Technical University of Denmark

Written examination: 19 December 2017, 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
D	C	C	B	C	A	B	B	A	D
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
B	C	A	C	A	D	A	C	D	B
21	**22**	**23**	**24**	**25**	**26**	**27**			
A	C	A	A	B	C	A			

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	Height (in feet)	Height
x_2	Weight (in pounds)	Weight
x_3	Percent of successful field goals (out of 100 attempted)	FG
x_4	Percent of successful free throws (out of 100 attempted)	FT
y	average points scored per game	PT

Table 1: The attributes of the Basketball dataset contains 54 observations of basketball players in terms of their height, weight and performance. The output y provides each player's average points scored per game.

Question 1. We will consider a basketball dataset containing 54 National Basketball Association (NBA) basketball players and their performance[1]. For brevity this dataset will be denoted the Basketball dataset in the following. In Table 1 the attributes of the data as well as the output attribute y defined by each player's average points scored per game are given. In Figure 1 is given a boxplot of the four attributes x_1–x_4 after standardizing the data, i.e. subtracting the mean of each attribute and dividing each attribute by its standard deviation. The worst performing player (i.e., the player with lowest value of y) is indicated by a black circle with an asteric inside. Considering the attributes described in Table 1 and the boxplot in Figure 1 which one of the following statements regarding the attributes x_1–x_4 and output variable y is correct?

A. The player with worst performance has Weight (i.e., x_2) that is between the 25th and 50th percentile.

B. The player with worst performance has FT (i.e., x_4) so low it is deemed an outlier.

C. The input attributes FT and FG (i.e., x_3 and x_4) are both ordinal variables.

D. The output y is ratio.

E. Don't know.

Solution 1. Inspecting the boxplot we observe that the value of x_2 is within the 50th and 75th percentiles.

[1]The dataset is taken from http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html

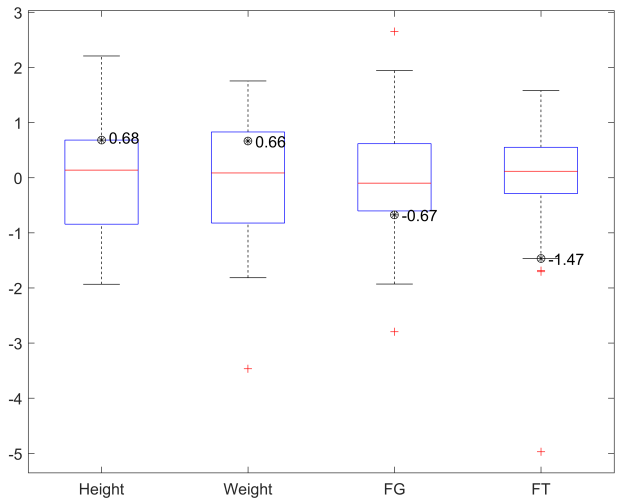

Figure 1: Boxplot of the four attributes x_1–x_4 after standardizing the data (i.e., subtracting the mean of each attribute and dividing each attribute by its standard deviation). The worst performing player (i.e., the player with lowest output value y) is indicated by a black circle with and asteric inside on top of each boxplot along with his standardized value of each attribute given as black digits.

The value of x_4 is positioned exactly at the smallest value for which the observation is still within the 25th percentile subtracted 1.5 times the interquartile range and therefore not an outlier as the lower whisker extends to this observation. In order to be ratio zero has to mean absence of what is being measured. As 0 percent of successful field goals and free throws implies absence of having scored these are ratio attributes and not just ordinal. We can here also talk about 25 % being half as frequent as 50 % etc. As zero average points scored per game implies absence of scoring this output variable is also ratio.

Question 2. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_4, forming the standardized matrix $\tilde{\boldsymbol{X}}$. The squared Frobenious norm of the standardized matrix is given by $\|\tilde{\boldsymbol{X}}\|_F^2 = 212$. A singular value decomposition is applied to the matrix $\tilde{\boldsymbol{X}}$ and we find that the first three singular values are $\sigma_1 = 11.1$, $\sigma_2 = 7.2$, $\sigma_3 = 5.2$. What is the value of the fourth singular value σ_4?

A. $\sigma_4 = 1.2$

B. $\sigma_4 = 2.3$

C. $\sigma_4 = 3.1$

D. $\sigma_4 = 9.9$

E. Don't know.

Solution 2. The variance explained by the i^{th} principal component is given by $\frac{\sigma_i^2}{\sum_{i'} \sigma_{i'}^2} = \frac{\sigma_i^2}{\|\tilde{\boldsymbol{X}}\|_F^2}$. Thus, $\sum_{i'} \sigma_{i'}^2 = \|\tilde{\boldsymbol{X}}\|_F^2$ and from this we know that $\sigma_4 = \sqrt{\|\tilde{\boldsymbol{X}}\|_F^2 - \sum_{i'=1}^3 \sigma_{i'}^2} = \sqrt{212 - 11.1^2 - 7.2^2 - 5.2^2} = 3.1$.

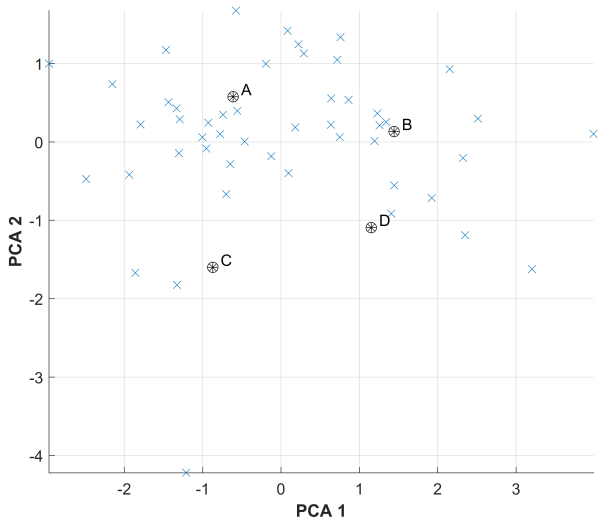

Figure 2: The Basketball data projected onto the first and second principal component. In the plot four observations are highlighted denoted A, B, C, and D of which one of these observation corresponds to the player with lowest output value y given by the circle with and asteric inside in Figure 1.

Question 3. From the singular value decomposition of $\tilde{\boldsymbol{X}}$ we further obtain the following $\boldsymbol{V}$ matrix:

$$\boldsymbol{V} = \begin{bmatrix} -0.60 & 0.02 & -0.41 & 0.69 \\ -0.61 & 0 & -0.33 & -0.72 \\ -0.46 & 0.46 & 0.76 & 0.04 \\ 0.25 & 0.89 & -0.39 & -0.04 \end{bmatrix}.$$

The data projected onto the first two principal components is given in Figure 2 including four observations denoted A, B, C, and D that are marked by a black circle with an asteric inside. One of these four observations corresponds to the worst performing player indicated by a similar black circle with an asteric inside in the boxplot of Figure 1. Which one of the four observations in Figure 2 corresponds to the worst performing player indicated in the boxplots of Figure 1?

