Technical University of Denmark

Written examination: May 26th 2021, 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.

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.

This exam only allows for electronic hand-in.
You hand in your answers at `https://eksamen.dtu.dk/`. To hand in your answers, write them in the file `answers.txt` (this file is available from the same place you downloaded this file). When you are done, upload the `answers.txt` file (and nothing else). Double-check that you uploaded the correct version of the file from your computer.

Do not change the format of `answers.txt`
The file is automatically parsed after hand-in. Do not change the file format of `answers.txt` to any other format such as `rtf`, `docx`, or `pdf`. Do not change the file structure. Only edit the portions of the file indicated by question marks.

No.	Attribute description	Abbrev.
x_1	Hour (0-23)	Hour
x_2	Temperature (Celcius)	Temperature
x_3	Humidity (percent)	Humidity
x_4	Wind speed (m/s)	Wind
x_5	Visibility (10m)	Visibility
x_6	Dew point temperature (Celcius)	Dewpoint
x_7	Solar Radiation (MJ/m2)	Solar
x_8	Rainfall (mm)	Rain
y_r	Bike rental/demand (bikes/hour)	Bike rental

Table 1: Description of the features of the Bicycle rental dataset used in this exam. Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. To ensure bikes are available at all times, it is important to forecast the number of bikes rented per hour y_r as a function of the time of day (measured by the hour attribute so that e.g. $x_1 = 15$ is 15:00-16:00) as well as other features. Visibility is the degree of visibility at 10m of distance (0 meaning no visibility at all) and humidity is measured in percentage of full water saturation (0 being completely dry air). The unit for solar radiation is mega joules per square meter. For classification, the attribute y_r is discretized to create the variable y, taking values $y = 1$ (corresponding to a low demand), $y = 2$ (corresponding to a medium demand), and $y = 3$ (corresponding to a high demand). There are $N = 8760$ observations in total.

Question 1. The main dataset used in this exam is the Bicycle rental dataset[1] described in Table 1. We will consider the type of an attribute as the highest level it obtains in the type-hierarchy (nominal, ordinal, interval, and ratio). Which of the following statements are true about the types of the attributes in the Bicycle rental dataset?

A. x_1 (*Hour*) is nominal, x_2 (*Temperature*) is ratio, x_4 (*Wind*) is ratio, and x_6 (*Dewpoint*) is interval

B. x_2 (*Temperature*) is nominal, x_4 (*Wind*) is nominal, x_7 (*Solar*) is ratio, and x_8 (*Rain*) is ratio

C. x_1 (*Hour*) is nominal, x_2 (*Temperature*) is interval, x_3 (*Humidity*) is ratio, and x_6 (*Dewpoint*) is interval

D. x_2 (*Temperature*) is interval, x_5 (*Visibility*) is ratio, x_6 (*Dewpoint*) is interval, and x_7 (*Solar*) is ratio

E. Don't know.

Solution 1. The problem is solved by simply thinking about what the attributes represent and comparing them to the definition in the different types. Recall that

- Nominal is a type that only allow comparison (equal or different)
- Ordinal allows ordering (but not differences)
- Interval allows differences but no (physically well-defined) zero
- Ratio is a type with a zero with a well-defined meaning

With these definitions, we see that

x_1 (***Hour***) is interval

x_2 (***Temperature***) is interval

x_3 (***Humidity***) is ratio

x_4 (***Wind***) is ratio

x_5 (***Visibility***) is ratio

x_6 (***Dewpoint***) is interval

x_7 (***Solar***) is ratio

x_8 (***Rain***) is ratio

y (***Bike rental***) is ratio

and therefore option D is correct.

[1] Dataset obtained from `https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand`

Question 2. A Principal Component Analysis (PCA) is carried out on the Bicycle rental dataset in Table 1 based on the attributes x_1 (HOUR), x_2 (TEMPERATURE), x_3 (HUMIDITY), x_6 (DEWPOINT), and x_7 (SOLAR).

