Technical University of Denmark

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

Course name: Introduction to machine learning and data mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

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

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

Answers:

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

Name: _____

Student number: _____

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

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

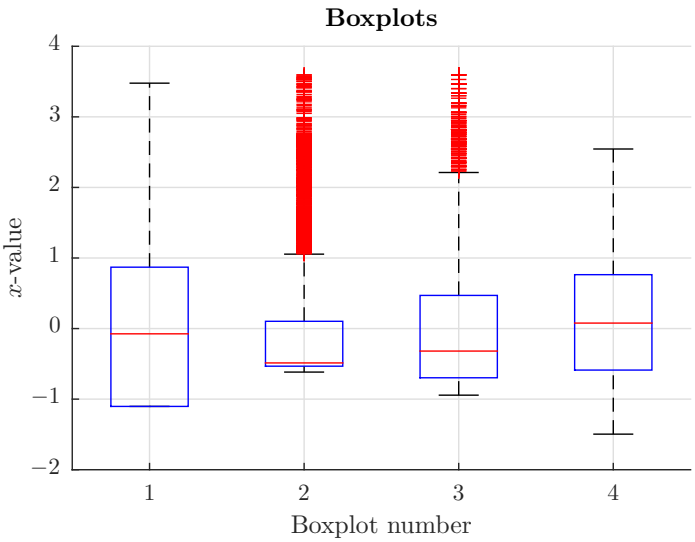

Figure 1: Boxplots corresponding to the variables x_1, x_2, x_3, x_4 of Figure 2 but not necessarily in that order. Notice the features have been standardized.

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

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

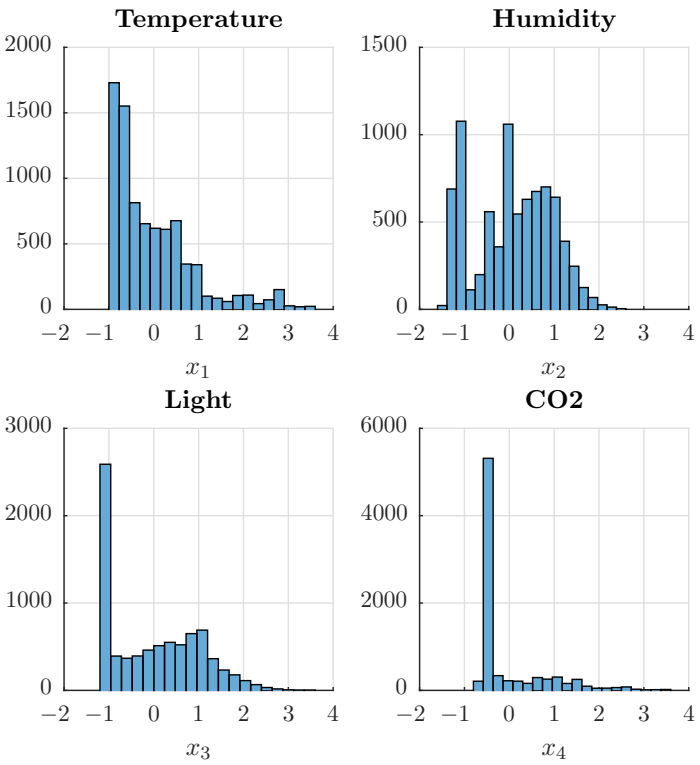

Figure 2: Plot of observations x_1, x_2, x_3, x_4 of the *Occupancy* dataset of Table 1 as histograms. Notice the features have been standardized.

