the Avila Bible dataset from Table 1. We wish to predict the copyist based on the attributes *upperm* and *mr/is*.

Therefore, suppose the attributes have been binarized such that $\tilde{x}_2 = 0$ corresponds $x_2 \leq -0.056$ (and otherwise $\tilde{x}_2 = 1$) and $\tilde{x}_{10} = 0$ corresponds $x_{10} \leq -0.002$ (and otherwise $\tilde{x}_{10} = 1$). Suppose the probability for each of the configurations of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on the copyist y are as given in Table 6. and the prior probability of the copyists is

$$p(y = 1) = 0.316, \; p(y = 2) = 0.356, \; p(y = 3) = 0.328.$$

Using this, what is then the probability an observation was authored by copyist 1 given that $\tilde{x}_2 = 1$ and $\tilde{x}_{10} = 0$?

A. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.25$

B. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.313$

C. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.262$

D. $p(y = 1|\tilde{x}_2 = 1, \tilde{x}_{10} = 0) = 0.298$

E. Don't know.

Feature(s)	Training RMSE	Test RMSE
none	3.429	4.163
x_1	3.043	3.252
x_5	3.303	4.52
x_6	3.424	4.274
x_8	3.399	4.429
x_9	2.866	5.016
x_1, x_5	3.001	3.44
x_1, x_6	3.031	3.423
x_5, x_6	3.297	4.641
x_1, x_8	3.017	3.42
x_5, x_8	3.299	4.485
x_6, x_8	3.396	4.519
x_1, x_9	2.644	4.267
x_5, x_9	2.645	5.495
x_6, x_9	2.787	5.956
x_8, x_9	2.71	5.536
x_1, x_5, x_6	2.988	3.607
x_1, x_5, x_8	3.0	3.453
x_1, x_6, x_8	3.007	3.574
x_5, x_6, x_8	3.292	4.61
x_1, x_5, x_9	2.523	4.704
x_1, x_6, x_9	2.562	5.184
x_5, x_6, x_9	2.544	6.552
x_1, x_8, x_9	2.517	4.686
x_5, x_8, x_9	2.628	5.532
x_6, x_8, x_9	2.629	6.569
x_1, x_5, x_6, x_8	2.988	3.614
x_1, x_5, x_6, x_9	2.425	5.725
x_1, x_5, x_8, x_9	2.491	4.734
x_1, x_6, x_8, x_9	2.433	5.687
x_5, x_6, x_8, x_9	2.53	6.597
x_1, x_5, x_6, x_8, x_9	2.398	5.766

Table 5: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict y in the avila dataset using different combinations of the features x_1, x_5, x_6, x_8, x_9.

$p(\tilde{x}_2, \tilde{x}_{10}\|y)$	$y = 1$	$y = 2$	$y = 3$
$\tilde{x}_2 = 0, \tilde{x}_{10} = 0$	0.19	0.3	0.19
$\tilde{x}_2 = 0, \tilde{x}_{10} = 1$	0.22	0.3	0.26
$\tilde{x}_2 = 1, \tilde{x}_{10} = 0$	0.25	0.2	0.35
$\tilde{x}_2 = 1, \tilde{x}_{10} = 1$	0.34	0.2	0.2

Table 6: Probability of observing particular values of $\tilde{x}_2$ and $\tilde{x}_{10}$ conditional on y.

Variable	$t=1$	$t=2$	$t=3$	$t=4$
y_1	1	2	2	2
y_2	1	2	2	1
y_3	2	2	2	1
y_4	1	1	1	2
y_5	1	1	1	1
y_6	2	2	2	1
y_7	1	2	2	1
y_8	2	1	1	2
y_9	2	2	2	2
y_{10}	1	1	2	2
y_{11}	2	2	1	2
y_{12}	2	1	1	2
y_1^{test}	2	1	1	2
y_2^{test}	2	2	1	2
ϵ_t	0.583	0.657	0.591	0.398
α_t	-0.168	-0.325	-0.185	0.207

Table 7: Tabulation of each of the predicted outputs of the AdaBoost classifiers, as well as the intermediate values α_t and ϵ_t, when the AdaBoost algorithm when evaluated for $T = 4$ steps. Note the table includes the prediction of the two test points in Figure 11.