The data is pre-processed by substracting the mean to obtain the centered data matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out to obtain the decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$

$$\boldsymbol{V} = \begin{bmatrix} 0.11 & -0.8 & 0.3 & -0.17 & -0.48 \\ -0.58 & -0.31 & 0.01 & -0.5 & 0.56 \\ 0.49 & 0.08 & -0.49 & -0.72 & -0.07 \\ 0.6 & -0.36 & 0.04 & 0.27 & 0.66 \\ -0.23 & -0.36 & -0.82 & 0.37 & -0.09 \end{bmatrix} \quad (1)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 126.15 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 104.44 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 92.19 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 75.07 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 53.48 \end{bmatrix}.$$

We let $\boldsymbol{u}_i$ denote the i'th column of $\boldsymbol{U}$ and $\boldsymbol{v}_i$ the i'th column of $\boldsymbol{V}$. Furthermore, suppose $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the first two unit vectors. The unit vectors are defined such that only coordinate 1 of $\boldsymbol{e}_1$ is 1 (and all other coordinates are zero) and only coordinate 2 of $\boldsymbol{e}_2$ is 1 and (and all other coordinates are zero), and it is assumed the dimensions of the unit vectors are such the matrix/vector multiplications below are possible. Finally, recall $\|\boldsymbol{X}\|_F$ is the Frobenius norm.

Which one of the following statements computes the variance explained by the first two principal components?

A. $\frac{(\boldsymbol{u}_1^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1)^2 + (\boldsymbol{u}_2^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_2)^2}{\|\boldsymbol{\Sigma}\|_F^2}$

B. $\frac{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}\boldsymbol{e}_1 + \boldsymbol{e}_2^\top \boldsymbol{\Sigma}\boldsymbol{e}_2}{\|\boldsymbol{\Sigma}\|_F}$

C. $\frac{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1 + \boldsymbol{e}_2^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_2}{\|\boldsymbol{\Sigma}\|_F}$

D. $\frac{(\boldsymbol{e}_1^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1)^2 + (\boldsymbol{e}_2^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_2)^2}{\|\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\|_F^2}$

E. Don't know.

Solution 2. The correct answer is A. To see this, recall the variance explained by a given component k of the PCA is given by

$$\frac{\sigma_k^2}{\sum_{j=1}^M \sigma_j^2}$$

where M is the number of attributes in the dataset being analyzed. The values of σ_k can be read off as entry $\sigma_k = \boldsymbol{\Sigma}_{kk}$ where $\boldsymbol{\Sigma}$ is the diagonal matrix of the SVD computed above. The denominator is therefore equal to the sum of the squares of all elements in $\boldsymbol{\Sigma}$ i.e. $\sum_{j=1}^M \sigma_j^2 = \|\boldsymbol{\Sigma}\|_F^2$. To obtain the numerator, note that since $\boldsymbol{U}, \boldsymbol{V}$ are orthonormal it holds that

$$\boldsymbol{u}_1^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_1 = \boldsymbol{e}_1^\top \boldsymbol{\Sigma}\boldsymbol{e}_1 = \Sigma_{11}.$$

Hence expressions of this form a suitable to compute the numerator, meaning that A are the correct answer.

Question 3. Consider again the PCA analysis for the Bicycle rental dataset, in particular the SVD decomposition of $\tilde{\boldsymbol{X}}$ in Equation (1). Which one of the following statements is true?

A. An observation with a low value of Temperature, a high value of Humidity, a high value of Dewpoint, and a low value of Solar will typically have a positive value of the projection onto principal component number 1.

B. An observation with a high value of **Hour**, a low value of **Humidity**, and a low value of **Solar** will typically have a negative value of the projection onto principal component number 3.

C. An observation with a high value of **Hour**, a low value of **Temperature**, and a low value of **Dewpoint** will typically have a positive value of the projection onto principal component number 5.

D. An observation with a low value of **Hour**, a low value of **Temperature**, a low value of **Dewpoint**, and a low value of **Solar** will typically have a negative value of the projection onto principal component number 2.

E. Don't know.

Solution 3. The correct answer is A. Focusing on the correct answer, note the projection onto principal component $\boldsymbol{v}_1$ (i.e. column one of $\boldsymbol{V}$) is

$$b_1 = \boldsymbol{x}^\top \boldsymbol{v}_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_6 & x_7 \end{bmatrix} \begin{bmatrix} 0.11 \\ -0.58 \\ 0.49 \\ 0.6 \\ -0.23 \end{bmatrix}$$

(we use these attributes since these were selected for the PCA). It is now a simple matter of observing that for this number to be (relatively large) and positive, this occurs if x_2, x_3, x_6, x_7 has large magnitude and the sign convention given in option A.

