Technical University of Denmark

Written examination: 16th December 2015, 9 AM - 1 PM. Page 1 of 17 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.

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 in the table below with one of the letters A, B, C, D, or E.

Please write your name and student number clearly. Only the present page (page 1) gives your answers to the written test. Other pages will not be considered.

Answers:

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

Name: _____

Student number: _____

ONLY THIS PAGE IS USED FOR THE EVALUATION.
ALL 17 PAGES MUST BE HANDED IN.

No.	Attribute description	Abbrev.
x_1	Age	AGE
x_2	Blood pressure	BP
x_3	Blood glucose random	BGR
x_4	Blood urea	BU
x_5	Serum creatinine	SC
x_6	Hemoglobin	HEMO
y	Chronic kidney disease (0: no 1: yes)	CKD

Table 1: Attributes of the *Chronic Kidney Disease* dataset. The dataset includes 6 attributes $(x_1, \ldots, x_6)$ of 294 person and whether they have a chronic kidney disease or not.

Question 1. Consider the *Chronic Kidney Disease* dataset with attributes given in Table 1[1]. Notice, the dataset has been pre-processed for this exam and only some of the attributes in the original data are presently considered whereas persons with missing data have been removed. A boxplot of the attributes are given in Figure 1. Which of the following statements is true?

A. Regression methods are more suitable than classification methods to predict the output y for this data.

B. BP is ratio whereas y is nominal.

C. All the attributes appear to be normal distributed.

D. From the boxplots it can be seen that there are many outliers in the data that have to be removed.

E. Don't know.

Solution 1.

A As y is nominal classification approaches are more well suited than regression methods.

B Indeed the BP attribute is ratio as zero would constitute an absence of what is being measured. As y is one or zero indicating if the person has a chronic kidney disease or not this is nominal. Thus, this answer option is correct.

C Not all attributes appear to be normal distributed. For instance BP seems to have the 50th and 75th percentiles coincide which would not be the case had the attribute been normal distributed.

[1]Dataset obtained from http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

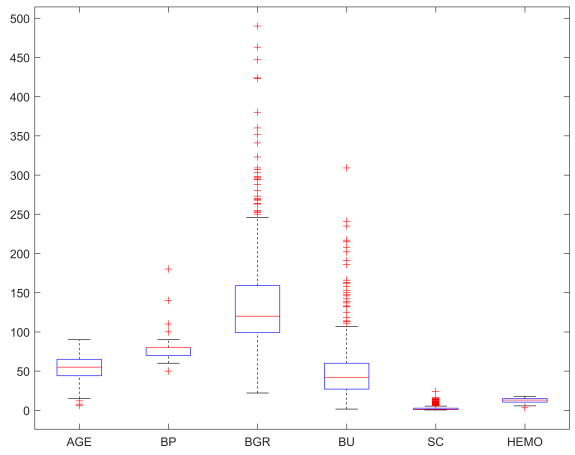

Figure 1: Boxplots of the six attributes of the Chronic Kidney Disease data.

D Even though outliers are indicated in red in the boxplot this is only because they are outside the 1.5 times interquartile range, but there is no reason to believe these values should not be correct.

Question 2. A principal component analysis is carried out on the *Chronic Kidney Disease* dataset based on the attributes $x_1, \ldots, x_6$ found in Table 1. We standardize the data, i.e. by subtracting the mean from each attribute and dividing each attribute by its standard deviation to form the standardized data matrix $\tilde{\boldsymbol{X}}$ of size 294 *subjects* $\times$ 6 *attributes* and apply a singular value decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$ where

$$\boldsymbol{S} = \begin{bmatrix} 27.9 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 18.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 15.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 14.5 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 11.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 8.4 \end{bmatrix}$$

$$\boldsymbol{V} = \begin{bmatrix} 0.26 & -0.57 & 0.40 & 0.66 & -0.07 & 0.01 \\ 0.29 & -0.19 & -0.89 & 0.25 & -0.12 & -0.05 \\ 0.27 & -0.62 & 0.06 & -0.70 & -0.21 & -0.00 \\ 0.51 & 0.34 & 0.16 & -0.02 & -0.22 & -0.74 \\ 0.50 & 0.37 & 0.11 & 0.01 & -0.42 & 0.66 \\ -0.51 & -0.04 & -0.01 & 0.09 & -0.85 & -0.13 \end{bmatrix}.$$

We note that the entries of the matrices above have been rounded. Which one of the following statements is true?

A. The first two principal components account for more than 60 % of the variance in the data.

B. The last principal component accounts for less than 2 % of the variance in the data.

C. The first principal component accounts for more than 50 % of the variance in the data.

D. The performed principal component analysis will mainly be driven by BGR as this attribute has the largest variance.

E. Don't know.

Solution 2. Recall that the variance of the first k components are

$$\text{var.} = \frac{\sum_{i=1}^{k} S_{ii}^2}{\sum_{j=1}^{6} S_{jj}^2}$$

Thus, the first two principal components account for $\frac{27.9^2+18^2}{27.9^2+18^2+15.8^2+14.5^2+11.1^2+8.4^2} = 0.6278$, whereas the last principal component accounts for $\frac{8.4^2}{27.9^2+18^2+15.8^2+14.5^2+11.1^2+8.4^2} = 0.0402$, and the first principal component for $\frac{27.9^2}{27.9^2+18^2+15.8^2+14.5^2+11.1^2+8.4^2} = 0.4433$ of the variance. Since the data has been standardized each attribute is given equal importance in the PCA. However, had the data not been standardized the analysis would be highly driven by BGR and it would be expected the first principal component would mainly capture variance along the direction of BGR.

