Technical University of Denmark

Written examination: 16 December 2016, 9 AM - 1 PM.

Course name: Introduction to Machine Learning and Data Mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

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

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

Answers:

1	2	3	4	5	6	7	8	9	10
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	Area	A
x_2	Perimeter	P
x_3	Length of kernel	L
x_4	Width of kernel	W
y	Seed type	

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

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

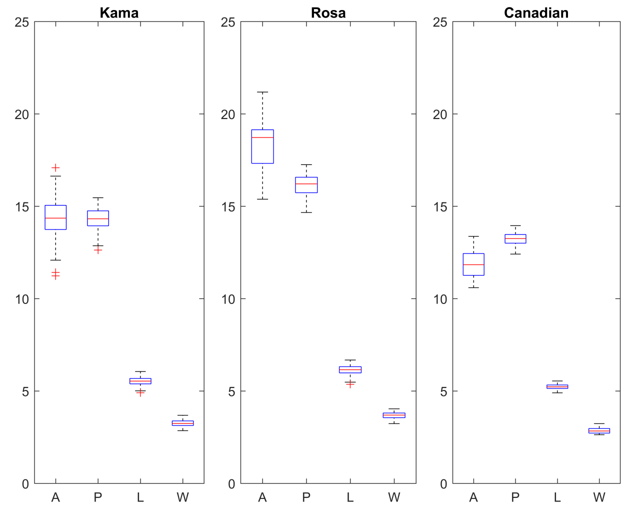

Figure 1: Boxplot of the data visualized separately for each of the three types of seeds; Kama, Rosa, and Canadian.

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

B. The output variable y is ordinal.

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

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

E. Don't know.

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

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

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

Which one of the following statements is *correct*?

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

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

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

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

E. Don't know.

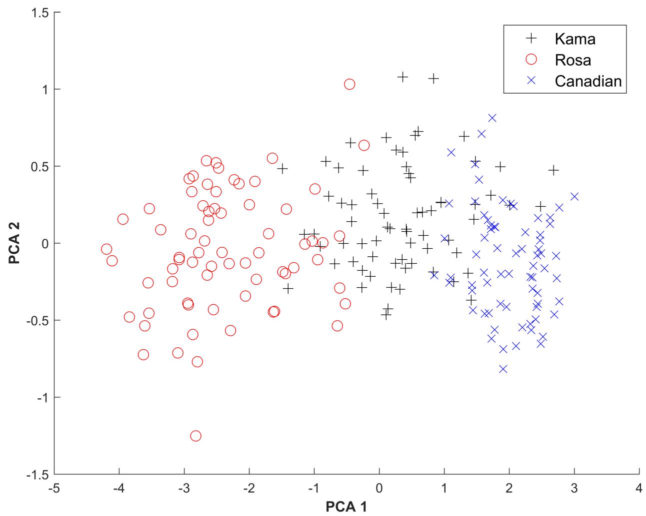

Figure 2: Data projected onto the first and second principal components.

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

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

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

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

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

E. Don't know.

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

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

What is the purity gain of this split?

A. -1.0148

B. 0.0148

C. 0.3333

D. 0.6666

E. Don't know.

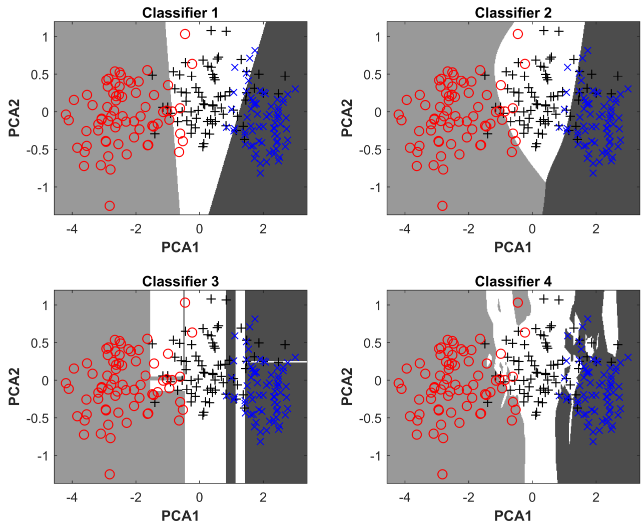

Figure 3: Decision boundaries for four different classifiers trained on the Seeds data projected onto the first two principal components.

