Technical University of Denmark

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

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	Number of cylinders	cyl
x_2	Horsepower	hp
x_3	Weight	wt
x_4	Transmission (0=automatic, 1=manual)	am
x_5	Number of forward gears	gear
y	Miles pr. gallon	mpg

Table 1: The attributes of the Motor Trend Car Road Tests dataset taken from https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv. The output y is given by the miles pr. gallon the car drives. The dataset has 32 observations and we presently consider the five input features x_1–x_5.

Question 1. We will consider the data of Motor Trend Car Road Tests based on 32 automobiles (observations) taken from https://vincentarelbundock.github.io/Rdatasets/csv/datasets/mtcars.csv for brevity denoted the Cars dataset in the following. The original data contains eleven attributes, however, we presently consider only five of these attributes given in Table 1 as well as the output attribute y given by how many miles the car drives pr. gallon of fuel.

Considering the attributes described in Table 1 which one of the following statements is *correct*?

A. The attribute x_5 is continuous.

B. The output variable y is ratio.

C. The attribute x_4 is ordinal.

D. The attribute x_1 is nominal.

E. Don't know.

Solution 1. As x_5 is given by the number of forward gears this attribute is discrete (i.e. defined by the integer numbers) and not continuous. As y has a meaningful zero value defining absence of miles pr. gallon y is ratio whereas we can talk about a car driving twice as far pr. gallon of fuel than another etc. Transmission is defined as either automatic or manual and thus categorical and not ordinal, (i.e., its is not meaningful to say that manual is better than automatic). x_1 is discrete ratio and thus not a nominal variable.

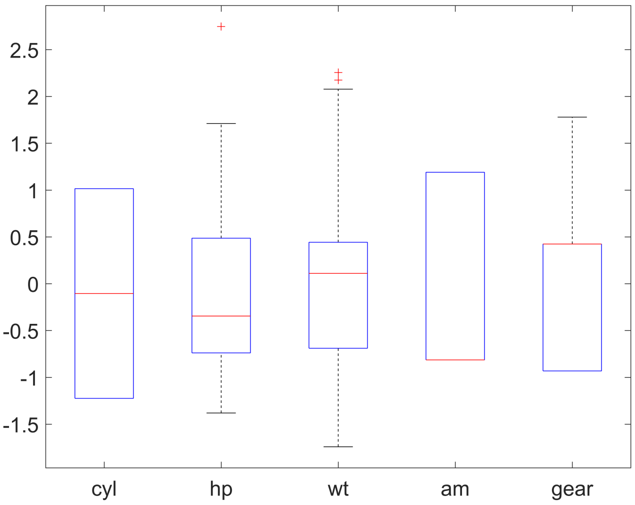

Figure 1: Boxplot of the five attributes x_1–x_5 after standardizing the data (i.e., subtracting the mean of each attribute and dividing the attribute by its standard deviation).

Question 2. In Figure 1 is given a boxplot of the five attributes x_1–x_5 after standardizing the data, i.e. subtracting the mean of each attribute and dividing each attribute by its standard deviation. Which one of the following statements is *correct*?

A. The majority of cars have automatic transmission.

B. The attribute x_5 (i.e., number of forward gears (gear)) appears to be normal distributed.

C. The attribute x_2 (i.e., horse power (hp)) has a clear outlier that should be removed.

D. From the boxplot it is clear that some of the attributes are highly correlated with each other.

E. Don't know.

Solution 2. As there are 32 cars and the median is placed at the lowest value the majority, i.e. at least 17 of cars indeed have to have automatic transmission. The attribute x_5 (gear) does not have a symmetric distribution and indeed seems far from normally distributed. In particular, its 50th and 75th percentile coincide. Even though the box plot indicates an outlier these should not be removed without very strong justification which is not provided here. Boxplots investigate each attribute separately and do not reveal any aspects in regards to the relationship between attributes, i.e. if they are correlated.

Question 3. A principal component analysis (PCA) is carried out on the standardized attributes x_1–x_5, forming the standardized matrix $\tilde{\boldsymbol{X}}$, resulting in the following $\boldsymbol{S}$ and $\boldsymbol{V}$ matrices obtained from a singular value decomposition of $\tilde{\boldsymbol{X}}$:

$$\boldsymbol{S} = \begin{bmatrix} 10.2 & 0 & 0 & 0 & 0 \\ 0 & 6.1 & 0 & 0 & 0 \\ 0 & 0 & 2.8 & 0 & 0 \\ 0 & 0 & 0 & 2.2 & 0 \\ 0 & 0 & 0 & 0 & 1.6 \end{bmatrix},$$
$$\boldsymbol{V} = \begin{bmatrix} 0.49 & -0.31 & 0.42 & -0.14 & 0.69 \\ 0.39 & -0.62 & 0.05 & -0.24 & -0.63 \\ 0.51 & -0.06 & -0.55 & 0.66 & 0.08 \\ -0.44 & -0.46 & 0.42 & 0.65 & -0.02 \\ -0.40 & -0.55 & -0.59 & -0.27 & 0.35 \end{bmatrix}.$$

The data projected onto the first two principal components are given in Figure 2. Which one of the following statements is *correct*?

A. The first principal component accounts for less than 70 % of the variance.

B. The two first principal components account for less than 90 % of the variance.

C. The fifth principal component accounts for less than 1% of the variance.

D. As can be observed in Figure 2 there is a positive correlation between the projection of the data to the first and second principal component.

E. Don't know.

Solution 3. The variance explained by the th principal component is given by $\frac{\sigma_i^2}{\sum_{i'} \sigma_{i'}^2}$. As such we find:

$$VarExpPC1 = \tfrac{10.2^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.6648$$
$$VarExpPC2 = \tfrac{6.1^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.2378$$
$$VarExpPC3 = \tfrac{2.8^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.0501$$
$$VarExpPC4 = \tfrac{2.2^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.0309$$
$$VarExpPC5 = \tfrac{1.6^2}{10.2^2+6.1^2+2.8^2+2.2^2+1.6^2} = 0.0164$$

As such, the first PC accounts for less than 70% of the variance, the first two principal components accounts for $0.6648 + 0.2378 = 0.9026$ which is not less than 90% of the variance. The fifth principal

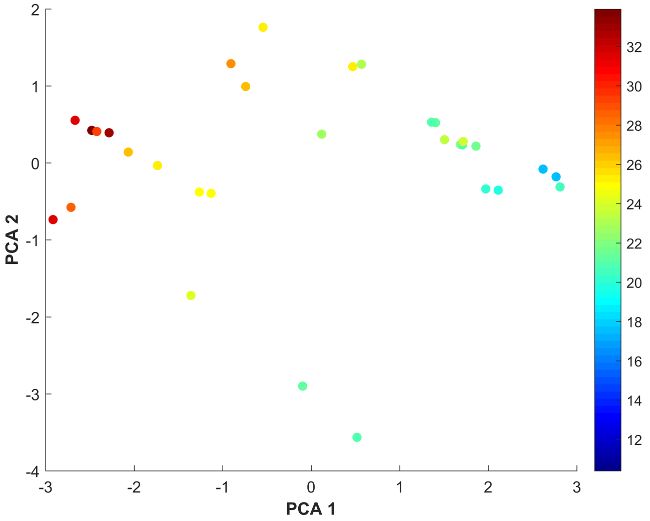

Figure 2: Data projected onto the first and second principal component. Each observation is color coded according to y, i.e. how many miles pr. gallon the car drives.

component accounts for 0.0164% which is not less than 1%. The data represented in the space of the PCA are uncorrelated, i.e. there is no correlation between the data projected onto the first and second principal components as $(\tilde{\boldsymbol{X}}\boldsymbol{v}_1)^\top(\tilde{\boldsymbol{X}}\boldsymbol{v}_2) = (\boldsymbol{U}_1S_{11})^\top(\boldsymbol{U}_2S_{22}) = S_{11}\boldsymbol{U}_1^\top\boldsymbol{U}_2S_{22} = S_{11}0S_{22} = 0$.