Question 21.

Consider again the Avila Bible dataset of Table 1. Suppose we limit ourselves to $N = 12$ observations from the original dataset and furthermore suppose we limit ourselves to class $y = 1$ or $y = 2$ and only consider the features x_6 and x_8. We wish to apply a KNN classification model ($K = 2$) to this dataset and apply AdaBoost to improve the performance. During the first $T = 4$ rounds of boosting, we obtain the decision boundaries shown in Figure 11. The figure also contains two test observations (marked by a cross and a square).

The prediction of the intermediate AdaBoost classifiers, as well as the values of α_t and ϵ_t, are given in Table 7. Given this information, how will the AdaBoost classifier, as obtained by combining the $T = 4$ weak classifiers, classify the two test observations?

A. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}1 & 1\end{bmatrix}$

B. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}2 & 1\end{bmatrix}$

C. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}1 & 2\end{bmatrix}$

D. $\begin{bmatrix}\tilde{y}_1^{\text{test}} & \tilde{y}_2^{\text{test}}\end{bmatrix} = \begin{bmatrix}2 & 2\end{bmatrix}$

E. Don't know.

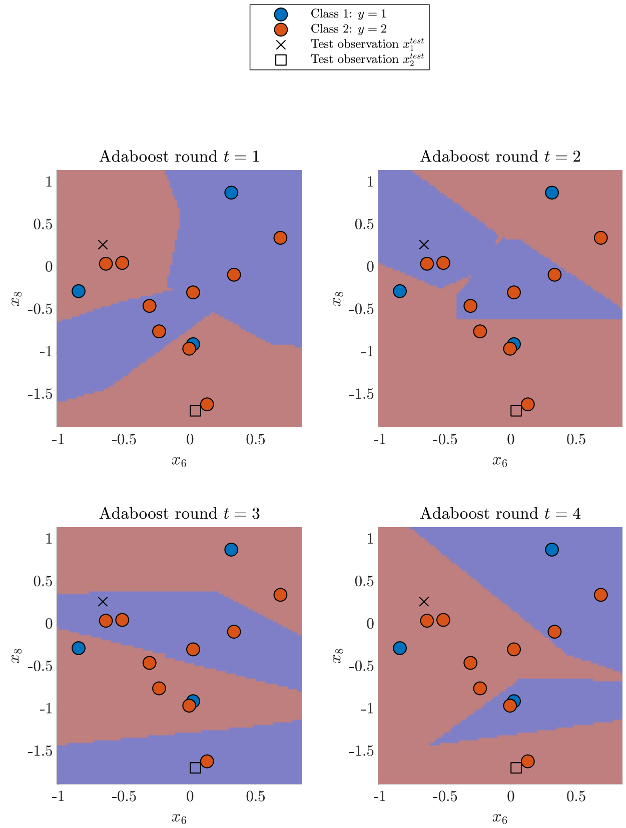

Figure 11: Decision boundaries for a KNN classifier for the first $T = 4$ rounds of boosting. Notice that in addition to the training data, the plot also indicate the location of two test points.