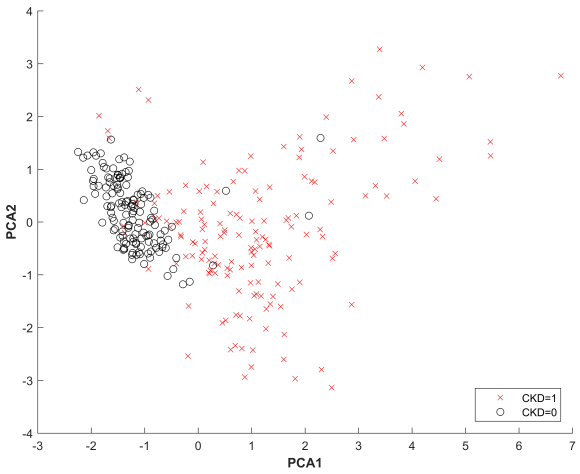

Figure 2: PCA of the chronic kidney disease data.

Question 3. In Figure 2 is given the data projected onto the first two principal components where observations corresponding to $CKD = 1$ is marked by red crosses (x) whereas observations corrresponding to $CKD = 0$ are makred by black circles (○). Which statement is correct?

A. The observations projected onto the first principal component direction can be found as the first column of the matrix $\tilde{\boldsymbol{X}}\boldsymbol{S}^\top$.

B. It seems to be more likely to have chronic kidney disease if a person has relative low values of AGE, BP, BGR, BU, SC and a relatively high value of $HEMO$.

C. Principal component three appears to mainly be discriminating old people with low blood pressure from young people with high blood pressure.

D. Principal component analysis identifies features that are optimal for discriminating between persons with chronic kidney disease from persons not having chronic kidney disease.

E. Don't know.

Solution 3. The data projected onto the first PCA direction is given by $X\boldsymbol{v}_1$. According to the data projected onto the first PCA given in figure fig. 2 it appears that high and not low values of AGE, BP, BGR, BU, SC and low not high values of $HEMO$ would result in a large projection onto PCA which is where most red crosses indicating observations with chronic kidney disease are found. Indeed it appears as if principal

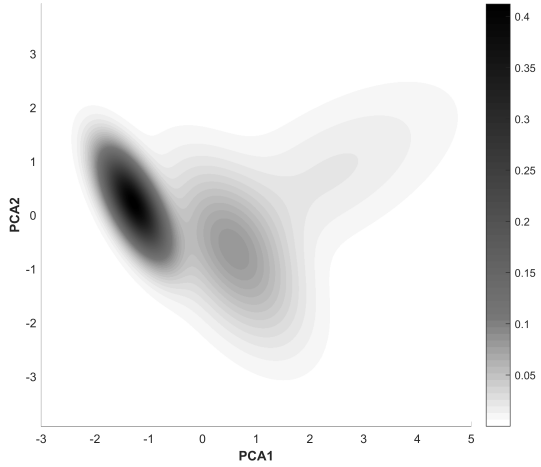

Figure 3: The density of a GMM fitted to the data projected onto the first two principal components given in Figure 2.

component three is mainly discriminating old people with low blood pressure from young people with high blood pressure as the first and second coefficients with largest magnitude pertain to x_1 and x_2 are 0.40 and -0.89 thus having opposing signs. PCA is optimized for accounting for variance and not for discrimination persons with and without chronic kidney disease.

Question 4. A Gaussian Mixture model is fitted to the data projected onto the first two principal components using K=3 mixture components. The estimated density is shown in Figure 3. Which of the following expressions corresponds to the estimated density?

A.

$$p(\boldsymbol{x}) = 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & -0.44 \\ -0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & -0.16 \\ -0.16 & 0.39 \end{bmatrix}).$$

B.

$$p(\boldsymbol{x}) = 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & -1.06 \\ -1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & 0.44 \\ 0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & 0.16 \\ 0.16 & 0.39 \end{bmatrix}).$$

C.

$$p(\boldsymbol{x}) = 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.16 & -0.16 \\ -0.16 & 0.39 \end{bmatrix}) + 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.64 & -0.44 \\ -0.44 & 1.13 \end{bmatrix}).$$

D.

$$p(\boldsymbol{x}) = 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & 0.44 \\ 0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & 0.16 \\ 0.16 & 0.39 \end{bmatrix}).$$

E. Don't know.

Solution 4. The density is:

$$p(x) = 0.18 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}, \begin{bmatrix} 2.42 & 1.06 \\ 1.06 & 1.24 \end{bmatrix}) + 0.34 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} 0.59 \\ -0.72 \end{bmatrix}, \begin{bmatrix} 0.64 & -0.44 \\ -0.44 & 1.13 \end{bmatrix}) + 0.48 \cdot \mathcal{N}(\boldsymbol{x}| \begin{bmatrix} -1.29 \\ 0.22 \end{bmatrix}, \begin{bmatrix} 0.16 & -0.16 \\ -0.16 & 0.39 \end{bmatrix}).$$

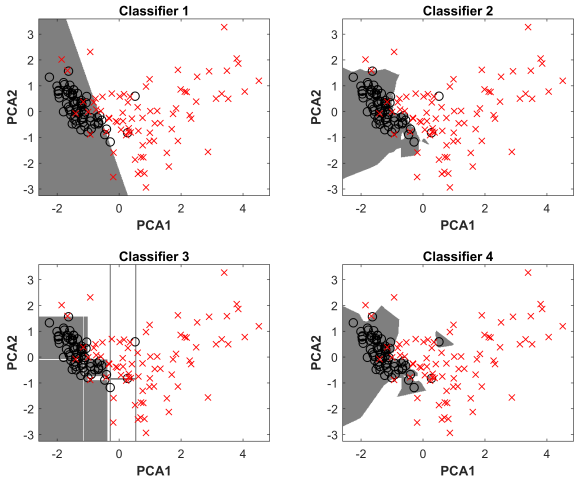

Figure 4: Four classifiers trained on half of the data using PCA1 and PCA2 as features for the classifier.