A. Observation A.

B. Observation B.

C. Observation C.

D. Observation D.

E. Don't know.

Solution 3. From the boxplot we know that the worst performing player in terms of the output value y has

the standardized observation vector $\tilde{x}^* = [0.68\ 0.66\ -0.67\ -1.47]$ and will thus have the projection onto the two first principal components given by

$$[0.68\ 0.66\ -0.67\ -1.47] \begin{bmatrix} -0.60 & 0.02 \\ -0.61 & 0 \\ -0.46 & 0.46 \\ 0.25 & 0.89 \end{bmatrix} = [-0.8699 - 1.6029].$$

Thus, the observation will in the projection be located at (-0.8699 , -1.6029) which corresponds to the observation denoted C.

Question 4. A least squares linear regression model is trained using different combinations of the four attributes x_1, x_2, x_3,and x_4 in order to predict the average points scored per game y. Table 2 provides the training and test root-mean-square error (RMSE=$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$) performance of the least squares linear regression model when trained using different combinations of the four attributes. Which one of the following statements is correct?

A. Forward selection will terminate when all features x_1–x_4 are included in the feature set.

B. The solution identified by Forward selection will be worse than the solution identified by Backward selection.

C. Forward selection will terminate using two feature in the feature set.

D. Backward selection will terminate using two features in the feature set.

E. Don't know.

Solution 4. Forward selection will select x_4 with performance 5.6845, including additional features will not improve performance and the procedure will therefore terminate with the feature set x_4. Backward selection will remove feature x_2 terminating at the feature set x_1 and x_3 and x_4 with performance 5.5099 since removing additional features only increases the test error.

Feature(s)	Training RMSE	Test RMSE
No features	5.8977	5.8505
x_1	5.8760	6.0035
x_2	5.8841	5.9037
x_3	5.1832	5.9272
x_4	5.8727	5.6845
x_1 and x_2	5.6272	7.4558
x_1 and x_3	5.1482	5.6409
x_1 and x_4	5.8451	5.8269
x_2 and x_3	5.0483	5.6656
x_2 and x_4	5.8660	5.7461
x_3 and x_4	5.1125	5.7390
x_1 and x_2 and x_3	4.9836	6.2823
x_1 and x_2 and x_4	5.6261	7.3888
x_1 and x_3 and x_4	5.0839	5.5099
x_2 and x_3 and x_4	5.0113	5.5605
x_1 and x_2 and x_3 and x_4	4.9645	6.0892

Table 2: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict average points scored per game y using different combinations of the four attributes (x_1–x_4).

Question 5. We will encode the output attribute y in terms of three different classes, i.e. low performing players having performance below the 33.3 percentile, mid-performing players having performance in the range of the 33.3 percentile to 66.6 percentile and high-performing players having performance above the 66.6 percentile (i.e, each of the three classes will contain approximately one third of the data). We presently consider only the attributes FG and FT. Consider the three Gaussian distributions given in Figure 3 where each Gaussian is fitted to each of the three classes separately. We recall that the multivariate Gaussian distribution is given by:
$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The three fitted covariances (in arbitrary order) are given by $\boldsymbol{\Sigma}_a = \begin{bmatrix} 0.0035 & 0.0003 \\ 0.0003 & 0.0030 \end{bmatrix}$, $\boldsymbol{\Sigma}_b = \begin{bmatrix} 0.0028 & -0.0013 \\ -0.0013 & 0.0191 \end{bmatrix}$, and $\boldsymbol{\Sigma}_c = \begin{bmatrix} 0.0020 & 0.0001 \\ 0.0001 & 0.0061 \end{bmatrix}$. The Gaussians are plotted in Figure 3 in terms of lines indicating where $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 5$, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 10$, and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 20$ thus defining ellipsoidal shapes at three different levels of the density functions. Which of the three classes correspond to which of the three covariance matrices $\boldsymbol{\Sigma}_a$, $\boldsymbol{\Sigma}_b$, and $\boldsymbol{\Sigma}_c$?

A. $\boldsymbol{\Sigma}_a$ corresponds to the low performing class, $\boldsymbol{\Sigma}_b$ corresponds to the mid performing class, $\boldsymbol{\Sigma}_c$ corresponds to the high performing class.

B. $\boldsymbol{\Sigma}_a$ corresponds to the low performing class, $\boldsymbol{\Sigma}_c$ corresponds to the mid performing class,

Solution 5.

Inspecting the covariance matrices indicated by the iso-line contours of each gaussian at $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 5$, $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 10$, and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 20$ we observe that the high performing (i.e., blue class) has negative covariance which is only the case for $\boldsymbol{\Sigma}_b$ Of the low performing and mid-performing classes we observe that the low performing has a larger variance in the second dimension (i.e., the x_4 direction) than the mid-performing, thus, as $\boldsymbol{\Sigma}_c(2,2) = 0.0061 > 0.0030 = \boldsymbol{\Sigma}_a(2,2)$ we have that $\boldsymbol{\Sigma}_c$ corresponds to the low performing and $\boldsymbol{\Sigma}_a$ to the mid performing classes.

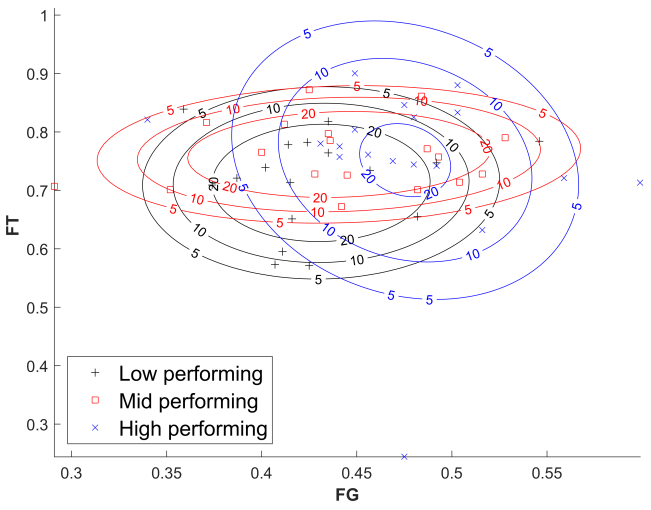

Figure 3: The 54 observations plotted in terms of percentage of successful field goals (FG) plotted against percentage of successful free throws (FT) for low, mid and high performing players respectively indicated by black plusses, red squares, and blue crosses. For each of these three classes a multivariate Gaussian distribution is fitted and the lines corresponding to density values of 5, 10, and 20 are plotted in black, red, and blue respectively.

Question 6. A decision tree is fitted to the data considering as output whether the basketball player was in the group performing low, mid, or high according to splitting the output value in terms of the 33.3 and 66.6 percentile as explained in the previous question. At the root of the tree it is considered to split according to Height (i.e., x_1), considering relatively short, medium, and tall players based on splitting x_1 also according to its 33.3 and 66.6 percentiles. For impurity we will use the Gini given by $I(v) = 1 - \sum_c p(c|v)^2$. Before the split, we have 18 low, 18 mid, and 18 high performing players and after the split we have

- Of the 18 short players we have that 6 have low, 9 have mid, and 3 have high performance.
- Of the 20 medium height players we have that 4 have low, 6 have mid, and 10 have high performance.
- Of the 16 tall players we have that 8 have low, 3 have mid, and 5 have high performance.

Which statement regarding the purity gain Δ of the split is correct?