Question 4. The data projected onto the two first principal components (as defined in Question 3) is given in Figure 2 where the output variable y is indicated by the color of each observation. which one of the following statements pertaining to the PCA is *correct*?

A. Cars with a relatively small number of cylinders, low horsepower, low weight, that have manual transmission, and many forward gears tend to drive longer pr. gallon of fuel.

B. The second principal component appears to provide a better description of how far a car drives pr. gallon of fuel than principal component direction one.

C. From the PCA plot it appears to be very difficult to predict fuel consumption based on the attributes x_1–x_5.

D. Cars with a relatively small number of cylinders, low horsepower, that are automatic, and with few forward gears will have a large negative projection onto the second principal component.

E. Don't know.

Solution 4. As the first three values of $\boldsymbol{v}_1$ are positive and the last two negative relatively small values of the first three attributes, i.e. cylinders, horsepower, and weight and relatively high value of the two last attribute, i.e. manual transmission and many forward gears will result in a negative projection onto the first principal component where most cars driving far pr. gallon of fuel are positioned, thus, this is correct. The second principal component direction does not appear to well characterize how far car drives pr. gallon when compared to the first principal component direction. From the plot it indeed seems feasible to predict fuel consumption. Finally, by inspecting $\boldsymbol{v}_2$ cars with a relatively small number of cylinders, low horsepower, that are automatic and with few forward gears will have a relatively large positive projection onto the second prinicipal component and not the reverse.

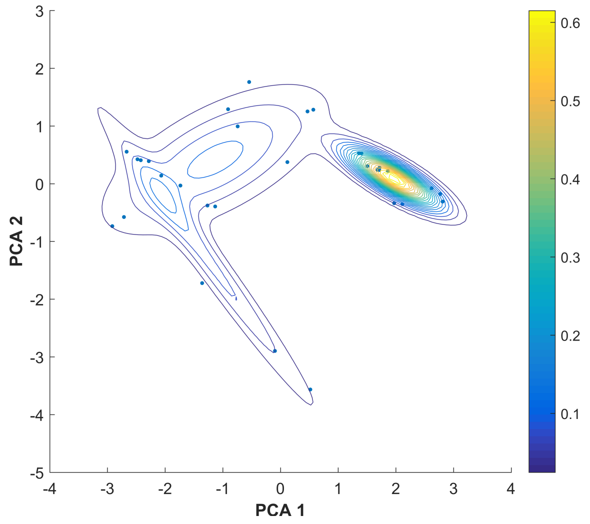

Figure 3: A Gaussian Mixture Model (GMM) with three clusters fitted to the Cars dataset projected onto the first two principal components.

Question 5. Consider the Gaussian Mixture Model (GMM) fitted to the Cars dataset projected onto the first two principal components for which the contours of the fitted GMM is given in Figure 3. We will in the following use: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$ to denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Which one of the following GMM densities corresponds to the fitted GMM?

A.

$$\begin{aligned} p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) \\ + 0.176 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) \\ + 0.375 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) \end{aligned}$$

B.

$$\begin{aligned} p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) \\ + 0.176 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) \\ + 0.375 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) \end{aligned}$$

C.

$$\begin{aligned} p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) \\ + 0.176 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) \\ + 0.375 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) \end{aligned}$$

D.

$$\begin{aligned} p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) \\ + 0.176 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) \\ + 0.375 \cdot \mathcal{N}(\boldsymbol{x}| &\begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) \end{aligned}$$

E. Don't know.

Solution 5. Inspecting the contour plot we see that the cluster located at $\begin{bmatrix} -1.1287 \\ 0.4168 \end{bmatrix}$ has positive covariance, the cluster located at $\begin{bmatrix} -1.2978 \\ -1.2612 \end{bmatrix}$ has a negative covariance with high variance and the cluster located at $\begin{bmatrix} 1.9574 \\ 0.0913 \end{bmatrix}$ a negative covariance with lower variance as the other cluster with negative covariance - in particular this cluster has more spread in the PCA_1 direction than the PCA_2 direction whereas the other negative covariance cluster at $\begin{bmatrix} -1.2978 \\ -1.2612 \end{bmatrix}$ has more spread in the PCA_2 direction than the PCA_1 . This property only holds for the following answer option:

$$\begin{aligned} p(\boldsymbol{x}) = 0.449 \cdot \mathcal{N}(\boldsymbol{x} | \begin{bmatrix} -1.129 \\ 0.417 \end{bmatrix}, \begin{bmatrix} 1.307 & 0.557 \\ 0.557 & 0.575 \end{bmatrix}) \\ + 0.176 \cdot \mathcal{N}(\boldsymbol{x} | \begin{bmatrix} -1.298 \\ -1.261 \end{bmatrix}, \begin{bmatrix} 1.458 & -1.982 \\ -1.982 & 2.771 \end{bmatrix}) \\ + 0.375 \cdot \mathcal{N}(\boldsymbol{x} | \begin{bmatrix} 1.957 \\ 0.091 \end{bmatrix}, \begin{bmatrix} 0.248 & -0.128 \\ -0.128 & 0.101 \end{bmatrix}) \end{aligned}$$

Question 6. A least squares linear regression model is trained using different combinations of the five attributes x_1, x_2, x_3, x_4, and x_5. Table 2 gives 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 five attributes. Which one of the following statements is *correct*?

A. Forward selection will select the same optimal feature set as backward selection.

B. For this problem backward selection will identify the optimal feature combination.

C. Forward selection will end up selecting all the features, i.e., terminate at the feature set x_1, x_2, x_3, x_4, x_5.

D. Forward selection will as first feature select x_3.

E. Don't know.

Solution 6. Forward selection will first select x_1 with performance 3.0343, then improve most by including x_3 with 2.7692, subsequently include x_2 with performance 2.4869. Including additional features to the set x_1, x_2, x_3 provides no improvements on the test set and the forward selection will thus terminate. Backward selection will first remove x_5 with performance 2.5095, subsequently x_1 with performance 2.3423 thereby identifying the correct optimal feature set being x_2, x_3, x_4.