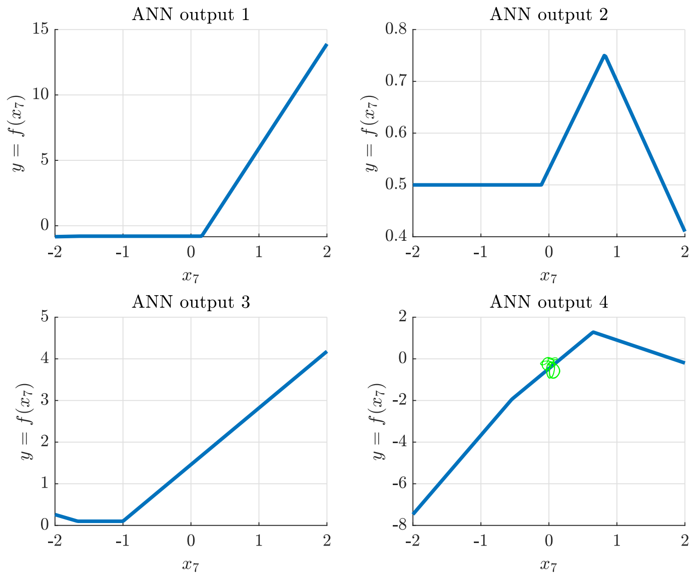

Figure 12: Suggested activation curves for an ANN applied to the feature x_7 from Avila Bible dataset.

Question 22.

We will consider an artificial neural network (ANN) applied to the Avila Bible dataset described in Table 1 and trained to predict based on just the feature x_7; that is, the neural network is a function that maps from a single real number to a single real number: $f(x_7) = y$

Suppose the neural network takes the form:

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

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer and the weights are given as:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -1.8 \\ -1.1 \end{bmatrix}$$
$$\boldsymbol{w}_2^{(1)} = \begin{bmatrix} -0.6 \\ 3.8 \end{bmatrix}$$
$$\boldsymbol{w}^{(2)} = \begin{bmatrix} -0.1 \\ 2.1 \end{bmatrix},$$
$$w_0^{(2)} = -0.8.$$

Which of the curves in Figure 12 will then correspond to the function f?

A. ANN output 4

B. ANN output 1

C. ANN output 3

D. ANN output 2

E. Don't know.

Question 23. Suppose a neural network is trained to translate documents. As part of training the network, we wish to select between four different ways to encode the documents (i.e., $S = 4$ models) and estimate the generalization error of the optimal choice. In the outer loop we opt for $K_1 = 3$-fold cross-validation, and in the inner $K_2 = 4$-fold cross-validation. The time taken to *train* a single model is 20 minutes, and this can be assumed constant for each fold. If the time taken to test a model is negligible, what is the total time required for the 2-level cross-validation procedure?

A. 1020 minutes

B. 2040 minutes

C. 300 minutes

D. 960 minutes

E. Don't know.

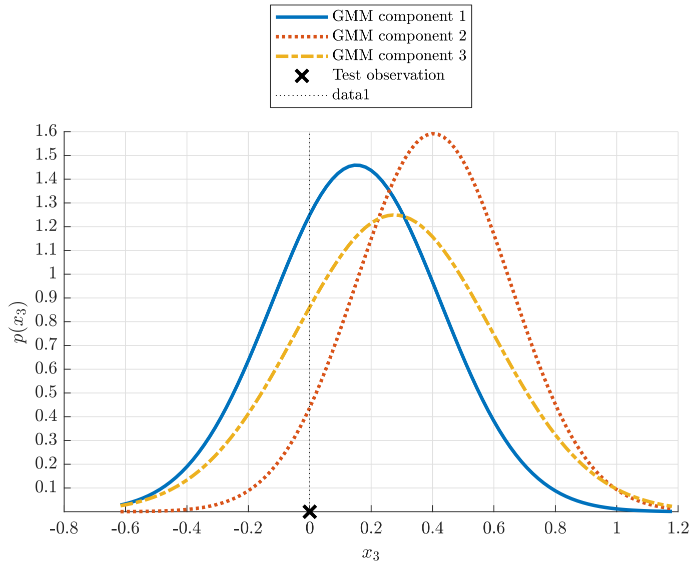

Figure 13: Mixture components in a GMM mixture model with $K = 3$.

Question 24.
We wish to apply the EM algorithm to fit a 1D GMM mixture model to the single feature x_3 from the Avila Bible dataset. At the first step of the EM algorithm, the $K = 3$ mixture components has densities as indicated by each of the curves in Figure 13 (i.e. each curve is a normalized, Gaussian density $\mathcal{N}(x; \mu_k, \sigma_k)$). In the figure, we have indicated the x_3-value of a single observation i from the dataset as a black cross.

Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step. We assume the weights of the components are

$$\boldsymbol{\pi} = \begin{bmatrix} 0.15 & 0.53 & 0.32 \end{bmatrix}$$

and the mean/variances of the components are those indicated in the figure.

According to the EM algorithm, what is the (approximate) probability the black cross is assigned to mixture component 3 (γ_{ik})?

A. 0.4

B. 0.86

C. 0.28

D. 0.58

E. Don't know.

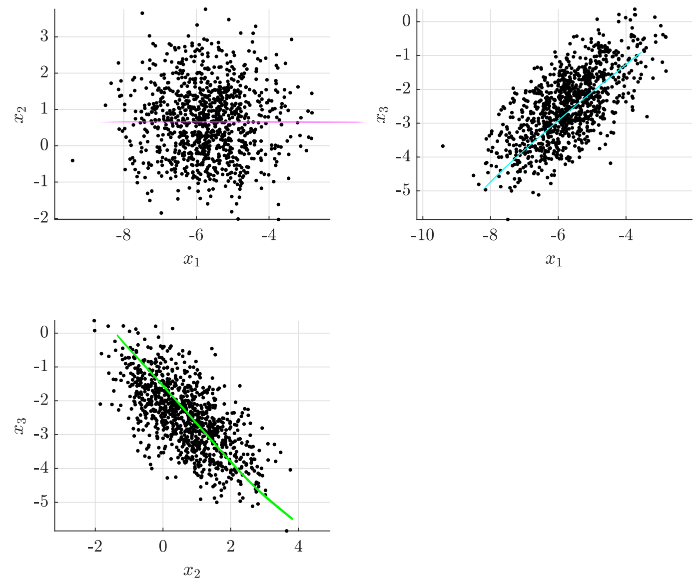

Figure 14: Scatter plot of each pairs of attributes of a vectors $\boldsymbol{x}$ drawn from a multivariate normal distribution of 3 dimensions.

Question 25. Consider a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ and mean μ and suppose we generate 1000 random samples from it:

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Plots of each pair of coordinates of the draws $\boldsymbol{x}$ is shown in Figure 14. What is the most plausible covariance matrix?

A. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.65 & -0.65 \\ 0.65 & 1.0 & 0.0 \\ -0.65 & 0.0 & 1.0 \end{bmatrix}$

B. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & 0.65 \\ 0.0 & 1.0 & -0.65 \\ 0.65 & -0.65 & 1.0 \end{bmatrix}$

C. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & -0.65 & 0.0 \\ -0.65 & 1.0 & 0.65 \\ 0.0 & 0.65 & 1.0 \end{bmatrix}$

D. $\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.0 & -0.65 \\ 0.0 & 1.0 & 0.65 \\ -0.65 & 0.65 & 1.0 \end{bmatrix}$