A. $\Delta = 0.0505$

B. $\Delta = 0.1667$

C. $\Delta = 0.3333$

D. $\Delta = 0.6667$

E. Don't know.

Solution 6. The purity gain is given by

$$\Delta = I(r) - \sum_{k=1}^{K} \frac{N(v_k)}{N} I(v_k),$$

where

$$I(v) = 1 - \sum_c p(c|v)^2.$$

Evaluating the purity gain for the split we have:

$$\begin{aligned}\Delta =&(1 - ((\tfrac{18}{54})^2 + (\tfrac{18}{54})^2 + (\tfrac{18}{54})^2))\\ &-[\tfrac{18}{54}(1 - ((\tfrac{6}{18})^2 + (\tfrac{9}{18})^2 + (\tfrac{3}{18})^2))\\ &+\tfrac{20}{54}(1 - ((\tfrac{4}{20})^2 + (\tfrac{6}{20})^2 + (\tfrac{10}{20})^2))\\ &+\tfrac{16}{54}(1 - ((\tfrac{8}{16})^2 + (\tfrac{3}{16})^2 + (\tfrac{5}{16})^2))]\\ &= 0.0505\end{aligned}$$

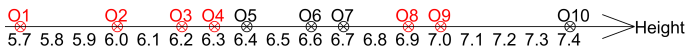

Figure 4: Considering the attribute Height for 10 observations in the Basketball data we inspect whether each of the 10 players here denoted O1,O2,...,O10 have a relatively high percentage of successful field goals (FG>45%) indicated in black and considered the positive class, i.e. observation O5,O6,O7, and O10 or a relatively low percentage of successful field goals (FG≤45%) indicated in red and considered the negative class, i.e. observations O1,O2,O3,O4,O8, and O9.

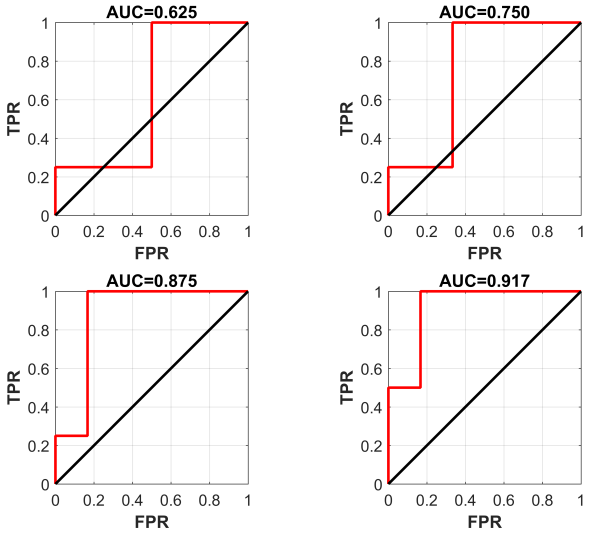

Figure 5: Four different receiver operator characteristic (ROC) curves and their corresponding area under curve (AUC) values.

Question 7. We suspect that a basketball player's Height (i.e., x_1) is predictive of whether the player is successful with his field goals FG (i.e., x_3). To quantify whether Height is predictive of FG we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature Height to discriminate between FG> 45% (positive class) or FG≤ 45% (negative class) considering the data given in Figure 4. Which one of the receiver operator characteristic (ROC) curves given in Figure 5 corresponds to the correct ROC curve?

A. The curve having AUC=0.625

B. The curve having AUC=0.750

C. The curve having AUC=0.875

D. The curve having AUC=0.917

E. Don't know.

Solution 7. There are a total of 4 positive and 6 negative observations. When lowering the threshold for predicting high performance based on the value of Height we observe that the first observation to be above the threshold is O10 which belongs to the positive class, thus TPR=1/4, FPR=0/6. Subsequently we get two observations from the negative class thus TPR=1/4, FPR=2/6 and then three positive observations being above the threshold, i.e. TPR=4/4, FPR=2/6. Lowering the threshold further we obtain the remaining negative observations such that TPR=4/4, FPR=6/6. The only curve having this property is the curve with AUC=0.750.

Question 8. We will consider the ten observations of the Basketball dataset given in Figure 4. We will cluster this data using k-means with Euclidean distance into two clusters (i.e., k=2) . Which one of the following solutions constitutes a converged solution in the k-means clustering procedure?

A. {O1}, {O2, O3, O4, O5, O6, O7, O8, O9, O10}.

B. {O1, O2, O3, O4, O5}, {O6, O7, O8, O9, O10}.

C. {O1, O2, O3, O4, O5, O6, O7}, {O8, O9, O10}.

D. {O1, O2, O3, O4, O5, O6, O7, O8, O9}, {O10}.

E. Don't know.

Solution 8. The solution {O1}, {O2, O3, O4, O5, O6, O7, O8, O9, O10} has centroids at 5.7 and (6.0+6.2+6.3+6.4+6.6+6.7+6.9+7.0+74)/9=6.5889. As such, O2 is closer to the centroid at 5.7 than 6.6111 and will thus be reassigned to this centroid hence this is not a converged solution. The solution {O1, O2, O3, O4, O5}, {O6, O7, O8, O9, O10} has centroids at (5.7+6.0+6.2+6.3+6.4)/5=6.12 and (6.6+6.7+6.9+7.0+7.4)/5=6.92. As such, O5 is closer to the centroid at 6.12 and O6 is closer to the centroid at 6.92. This will thus form a converged solution The solution {O1, O2, O3, O4, O5, O6, O7}, {O8, O9, O10} has centroids at (5.7+6.0+6.2+6.3+6.4+6.6+6.7)/7= 6.2714 and (6.9+7.0+7.4)/3=7.1. As such, O7 is closer to the centroid at 7.1 than the one at 6.2714 and will thus be reassigned to this centroid, hence, this is also not a converged solution. The solution {O1, O2, O3, O4, O5, O6, O7, O8, O9}, {O10} has centroid at 6.4222 and 7.4. As such, O9 is closer to 7.4 than 6.4222 and will thus be reassigned to this centroid, hence, this is also not a converged solution.

Question 9. We suspect that observation O10 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 Euclidean distance and the observations given in Figure 4 only. We recall that the KNN density and average relative density (ard) for the observation $\boldsymbol{x}_i$ are given by:

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

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

where $N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}_i$ excluding the i'th observation, and $\text{ard}_{\boldsymbol{X}}(\boldsymbol{x}_i, K)$ is the average relative density of $\boldsymbol{x}_i$ using K nearest neighbors (if observations are tied in terms of their distances to an observation, the observation with smallest observation number will be selected). Based on considering only the attribute Height and the ten observations in Figure 4, what is the average relative density for observation O10 for $K = 3$ nearest neighbors?