Feature(s)	Training	Test
	RMSE	RMSE
x_1	3.2522	3.0343
x_2	3.1721	5.5072
x_3	2.907	3.4486
x_4	4.4608	5.8157
x_5	4.4637	9.0155
x_1 and x_2	3.043	3.8668
x_1 and x_3	2.4192	2.7692
x_1 and x_4	2.8918	3.3297
x_1 and x_5	3.2465	3.3177
x_2 and x_3	2.5325	2.5853
x_2 and x_4	2.8066	3.2509
x_2 and x_5	3.1646	4.6723
x_3 and x_4	2.8601	3.7105
x_3 and x_5	2.8230	4.3690
x_4 and x_5	3.9742	8.4346
x_1 and x_2 and x_3	2.4098	2.4869
x_1 and x_2 and x_4	2.7253	2.6258
x_1 and x_2 and x_5	3.0341	4.6252
x_1 and x_3 and x_4	2.3834	2.9486
x_1 and x_3 and x_5	2.3709	2.9676
x_1 and x_4 and x_5	2.7956	3.4894
x_2 and x_3 and x_4	2.4703	2.3423
x_2 and x_3 and x_5	2.5319	2.7017
x_2 and x_4 and x_5	2.7847	4.2531
x_3 and x_4 and x_5	2.8028	4.4864
x_1 and x_2 and x_3 and x_4	2.3679	2.5095
x_1 and x_2 and x_3 and x_5	2.3623	3.0234
x_1 and x_2 and x_4 and x_5	2.6249	4.3059
x_1 and x_3 and x_4 and x_5	2.2937	3.0251
x_2 and x_3 and x_4 and x_5	2.4561	2.8221
x_1 and x_2 and x_3 and x_4 and x_5	2.2759	3.1368

Table 2: Root-mean-square error (RMSE) for the training and test set when using least squares regression to predict fuel consumption of a car, i.e., mpg, using different combinations of the five attributes (x_1–x_5).

Question 7. Using the 32 observations of the Cars dataset we would like to predict the fuel consumption of cars (y) based on the five features ($x_1 - x_5$). For this purpose we consider regularized least squares regression which minimizes with respect to $\boldsymbol{w}$ the following cost function:

$$E(\boldsymbol{w}) = \sum_n (y_n - [1\ x_{n1}\ x_{n2}\ x_{n3}\ x_{n4}\ x_{n5}]\boldsymbol{w})^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w},$$

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 nine 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 4 where the optimal value of lambda is found to be $\lambda = 10^{-1.75}$ indicated with a black cross in the figure. Which one of the following statements is correct?

A. The identified test performance having RMSE=2.8 is an unbiased estimator of how the optimal model identified will generalize to new data.

B. To create the test error curve given in Figure 4 requires training 288 regularized least squares regression models.

C. Leave-one-out cross-validation is computationally more efficient than 10-fold cross-validation.

D. As λ increases the fitted models will have smaller and smaller bias.

E. Don't know.

Solution 7. The identified test performance is not an unbiased estimator of the optimal identified models generalization. To quantify the generalization performance of the optimally selected model would require two-layer cross-validation. In order to generate the test curve we need for each value of λ (9 values) to fit a model corresponding to each of the 32 observations left out which would require fitting $9 \cdot 32 = 288$ models, thus, this statement is correct. Leave-one-out cross-validation is not more computationally efficient as we have to fit more models than in 10-fold cross-validation and each of these models would also include more data for training. As we increase λ we also increase the models bias, i.e. eventually the model will predict everything as 0.

Question 8. We will build a model to determine if a car has a relatively high or low fuel consumption using the original data (i.e., the data is no longer standardized). For this purpose we will split the output y in two classes forming the new output variable z defined by thresholding at the median value of y, i.e. if $y_i > \text{median}(y)$ then $z_i = 1$, otherwise $z_i = 0$. We fit a logistic regression model using the features x_1–x_5 and use as output for the logistic regression model the binary variable z indicating if a car has relatively high ($z = 0$) or low ($z = 1$) fuel consumption. The predicted output probability is given by:

$$\begin{aligned}\hat{z} = \sigma(&1257.6 - 46.8x_1 + 0.6x_2 \\ &- 271.1x_3 - 31.9x_4 - 44.7x_5),\end{aligned}$$

where $\sigma(\cdot)$ is the logistic sigmoid function. Which one of the following statements regarding the logistic regression model is correct?

A. x_2 is the least important attribute for defining whether a car has high (i.e., $z = 0$) or low (i.e., $z = 1$) fuel consumption.

B. According to the model increasing a car's number of forward gears will make it more likely to have high fuel consumption (i.e., $z = 0$).

C. A new car with the following observation vector: $\boldsymbol{x}^* = [6\ 120\ 3.2\ 0\ 4]$ will be more likely to have high (i.e., $z = 0$) than low (i.e., $z = 1$) fuel consumption.

D. Logistic regression is not very suited for classification as it is a regression method.

E. Don't know.

Solution 8. We cannot interpret importance with respect to the amplitude of the coefficients as the scale of each attribute is very different and the attributes correlated. As x_5 has a negative coefficient increasing the number of forward gears will decrease the probability of low (i.e., $z = 1$) fuel consumption and thereby increase the probability of high (i.e., $z = 0$) fuel consumption. Thus, this statement is correct. We have for the probability according to the logistic regression model of the observation $\boldsymbol{x}^* = [6\ 120\ 3.2\ 0\ 4]$:

$$\begin{aligned}\hat{z}^* &= \frac{1}{1+\exp(-(1257.6-46.8\cdot 6+0.6\cdot 120-271.1\cdot 3.2-31.9\cdot 0-44.7\cdot 4))} \\ &= 0.9227.\end{aligned}$$

Thus, this car is likely to have low (i.e, $z = 1$) fuel consumption. Logistic regression is a classification procedure and designed for this purpose.

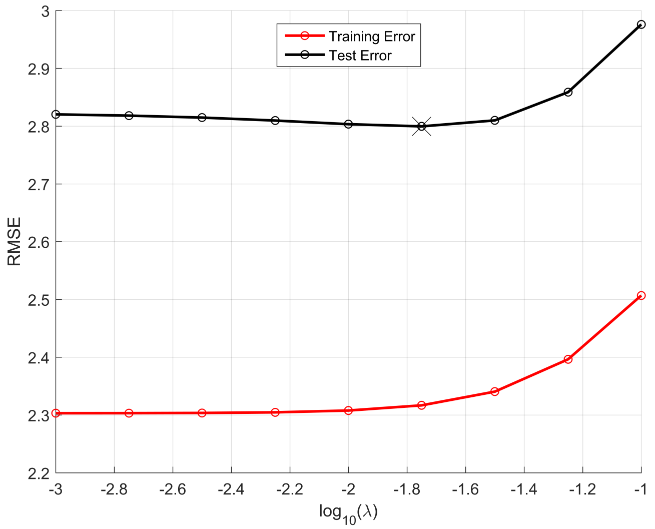

Figure 4: Regularized least squares regression applied to the Cars data in order to predict fuel consumption. Given is training and test performance as quantified by the root mean square error (RMSE) in red and black respectively as a function of the value of λ based on leave-one-out cross-validation.