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

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

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

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

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

E. Don't know.

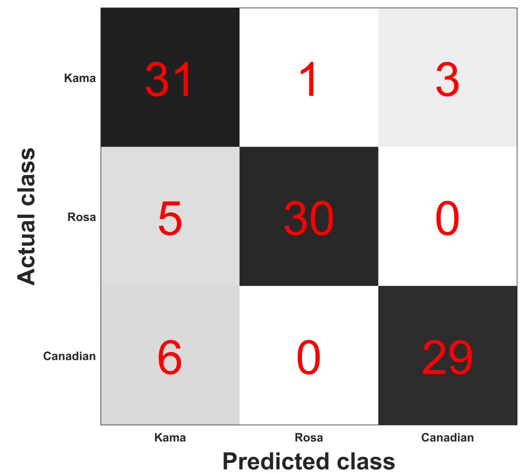

Figure 4: The confusion matrix of a 3-nearest neighbor classifier used to predict the Seeds data.

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

A. 0.1429

B. 0.8571

C. 0.8911

D. 0.9574

E. Don't know.

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

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

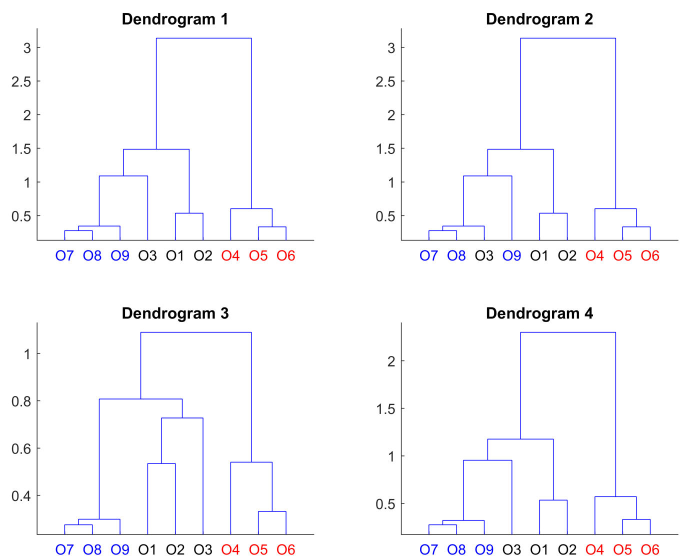

Figure 5: Four different dendrograms derived from the distances between the nine observation in Table 2.

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

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

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

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

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

A. 0.7500

B. 0.7778

C. 0.8611

D. 1.0000

E. Don't know.

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

A. All the observations will be correctly classified.

B. One of the observations will be misclassified.

C. Two of the observations will be misclassified.

D. Three of the observations will be misclassified.

E. Don't know.

Question 10. We suspect that observation O4 may be an outlier. In order to assess if this is the case we would like to calculate the average relative KNN density based on the observations given in Table 2 only. We recall that the KNN density and average relative density for the observation $\boldsymbol{x}_i$ are given by:

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

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

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors. Based on the data in Table 2, what is the average relative density for observation O4 for $K = 1$ nearest neighbors?

A. 0.54

B. 0.61

C. 1.63

D. 1.85

E. Don't know.

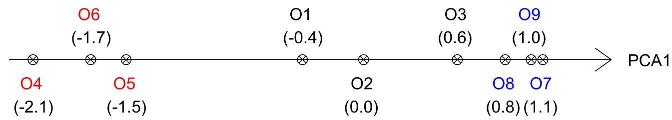

Figure 6: The nine observations considered in Table 2 projected onto the first principal component (the location of the projection is given in parenthesis).