A. 0.409

B. 0.500

C. 0.533

D. 1.875

E. Don't know.

Solution 9.

$$\text{density}(\boldsymbol{x}_{O10}, 3) = (\frac{1}{3}(0.4 + 0.5 + 0.7))^{-1} = 1.8750$$

$$\text{density}(\boldsymbol{x}_{O9}, 3) = (\frac{1}{3}(0.1 + 0.3 + 0.4))^{-1} = 3.7500$$

$$\text{density}(\boldsymbol{x}_{O8}, 3) = (\frac{1}{3}(0.1 + 0.2 + 0.3))^{-1} = 5$$

$$\text{density}(\boldsymbol{x}_{O7}, 3) = (\frac{1}{3}(0.1 + 0.2 + 0.3))^{-1} = 5$$

$$\text{a.r.d.}(\boldsymbol{x}_{O10}, 3)) = \frac{\text{density}(\boldsymbol{x}_{O10}, 3)}{\frac{1}{3}(\text{density}(\boldsymbol{x}_{O9}, 3) + \text{density}(\boldsymbol{x}_{O8}, 3) + \text{density}(\boldsymbol{x}_{O7}, 3))} = \frac{1.8750}{\frac{1}{3}(3.75 + 5 + 5)} = 0.4091$$

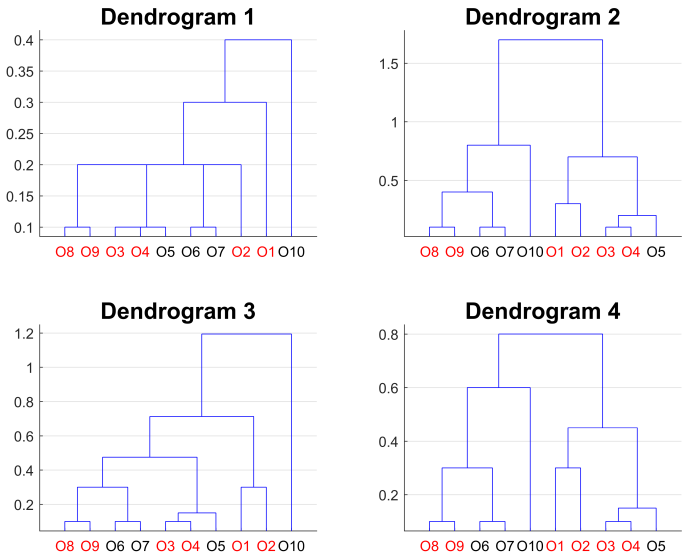

Figure 6: Four different dendrograms derived using the Euclidean distance between the 10 observations based on the attribute Height only. The value of Height for each of the 10 observations can be found in Figure 4. Red observations correspond to low values of FG and black observations to high values of FG.

Question 10. We will consider the Euclidean distance between the observations in Figure 4 based on the attribute Height only, (i.e., the Euclidean distance between observation O1 and O2 is $\sqrt{(5.7-6.0)^2} = 0.3$). A hierarchical clustering is used to cluster the observations based on their distances to each other using average linkage (when ties in the agglomerative procedure clusters containing the smallest observation numbers will merge first). Which one of the dendrograms given in Figure 6 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 10. Initially, O3, O4 will merge and O6, O7 will merge and O8, O9 will merge at the level of 0.1. Subsequently, O5 will merge onto {O3,O4} at the level of (0.1+0.2)/2=0.15. Next {O6,O7} will merge with {O8,O9} at the level of (0.3+0.3+0.2+0.3)/4=0.275. Next, O1,O2 will merge at 0.3 and subsequently {O1,O2} with {O3,O4,O5} at the level of (0.5+0.6+0.7+0.2+0.3+0.4)/6=0.45 and then O10 will merge with {O6,O7,O8,O9} at the level of (0.8+0.7+0.5+0.4)/4=0.6. The only dendrogram having these properties is dendrogram 4 and we can thus rule out the other dendrograms.

Question 11. We will cut dendrogram 2 at the level of two clusters and evaluate this clustering in terms of its correspondence with the class label information in which O1, O2, O3, O4, O8, and O9 correspond to low values of FG whereas O5, O6, O7, and O10 correspond to high values of FG. We recall that the Rand index also denoted the simple matching coefficient (SMC) between the true labels and the extracted clusters is given by $R = \frac{f_{11}+f_{00}}{K}$, where f_{11} is the number of object pairs in same class assigned to same cluster, f_{00} is the number of object pairs in different class assigned to different clusters, and $K = N(N-1)/2$ is the total number of object pairs, where N is the number of observations considered. What is the value of R between the true labeling of the observations in terms of high and low FG values and the two clusters?

A. 0.3226

B. 0.5333

C. 0.5778

D. 0.6222

E. Don't know.

Solution 11. The cluster indices are given by the vector: $[2\ 2\ 2\ 2\ 2\ 1\ 1\ 1\ 1\ 1]^\top$, whereas the true class labels are given by the vector $[1\ 1\ 1\ 1\ 2\ 2\ 2\ 1\ 1\ 2]^\top$. From this, we obtain: Total number of object pairs is:
$K = 10(10-1)/2 = 45$
$f_{00} = 4 \cdot 3 + 1 \cdot 2 = 14$
$f_{11} = 4{\cdot}(4{-}1)/2{+}1{\cdot}(1{-}1)/1{+}3{\cdot}(3{-}1)/2{+}2{\cdot}(2{-}1)/2 = 10$
$R = \frac{f_{11}+f_{00}}{K} = \frac{10+14}{45} = 24/45.$

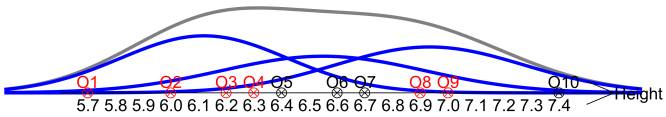

Figure 7: A Gaussian mixture model (GMM) with three clusters fitted to the 10 observations based only on the attribute Height. The overall probability density is given in gray and in blue the contribution from each of the three clusters to the density.

Question 12. We will fit a Gaussian mixture model (GMM) with three clusters to the 10 observations given in Figure 4. The fitted density is given in Figure 7 in which the overall density is given in gray and the contribution of each Gaussain given in blue. We recall that the Gaussian mixture model for 1-dimensional data is given by: $p(x) = \sum_k w_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$.
For the clustering we have:
$w_1 = 0.37, w_2 = 0.29, w_3 = 0.34,$
$\mu_1 = 6.12, \mu_2 = 6.55, \mu_3 = 6.93,$
$\sigma_1^2 = 0.09, \sigma_2^2 = 0.13, \sigma_3^2 = 0.12.$
What is the probability that observation O8 is assigned to cluster 2 according to the GMM?

A. 0.20

B. 0.29

C. 0.33

D. 0.37

E. Don't know.

Solution 12. The probability that the k'th cluster generated the observation O8 is given by $w_k N(6.9|\mu_k, \sigma_k^2)$ and we thus have:
$p(x_8 = 6.9, z_8 = 1) = 0.37 \frac{1}{\sqrt{2\pi 0.09}} \exp\left(-\frac{1}{2{\cdot}0.09}(6.9-6.12)^2\right) = 0.0168$
$p(x_8 = 6.9, z_8 = 2) = 0.29 \frac{1}{\sqrt{2\pi 0.13}} \exp\left(-\frac{1}{2{\cdot}0.13}(6.9-6.55)^2\right) = 0.2003$
$p(x_8 = 6.9, z_8 = 3) = 0.34 \frac{1}{\sqrt{2\pi 0.12}} \exp\left(-\frac{1}{2{\cdot}0.12}(6.9-6.93)^2\right) = 0.3901$
$p(z_8 = 2|x_8 = 6.9) = \frac{p(x_8=6.9, z_8=2)}{\sum_{k'} p(x_8=6.9, z_8=k')}$
$= \frac{0.2003}{0.0168+0.2003+0.3901} = 0.33.$

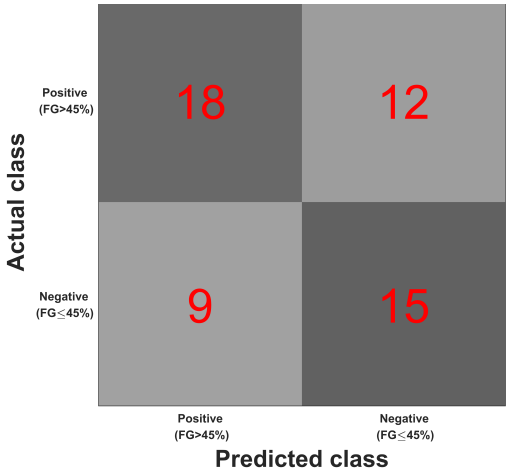

Figure 8: Confusion matrix based on a classifier's predictions of high or low success rate of field goals, (i.e., FG>45% considered the positive class or FG≤45% considered the negative class respectively).