Question 9. In order to improve the performance of the logistic regression model we will use ensembling based on the Adaboost algorithm using the 32 observations of the Cars dataset (note that for the Adaboost algorithm $\log(\cdot)$ is based on the natural logarithm). The first trained logistic regression classifier (i.e., for boosting round $t = 1$) has an error rate of $1/16$. What will be the updated weight for each of the correctly classified observations?

A. 0.0081

B. 0.0167

C. 0.0332

D. 0.2500

E. Don't know.

Solution 9. In the first round (t=1) all samples are weighted equally thus $w_1 = w_2 = \ldots = w_{32} = 1/32$. As a result $\epsilon_1 = \sum_{i=1}^{N} w_i(1 - \delta_{f_t(x_i),y_i}) = \frac{1}{32}\sum_{i=1}^{N}(1 - \delta_{f_t(x_i),y_i}))$ which is the error rate, i.e. number of misclassified observations divided by the

	3 gears	4 gears	5 gears
	($x_5 = 3$)	($x_5 = 4$)	($x_5 = 5$)
Low mpg ($z = 0$)	13	2	2
High mpg ($z = 1$)	2	10	3

Table 3: Number of low mpg and high mpg cars (i.e. $z = 0$ and $z = 1$) according to the number of gears, i.e. $x_5 = 3$, $x_5 = 4$, or $x_5 = 5$.

total number of observations (i.e., divided by 32) which is 1/16 as explained in the text. We then have $\alpha_1 = \frac{1}{2}\log(\frac{1-\epsilon_1}{\epsilon_1}) = 0.5\log(15)$. Thus $\tilde{w}_m = \frac{1}{32}exp(0.5\log(15))$ for a mis-classified observation and $\tilde{w}_c = \frac{1}{32}exp(-0.5\log(15))$ for a correctly classified observation. As two observations are misclassified and 30 correctly classified (i.e. the error rate is 1/16) we have $w_c = \frac{\frac{1}{32}exp(-0.5\log(15))}{2\cdot\frac{1}{32}exp(0.5\log(15))+30\cdot\frac{1}{32}exp(-0.5\log(15))} = 0.0167$

Question 10. A decision tree is fitted to the data considering as output whether the car has a relatively high (i.e., $z = 0$) or low (i.e., $z = 1$) fuel consumption. At the root of the tree three different splits according to the number of forward gears are considered based on the data given in Table 3. For impurity we will use the classification error given by $I(v) = 1 - \max_c p(c|v)$. For the three considered splits we have:

- Split A: 3 gear vs. 4 or 5 gears.
- Split B: 3 or 4 gears vs. 5 gears.
- Split C: 3 gears vs. 4 gears vs. 5 gears.

Thus, split A and B have two branches whereas split C has three branches. Which statement about the splits is correct?

A. Split B provides a higher purity gain than split A.

B. Split C provides a higher purity gain than split A.

C. The best obtainable purity gain is 9/32.

D. Split B provides a higher purity gain than split C.

E. Don't know.

Solution 10. The purity gain is given by

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

where

$$I(v) = 1 - \max_c p(c|v).$$

Evaluating the purity gain for split A we have:

$$\begin{aligned}\Delta =&(1-(\tfrac{17}{32}))\\ &-[\tfrac{15}{32}(1-(\tfrac{13}{15}))\\ &+\tfrac{17}{32}(1-(\tfrac{13}{17}))]\\ &=\frac{15}{32}-\frac{6}{32}=9/32.\end{aligned}$$

Evaluating the purity gain for split B we have:

$$\begin{aligned}\Delta =&(1-(\tfrac{17}{32}))\\ &-[\tfrac{27}{32}(1-(\tfrac{15}{27}))\\ &+\tfrac{5}{32}(1-(\tfrac{3}{5}))]\\ &=\frac{15}{32}-\frac{14}{32}=1/32.\end{aligned}$$

Evaluating the purity gain for split C we have:

$$\begin{aligned}\Delta =&(1-(\tfrac{17}{32}))\\ &-[\tfrac{15}{32}(1-(\tfrac{13}{15}))\\ &+\tfrac{12}{32}(1-(\tfrac{10}{12}))\\ &+\tfrac{5}{32}(1-(\tfrac{3}{5}))]\\ &=\frac{15}{32}-\frac{6}{32}=9/32.\end{aligned}$$

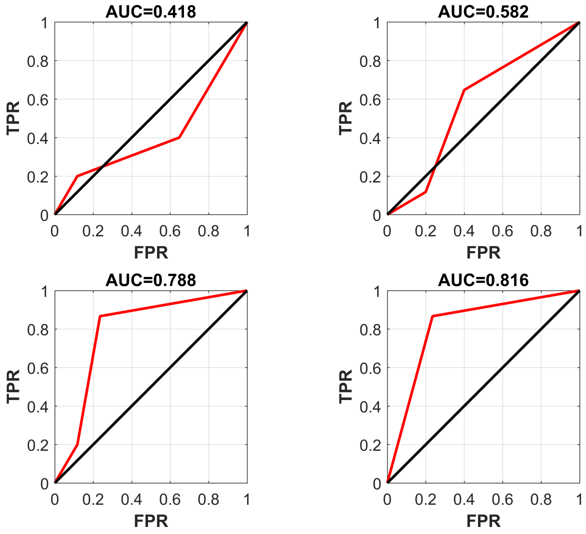

Figure 5: Four different receiver operator characteristic (ROC) curves and their area under curve (AUC) value.

Question 11. We will evaluate the feature x_5 (gear) in its ability to discriminate low mpg, i.e., $z = 0$, (considered the negative class) from high mpg, i.e. $z = 1$, (considered the positive class) based on the data given in Table 3. For this purpose, we will evaluate the area under curve (AUC) of the receiver operator characteristic (ROC) using the feature x_5 to discriminate cars with high (i.e., $z = 0$) from cars with low (i.e., $z = 1$) fuel consumption. Which one of the ROC curves given in Figure 5 corresponds to using x_5 to discriminate between high fuel consumption ($z = 0$) and low fuel consumption ($z = 1$)?

A. The curve having AUC=0.418

B. The curve having AUC=0.582

C. The curve having AUC=0.788

D. The curve having AUC=0.816

E. Don't know.

Solution 11. The ROC curve can be calculated by lowering the threshold, as no cars have more than 5 forward gears a threshold above 5 will result in the point (0,0). Lowering the threshold we find at the value 5 that 2/17 of the low mpg cars (FPR) are at 5 and 3/15 of the high mpg cars (TPR) are at 5 corresponding to the point (2/17,3/15). When lowering to a threshold of 4 gears or more we are at the point (4/17,13/15) and at a threshold at 3 gears or more we have (1,1). Thus, this curve corresponds to the curve having AUC=0.788.