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

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

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

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

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

E. Don't know.

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

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

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

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

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

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

D. Forward selection will select only one feature.

E. Don't know.

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

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

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

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

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

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

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

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

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

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

E. Don't know.

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

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

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

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

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

A. 15.83 %

B. 32.20 %

C. 65.52 %

D. 98.82%

E. Don't know.

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

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

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

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

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

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

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

E. Don't know.

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

A. 0.0%

B. 7.1 %

C. 21.4%

D. 100.0 %

E. Don't know.

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

A. 1/14

B. 3/14

C. 1/2

D. 11/19

E. Don't know.

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

A. The curve with AUC=0.636.

B. The curve with AUC=0.864.

C. The curve with AUC=0.909.

D. The curve with AUC=1.000.

E. Don't know.

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

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

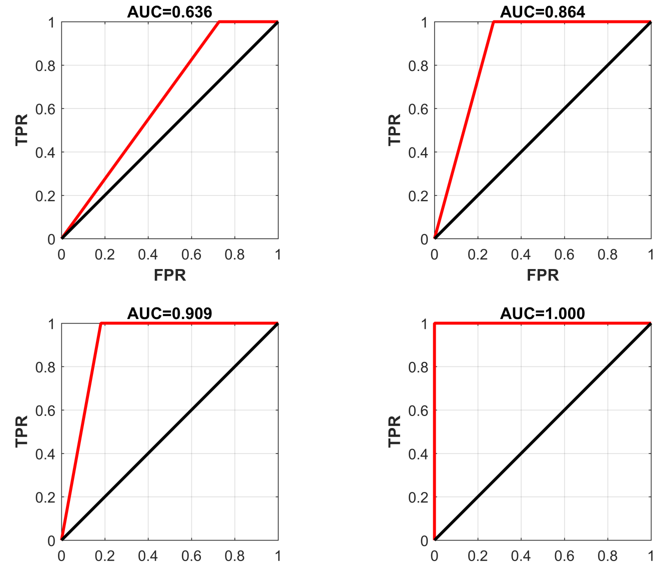

Figure 7: Four different receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC).

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

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

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

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

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

E. Don't know.

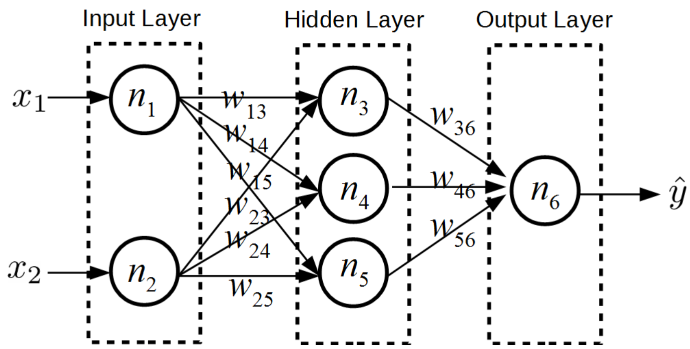

Figure 8: The architecture of the considered neural network having one hidden layer.