Question 13. We will consider a simple classifier that predicts the 54 basketball players as having high success rate of field goals (FG>45%, considered the positive class) if they are taller than 6.65 foot and low otherwise ($FG \leq 45\%$, considered the negative class). The confusion matrix of the classifier is given in Figure 8. Which statement regarding the classifier is correct?

A. The recall of the classifier is 60.0 %.

B. The precision of the classifier is 61.1 %.

C. The accuracy of the classifier is 66.7 %.

D. The dataset is perfectly balanced.

E. Don't know.

Solution 13. The recall of the classifier is TP/(TP+FN)=18/(18+12)= 60.0%. The precision of the classifier is TP/(TP+FP)=18/(18+9)=66.7%. The accuracy rate of the classifier is (TP+TN)/(TP+FP+TN+FN)=(18+15)/54=61.1%. There are 30 positive examples and 24 negative examples in the test set.

Question 14. The National Basketball Association (NBA) is the top basketball league in USA and all males playing in the NBA earns more than several million dollars a year or more making them all have a very high salary. In USA we will assume approximately 0.2 % of the male population that are not playing in the NBA makes such similar very high salary. Furthermore, we will assume two out of a million American males are playing in the NBA. Assuming the above, what is the probability that a male in USA making such very high salary plays in the NBA?

A. 0.0002%

B. 0.0010%

C. 0.0999%

D. 0.2002%

E. Don't know.

Solution 14. What we are interested in is $P(\text{NBA}|\text{Very high salary}) = \frac{P(\text{Very high salary}|\text{NBA})P(NBA)}{P(\text{Very high salary}|\text{NBA})P(NBA)+P(\text{Very high salary}|\text{not NBA})P(\text{not NBA})} = \frac{\frac{1\cdot 2/1000000}{1\cdot 2/1000000+0.002\cdot(1-2/1000000)}}{} = \frac{2}{2+0.002*999998} = 0.0999\%$.

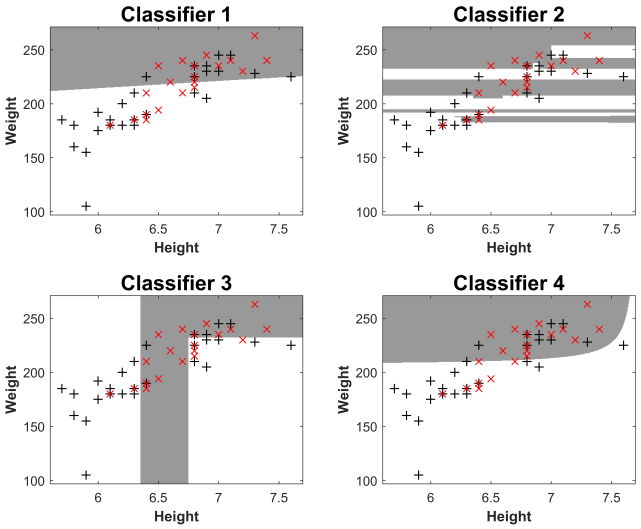

Figure 9: Decision boundaries for four different classifiers trained on the Basketball dataset considering the features Height and Weight. Gray regions classify into red crosses whereas white regions into black plusses.

Question 15. Four different classifiers are trained on the Basketball dataset considering the features Height and Weight in order to predict if the percentage of successful field goals is high (FG$>$ 45%) or low (FG$\leq$ 45%). The decision boundary for each of the four classifiers is given in Figure 9. Which one of the following statements is correct?

A. Classifier 1 corresponds to logistic regression considering as input x_1 and x_2, Classifier 2 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 3 is a decision tree including three decisions, Classifier 4 correponds to a logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$.

B. Classifier 1 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 2 is a decision tree including three decisions, Classifier 3 corresponds to logistic regression considering as input x_1 and x_2, Classifier 4 corresponds to a logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$.

C. Classifier 1 corresponds to a logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$, Classifier 2 is a 3-nearest neighbor classifier using Euclidean distance, Classifier 3 is a decision tree including three decisions, Classifier 4 corresponds to logistic regression considering as input x_1 and x_2,

D. Classifier 1 is a decision tree including three decisions, Classifier 2 corresponds to logistic regression considering as input x_1, x_2, and $x_1 \cdot x_2$, Classifier 3 corresponds to a logistic regression considering as input x_1 and x_2, Classifier 4 is a 3-nearest neighbor classifier using Euclidean distance.

E. Don't know.

Solution 15. The decision boundary of classifier 1 is a straight line thus conforms to a logistic regression model using only the features x_1 and x_2 as inputs. Classifier 2 is a 3-nearest neighbor classifier and when using Euclidean distance the scale of Weight has much more variance than that of height thereby heavily influencing the distance measure. Classifier 3 is a decision tree with two vertical and one horizontal line corresponding to three decisions. Classifier 4 is non-linear and smooth corresponding to a logistic regression including a transformed variable, i.e. $x_1 \cdot x_2$.

Question 16. We will consider an artificial neural network (ANN) trained to predict the average score of a player (i.e., y). The ANN is based on the model:

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

where $h^{(1)}(x) = \max(x, 0)$ is the rectified linear function used as activation function in the hidden layer (i.e., positive values are returned and negative values are set to zero). We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix},$$

and $w_0^{(2)} = 2.84$, $w_1^{(2)} = 3.25$, and $w_2^{(2)} = 3.46$.
What is the predicted average score of a basketball player with observation vector $\boldsymbol{x}^* = [6.8\ 225\ 0.44\ 0.68]$?

A. 1.00

B. 3.74

C. 8.21

D. 11.54

E. Don't know.

Solution 16. The output is given by:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{w}) &= 2.84 \\ &+ 3.25 \cdot \max([1\ 6.8\ 225\ 0.44\ 0.68] \cdot \begin{bmatrix} 21.78 \\ -1.65 \\ 0 \\ -13.26 \\ -8.46 \end{bmatrix}, 0) \\ &+ 3.46 \max([1\ 6.8\ 225\ 0.44\ 0.68] \cdot \begin{bmatrix} -9.60 \\ -0.44 \\ 0.01 \\ 14.54 \\ 9.50 \end{bmatrix}, 0) \\ &= 2.84 + 3.25 \cdot \max(-1.027, 0) + 3.46 \max(2.516, 0) \\ &= 11.54 \end{aligned}$$

Question 17. Which statement regarding cross-validation is correct?

A. An advantage of five-fold cross-validation over three-fold cross-validation is that the datasets used for training are larger.