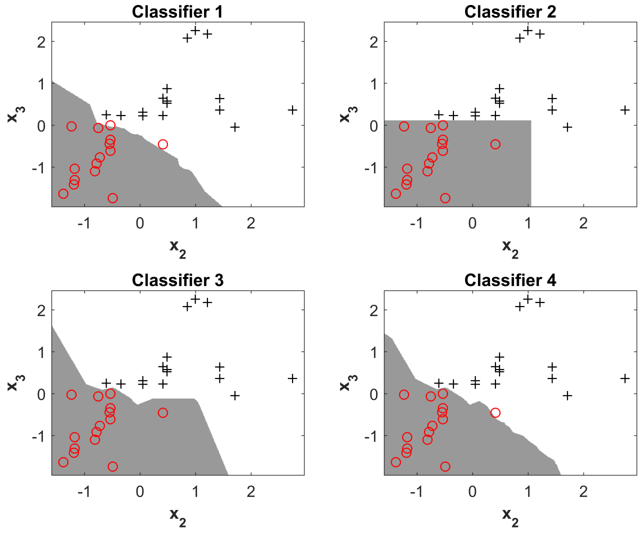

Figure 6: Decision boundaries for four different classifiers trained on the Cars dataset using only the two features x_2 and x_3.

Question 12. Four different classifiers are trained on the Cars dataset using only x_2 and x_3 as features (the features have been standardized) and the decision boundary for each of the four classifiers is given in Figure 6. Which one of the following statements is *correct*?

A. Classifier 1 is an artificial neural network with one hidden unit in the hidden layer.

B. Classifier 2 is a 3-nearest neighbor classifier.

C. Classifier 3 is a 1-nearest neighbor classifier.

D. Classifier 4 is a logistic regression classifier.

E. Don't know.

Solution 12. The decision boundary of classifier 1 cannot be a neural network with one hidden unit as this would correspond to linear decision boundary similar to logistic regression. Classifier 2 has straight vertical and horizontal lines resembling a decision tree and not a 3-nearest neighbor classifier. Classifier 3 is indeed a 1-nearest neighbor classifier as can be seen by the decision boundary following the most close-by observation. Classifier 4 cannot be a logistic regression classifier as this would require a decision boundary formed by a straight line (as for a neural network with one hidden unit).

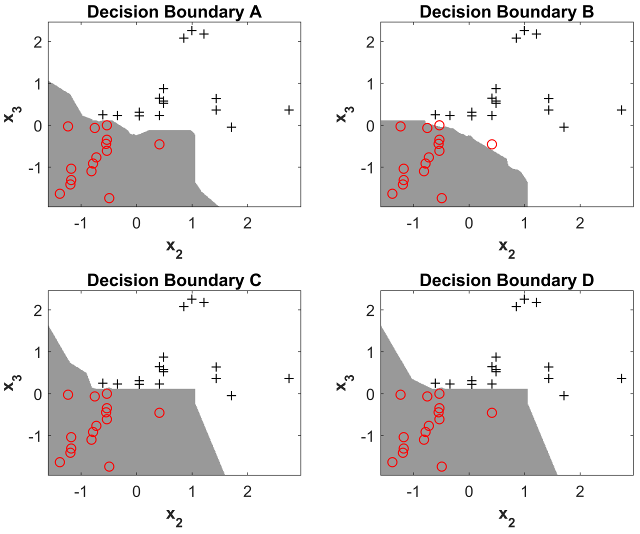

Figure 7: Decision boundaries for which one of the four decision boundaries corresponds to combining Classifier 1, Classifier 2, and Classifier 3 in Figure 6 using majority voting.

Question 13. In an attempt to make a stronger classifier the first three classifiers (i.e., Classifier 1, Classifier 2, and Classifier 3) are combined using majority voting. Which one of the decision boundaries given in Figure 7 corresponds to the combined classifier?

A. Decision boundary A.

B. Decision boundary B.

C. Decision boundary C.

D. Decision boundary D.

E. Don't know.

Solution 13. When combining the three classifiers the majority class is the class receiving two or more votes. Combining the three decision boundaries of the three first classifiers in Figure 6 to form a white region thus requires two of the classifiers being white in that region and vice versa for the gray region. This only holds for decision boundary A.

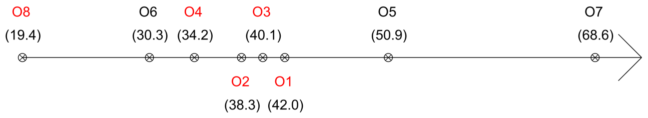

Figure 8: The eight first observations of the Cars dataset considered in regards to the feature $q = x_2/x_3$ (the value of q is given in parenthesis).

Question 14. We will consider the first eight observations of the Cars dataset and the new feature defined by the ratio of horse power to weight, i.e. defined by $q = x_2/x_3$. In Figure 8 is shown the value of q for the first eight observations that are colored according to low (red) and high (black) fuel consumption. We will cluster this data using k-means with Euclidean distance into two clusters (i.e., k=2) and initialize the k-means algorithm with centroids located at observation O8, and O6. Which one of the following statements is *correct*?

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

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

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

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

E. Don't know.

Solution 14. With the described initialization, observation O8 will be assigned to the cluster located at O8, and the remaining observation will be assigned to the cluster located at O6, i.e. {O1, O2, O3, O4, O5, O6, O7}. Thus, only cluster located at O6 will change location and the location updated to $\frac{30.3+34.2+38.3+40.1+42.0+50.9+68.6}{7} = 43.5$. For this new location O6 is closer to cluster located at O8 than the cluster located at 43.5, resulting in the updated clustering {O6,O8}, {O1, O2, O3, O4, O5, O7}. Thus, the first cluster will change location to $\frac{19.4+30.3}{2} = 24.85$ and the second cluster to $\frac{34.2+38.3+40.1+42.0+50.9+68.6}{6} = 45.68$. Subsequently, the first cluster will be updated to contain {O4,O6,O8} with centroid at $\frac{19.4+30.3+34.2}{3} = 27.97$ and the second cluster with {O1, O2, O3, O5, O7} will have centroid located at $\frac{38.3+40.1+42.0+50.9+68.6}{5} = 47.98$ which will form a converged solution as no observation change assignment.

Question 15. We will consider a clustering given by the two clusters {O2, O3, O4, O6, O8},{O1, O5, O7}. We will evaluate this clustering in terms of its correspondence with the class label information in which O1,O2,O3,O4, and O8 correspond to cars with low fuel consumption and O5, O6, and O7 correspond to cars with high fuel consumption. We recall that the Jaccard coefficient between the true labels and the extracted clusters is given by:

$$J = \frac{f_{11}}{K - f_{00}},$$

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 above value of J between the true labeling of the observations in terms of high and low fuel consumption and the two clusters?