Question 4. Consider again the Bicycle rental dataset and the PCA decomposition described in Equation (1). Recall the PCA decomposition is obtained by first forming the centered data matrix $\tilde{\boldsymbol{X}}$ by subtracting the column-wise mean

$$\boldsymbol{\mu} = \begin{bmatrix} 12.9 \\ 58.2 \\ 1.7 \\ 1436.8 \\ 4.1 \end{bmatrix}$$

from the data matrix $\boldsymbol{X}$. Assume an observation has coordinates

$$\boldsymbol{x} = \begin{bmatrix} 15.5 \\ 59.2 \\ 1.4 \\ 1438.0 \\ 5.3 \end{bmatrix}.$$

Which coordinates in the coordinate system spanned by the principal component vectors corresponds to $\boldsymbol{x}$?

A. $\boldsymbol{b} = \begin{bmatrix} 0.0 & -3.2 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

B. $\boldsymbol{b} = \begin{bmatrix} 0.0 & 1.2 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

C. $\boldsymbol{b} = \begin{bmatrix} 0.0 & 1.5 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

D. $\boldsymbol{b} = \begin{bmatrix} 0.0 & -1.6 & 0.0 & 0.0 & 0.0 \end{bmatrix}^\top$

E. Don't know.

Solution 4. The simplest way to solve this problem is to begin with one of the possible values of $\boldsymbol{b}$ and check if it corresponds to $\boldsymbol{x}$ or not. Recall that this is done by computing:

$$\boldsymbol{x} = \boldsymbol{V}\boldsymbol{b} + \boldsymbol{\mu}. \tag{2}$$

We will only compute the value of $\boldsymbol{x}$ at the first coordinate, i.e. x_1. Since most entries in $\boldsymbol{b}$ are zero this can be done as simply:

$$x_1 = V_{12} b_2 + \mu_1. \tag{3}$$

Knowing that $x_1 = 15.5$ we observe that A is the correct answer.

Question 5. Consider again the Bicycle rental dataset. The empirical covariance matrix of the first 5 attributes $x_1, \ldots, x_5$ is given by:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 143.0 & 39.0 & -0.0 & 253.0 & 142.0 \\ 39.0 & 415.0 & -7.0 & -6727.0 & 143.0 \\ -0.0 & -7.0 & 1.0 & 108.0 & -2.0 \\ 253.0 & -6727.0 & 108.0 & 370027.0 & -1403.0 \\ 142.0 & 143.0 & -2.0 & -1403.0 & 171.0 \end{bmatrix}.$$

What is the empirical correlation of x_2 (TEMPERATURE) and x_3 (HUMIDITY)?

A. -0.12987

B. -0.01687

C. -0.34362

D. -2.64575

E. Don't know.

Solution 5. Recall the correlation is defined as

$$\text{cor}[x, y] = \frac{\text{cov}[x, y]}{\text{std}[x]\,\text{std}[y]}$$

Next, by definition the diagonal elements of the covariance matrix are estimates of the variance and the off-diagonal elements are estimates of the covariance, i.e. for $i \neq j$:

$$\hat{\Sigma}_{ii} = \text{Var}[x_i], \quad \hat{\Sigma}_{ij} = \text{cov}[x_i, x_j]$$

Therefore we get:

$$\text{cor}[x_i, y_j] = \frac{\hat{\Sigma}_{ij}}{\sqrt{\hat{\Sigma}_{ii}\hat{\Sigma}_{jj}}}.$$

By simple insertion, we see option C is correct.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	5.0	7.7	6.1	4.2	11.0	7.3	9.0	11.3	1.4
o_2	5.0	0.0	5.4	4.0	7.5	7.9	5.3	6.8	11.9	3.5
o_3	7.7	5.4	0.0	5.2	7.2	6.1	7.8	6.7	12.9	6.4
o_4	6.1	4.0	5.2	0.0	5.1	5.4	8.4	3.3	8.1	4.8
o_5	4.2	7.5	7.2	5.1	0.0	8.7	8.8	6.6	7.7	4.1
o_6	11.0	7.9	6.1	5.4	8.7	0.0	12.0	4.2	9.3	9.8
o_7	7.3	5.3	7.8	8.4	8.8	12.0	0.0	11.0	16.3	6.7
o_8	9.0	6.8	6.7	3.3	6.6	4.2	11.0	0.0	6.2	7.8
o_9	11.3	11.9	12.9	8.1	7.7	9.3	16.3	6.2	0.0	10.4
o_{10}	1.4	3.5	6.4	4.8	4.1	9.8	6.7	7.8	10.4	0.0