This can be observed sa the cluster with center at $\begin{bmatrix} 2.39 \\ 0.78 \end{bmatrix}$ has positive covariance between PCA1 and PCA2 whereas the remaining two clusters have negative covariance. While the first cluster also have the lowest density values, i.e. lowest value of the mixing proportions. This is only the case for this density.

Question 5. Using the data projected onto the first two principal components four different classifiers are trained on half of the data and their decision boundaries given in Figure 4. Which one of the following statements is true?

A. Classifier 1 is a decision tree, Classifier 2 is a one-nearest-neighbour classifier, Classifier 3 is a logistic regression model, and Classifier 4 is an artificial neural network with 5 hidden units.

B. Classifier 1 is a logistic regression, Classifier 2 is a one-nearest-neighbour classifier, Classifier 3 is a decision tree, and Classifier 4 is an artificial neural network with 5 hidden units.

C. Classifier 1 is a three-nearest-neighbour classifier, Classifier 2 is a decision tree, Classifier 3 is a logistic regression, and Classifier 4 is an artificial neural network with 5 hidden units.

D. Classifier 1 is a logistic regression, Classifier 2 is a three-nearest-neighbour classifier, Classifier 3 is a decision tree, and Classifier 4 is a one-nearest neighbor classifier.

E. Don't know.

Solution 5. Classifier 1 is a logistic regression as the boundary is defined by a straight line, Classifier 2 is a three-nearest neighbor classifier since single observations are not surrounded by a decision boundary for their classes, Classifier 3 is a decision tree which is observed from the horizontal and vertical lines, and Classifier 4 is a one-nearest-neighbour classifier as it can be observed that single observations are surrounded by decision boundaries pertaining to them .

Feature(s)	Training ErrorRate	Test ErrorRate
x_1	0.3537	0.3061
x_2	0.4286	0.4422
x_3	0.3605	0.2517
x_4	0.2993	0.3061
x_1 and x_2	0.3265	0.3401
x_1 and x_3	0.3401	0.2653
x_1 and x_4	0.2517	0.2381
x_2 and x_3	0.2857	0.2653
x_2 and x_4	0.2245	0.2449
x_3 and x_4	0.1701	0.1497
x_1 and x_2 and x_3	0.2653	0.2449
x_1 and x_2 and x_4	0.2041	0.2313
x_1 and x_3 and x_4	0.1701	0.1429
x_2 and x_3 and x_4	0.1769	0.1565
x_1 and x_2 and x_3 and x_4	0.1701	0.1633

Table 2: Error rate for the training and test set when using a logistic regression to predict kidney disease based only on the four attributes x_1, x_2, x_3 and x_4.

Question 6. A logistic regression classifier is trained using only combinations of x_1, x_2, x_3 and x_4 and the error rate on the training and test data where half the data again is used for training and the other half for testing is given in Table 2. Which one of the following statements is correct?

A. Forward selection will select the feature set x_3 and x_4.

B. Forward selection will select the feature set x_1 and x_3 and x_4.

C. Backward selection will select the features set x_1 and x_2 and x_3 and x_4.

D. Backward selection will select the feature set x_3 and x_4.

E. Don't know.

Solution 6. Using forward selection, x_3 will initially be selected according to the test error rate, next x_4 and subsequently x_1. As the error rate does not improve by selecting x_2 the forward selection procedure will terminate choosing x_1, x_3, x_4. Backward selection will also terminate at this features set.

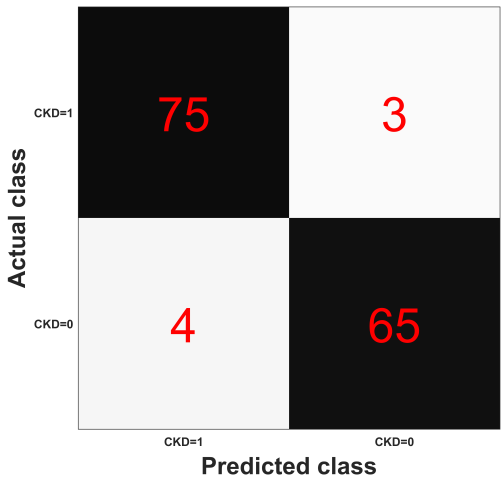

Figure 5: The confusion matrix of a logistic regression classifier evaluated on the test data.

Question 7. The confusion matrix of a logistic regression classifier evaluated on the test data is given in Figure 5. To generate the confusion matrix data has been split in two, half of the data is used for training the classifier and the other half for testing. We will presently consider CKD=1 as the positive class and CKD=0 as the negative class of the classifier. Which one of the following statements is true?

A. The accuracy of the classifier is 7/147 and the error rate is 140/147.

B. The Precision of the classifier is 75/78.

C. The used procedure corresponds to two-fold cross-validation.

D. The used procedure corresponds to the holdout method.

E. Don't know.

Solution 7. The accuracy is 140/147 and the error rate 7/147, not the reverse. The precision of the classifier is 75/(75+4), see also p. 297. The used procedure corresponds to the hold-out method where 50% of the data has been hold out and not two fold-cross validation as only one and not two models are trained.

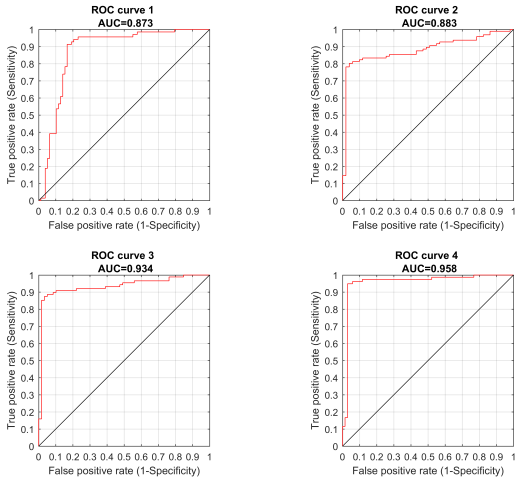

Figure 6: The Receiver Operator Characteristic (ROC) curve for four different classifiers.