B. The more data used for training a model the more we can expect the model to overfit to the training data.

C. 10-fold cross-validation is more accurate but also more computationally expensive than leave-one-out cross-validation.

D. When upsampling data in order to avoid class imbalance issues the same observations should be included in the training and test set such that the training and test sets reflect the same properties.

E. Don't know.

Solution 17. When using k-fold cross-validation 1/k of the data is used for testing and (k-1)/k of the data for training during each fold. As such, five-fold cross-validation uses larger training sets than three-fold cross-validation. The more data used for training a model the less we can expect overfitting to occur as the model will be less prone to fit to specific aspects of the training set. 10-fold cross-validation should not be more accurate and in particular, it is not more expensive than leave-one-out cross-validation. Leave-one-out cross-validation is more expensive as we have to use as many folds as we have observations and each model is trained on a larger training set size. When upsampling the data it is important that the same observations do not occur in both the training and test set as we otherwise are training the model on parts of the test set and thereby fitting the model also to test data.

	H_L	H_H	W_L	W_H	$FG_{\leq 45\%}$	$FG_{>45\%}$	$FT_{<75\%}$	$FT_{>75\%}$
O1	1	0	1	0	1	0	1	0
O2	1	0	1	0	1	0	1	0
O3	1	0	1	0	1	0	1	0
O4	1	0	1	0	1	0	0	1
O5	1	0	1	0	0	1	0	1
O6	1	0	0	1	0	1	1	0
O7	0	1	1	0	0	1	0	1
O8	0	1	1	0	1	0	0	1
O9	0	1	0	1	1	0	1	0
O10	0	1	0	1	0	1	1	0

Table 3: The ten considered observations of the Basketball dataset binarized considering the attribute x_1–x_4. The attributes x_1 and x_2 are binarized according to whether they are below or above the median value of the attribute for the entire dataset of 54 observations. x_3 and x_4 are respectively threshold at a success rate of 45% for field goals (FG) and a success rate of 75% for free throw (FT). The ten observations are color coded in terms of average points scored per game (y) being in the low range {O2, O3, O6, O9} mid-range {O1, O4} and high range {O5, O7, O8, O10}.

Question 18. Considering the dataset in Table 3 as a market basket problem with observation O1–O10 corresponding to customers and H_L, H_H, W_L, W_H, $FG_{\leq 45\%}$, $FG_{>45\%}$, $FT_{\leq 75\%}$, and $FT_{>75\%}$ corresponding to items. What are all frequent itemsets with support greater than 35%?

A. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, and $\{FT_{>75\%}\}$.

B. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, and $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$.

C. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$, and $\{H_L,W_L,FG_{\leq 45\%}\}$.

D. $\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$, $\{H_L,W_L,FG_{\leq 45\%}\}$, and $\{H_L,W_L,FT_{\leq 75\%}\}$.

E. Don't know.

Solution 18. For a set to have support more than 35% the set must occur at least $0.35 \cdot 10 = 3.5$, i.e. 4 out of the 10 times. All the itemsets that have this property are:
$\{H_L\}$, $\{H_H\}$, $\{W_L\}$, $\{FG_{\leq 45\%}\}$, $\{FG_{>45\%}\}$, $\{FT_{\leq 75\%}\}$, $\{FT_{>75\%}\}$, $\{H_L,W_L\}$, $\{H_L,FG_{\leq 45\%}\}$, $\{H_L,FT_{\leq 75\%}\}$, $\{W_L,FG_{\leq 45\%}\}$, $\{W_L,FT_{>75\%}\}$, $\{FG_{\leq 45\%},FT_{\leq 75\%}\}$, and $\{H_L,W_L,FG_{\leq 45\%}\}$.

Question 19. We consider again the data in Table 3 as a market basket problem. What is the confidence of the association rule $\{\mathrm{H}_L, \mathrm{W}_L\} \rightarrow \{\mathrm{FG}_{\leq 45\%}, \mathrm{FT}_{\leq 75\%}\}$?

A. 30 %

B. 40 %

C. 50 %

D. 60 %

E. Don't know.

Solution 19. The confidence is given as

$$P(\mathrm{FG}_{\leq 45\%}, \mathrm{FT}_{\leq 75\%} | \mathrm{H}_L, \mathrm{W}_L) = \frac{\mathrm{FG}_{\leq 45\%}, \mathrm{FT}_{\leq 75\%}, \mathrm{H}_L, \mathrm{W}_L)}{P(\mathrm{H}_L, \mathrm{W}_L)} = \frac{3/10}{5/10} = 3/5 = 60\%$$

Question 20. We would like to predict whether a basketball player has a high average score using the data in Table 3. We will apply a Naïve Bayes classifier that assumes independence between the attributes given the class label (i.e., the class label is given by the average points scored per game being low (black color), in the mid-range (red color) or high (blue color) respectively in the table). Given that a basketball player is relatively tall ($\mathrm{H}_H = 1$) relatively light weight ($\mathrm{W}_L = 1$) what is the probability that the basketball player will have a high average score according to the Naïve Bayes classifier derived from the data in Table 3?

A. 9/16

B. 9/11

C. 3/4

D. 1

E. Don't know.

Solution 20. Let HAS denote high average score. According to the Naïve Bayes classifier we have

$$P(HAS|\mathrm{H}_H = 1, \mathrm{W}_L = 1) = \frac{\begin{pmatrix} P(\mathrm{H}_H = 1||HAS)\times \\ P(\mathrm{W}_L = 1|HAS)\times \\ P(HAS) \end{pmatrix}}{\begin{pmatrix} P(\mathrm{H}_H = 1||LAS)\times \\ P(\mathrm{W}_L = 1|LAS)\times \\ P(LAS) + P(\mathrm{H}_H = 1|MAS)\times \\ P(\mathrm{W}_L = 1|MAS)\times \\ P(MAS) + P(\mathrm{H}_H = 1||HAS)\times \\ P(\mathrm{W}_L = 1|HAS)\times \\ P(HAS) \end{pmatrix}} = \frac{3/4 \cdot 3/4 \cdot 4/10}{1/4 \cdot 2/4 \cdot 4/10 + 0 \cdot 2/2 \cdot 2/10 + 3/4 \cdot 3/4 \cdot 4/10} = \frac{9/40}{2/40 + 0 + 9/40} = 9/11.$$

Question 21. Considering the data in Table 3, we will use a 3-nearest neighbor classifier to classify observation O10 (i.e., with binary observation vector [0 1 0 1 0 1 1 0]) based on observation O1–O9. We will classify according to the three neighboring observations with *highest* similarity according to the Jaccard (J) measure of similarity given by $J(\boldsymbol{a}, \boldsymbol{b}) = \frac{f_{11}}{M - f_{00}}$, where f_{11} and f_{00} are the number of one matches and zero matches respectively and M the total number of binary features. Which one of the following statements is correct?

A. O10 will be classified as black.

B. O10 will be classified as blue.

C. O10 will be classified as red.

D. The classifier will be tied between the classes black and blue.

E. Don't know.

Solution 21. For $O10$ we have:

$$J(O10, O1) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O2) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O3) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O4) = \frac{0}{8-0} = 0$$
$$J(O10, O5) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O6) = \frac{3}{8-3} = \frac{3}{5}$$
$$J(O10, O7) = \frac{3}{8-3} = \frac{3}{5}$$
$$J(O10, O8) = \frac{1}{8-1} = \frac{1}{7}$$
$$J(O10, O9) = \frac{3}{8-3} = \frac{3}{5}$$

Hence the three nearest neighbors are O6,O7, and O9 with two black and one blue observation thus according to majority voting the observation will be classified as black.