Table 2: The pairwise cityblock distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{p=1} = \sum_{k=1}^{M} |x_{ik} - x_{jk}|$ between 10 observations from the Bicycle rental dataset (recall that $M = 8$. Each observation o_i corresponds to a row of the data matrix $\boldsymbol{X}$ of Table 1. The colors indicate classes such that the black observations $\{o_1, o_2\}$ belong to class C_1 (corresponding to a low demand), the red observations $\{o_3, o_4, o_5, o_6\}$ belong to class C_2 (corresponding to a medium demand), and the blue observations $\{o_7, o_8, o_9, o_{10}\}$ belong to class C_3 (corresponding to a high demand). To avoid single features to dominate, the dataset was standardized by subtracting the mean and dividing by the standard deviation.

Question 6. To examine if observation o_3 may be an outlier, we will calculate the average relative density using the cityblock distance based on the observations given in Table 2 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. What is the average relative density for observation o_3 for $K = 2$ nearest neighbors?

A. 0.7

B. 0.4

C. 0.63

D. 0.19

E. Don't know.

Solution 6.

To solve the problem, first observe the $k = 2$ neighborhood of o_3 and density is:

$$N_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = \{o_4, o_2\}, \quad \text{density}_{\boldsymbol{X}_{\setminus 3}}(\boldsymbol{x}_3) = 0.189$$

For each element in the above neighborhood we can then compute their $K = 2$-neighborhoods and densities to be:

$$N_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = \{o_8, o_2\}, \quad N_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = \{o_{10}, o_4\}$$

and

$$\text{density}_{\boldsymbol{X}_{\setminus 4}}(\boldsymbol{x}_4) = 0.274, \text{density}_{\boldsymbol{X}_{\setminus 2}}(\boldsymbol{x}_2) = 0.267.$$

From these, the ARD can be computed by plugging in the values in the formula given in the problem.

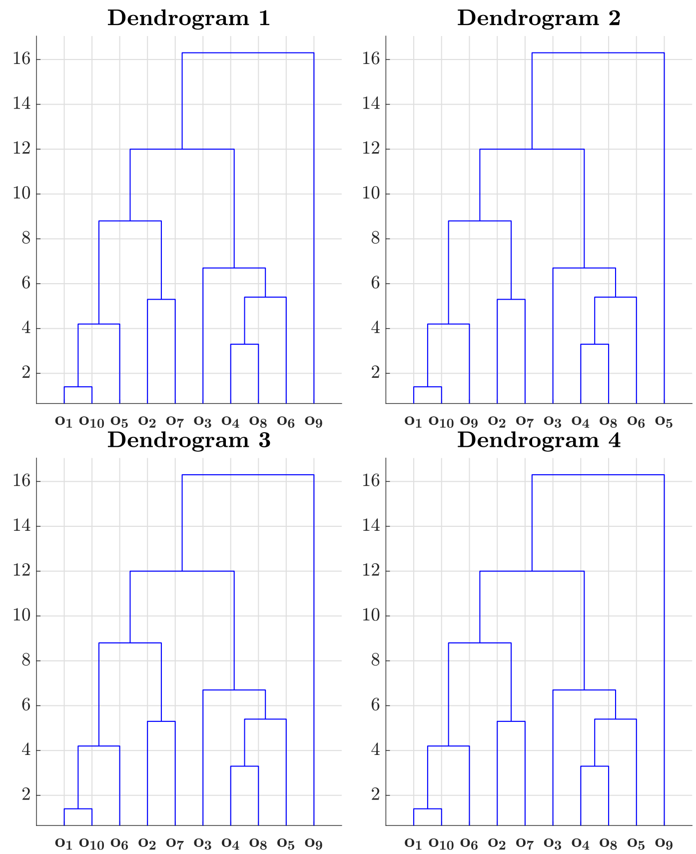

Figure 1: Proposed hierarchical clustering of the 10 observations in Table 2.