Question 8. The performance of four different classifiers are given in terms of their Receiver Operator Characteristic (ROC) curve in Figure 6. One of the classifiers correspond to the classifier with confusion matrix given in Figure 5, which one?

A. ROC curve 1.

B. ROC curve 2.

C. ROC curve 3.

D. ROC curve 4.

E. Don't know.

Solution 8. According to the confusion matrix there exist a threshold in which the true positive rate TPR=75/(75+3)=0.9615 and false positive rate FPR=4/(4+65)=0.0580. This only holds for ROC curve 4.

	O1	O2	O3	O4	O5	O6	O7
O1	0	69	55	117	50	326	36
O2	69	0	36	128	104	303	85
O3	55	36	0	129	94	314	78
O4	117	128	129	0	85	220	91
O5	50	104	94	85	0	303	23
O6	326	303	314	220	303	0	307
O7	36	85	78	91	23	307	0

Table 3: Pairwise Euclidean distance, i.e $d(Oa, Ob) = ||\boldsymbol{x}_a - \boldsymbol{x}_b||_2 = \sqrt{\sum_m (x_{am} - x_{bm})^2}$, between the first four subjects with and first three subjects without chronic kidney disease respectively. Red observations (i.e., O1, O2, O3, and O4) correspond to the four subjects having chronic kidney disease (CKD=1) whereas black observations (i.e., O5, O6, O7) correspond to the four subjects without (CKD=0).

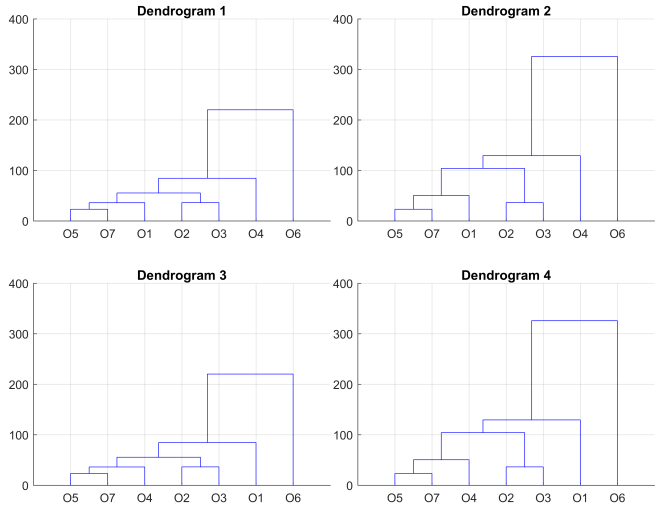

Figure 7: Hierarchical clustering of the seven observations considered in Table 3.

Question 9. In Table 3 is given the pairwise distances between the first four subjects with and the first three subjects without chronic kidney disease. A hierarchical clustering is used to cluster these seven observations using complete (i.e., maximum) linkage. Which one of the dendrograms given in Figure 7 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 9. In complete linkage the observations that is the furthest between the clusters define the level in

which they merge. Initially, O5 and O7 will merge at 23 and O2 and O3 at 36. Next O1 will merge with O5, O7 at 50 and subsequently O2, O3 will merge with O1, O5, O7 at 104, then O4 with O1, O2, O3, O5, O7 at 129, and finally 06 will merge with O1, O2, O3, O4, O5, O7 at 326. This corresponds to dendrogram 2.

Question 10. In order to predict if an observation corresponds to a subject having chronic kidney disease or not we will use a k-nearest neighbor (KNN) classifier based on the Euclidean distance between the seven observations given in Table 3. We will use leave-one-out cross-validation for the KNN in order to classify whether the seven considered observations constitute subjects with or without chronic kidney disease (CKD=1 given in red, i.e. observation O1, O2, O3, O4) and (CKD=0 given in black, i.e. observation O5, O6, O7) using a five-nearest neighbor classifier, i.e. $K = 5$. The analysis will be based only on the data given in Table 3. Which one of the following statements is *correct*?

A. The error rate of the classifier will be 3/7.

B. The error rate of the classifier will be 4/7.

C. All subjects without chronic kidney disease will be correctly classified.

D. All subjects will be correctly classified.

E. Don't know.

Solution 10. All subjects with chronic kidney disease, i.e. O1, O2, O3, and O4 will be correctly classified as there neighbors will be the other observations except observation O6. As there are only two observations that do not have chronic kidney disease available when using leave-one-out the majority will be having chronic kidney disease thus observation O5, O6 and O7 will be miss-classified.

Question 11. We suspect that observation O6 in Table 3 is an outlier. Which procedure would *not* indicate that O6 is the strongest candidate of being an outlier of the seven observations O1–O7?

A. Using the inverse average distance to K-nearest neighbor as density with K=1.

B. Using the average relative density with K=3.

C. Hierarchical clustering using single linkage when inspecting the top split in the dendrogram.

D. Using the density obtained from a Gaussian Mixture Model (GMM) having two clusters with the first cluster mean fixed at observation O1 and the second cluster mean fixed at observation O6 (i.e., the mean values are not updated during the M-step).

E. Don't know.

Solution 11. Using the inverse average distance to K-nearest neighbours as density measure, average relative density, and hierarchical clustering will all clearly point to observation O6 being an outlier. However, the Gaussian Mixture Model with two components, one centered on observation O1 and one centered on observation O6 would create a covariance matrix for the second cluster that would be very small therefore providing a very high density of observation O6 that would therefore not be indicated to be an outlier from the density value.