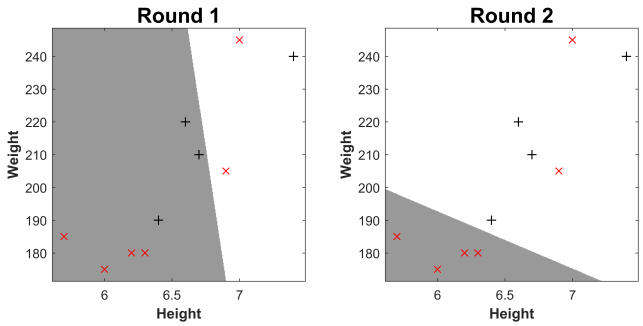

Figure 10: Decision boundaries for two rounds of boosting considering a logistic regression model using the features Height and Weight and the 10 observations also considered previously in Figure 4. Gray region indicates that the observation will be classified as red crosses, white regions that the observation will be classified as black plusses.

Question 22. We will consider classifying the 10 observations considered in Figure 4 using logistic regression and boosting by Adaboost (notice, the Adaboost algorithm uses the natural logarithm). For this purpose we include only two boosting rounds considering only the features Height (i.e., x_1) and Weight (i.e., x_2) as inputs. In the first round the data is sampled with equal probability $w_i = 1/10$ for $i = \{1, \ldots, 10\}$ and the logistic regression model with decision boundary given to the left of Figure 10 trained. A new dataset is subsequently sampled and a new logistic regression classifier given to the right of the Figure 10 trained. Based on these two rounds of the Adaboost algorithm what will an observation located at $x_1 = 6$ and $x_2 = 240$ be classified as?

A. The two classes will be tied for the Adaboost procedure.

B. The observation will be classified as red cross.

C. The observation will be classified as black plus.

D. The weights $w_1, \ldots, w_{10}$ are changed in round 1 of the Adaboost procedure.

E. Don't know.

Solution 22. We have for the first round that the weighted error rate $\epsilon_1 = 5/10$ with associated $\alpha_1 = \frac{1}{2}\log\frac{1-\epsilon_1}{\epsilon_1} = \frac{1}{2}\log 1 = 0$. The updated weights will thus be unchanged as $e^0 = 1$. In the next round

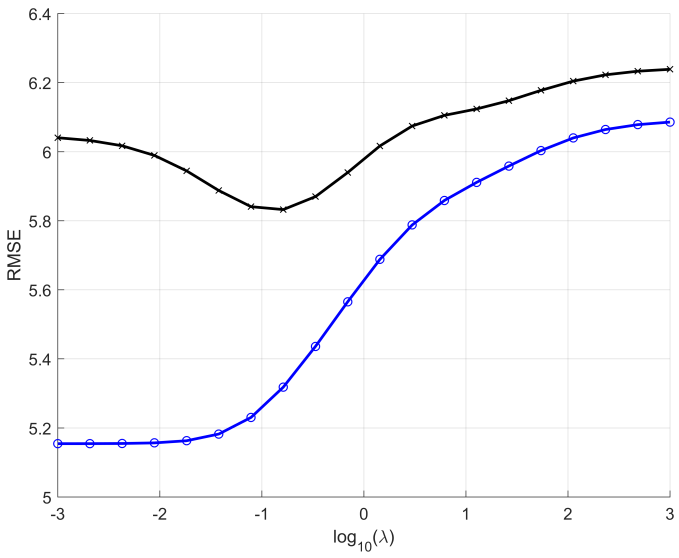

Figure 11: Root-mean square error (RMSE) curves as function of the regularization strenght λ for regularized least square regression predicting the average points scored per game y based on the attributes x_1–x_4.

$\epsilon_2 = 2/10$ and thus $\alpha_2 = \frac{1}{2}\log\frac{1-2/10}{2/10} = \frac{1}{2}\log 4 = 0.693$. When determining the class we weight each classifier by the importance α_t of the round t. However, as the first round has $\alpha_1 = 0$ this round has zero weight in the voting and thus the classifier will solely be based on the second round classifier for which $\alpha_2 = 0.693$. This classifier will deem the observation at $x_1 = 6$ and $x_2 = 240$ to belong to the class of black plusses as the decision region is white.

Question 23. Using the 54 observations of the Basketball dataset we would like to predict the average points scored per game (y) based on the four features (x_1–x_4). For this purpose we consider regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function:
$E(\boldsymbol{w}) = \sum_n (y_n - [1\ x_{n1}\ x_{n2}\ x_{n3}\ x_{n4}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$,
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 consider 20 different values of λ and use leave-one-out cross-validation to quantify the performance of each of these different values of λ. The results of the leave-one-out cross-validation performance is given in Figure 11. Inspecting the model for the value of $\lambda = 0.6952$ the following model is identified:
$f(\boldsymbol{x}) = 2.76 - 0.37x_1 + 0.01x_2 + 7.67x_3 + 7.67x_4$.
Which one of the following statements is correct?

A. In Figure 11 the blue curve with circles corresponds to the training error whereas the black curve with crosses corresponds to the test error.

B. According to the model defined for $\lambda = 0.6952$ increasing a players height will increase his average points scored per game.

C. There is no optimal way of choosing λ since increasing λ reduces the variance but increases the bias.

D. As we increase λ the 2-norm of the weight vector $\boldsymbol{w}$ will also increase.

E. Don't know.

Solution 23. The blue curve monotonically increases with λ reflecting a worse fit to the training set as we increase λ using regularization we can reduce the variance by introducing bias and the black curve indicates that an optimal tradeoff at around $10^{-0.8}$ as reflected by the test error indicated in the black curve being minimal. As we increase λ we will penalize the weights according to the squared 2-norm more and more and thus the 2-norm will be reduced. Finally, according to the fitted model we observe that the coefficient in front of x_1 (Height) is negative thus indicating that an increase in height will reduce the models prediction of average points scored per game.

Question 24. We will again consider the ridge regression described in the previous question. Which one of the following statements is correct?

A. Exhaustively evaluating all combinations of features would require the fitting of less models than the proposed ridge-regression procedure.

B. To generate the test curve we need to make predictions from a total of 1060 different models.

C. We can obtain an unbiased estimate of the generalization error of the best performing model from Figure 11.

D. The ridge regression model will be non-linear as the model includes regularization.

E. Don't know.

Solution 24. Exhaustively evaluating all feature combinations of four features would require evaluating $2^4 = 16$ models whereas we currently consider 20 different models. For each of the 20 values of λ we have to estimate 54 models according to the leave-one-out procedure, (i.e., 54 times we leave out an observation as the dataset contains 54 observations.) thus we need a total of $20 \cdot 54 = 1080$ different models for making the necessary predictions. In order to get an unbiased estimate of generalization of the best model we would need two-layer cross-validation (we currently only have one-level cross-validation). The ridge regression model will not be non-linear but still linear in the input data regardless of the regularization.

Question 25. Consider the clustering problem given in Figure 12. Which clustering approach is *most* suited for correctly separating the data into the four groups indicated by black crosses, red circles, magenta plusses, and blue asterics?