Question 7. A hierarchical clustering is applied to the 10 observations in Table 2 using *maximum* linkage. Which one of the dendrograms shown in Figure 1 corresponds to the distances given in Table 2?

A. Dendrogram 1

B. Dendrogram 2

C. Dendrogram 3

D. Dendrogram 4

E. Don't know.

Solution 7. The correct solution is A. We can rule out the other solutions by observing the first merge operation at which they diverge from the correct solution.

- In dendrogram 2, merge operation number 3 should have been between the sets $\{f_5\}$ and $\{f_1, f_{10}\}$ at a height of 4.2, however in dendrogram 2 merge number 3 is between the sets $\{f_9\}$ and $\{f_1, f_{10}\}$.

- In dendrogram 3, merge operation number 3 should have been between the sets $\{f_5\}$ and $\{f_1, f_{10}\}$ at a height of 4.2, however in dendrogram 3 merge number 3 is between the sets $\{f_6\}$ and $\{f_1, f_{10}\}$.
- In dendrogram 4, merge operation number 3 should have been between the sets $\{f_5\}$ and $\{f_1, f_{10}\}$ at a height of 4.2, however in dendrogram 4 merge number 3 is between the sets $\{f_6\}$ and $\{f_1, f_{10}\}$.

Question 8. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two binary vectors of (even) dimension M such that the first two elements of $\boldsymbol{x}_1$ are 1 (and the rest are 0) and the first $\frac{M}{2}$ elements of $\boldsymbol{x}_2$ are 1 (and the rest are 0).

Which of the following expressions computes the Jaccard similarity of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ when $M \geq 4$?

A. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{4}{M}$

B. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\frac{1}{2}M}{\frac{1}{2}M+2}$

C. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{2}{M}$

D. $J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{2}{\frac{1}{2}M-2}$

E. Don't know.

Solution 8. The Jaccard similarity is given as

$$J(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{n_{11}}{M - n_{00}}.$$

Given the information in the problem the number of 0-0 matches is $n_{00} = \frac{M}{2}$ while the number of 1-1 matches is $n_{11} = 2$. The solution is therefore A.

Question 9. Consider again the Bicycle rental dataset in Table 1. We apply backward selection to find an interpretable linear regression model which uses a subset of the $M = 8$ attributes to predict the bike rental y_r. Recall backward selection chooses models based on the test error as determined by cross-validation, and in our case we use the hold-out method to generate a single test/training split.

Suppose backward selection ends up selecting the attributes x_1, x_3, x_4, x_5, x_6, x_7, and x_8, what is the minimal number of models which were *tested* in order to obtain this result?

A. 15 models

B. 18 models

C. 16 models

D. 8 models

E. Don't know.

Solution 9.

Note the solution selected all variables except one. Since we use backward selection, we first have to evaluate a single model with all features. Then we evaluated all models with a single missing feature giving an additional M models. One of these models were selected and variable selection proceeded at the next level where an additonal $M-1$ models were evaluated. However, since all had a higher cost, none were selected and the method terminated. This gives

$$1 + M + M - 1$$

evaluations and so C is correct.

Question 10. We wish to predict which of the three classes an observation $\boldsymbol{x}$ belong to in the Bicycle rental dataset described in Table 1. To accomplish this we apply a Naïve-Bayes classifier where we model each of the $M = 8$ features using a 1-dimensional normal distribution. The classifier will be used in an embedded setting where model prediction speed is paramount. Therefore, consider a single model evaluation:

$$p(y = \text{LOW DEMAND}|\boldsymbol{x}).$$

What is the minimum number of evaluations of the normal density function $\mathcal{N}(x|\mu, \sigma^2)$ we have to perform to compute this quantity?

A. 24

B. 27

C. 36

D. 32

E. Don't know.

Solution 10. Recall the formula for Naïve-Bayes is

$$p(y|\boldsymbol{x}) = \frac{p(y) \prod_{k=1}^{M} p(x_k|y)}{\sum_{y'} p(y') \prod_{k=1}^{M} p(x_k|y')}.$$

The total number of evaluations is equal to the total number of evaluations of terms $p(x_k|y)$. Note that once we have evaluated the denominator, we will also have evaluated the numerator since they share the same terms $p(x_k|y)$, cutting down on computations. The total number of evaluations is therefore simply CM where C is the number of classes and therefore answer A is correct.