No.	Attribute description	Abbrev.
x_1	Red blood cells	RBC
x_2	Pus cell	PC
x_3	Pus cell clumps	PCC
x_4	Hypertension	HTN
x_5	Diabetes mellitus	DM
x_6	Coronary artery disease	CAD
x_7	Pedal edema	PE
y	Chronic kidney disease (0: no 1: yes)	CKD

Table 4: Binary attributes of the *Chronic Kidney Disease* dataset. The dataset includes seven attributes $(x_1, \ldots, x_7)$ of 232 persons and whether they have a chronic kidney disease or not.

Question 12. We will now consider some of the Binary attributes also available in the original Chronic Kidney Disease dataset[2]. These binary attributes are given in Table 4. Using these attributes it is found that:

- 56.27 % of the subjects have chronic kidney disease (CKD=1).
- 49.46 % of the subjects with chronic kidney disease (CKD=1) have coronary artery disease (CAD=1).
- 0.7 % of the subjects without chronic kidney disease (CKD=0) have coronary artery disease (CAD=1).

According to this data what is the probability that a subject that has coronary artery disease also has chronic kidney disease?

A. 27.83 %

B. 56.27 %

C. 87.89 %

D. 98.91 %

E. Don't know.

Solution 12. Using Bayes theorem we have :

$$P(CKD=1|CAD=1) = \frac{P(CAD=1|CKD=1)P(CKD=1)}{P(CAD=1)}$$
$$= \frac{P(CAD=1|CKD=1)P(CKD=1)}{P(CAD=1|CKD=1)P(CKD=1)+P(CAD=1|CKD=0)P(CKD=0)}$$
$$= \frac{0.4946 \cdot 0.5627}{0.4946 \cdot 0.5627 + 0.007 \cdot (1-0.5627)}$$
$$= 0.9891 \approx 98.91\%$$

[2]Dataset obtained from http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

	RBC	PC	PCC	HTN	DM	CAD	PE
O_1	0	0	0	0	1	0	0
O_2	0	1	1	1	0	0	1
O_3	0	0	0	0	0	0	0
O_4	0	1	0	0	1	0	1
O_5	0	1	1	1	1	0	0
O_6	1	1	1	1	1	0	0
O_7	1	1	1	1	1	0	1
O_8	0	1	1	1	1	1	1
O_9	0	1	0	1	0	0	0
O_{10}	1	1	0	0	0	1	0
O_{11}	0	0	0	0	1	0	0
O_{12}	0	0	0	0	0	0	0
O_{13}	0	0	0	0	0	0	0
O_{14}	0	0	0	0	0	0	0
O_{15}	0	0	0	0	0	0	0

Table 5: For each observation there are $M = 7$ binary features and $N = 15$ observations $O_1, \ldots, O_{15}$ belonging to two categories (i.e., CKD=1 for $O_1, \ldots, O_9$ and CKD=0 for $O_{10}, \ldots, O_{15}$).

Question 13. We consider the fifteen subjects given in Table 5. We will consider this data set a market basket problem in which the fifteen subjects have various combinations of the seven items denoted RBC, PC, PCC, HTN, DM, CAD, PE. Which one of the proposed solutions below includes *all* the frequent itemsets with support of more than 30 %?

A. {PC}, {PCC}, {HTN}, and {DM}.

B. {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, and {PCC, HTN}.

C. {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, {PC, DM}, and {PCC, HTN}.

D. {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, {PC, DM}, {PCC, HTN}, {PC, PCC, HTN}.

E. Don't know.

Solution 13. Support of more than 30 % implies that there has to at least be 5 observations in an itemset (i.e. 33% of observations). This is the case for the itemsets: {PC}, {PCC}, {HTN}, {DM}, {PC, PCC}, {PC, HTN}, {PC, DM}, {PCC, HTN}, {PC, PCC, HTN}.

Question 14. What is the support and confidence for the decision rule {RBC, PC} → {CAD}?

A. The support is 1/15 and the confidence is 1/3.

B. The support is 1/3 and the confidence is 1/15.

C. The support is 1/3 and the confidence is 1/2.

D. The support is 1/3 and the confidence is 2/15.

E. Don't know.

Solution 14. The support is given by the support of the itemset {RBC, PC, CAD} which is 1/15. The confidence is given as the number of transactions with {RBC, PC, CAD} divided by the number of transactions with {RBC, PC} which is 1/3.

Question 15. We will use a one-nearest neighbor classifier to classify observations as having chronic kidney disease (CKD=1) or not having chronic kidney disease (CKD=0) respectively using the data in Table 5. For the one-nearest neighbor we will use as distance measure $d(O_x, O_y) = 1/SMC(O_x, O_y)$, i.e. high $SMC(O_x, O_y)$ value implies a low $d(O_x, O_y)$ value. If several observations are equally close we will use majority voting to determine the class. Which one of the following statements is correct?

A. The first observation O_1 will be *correctly* classified.

B. The third observation O_3 will be *correctly* classified.

C. The fifth observation O_5 will be *correctly* classified.

D. The twelfth observation O_{12} will be *incorrectly* classified.

E. Don't know.

Solution 15. Recalling that the simple matching coefficient between two binary observations is $SMC(O_x, O_y) = \frac{f_{00}(O_x,O_y)+f_{11}(O_x,O_y)}{M}$ we have that the inverse of the SMC similarity is $SMC^{-1}(O_x, O_y) = \frac{M}{f_{11}(O_x,O_y)+f_{00}(O_x,O_y)}$. Thus:
O_1 is closest to O_{11} and will be miss-classified
O_3 is equally close to O_{12}, O_{13}, O_{14} and will be miss-classified
O_5 is closest to O_6 and will be correctly classified.

Question 16. Nine of the fifteen observations in Table 5 have chronic kidney disease (i.e., O_1–O_9 given in red) whereas six of the observations do not have chronic kidney disease (i.e., O_{10}–O_{15} given in black). We would like to predict whether a subject has chronic kidney disease or not using the data in Table 5 and the attributes RBC, PC, DM, and CAD. We will apply a Naïve Bayes classifier that assumes independence between the four attributes. Given that a subject has these four attributes (i.e., $RBC = 1$, $PC = 1$, $DM = 1$, and $CAD = 1$) what is the probability that the person has chronic kidney disease, i.e., what is $P(CKD = 1|RBC = 1, PC = 1, DM = 1, CAD = 1)$ according to the Naïve Bayes classifier?