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

E. Don't know.

Solution 2. Recall the variance of a given component is

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

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

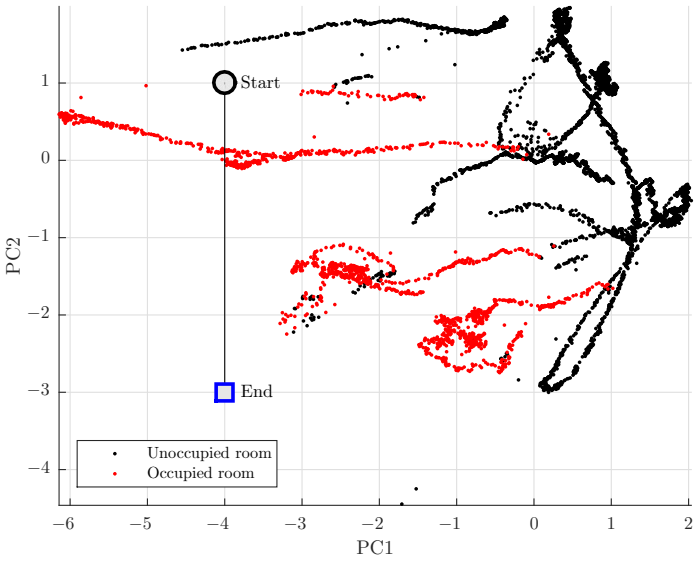

Figure 3: Plot of the 8143 observations from the *Occupancy* dataset of Table 1 projected onto the first two principal directions.

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

from which it follows

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

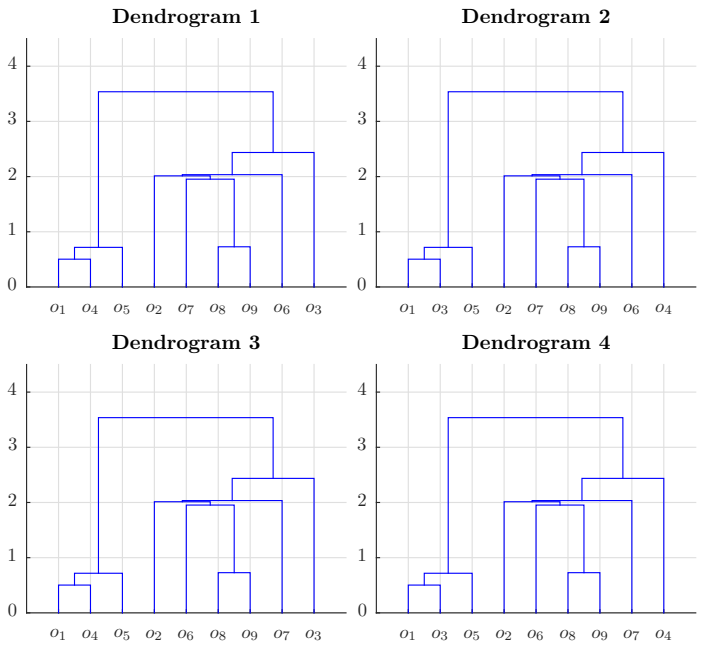

Figure 4: Proposed hierarchical clustering of the 9 observations considered in Table 2

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

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

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

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

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

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

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

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

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

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

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

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

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

E. Don't know.

Solution 8. Notice the inner product can be recovered as

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

and the definition of Cosine similarity is

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

Thus the true answer is -0.862763157895

Question 9.

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

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

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

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

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

E. Don't know.

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

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

And we compute:

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

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

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

Question 10.

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

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

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

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

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

E. Don't know.

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

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

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

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

where

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

E. Don't know.

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

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

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

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

Question 13.

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

option A is correct.

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

following statements is true?