Question 11. Consider the Bicycle rental dataset from Table 1 consisting of $N = 8760$ observations, and suppose the attribute Humidity has been binarized into low and high values. We still consider the goal to predict the bike rental and are given the following information

- Of the 3285 observations with low demand, 1327 had a low value of Humidity.
- Of the 2190 observations with medium demand, 1718 had a low value of Humidity.
- Of the 3285 observations with high demand, 2344 had a low value of Humidity.

Suppose a particular observation has a high value of Humidity, what is the probability of observing high demand?

A. 0.279

B. 0.286

C. 0.04

D. 0.487

E. Don't know.

Solution 11. The problem is solved by applying Bayes rule. Introducing the binary variable x such that $x = 1$ if an observation has a high value of Humidity (and otherwise $x = 0$) the question asked is equivalent to computing $p(y = 3|x = 1)$. Applying Bayes' theorem we get:

$$p(y = 3|x = 1) = \frac{p(x = 1|y = 3)p(y = 3)}{\sum_{k=1}^{3} p(x = 1|y = k)p(y = k)}$$

Recall that $p(x = 1|y) = 1 - p(x = 0|y)$, we can obtain the required probabilities from each of the three bullet points above. We obtain:

- $p(y = 1) = \frac{3285}{N}$ and $p(x = 0|y = 1) = \frac{1327}{3285}$.
- $p(y = 2) = \frac{2190}{N}$ and $p(x = 0|y = 2) = \frac{1718}{2190}$.
- $p(y = 3) = \frac{3285}{N}$ and $p(x = 0|y = 3) = \frac{2344}{3285}$.

Plugging these into Bayes theorem, and using that $p(x = 0|y) = 1 - p(x = 1|y)$ because x is binary, we see $p(y = 3|x = 1) = 0.279$ and hence that option A is correct.

Question 12. Consider the Bicycle rental dataset described in Table 1. Suppose we apply a market basket analysis to the dataset in the usual fashion: We first binarize each of the attributes, thereby obtaining $M = 8$ items, and consider each of the $N = 8760$ observations as a transaction containing a (subset) of the binarized attributes. We will let $C(\{I_1, \ldots, I_k\})$ be the number of the $N = 8760$ transactions containing the itemset $\{I_1, \ldots, I_k\}$. For this problem we focus on just three items and are given the information:

- $C(\{\text{Visibility}\}) = 4091$.
- $C(\{\text{Humidity}\}) = 3637$.
- $C(\{\text{Dewpoint}\}) = 3459$.

Finally, consider the itemset:

$$I : \{\text{Visibility}, \text{Humidity}\}.$$

Which of the following options indicate the *highest possible* support of the itemset I which is consistent (i.e., obtainable) given the information in the bullet list above?

A. **supp**$(I) = 0.415$

B. supp$(I) = 0.441$

C. supp$(I) = 0.217$

D. supp$(I) = 0.467$

E. Don't know.

Solution 12. For a transaction to include an itemset, it must (by the downwards closure property) include all subsets, and hence the maximal number of transactions which contain $\{\text{Visibility}, \text{Humidity}\}$ is upper-bounded by the number of transactions which contain either of the items, i.e. $\min\{C(\{\text{Visibility}\}), C(\{\text{Humidity}\})\}$:

$$C(\{\text{Visibility}, \text{Humidity}\}) \leq \min\{C(\{\text{Visibility}\}), C(\{\text{Humidity}\})\} = 3637.$$

Dividing by N we see answer A is correct.

	1	2	3	4	5	6	7	8
x_1	-1.1	-0.8	0.08	0.18	0.34	0.6	1.42	1.68
y_r	12	5	10	23	6	17	14	13

Table 3: Values of x_1 and the corresponding value of y_r.