A. 2.56 %

B. 96.14 %

C. 98.03 %

D. 100 %

E. Don't know.

Solution 16. According to the Naïve Bayes classifier we have

$$P(CKD = 1|RBC = 1, PC = 1, DM = 1, CAD = 1) = \frac{\begin{pmatrix} P(RBC = 1|CKD = 1)\times \\ P(PC = 1|CKD = 1)\times \\ P(DM = 1|CKD = 1)\times \\ P(CAD = 1|CKD = 1)\times \\ P(CKD = 1) \end{pmatrix}}{\begin{pmatrix} P(RBC = 1|CKD = 1)\times \\ P(PC = 1|CKD = 1)\times \\ P(DM = 1|CKD = 1)\times \\ P(CAD = 1|CKD = 1)\times \\ P(CKD = 1) \end{pmatrix} + \begin{pmatrix} P(RBC = 1|CKD = 0)\times \\ P(PC = 1|CKD = 0)\times \\ P(DM = 1|CKD = 0)\times \\ P(CAD = 1|CKD = 0)\times \\ P(CKD = 0) \end{pmatrix}}$$

$$= \frac{2/9 \cdot 7/9 \cdot 6/9 \cdot 1/9 \cdot 9/15}{2/9 \cdot 7/9 \cdot 6/9 \cdot 1/9 \cdot 9/15 + 1/6 \cdot 1/6 \cdot 1/6 \cdot 1/6 \cdot 6/15} = 0.9614.$$

Question 17. We will consider a Bayes classifier using the attributes RBC, PC, and DM in Table 5 (i.e., we no longer consider the attribute CAD). What is $P(CKD = 1|RBC = 1, PC = 1, DM = 1)$ according to a Bayes classifier (i.e. we are no longer imposing independence as in the Naïve Bayes classifier)?

A. 26.67 %

B. 97.07 %

C. 98.03 %

D. 100 %

E. Don't know.

Solution 17. According to the Bayes classifier we have

$$P(CKD = 1|RBC = 1, PC = 1, DM = 1) = \frac{\left(\begin{array}{r} P(RBC = 1, PC = 1, DM = 1|CKD = 1)\times \\ P(CKD = 1) \end{array}\right)}{\left(\begin{array}{r} P(PRBC = 1, PC = 1, DM = 1|CKD = 1)\times \\ P(CKD = 1) \end{array}\right) + \left(\begin{array}{r} P(RBC = 1, PC = 1, DM = 1|CKD = 0)\times \\ P(CKD = 0) \end{array}\right)} = \frac{2/9 \cdot 9/15}{2/9 \cdot 9/15 + 0/6 \cdot 6/15} = 1$$

Question 18. We will use the data in Table 5 to build a decision tree. We will consider splitting at the root of the tree according to the attribute PC (i.e., according to whether $PC = 0$ or $PC = 1$). What is the purity gain, Δ, of splitting according to PC using the classification error as impurity measure $I(t)$, (i.e., $I(t) = 1 - \max_i[p(i|t)]$)?

A. 2/15

B. 1/5

C. 4/15

D. 2/5

E. Don't know.

Solution 18. The purity gain is given by $\Delta = I(Parent) - (N_{Left}/N \cdot I(left) + N_{Rigth}/N \cdot I(Right))$ where $I = 1 - max_c p(c|j)$. At the root of the tree we have:
$I(Parent) = 1 - 9/15 = 6/15$. For the attribute conditions we obtain:
PC: $N_{Left}/N \cdot I(left) + N_{Right}/N \cdot I(Right) = 7/15 \cdot (1 - 5/7) + 8/15 \cdot (1 - 7/8) = 3/15$ Thus, the purity gain is given as $\Delta = 6/15 - 3/15 = 3/15$.

Question 19. We cluster the binary data considering only the attribute RBC according to K-means using euclidean distance with two clusters having centroids in 0 and 1 respectively. The purity of the clustering is given as:

$$Purity = \sum_{i=1}^{K} \frac{m_i}{m} p_i, \text{ where } p_i = \max_j m_{ij}/m_i$$

where m_i is the number of observation in cluster i, m the total number of observations in all clusters and m_{ij} the number of observations in cluster i of class j. What is the Purity of the clustering?

A. Purity=1/5

B. Purity=3/5

C. Purity=2/3

D. Purity=12/15

E. Don't know.

Solution 19. As cluster 1 containing the observations having $RBC = 0$ has $O_1, O_2, O_3, O_4, 0_5, O_8, O_9, O_{11}, O_{12}, O_{13}, O_{14}, O_{15}$ with seven chronic kidney disease observations, we have that $p_1 = \max\{7/12, 5/12\} = 7/12$. Whereas cluster 2 containing the observations having $RBC = 1$ are O_6, O_7, O_{10} with two chronic kidney disease observations, we have that $p_1 = \max\{2/3, 1/3\} = 2/3$. Thus $Purity = 12/15 \cdot 7/12 + 3/15 \cdot 2/3 = 9/15 = 3/5$

Question 20. Using all 232 observations of the binary data in Table 4 we would like to investigate the generalization performance of logistic regression using two-level cross-validation where we in the outer folds use leave-one-out cross validation and in the inner folds use five-fold cross-validation. In our inner cross-validation we determine the optimal feature combination from all potential combinations of the seven attributes (RBC, PC, PCC, HTN, DM, CAD, and PE) providing $2^7 = 128$ different logistic regression models. Once we from the inner folds have determined the optimal feature combination we train a model using this feature combination on all the training data from the outer fold and estimate the generalization error on the test data of the outer fold. Which one of the following statements is correct?