A. A well-separated clustering approach.

B. A contiguity-based clustering approach.

C. A center-based clustering approach.

D. A conceptual clustering approach.

E. Don't know.

Solution 25. As the observation in each cluster is at least closest to one other observation in its cluster than to an observation in another cluster a contiguity based approach is most suited.

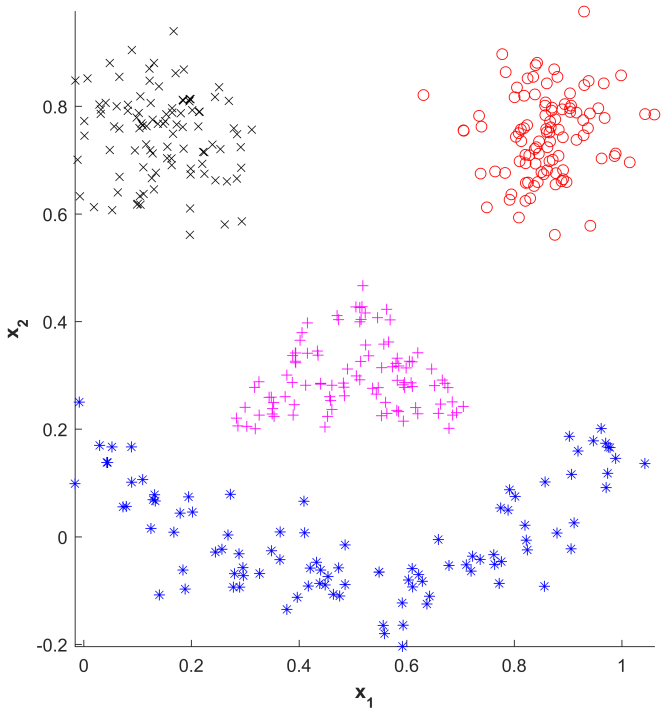

Figure 12: A clustering problem containing four clusters indicated by black crosses, red circles, magenta plusses and blue stars.

Question 26. Which one of the following statements is correct?

A. Multinomial regression can only handle classification problems where the problem is to classify between two classes.

B. Decision trees return the probability that an observation is in a given class.

C. k-means, Gaussian Mixture Models (GMM) and Artificial Neural Networks (ANN) are all prone to local minima issues and thus it is recommended to run the procedures using multiple initializations.

D. The accuracy is a good performance measure when facing severe class imbalance issues in a two class classification problem.

E. Don't know.

Solution 26. Multinomial regression is a generalization of two class logistic regression to handle multiple

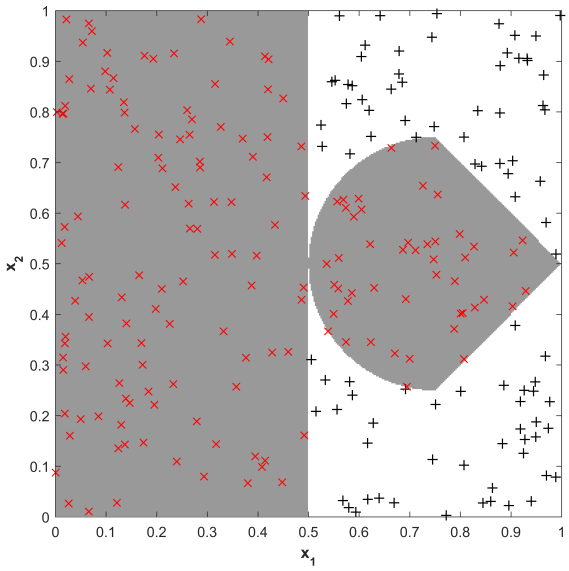

Figure 13: A two class classification problem with red crosses (i.e., x) and black plusses (i.e., +) constituting the two classes as well as the associated decision boundaries of the two classes indicated respectively by gray and white regions.

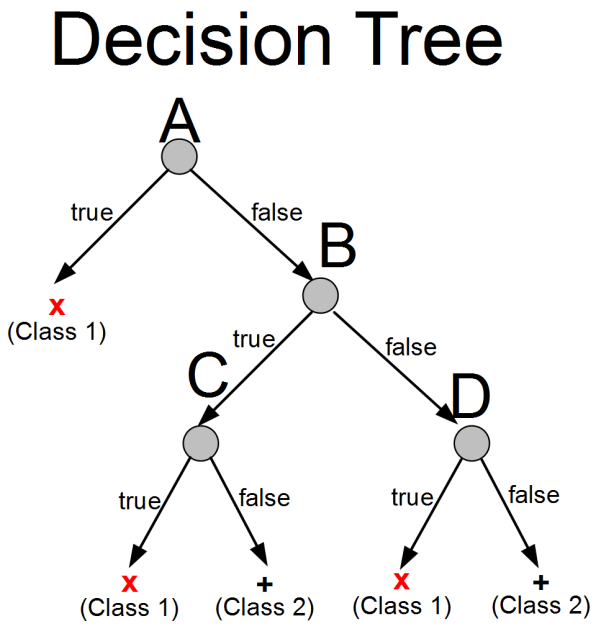

Figure 14: A decision tree with four decisions (A,B,C, and D) forming the decision boundaries given in Figure 13 if adequately defined.

Question 27. We will consider the two class classification problem given in Figure 13 in which the goal is to separate red crosses (i.e., x) from black plusses (i.e., +) based on the decision boundaries in gray and white indicated in the top panel of the figure. Which one of the following procedures based on the decision tree given in Figure 14 will perfectly separate the two classes?

A. **A=**$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$**,**
B=$x_1 < 0.75$**,**
C= $\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_2 < 0.25$**,**
D=$\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_1 < 0.25$**.**

B. A=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_1 < 0.5$,
B=$x_1 < 0.75$,
C= $\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_2 < 0.25$,
D=$\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_\infty < 0.25$.

C. A=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$,
B=$x_1 < 0.75$,
C= $\|\boldsymbol{x} - \begin{bmatrix} 0.5 \\ 0.75 \end{bmatrix}\|_2 < 0.25$,
D=$\|\boldsymbol{x} - \begin{bmatrix} 0.5 \\ 0.75 \end{bmatrix}\|_1 < 0.25$.

D. A=$x_1 < 0.75$,
B=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$,

classes. Decision trees do not return probabilities of being in each class but hard assigns observations to the classes based on majority voting in each terminal leaf. K-means, Gaussian Mixture Models and Artificial Neural Networks (ANN) are indeed all prone to local minima and it is therefore advised to use multiple restarts selecting the initialization with best solution. Accuracy is not a good performance measure when facing severe class-imbalance issues as we may trivially obtain a very high accuracy simply by classifying by chance. The AUC of the receiver operator characteristic would here be more appropriate as it is not influenced by the relative sizes of the two classes.

Solution 27. All observations for which $x_1 < 0.5$ are red crosses which can be captured by the initial decision A=$\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}\|_\infty < 0.5$. For the remaining observations it appears two different norms are at play depending on whether $x_1 < 0.75$ or not, thus B=$x_1 < 0.75$. If $x_1 < 0.75$ we observe that decision C should have a circular shape defined by C= $\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_2 < 0.25$ whereas if $x_1 \geq 0.75$ we have a diamond shape defined by D=$\|\boldsymbol{x} - \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}\|_1 < 0.25$. The other solutions will not similarly correctly define the decision boundaries.