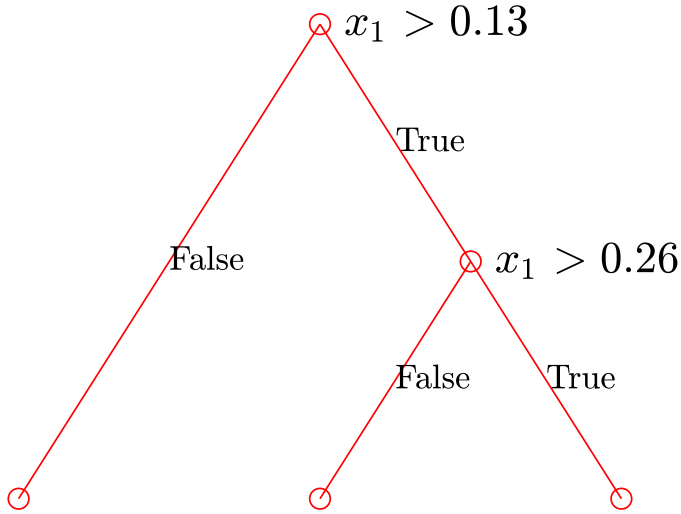

Figure 2: Structure of a regression tree. The nodes show the decision rules which determine how the observations are propagated towards the leafs of the tree.

Question 13. We will consider the first 8 observations of the Bicycle rental dataset shown in Table 2. Table 3 shows their corresponding value of x_1 and y_r. We fit a small regression tree to this dataset. The structure (and binary splitting rules) is depicted in Figure 2. Which one of the prediction rules (i.e., the model output $\hat{y}_r$ as a function of x_1) shown in Figure 3 corresponds to the tree?

A. Prediction rule 1

B. Prediction rule 2

C. Prediction rule 3

D. Prediction rule 4

E. Don't know.

Solution 13.

The problem is easiest solved by selecting a lucky x-value and using it to rule out the wrong plot. In our case, we select $x_1 = 0.4$. The predicted value for a given input is computed as the average y-value of those

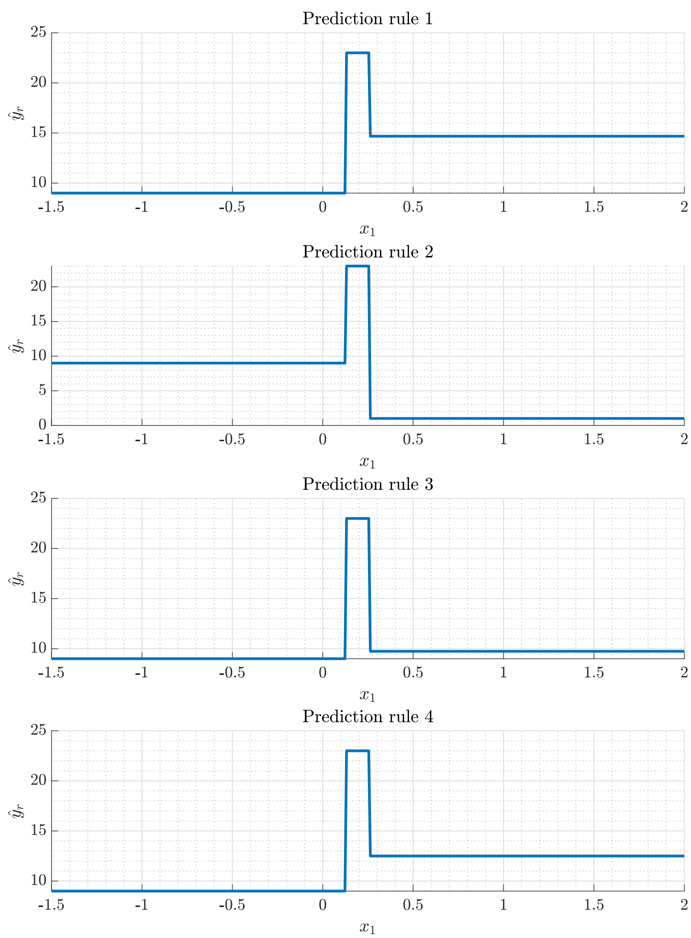

Figure 3: Possible model predictions of $\hat{y}_r$ as a function of x_1 for the decision tree illustrated in Figure 2.

observations in the training set which is assigned to the same leaf node v as the input, i.e.