A. In total 29696 logistic regression models will be trained.

B. In total 148480 logistic regression models will be trained.

C. In total 148712 logistic regression models will be trained.

D. In total 178176 logistic regression models will be trained.

E. Don't know.

Solution 20. As we use leave-one-out-cross validation in the outer fold we have 232 splits into training and testing. For each of these splits we have 5 folds where we evaluate for each of these 5 folds 128 different model settings which gives 640 models. Once we have found the optimal setting we train one additional model on all the training data which provide one additional model to fit, i.e. a total of 641 models fitted 232 times (leave-one-cross validation provides 232 folds) which gives $641 \cdot 232 = 148712$ models. endsolution

Question 21. In Figure 8 is given a boxplot of an attribute denoted A. Which of the following statements regarding this attribute is correct?

A. The mean value of the attribute A is 0.

B. The mode of the attribute A is 0.

C. The range of the attribute A is 5.

D. The attribute A appears to be normal distributed.

E. Don't know.

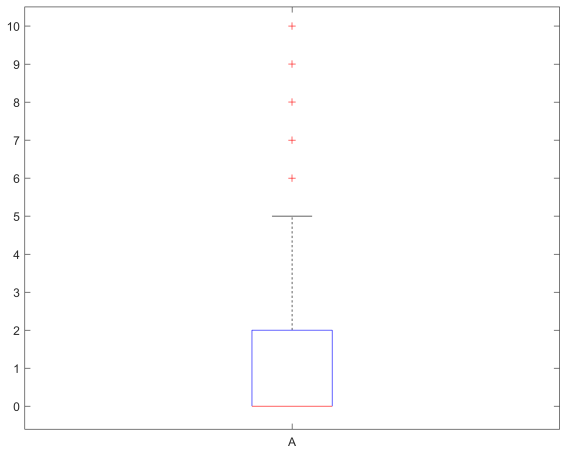

Figure 8: Boxplot of an attribute denoted A.

Solution 21. The mean value of the attribute A is greater than zero as the smallest value of A is zero and there are several values larger than zero, thus the mean will also be larger than zero. The mode is zero as the 50th percentile is located at 0 with at least 50% of the observations therefore taking the value 0. The range of A is 10 with smallest value of zero and largest of 10. The attribute A does not appear to be normally distributed as the distribution is highly asymmetric/skewed.

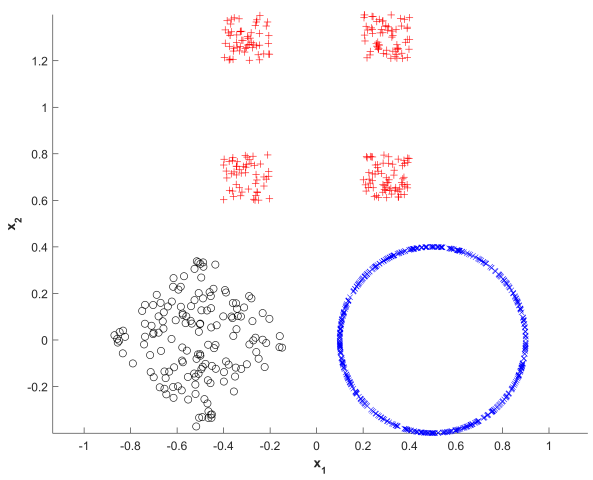

Figure 9: A dataset with two attributes x_1 and x_2 and three clusters given by red plus (+), black circles ($\circ$) and blue crosses (x).

Question 22. We will use the decision tree given in Figure 10 to attempt to separate the observations into the three classes (i.e. red pluses, black circles, and blue crosses) given in Figure 9. Which one of the following choices for the two decisions A, and B in the decision tree would be the most well suited to separate the three classes?

A. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_2 < 0.4$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \|_1 < 1$

B. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_2 < 0.4$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \|_2 < 0.5$

C. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_\infty < 0.5$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \|_1 < 0.5$

D. A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_1 < 0.5$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} \|_2 < 0.5$

E. Don't know.

Solution 22. The three classes would be well separated by the decisions

A=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} \|_\infty < 0.5$
B=$\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} -0.5 \\ 0 \end{bmatrix} \|_1 < 05$
Decision A would capture the red pluses within a square box centered at (0,1) and radius 0.5. Decision B will separate the black circles from blue croses as the black circles are well contained within a diamond shape with radius 0.5 centered at (-0.5, 0) forming the 1-norm.

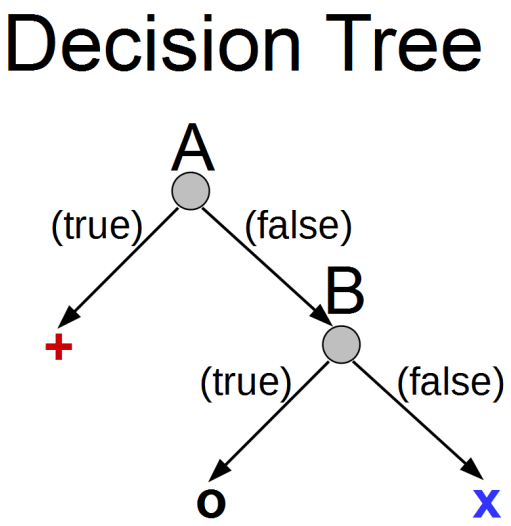

Figure 10: A decision tree with two decisions denoted A and B resulting in a separation into the three clusters given in Figure 9.

Question 23. We again consider the data given in Figure 9 containing the three classes given by red pluses (+), black circles (∘) and blue crosses(x). Which one of the following clustering approaches is most suited to separate the data into the three classes?