A. 0.1665

B. 0.3684

C. 0.5714

D. 0.7500

E. Don't know.

Solution 15. The cluster indices are given by the vector: $[2\ 1\ 1\ 1\ 2\ 1\ 2\ 1]^\top$, whereas the true class labels are given by the vector $[1\ 1\ 1\ 1\ 2\ 2\ 2\ 1]^\top$. From this, we obtain: Total number of object pairs is: $K = 8(8-1)/2 = 28$
$f_{00} = 4 \cdot 2 + 1 \cdot 1 = 9$
$f_{11} = 4 \cdot (4-1)/2 + 1 \cdot (1-1)/1 + 1 \cdot (1-1)/2 + 2 \cdot (2-1)/2 = 7$
$J = \frac{f_{11}}{K - f_{00}} = \frac{7}{28-9} = 0.3684$

Question 16. According to the Cars dataset we have that 59.38% of the cars have automatic transmission. Furthermore, 63.16% of the cars that have automatic transmission have eight cylinders, whereas 15.38% of the cars that have manual transmission have eight cylinders. According to the Cars dataset what is the probability that a car that has eight cylinders, i.e. $x_1 = 8$ will have automatic transmission, i.e. $x_4 = 0$?

A. 16.7 %

B. 37.5 %

C. 63.2%

D. 85.7 %

E. Don't know.

Solution 16. According to Bayes' theorem we have:

$$\begin{aligned} P(x_4 = 0|x_1 = 8) &= \frac{P(x_1=8|x_4=0)P(x_4=0)}{P(x_1=8)} \\ &= \frac{P(x_1=8|x_4=0)P(x_4=0)}{P(x_1=8|x_4=0)P(x_4=0)+P(x_1=8|x_4=1)P(x_4=1)} \\ &= \frac{0.6316 \cdot 0.5938}{0.6316 \cdot 0.5938 + 0.1538 \cdot (1-0.5938)} = 85.7\% \end{aligned}$$

Question 17. We will consider an artificial neural network (ANN) trained to predict mpg (i.e., y) of a car. The ANN is based on the model:

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

where $h^{(1)}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ is the hyperbolic tangent function used as activation function in the hidden layer. We will consider an ANN with two hidden units in the hidden layer defined by:

$$\boldsymbol{w}_1^{(1)} = \begin{bmatrix} -4 \\ 1 \\ 0.01 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \boldsymbol{w}_2^{(1)} = \begin{bmatrix} -10 \\ 1 \\ -0.02 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

and $w_0^{(2)} = 7$, $w_1^{(2)} = 8$, and $w_2^{(2)} = 9$.
What is the predicted fuel consumption of a car with observation vector $\boldsymbol{x}^* = [6\ 120\ 3.2\ 0\ 4]$?

A. 17.00

B. 20.85

C. 24.00

D. 33.40

E. Don't know.

Solution 17. The output is given by:

$$8 \cdot tanh(-4 + 1 \cdot 6 + 0.01 \cdot 120 + 1 \cdot 3.2 - 1 \cdot 0 - 1 \cdot 4)$$
$$+9 \cdot tanh(-10 + 1 \cdot 6 - 0.02 \cdot 120 + 1 \cdot 3.2 + 1 \cdot 0 + 1 \cdot 4)$$
$$+7 = 20.85$$

	hp$_L$	hp$_H$	wt$_L$	wt$_H$	am=0	am=1
O1	1	0	1	0	0	1
O2	1	0	1	0	0	1
O3	1	0	1	0	0	1
O4	1	0	1	0	1	0
O5	0	1	0	1	1	0
O6	1	0	0	1	1	0
O7	0	1	0	1	1	0
O8	1	0	1	0	1	0

Table 4: The first eight observations of the Cars dataset binarized considering the attribute x_2, x_3, and x_4 such that x_2 is split according to the median value in terms of low and high horse power (i.e. hp$_L$ and hp$_H$), low and high weight (i.e. wt$_L$ and wt$_H$), as well as whether the car has automatic (i.e., am=0) or manual (i.e., am=1) transmission. The eight observations are color coded in terms of low fuel consumption {O1, O2, O3, O4, O8} and high fuel consumption {O5, O6, O7}.

Question 18. Considering the dataset in Table 4 as a market basket problem with observation O1–O8 corresponding to customers and hp$_L$, hp$_H$,wt$_L$, wt$_H$, am=0, am=1 corresponding to items, what is the confidence of the association rule {wt$_H$,am=0}→{hp$_H$}?

A. 0.0%

B. 25.0

C. 66.7%

D. 100.0 %

E. Don't know.

Solution 18. The confidence is given as

$$P(\text{hp}_H = 1 | \text{wt}_H = 1, am = 0) = \frac{P(\text{hp}_H = 1, \text{wt}_H = 1, am = 0)}{P(\text{wt}_H = 1, am = 0)}$$
$$= \frac{2/8}{3/8} = 2/3 = 66.7\%$$

Question 19. We will again consider the data in Table 4. What are all frequent itemsets with support greater than 30%?

A. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}.

B. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}.

C. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}, {hp$_L$, wt$_L$, am=1},

D. {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}, {hp$_L$, wt$_L$, am=1}, {hp$_L$, wt$_L$, am=0},

E. Don't know.

Solution 19. For a set to have support more than 30% the set must occur at least $0.3 \cdot 8 = 2.4$, i.e. 3 out of the 8 times. All the itemsets that have this property are {hp$_L$}, {wt$_L$}, {wt$_H$}, {am=0}, {am=1}, {hp$_L$, wt$_L$}, {hp$_L$, am=0}, {hp$_L$, am=1}, {wt$_L$, am=1}, {wt$_H$, am=0}, {hp$_L$, wt$_L$, am=1}.

Question 20. Considering the data in Table 4, we will calculate the similarity as well as distance between O1 given by the vector $\boldsymbol{a} = [1\ 0\ 1\ 0\ 0\ 1]$ and O4 given by the vector $\boldsymbol{b} = [1\ 0\ 1\ 0\ 1\ 0]$ using respectively the Jaccard (J), simple matching coefficient (SMC), cosine similarity (cos) and p-norm ($\|\cdot\|_p$) given by

$$\begin{aligned} J(\boldsymbol{a},\boldsymbol{b}) &= \frac{f_{11}}{M - f_{00}}, \\ SMC(\boldsymbol{a},\boldsymbol{b}) &= \frac{f_{11} + f_{00}}{M}, \\ cos(\boldsymbol{a},\boldsymbol{b}) &= \frac{f_{11}}{\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2}, \\ \|\boldsymbol{a} - \boldsymbol{b}\|_p &= (\sum_{m=1}^{M} |a_m - b_m|^p)^{1/p}. \end{aligned}$$

Which one of the following statements is correct?