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

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

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

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

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

A. 0

B. 0.25

C. 0.75

D. 1

E. Don't know.

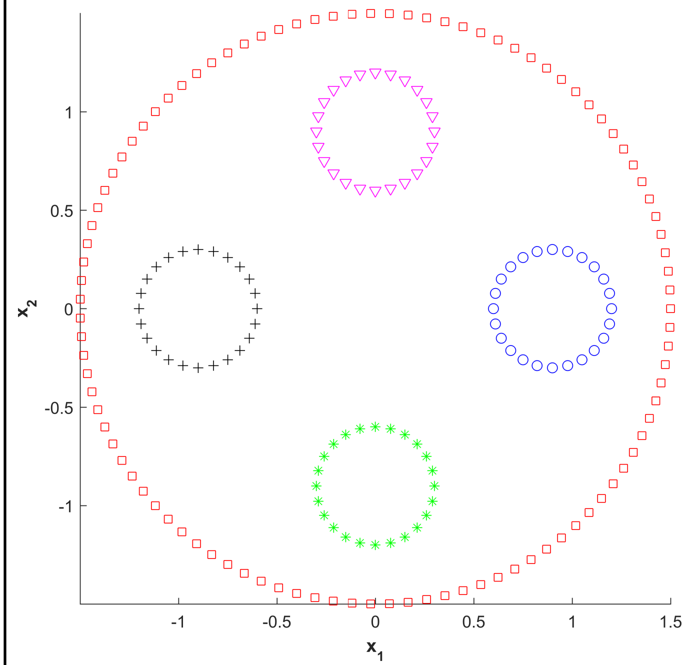

Figure 9: A dataset with five classes given respectively by a large circle and four smaller circles.

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

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

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

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

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

E. Don't know.

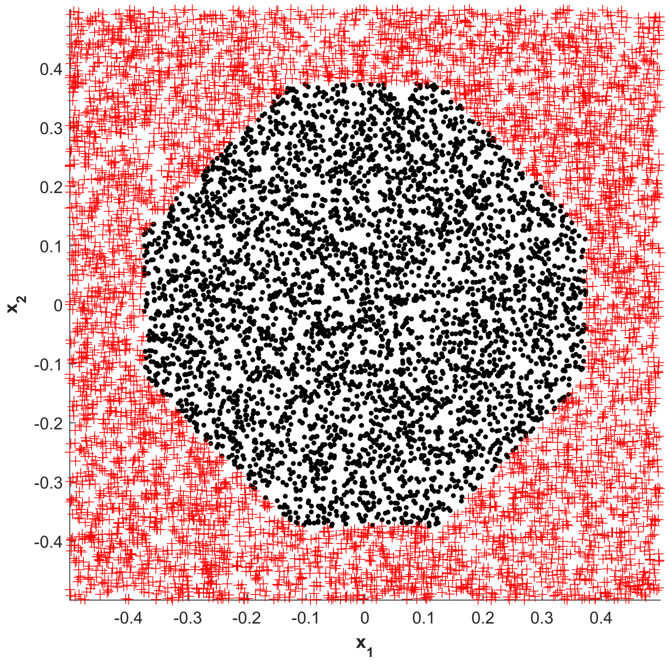

Figure 10: A two class classification problem.

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

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

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

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

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

E. Don't know.

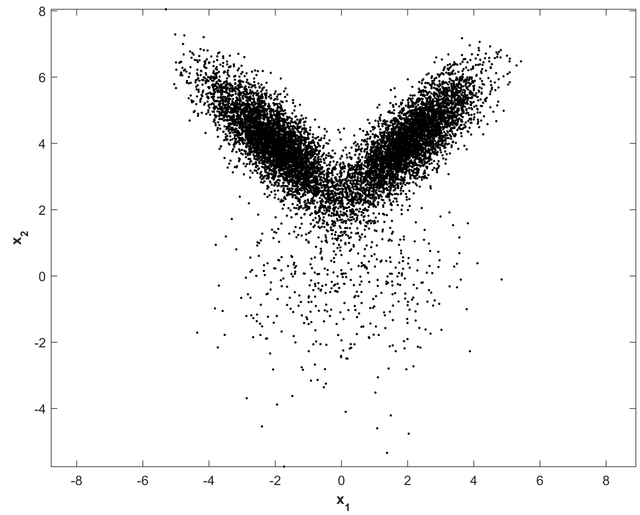

Figure 11: 10.000 data observations drawn from a Gaussian Mixture Model (GMM).

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

A.

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

B.

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

C.

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

D.

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

E. Don't know.

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

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

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

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

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

E. Don't know.

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

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

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

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

A. 0.02

B. 0.025

C. 0.08

D. 0.1

E. Don't know.

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

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

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

C. Select the observations that minimize the training error.

D. Select the number of neighbors in KNN classification.

E. Don't know.

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

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

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

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

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

E. Don't know.