A. Well-separated.

B. Center-based.

C. Hierarchical clustering using single linkage.

D. Density-based.

E. Don't know.

Solution 23. For Well-separated we have: Each point is closer to all points in its cluster than any point in another cluster.
Center-based: Each point is closer to the center of its cluster than to the center of any other cluster.
Hierarchical clustering using single linkage: Clusters are merged according to their minimal distance between two points.
Density-based: Clusters are regions of high density separated by regions of low density.
The center-based definition would adequately separate the three classes into clusters as they would all be closer to their center than the center of other clusters. Well-separated would not work since for instance the lower right red observations are closer to some of the blue observations. Contiguity based would at first seem reasonable but it would fail when merging the lower right red cluster as this is closer to the some blue observations. Density based requires clusters be regions of high density separated by regions of low density. However, the red cluster itself contains low-density regions and would therefore not be adequately defined.

Question 24. Which one of the following statements regarding ensemble methods is correct?

A. Ensemble methods aim at combining weak classifiers that each are as similar to each other as possible in terms of how they make predictions.

B. Random Forest corresponds to combining multiple decision trees where each decision tree is trained using features selected according to the AdaBoost sampling procedure.

C. In each Bagging round the same observation can occur more than once in the training set.

D. Bagging puts more emphasis on miss-classified observations.

E. Don't know.

Solution 24. Ensemble methods aim at combining weak classifiers that are independent of each other and not as similar to each other as possible. Random forest randomly sub-samples when training each decision tree however it generally considers more than one attribute for each tree. Bagging uses sampling with replacement, thus, the same observation can occur more than once in the training set. However, Bagging does not put more emphasis on miss-classified observations - this is the strategy of Boosting.

X	2	4	8	11	15	19	20	27	30	31

Table 6: Simple 1-dimensional dataset with $N = 10$ observations.

Question 25. Consider the 1-dimensional data set having $N = 10$ observations shown in table 6. We will cluster the data using K-means based on Euclidean distance and initialized with centroids positioned at the first three observations, i.e. cluster one is located at $x_1 = 2$, cluster two at $x_2 = 4$ and cluster three at $x_3 = 8$. We will use the basic K-means algorithm described in the book and lecture slides. What will be the converged solution of the K-means procedure? (Note: This exercise cannot be solved using computer implementations of K-means as they may use update schemes that are not the same as the basic algorithm given in the book and lecture slides.)

A. {2, 4}, {8, 11, 15}, {19, 20, 27, 30, 31}.

B. {2, 4}, {8, 11, 15, 19, 20}, {27, 30, 31}.

C. {2, 4, 8}, {11, 15, 19, 20}, {27, 30, 31}.

D. {2, 4, 8, 11},{15, 19, 20},{ 27, 30, 31}.

E. Don't know.

Solution 25. Initially only cluster 3 will be closest to the remaining observations and its centroid updated to 20.1250. Subsequently, cluster 2 will be closer to 8 and 11 and cluster 2 and cluster 3 therefore updated to: Cluster 1: 3, cluster 2: 7.66, cluster 3: 23.66
Subsequently the centroids will be updated to: Cluster 1: 3, cluster 2: 11.33 cluster 3: 25.40
After which no more change in assignment will occur and the procedure therefore converge.

Question 26. Which one of the following approaches will not be sensitive to the initialization of the parameters of the model?

A. K-means with five clusters.

B. An Artificial Neural Network (ANN) with 3 hidden units.

C. An Artificial Neural Network (ANN) with 10 hidden units.

D. A one cluster Gaussian Mixture Model (GMM).

E. Don't know.

Solution 26. K-means and artificial neural networks are prone to local minima as is Gaussian Mixture Models. However, when there are only one cluster the expectation step will always be giving probability one of all observations belonging to the same cluster and therefore the GMM with one component will converge to the same solution regardless of initialization.

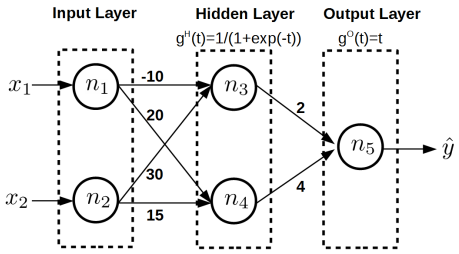

Figure 11: A neural network.

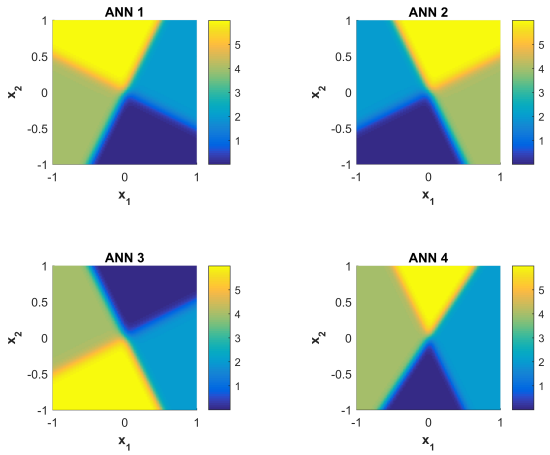

Figure 12: The predictions made by four different neural networks, denoted ANN 1, ANN 2, ANN 3, and ANN 4.

Question 27. Consider the artificial neural network in Figure 11 that has a logistic function as non-linearity in the hidden layer and a linear function in the output layer. I.e., the transfer function for n_3 and n_4 is $g^H(t) = 1/(1 + exp(-t))$ whereas the transfer function for n_5 is $g^O(t) = t$. There are no biases in the network, i.e. the bias for all neurons are 0. Which one of the four artificial networks in Figure 12 corresponds to the ANN in Figure 11?

A. ANN 1

B. ANN 2

C. ANN 3

D. ANN 4

E. Don't know.

Solution 27. Propagating $x_1 = -1$ and $x_2 = -1$ we propagate something that is almost zero through each hidden unit to obtain an output close to zero. This is only the case for ANN 2.