A. $\|\boldsymbol{a} - \boldsymbol{b}\|_2 = 2$.

B. $\|\boldsymbol{a} - \boldsymbol{b}\|_1 < \|\boldsymbol{a} - \boldsymbol{b}\|_2$.

C. $J(\boldsymbol{a},\boldsymbol{b}) = 2/3$

D. $cos(\boldsymbol{a},\boldsymbol{b}) = SMC(\boldsymbol{a},\boldsymbol{b})$**.**

E. Don't know.

Solution 20. For $\boldsymbol{a}$ and ,$\boldsymbol{b}$ we have:

$$\begin{aligned} \|\boldsymbol{a} - \boldsymbol{b}\|_2 &= \sqrt{0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2} = \sqrt{2}, \\ \|\boldsymbol{a} - \boldsymbol{b}\|_1 &= 0 + 0 + 0 + 0 + 1 + 1 = 2, \\ J(\boldsymbol{a},\boldsymbol{b}) &= \frac{f_{11}}{M - f_{00}} = 2/(6-2) = 1/2, \\ SMC(\boldsymbol{a},\boldsymbol{b}) &= \frac{f_{11} + f_{00}}{M} = 4/6 = 2/3, \\ cos(\boldsymbol{a},\boldsymbol{b}) &= \frac{f_{11}}{\|\boldsymbol{r}\|_2 \|\boldsymbol{s}\|_2} = 2/(\sqrt{3}\sqrt{3}) = 2/3. \end{aligned}$$

Hence, $cos(\boldsymbol{a},\boldsymbol{b}) = SMC(\boldsymbol{a},\boldsymbol{b})$ is correct.

	O1	O2	O3	O4	O5	O6	O7	O8
O1	0	0.2606	1.1873	2.4946	2.9510	2.5682	3.4535	2.4698
O2	0.2606	0	1.2796	2.4442	2.8878	2.4932	3.3895	2.4216
O3	1.1873	1.2796	0	2.8294	3.6892	2.9147	4.1733	2.2386
O4	2.4946	2.4442	2.8294	0	1.4852	0.2608	2.2941	1.8926
O5	2.9510	2.8878	3.6892	1.4852	0	1.5155	1.0296	3.1040
O6	2.5682	2.4932	2.9147	0.2608	1.5155	0	2.3316	1.8870
O7	3.4535	3.3895	4.1733	2.2941	1.0296	2.3316	0	3.7588
O8	2.4698	2.4216	2.2386	1.8926	3.1040	1.8870	3.7588	0

Table 5: Pairwise Euclidean distance between the first eight observations in the Cars dataset. Red observations (i.e., O1, O2, O3, O4, O8) are observations corresponding to low fuel consumption, whereas black observations (i.e., O5, O6, O7) are observations with high fuel consumption.

Question 21. We would like to predict whether a car has high fuel or low fuel consumption using the data in Table 4. We will apply a Naïve Bayes classifier that assumes independence between the attributes given the class label (i.e. high or low fuel consumption corresponding to observation indicated in black and red respectively in the table). Given that a car has $\text{hp}_L = 1$, and am=0 what is the probability that the car will have high fuel consumption according to the Naïve Bayes classifier derived from the data in Table 4?

A. 3/80

B. 1/8

C. 1/3

D. 3/5

E. Don't know.

Solution 21. We will let high fuel consumption be denoted by $z = 0$ and low fuel consumption by $z = 1$. According to the Naïve Bayes classifier we have

$$P(z=0|\text{hp}_L=1, am=0) = \frac{\left(\begin{array}{r} P(\text{hp}_L=1|z=0)\times \\ P(am=0|z=0)\times \\ P(z=0) \end{array}\right)}{\left(\begin{array}{r} P(\text{hp}_L=1|z=0)\times \\ P(am=0|z=0)\times \\ P(z=0) \end{array}\right) + \left(\begin{array}{r} P(\text{hp}_L=1|z=1)\times \\ P(am=0|z=1)\times \\ P(z=1) \end{array}\right)}$$

$$= \frac{1/3 \cdot 3/3 \cdot 3/8}{1/3 \cdot 3/3 \cdot 3/8 + 5/5 \cdot 2/5 \cdot 5/8} = \frac{1/8}{1/8 + 2/8} = 1/3.$$

Question 22. To determine whether the fuel consumption of a car is high or low we will use a k-nearest neighbor (KNN) classifier to predict each of the eight observations based on the Euclidean distance between the observations given in Table 5. We will use leave-one-out cross-validation for the KNN in order to classify the eight considered observations using a one-nearest neighbor classifier, i.e. $K = 1$. The analysis will be based only on the data given in Table 5. Which one of the following statements is *correct*?

A. None of the observations will be misclassified.

B. One of the observations will be misclassified.

C. Two the observations will be misclassified.

D. Three of the observations will be misclassified.

E. Don't know.

Solution 22. $N(O1, 1) = \{O2\}$ as O2 is closest it will be correctly classified as having low fuel consumption.
$N(O2, 1) = \{O1\}$ as O1 is closest it will be correctly classified as having low fuel consumption.
$N(O3, 1) = \{O1\}$ as O1 is closest it will be correctly classified as having low fuel consumption.
$N(O4, 1) = \{O6\}$ as O6 is closest it will be incorrectly classified as having high fuel consumption.
$N(O5, 1) = \{O7\}$ as O7 is closest it will be correctly classified as having high fuel consumption.
$N(O6, 1) = \{O4\}$ as O4 is closest it will be incorrectly classified as having low fuel consumption.
$N(O7, 1) = \{O5\}$ as O5 is closest it will be correctly classified as having high fuel consumption.
$N(O8, 1) = \{O6\}$ as O6 is closest it will be incorrectly classified as having high fuel consumption.
Thus, three out of the eight observations will be misclassified.

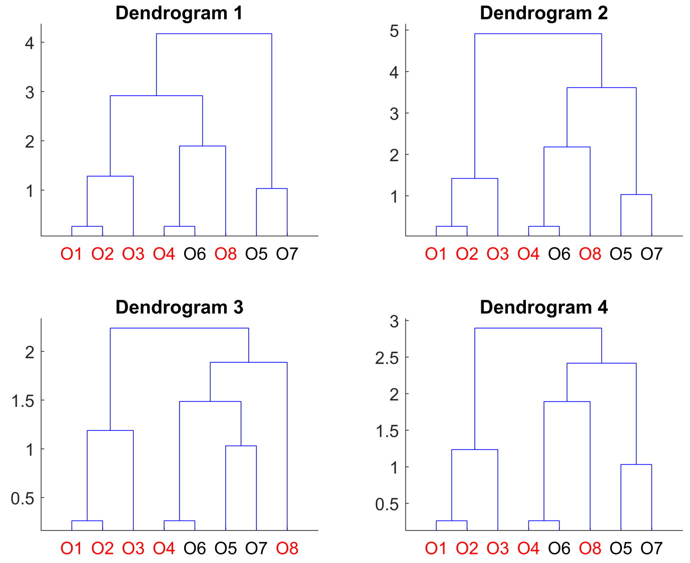

Figure 9: Four different dendrograms derived from the distances between the eight first cars given in Table 5.

Question 23. In Table 5 is given the pairwise Euclidean distances between the first eight observations of the Cars data. A hierarchical clustering is used to cluster these observations using average linkage. Which one of the dendrograms given in Figure 9 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 23. Using average linkage clusters are merged according to their average distance between the observations from each cluster. The dendrogram grows by first merging O1 and O2 at 0.2606, then O4 and O6 at 0.2608, then O5 and O7 at 1.0296, then {O1,O2} with O3 at (1.1873+1.2796)/2=1.2334, then {O4,O6} with O8 at (1.8926+1.8870)/2= 1.8898. Subsequently {O5, O7} merge with {O4,O6,O8} at (1.4852+2.2941+1.5155+ 2.3316+ 3.1040+ 3.7588)/6=2.4149, only dendrogram 4 has this property.