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

B. x_2 is ratio and x_1 is nominal

C. x_1, x_4 are ordinal

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

E. Don't know.

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

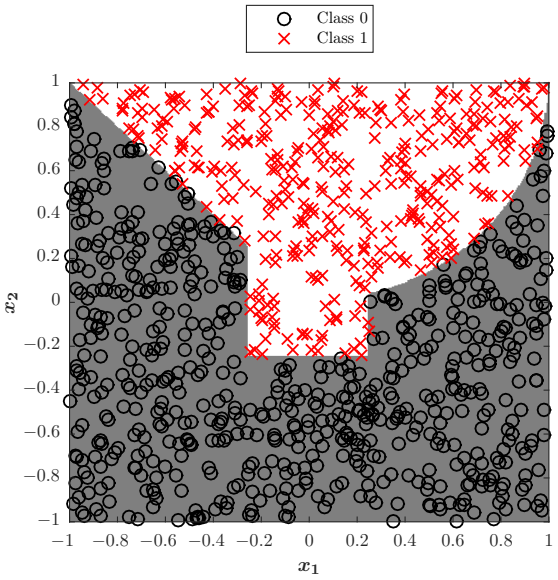

Figure 5: Two-class classification problem

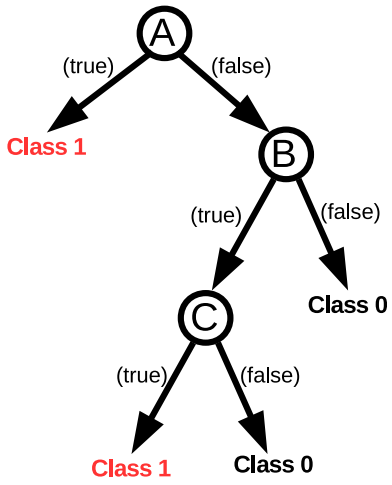

Figure 6: Decision tree with 3 nodes A, B and C

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

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

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

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

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

E. Don't know.

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

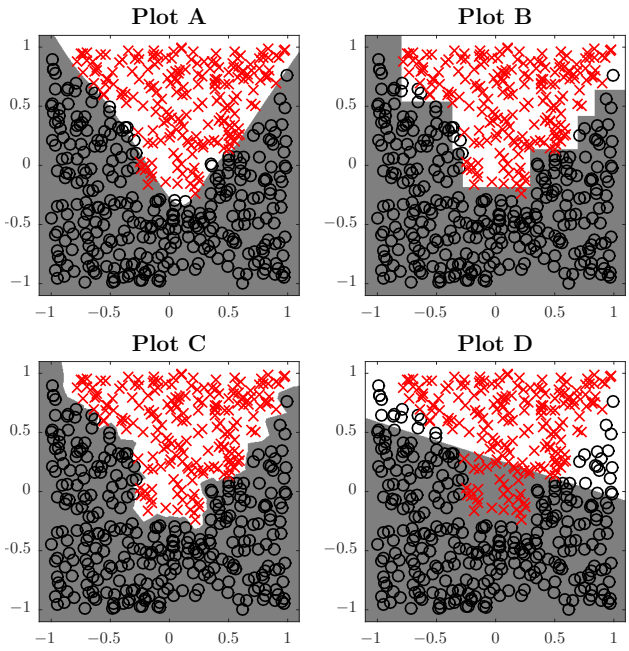

Figure 7: Four classifiers applied to a two-class classification problem

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

(1NN) A 1-nearest neighbour classifier

(TREE) A decision tree

(LREG) Logistic regression

(NNET) An artificial neural network with four hidden units

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

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