E. Don't know.

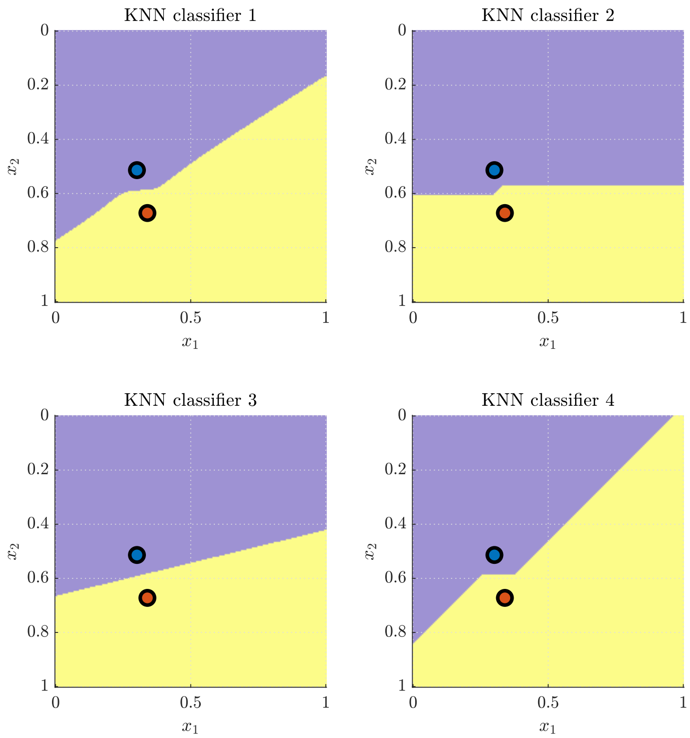

Figure 15: Decision boundaries for a KNN classifier, $K = 1$, computed for the two observations marked by circles (the colors indicate class labels), but using four different p-distances $d_p(\cdot, \cdot)$ to compute k-neighbors.

Question 26.
We consider a K-nearest neighbor (KNN) classifier with $K = 1$. Recall in a KNN classifier, we find the nearest neighbors by computing the distances using a distance measure $d(\boldsymbol{x}, \boldsymbol{y})$. For this problem, we will consider KNN classifiers based on different distance measures based on p-norms

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{j=1}^{M} |x_j - y_j|^p \right)^{\frac{1}{p}}, p \geq 1$$

and what decision surfaces they induce.

In Figure 15 are shown four different decision boundaries obtained by training the KNN ($K = 1$) classifiers using the training observations (marked by the two circles in the figure):

$$\boldsymbol{x}_1 = \begin{bmatrix} 0.301 \\ 0.514 \end{bmatrix}, \quad \boldsymbol{x}_2 = \begin{bmatrix} 0.34 \\ 0.672 \end{bmatrix}$$

and with corresponding class labels $y_1 = 0$ and $y_2 = 1$, but with distance measures based on $p = 1,\ 2,\ 4,\ \infty$ (not necessarily plotted in that order).

Which norms were used in the four KNN classifiers?

A. KNN classifier 1 corresponds to $p = \infty$, KNN classifier 2 corresponds to $p = 2$, KNN classifier 3 corresponds to $p = 4$, KNN classifier 4 corresponds to $p = 1$

B. KNN classifier 1 corresponds to $p = 4$, KNN classifier 2 corresponds to $p = 2$, KNN classifier 3 corresponds to $p = 1$, KNN classifier 4 corresponds to $p = \infty$

C. KNN classifier 1 corresponds to $p = 4$, KNN classifier 2 corresponds to $p = 1$, KNN classifier 3 corresponds to $p = 2$, KNN classifier 4 corresponds to $p = \infty$

D. KNN classifier 1 corresponds to $p = \infty$, KNN classifier 2 corresponds to $p = 1$, KNN classifier 3 corresponds to $p = 2$, KNN classifier 4 corresponds to $p = 4$

E. Don't know.

Question 27. Consider a small dataset comprised of $N = 9$ observations

$$x = \begin{bmatrix} 0.1 & 0.3 & 0.5 & 1.0 & 2.2 & 3.0 & 4.1 & 4.4 & 4.7 \end{bmatrix}.$$

Suppose a k-means algorithm is applied to the dataset with $K = 4$ and using Euclidian distances. At a given stage of the algorithm the data is partitioned into the blocks:

$$\{0.1, 0.3\}, \{0.5, 1\}, \{2.2, 3, 4.1\}, \{4.4, 4.7\}$$

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

A. $\{0.1, 0.3, 0.5, 1\}$, $\{2.2\}$, $\{\}$, $\{3, 4.1, 4.4, 4.7\}$

B. $\{0.1, 0.3\}$, $\{0.5, 1\}$, $\{2.2, 3\}$, $\{4.1, 4.4, 4.7\}$

C. $\{0.1, 0.3\}$, $\{0.5\}$, $\{1, 2.2\}$, $\{3, 4.1, 4.4, 4.7\}$

D. $\{0.1, 0.3\}$, $\{0.5, 1, 2.2, 3\}$, $\{4.1, 4.4\}$, $\{4.7\}$

E. Don't know.