Question 24. We suspect that observation O8 may be an outlier. In order to assess if this is the case we would like to calculate the average relative KNN density based on the observations given in Table 5 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. Based on the data in Table 5, what is the average relative density for observation O8 for $K = 2$ nearest neighbors?

A. 0.4660

B. 0.4800

C. 0.5292

D. 1.8898

E. Don't know.

Solution 24.

$$\text{density}(\boldsymbol{x}_{O8}, 2) = (\frac{1}{2}(1.8870 + 1.8926))^{-1} = 0.5292$$

$$\text{density}(\boldsymbol{x}_{O6}, 2) = (\frac{1}{2}(0.2608 + 1.5155))^{-1} = 1.1259$$

$$\text{density}(\boldsymbol{x}_{O4}, 2) = (\frac{1}{2}(1.4852 + 0.2608))^{-1} = 1.1455$$

$$\text{a.r.d.}(\boldsymbol{x}_{O8}, 2) = \frac{\text{density}(\boldsymbol{x}_{O8}, 2)}{\frac{1}{2}(\text{density}(\boldsymbol{x}_{O6}, 2) + \text{density}(\boldsymbol{x}_{O4}, 2))}$$

$$= \frac{0.5292}{\frac{1}{2}(1.1259 + 1.1455)} = 0.4660$$

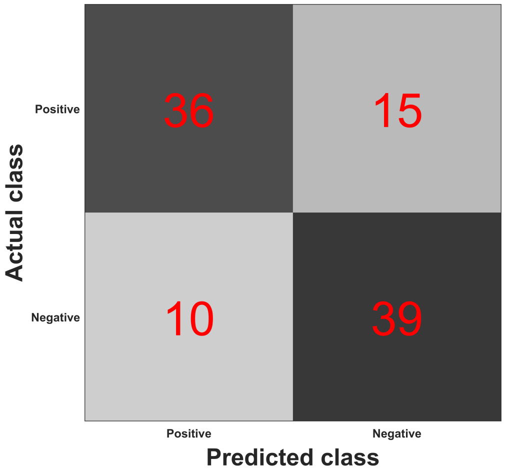

Figure 10: Confusion matrix of a classifier discriminating between 100 positive and negative test observations.

Question 25. We will consider a classifier classifying a dataset with 100 test observations into two classes (positive and negative) with confusion matrix given in Figure 10. Which statement regarding the classifier is correct?

A. The error rate of the classifier is 33.3 %.

B. The precision of the classifier is 75.0 %.

C. The recall of the classifier is 70.6 %.

D. There are more negative than positive examples in the test set.

E. Don't know.

Solution 25. The error rate of the classifier is (FP+FN)/(TP+FP+TN+FN)=(10+15)/100=25.0%. The Precision of the classifier is TP/(TP+FP)=36/(36+10)=78.3%. The Recall of the classifier is TP/(TP+FN)=36/(36+15)= 70.6%. There are 51 positive examples and 49 negative examples in the test set.

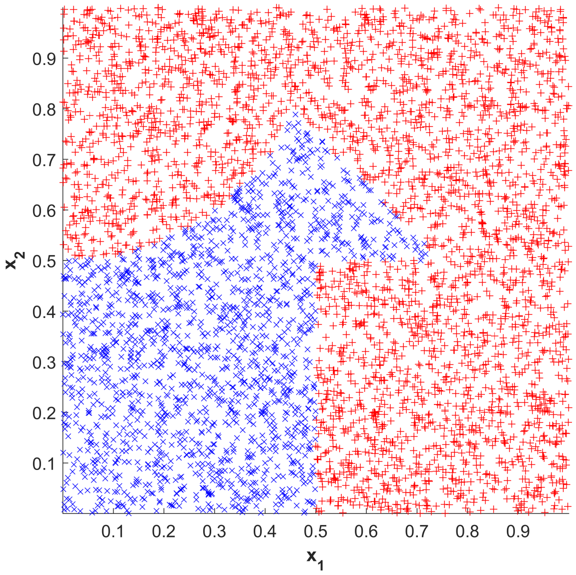

Figure 11: A two class classification problem with red plusses (i.e., +) and blue crosses (i.e., x) constituting the two classes.

Question 26. We will consider the two class classification problem given in Figure 11 in which the goal is to separate red plusses (i.e., +) from blue crosses (i.e., x). Which one of the following procedures will perfectly separate the two classes?

A. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_2 > 0.5$ **and** $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_\infty > 0.5$ **and** $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_1 > 0.75$ **then blue cross, otherwise red plus.**

B. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_\infty > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_2 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_1 > 0.75$ then blue cross, otherwise red plus.

C. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_2 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_1 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_\infty > 0.75$ then blue cross, otherwise red plus.

D. $\|\boldsymbol{x}-\begin{bmatrix}0\\1\end{bmatrix}\|_\infty > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\0\end{bmatrix}\|_1 > 0.5$ and $\|\boldsymbol{x}-\begin{bmatrix}1\\1\end{bmatrix}\|_\infty > 0.75$ then blue cross, otherwise red plus.

E. Don't know.

Solution 26. The blue crosses are more than 0.5 in radius from the point $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Furthermore, they are more than 0.5 using the infinite norm (forming a box) from $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ (and more than 0.75 from $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ using the 1-norm). Thus, the solution is given by: $\|\boldsymbol{x} - \begin{bmatrix} 0 \\ 1 \end{bmatrix}\|_2 > 0.5$ and $\|\boldsymbol{x} - \begin{bmatrix} 1 \\ 0 \end{bmatrix}\|_\infty > 0.5$ and $\|\boldsymbol{x} - \begin{bmatrix} 1 \\ 1 \end{bmatrix}\|_1 > 0.75$ then blue cross, otherwise red plus.

Question 27. Which one of the following statements is correct?

A. Unsupervised learning differs from supervised learning in that unsupervised learning both uses the input data and the outputs for training whereas supervised learning only uses the input data.

B. When using Gaussian Mixture Models (GMM) for outlier detection it is important that the observations evaluated for being outliers are included in the training of the GMM.

C. When training an artificial neural network for a dataset with very few observations it is important to include many hidden units in order to avoid overfitting.

D. Cross-validation can both be used for supervised and unsupervised learning.

E. Don't know.

Solution 27. Unsupervised learning differs from supervised learning in not having access to the output data y and not the reverse. It is important when evaluating outliers using a GMM to not include these data in the fitted density as the model may otherwise fit the density to the outliers and thereby not adequately identify these observations as outliers when they are included in defining the density. When having few observations an artificial neural network is very prone to overfitting if many hidden units are included and not the reverse. Indeed cross-validation can be used both for supervised and unsupervised learning. We used extensively cross-validation for supervised learning, i.e. classification and regression - however, we also used cross-validation in unsupervised learning in order to determine the number of clusters in a GMM and to quantify the kernel width in kernel density estimation.