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

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

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

E. Don't know.

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

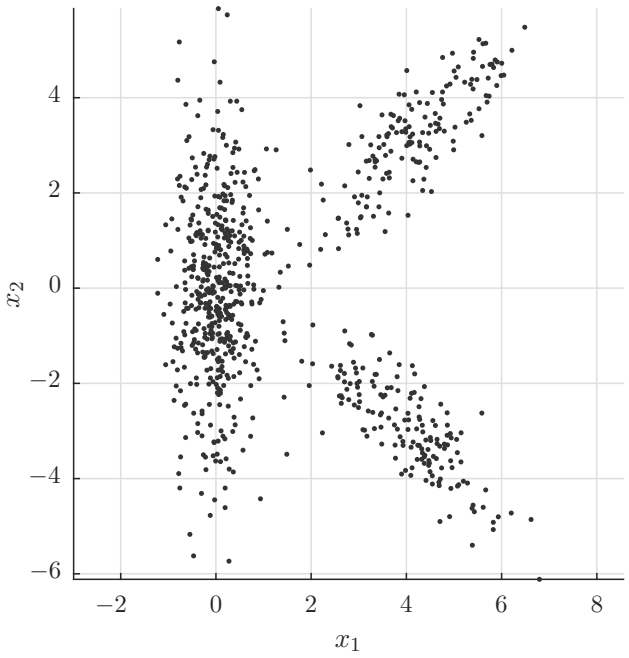

Figure 8: Scatter plot of observations generated from a Gaussian mixture-model

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

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

A. The density is:

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

B. The density is:

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

C. The density is:

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

D. The density is:

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

E. Don't know.

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

Question 20. Consider a 1D GMM mixture model

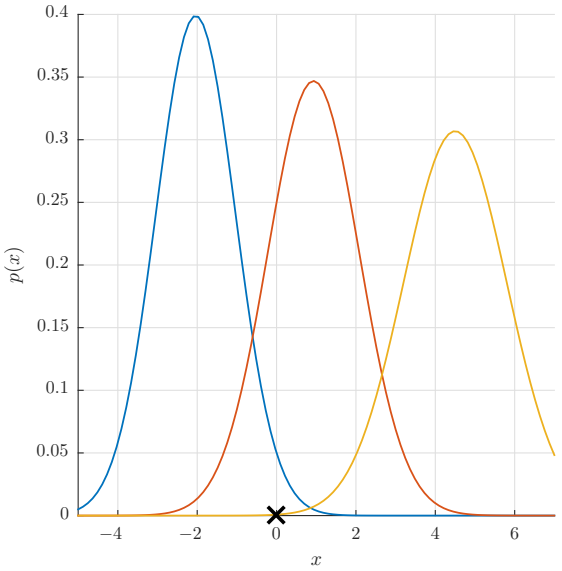

Figure 9: Mixture components in a GMM mixture model with $K = 3$

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

A. 0.05

B. 0.17

C. 0.25

D. 0.02

E. Don't know.

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

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

Question 21.

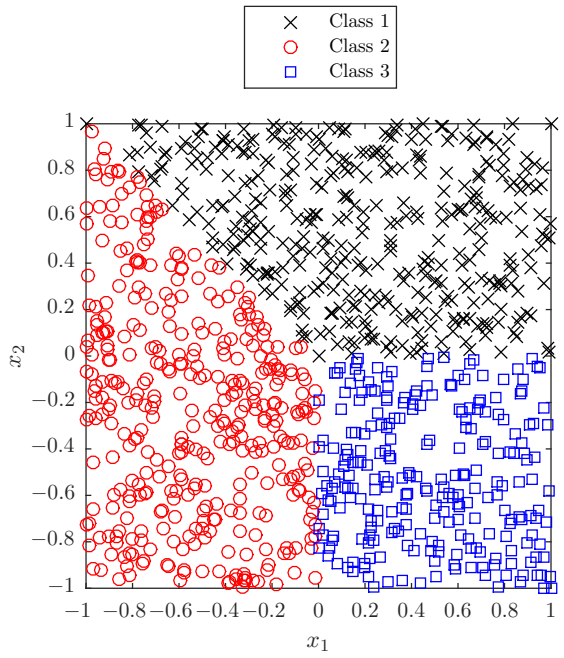

Figure 10: Observations labelled with the most probable class

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

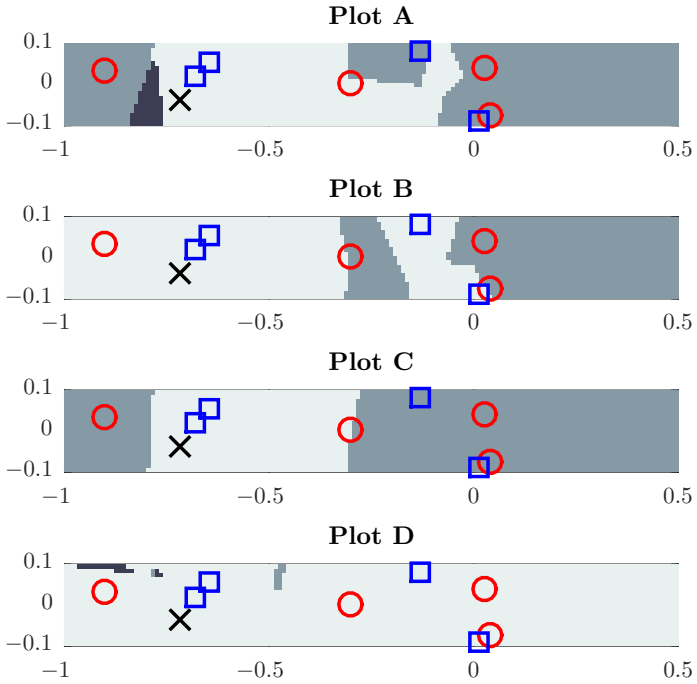

Figure 11: Decision boundaries for four KNN classifiers. The dataset contains three classes indicated by the black crosses (class 1), red circles (class 2) and blue squares (class 3).

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

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

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

X	1	3	4	6	7	8	13	15	16	17

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

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

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

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

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

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

E. Don't know.

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

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

It is easy to check the third option has converged.

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

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

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

A. s is scale invariant

B. s is translation invariant

C. s is both translation and scale invariant

D. s is neither translation or scale invariant

E. Don't know.

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

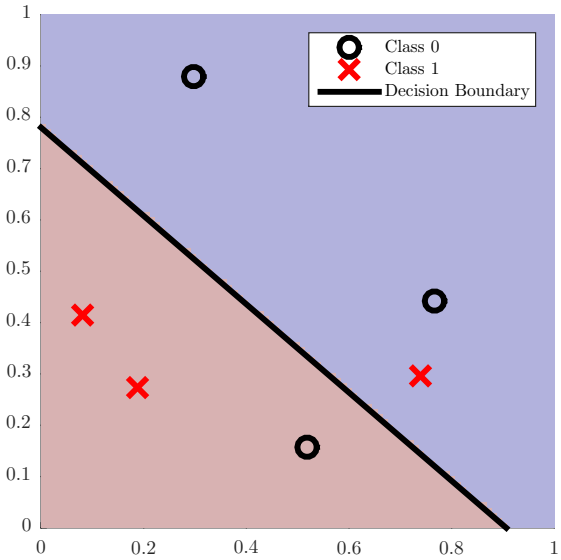

Figure 12: A binary classification problem and the decision boundary obtained by logistic regression. Observations left of the boundary are classified as belonging to the positive class 1 (red crosses) and observations right of the boundary to the negative class 0 (black circles)

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

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

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

and normalizing:

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

accordingly option B is correct.

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

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

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

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

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

E. Don't know.

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

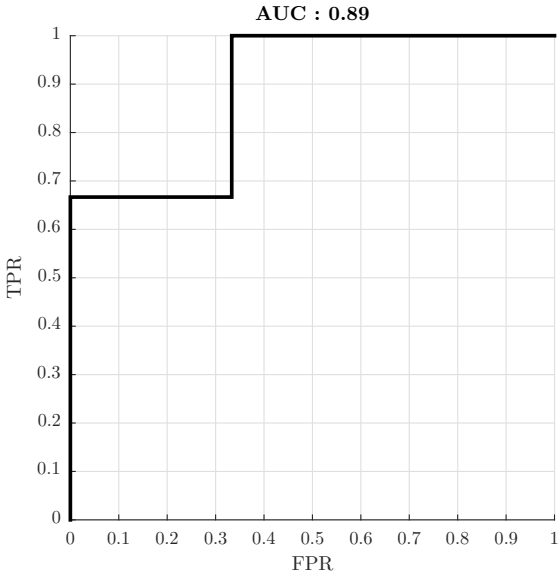

Figure 13: AUC scores computed by plotting the TPR and FPR of Figure 12 against each other. The rates are obtained by translating the decicion boundary horizontally starting from the far right (everything is in the positive class).

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

$$3, 4, 5.$$

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

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

A. $K_1 = 6, K_2 = 5$

B. $K_1 = 3, K_2 = 11$

C. $K_1 = 14, K_2 = 2$

D. $K_1 = 4, K_2 = 9$

E. Don't know.

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

$$K_1(K_2L + 1).$$

This gives for each of the options:

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

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