$$y(v) = \frac{1}{N(v)} \sum_{i \in v} y_i$$

(see the section on regression trees in lecture notes). Therefore, we first need to find out which leaf node the observation is assigned to. To do this, start at the root and compare $x_1 = 0.4$ to the rule in the split

$$x_1 > 0.13$$

and we continue down the right branch. Continuing in this manner, we see $x_1 = 0.4$ is classified to leaf number three from the left. Then, proceeding in the same manner with the x_1 observations in Table 3, we see the observations o_5, o_6, o_7, and o_8 are also assigned to leaf three (counted from the left). According to the above the prediction is then simply the average of their y-value

$$\hat{y} = \frac{1}{4}(6 + 17 + 14 + 13)$$

or $\hat{y} = 12.5$. we compare this information in Figure 3 and note this allows us to rule out all the wrong options. Therefore, D is correct.

Question 14. In this problem, we will again consider the 8 observations from the Bicycle rental dataset shown in Table 3. Recall that Figure 2 shows the structure of the small regression tree fitted to this dataset using Hunt's algorithm along with the thereby obtained binary splitting rules. What was the purity gain Δ of the **second** split Hunt's algorithm accepted?

A. $\Delta = 101.2$

B. $\Delta = 30.64$

C. $\Delta = 17.64$

D. $\Delta = 13.0$

E. Don't know.

Solution 14.

The second split Hunt's algorithm accepted must be the non-root split, i.e. corresponding to the rule

$$x > 0.26.$$

This node, which we will call the base node of the split, partitions the observations:

$$v_0 = \{4, 5, 6, 7, 8\}$$

into the two sets

$$v_1 = \{4\}, \quad v_2 = \{5, 6, 7, 8\}$$

along the two legs of the split. The impurity of these two sets, and the impurity of all y-values at the root, is computed using the impurity measure appropriate for regression trees

$$I(v) = \frac{1}{N(v)} \sum_{i \in v_1} (y_i - y(v))^2$$

where $y(v)$ is the average of the y-values in v_i. Specifically

$$y(v_1) = 23.0, \quad y(v_2) = 12.5$$

At the base node v_0 of the split we consider we perform a similar calculation for the 5 observations and find they have a mean y-value of:

$$y(v_0) = 12.5$$

Therefore:

$$I(v_0) = 30.64, \quad I(v_1) = 0.0, \quad I(v_2) = 16.25$$

these are finally combined to the impurity gain as

$$\Delta = I(v_0) - \sum_{k=1}^{2} \frac{N(v_k)}{N} I(v_k)$$

where for instance $N(v_1) = 1$ are the number of observations in branch 1. We find by insertion that $\Delta = 17.64$ and hence C is correct.

Question 15. Consider again the Bicycle rental dataset of Table 1. Suppose we wish to predict the class label y using a multivariate regression model, and to improve performance we wish to apply Adaboost. Recall the first steps of adaboost consists of: (i) Initialize weights, (ii) select a training set (iii) fit a model to the training set. In the first round of boosting, the fitted model has an error rate ϵ when evaluated on the full dataset, and it made a correct prediction of the class membership of observation $i = 5$ and an incorrect prediction of the class membership of observation $i = 1$.

After the first round of boosting, which of the following expressions will compute the ratio of weights of observation 1, w_1 and observation 5, w_5?

A. $\frac{w_1}{w_5} = \exp\left(\frac{1-\epsilon}{\epsilon}\right)$

B. $\frac{w_1}{w_5} = \frac{1-\epsilon}{\epsilon}$

C. $\frac{w_1}{w_5} = \frac{\exp\left(\frac{1-\epsilon}{\epsilon}\right)}{\exp\left(-\frac{1-\epsilon}{\epsilon}\right)}$

D. $\frac{w_1}{w_5} = \sqrt{\frac{1-\epsilon}{\epsilon}}$

E. Don't know.

Solution 15. Recall the weights in the Adaboost algorithm, prior to normalization, are computed as $w_i(1)e^{\pm\alpha_i}$ and they are initialized as $w_i(1) = \frac{1}{N}$. Falsely classified observations are boosted up, and correctly are boosted down. Hence:

$$\frac{w_1}{w_5} = \frac{\frac{1}{N}e^{\alpha}}{\frac{1}{N}e^{-\alpha}} = e^{2\alpha}$$

Using that $\alpha = \frac{1}{2}\log\frac{1-\epsilon}{\epsilon}$ we see that option B is correct.

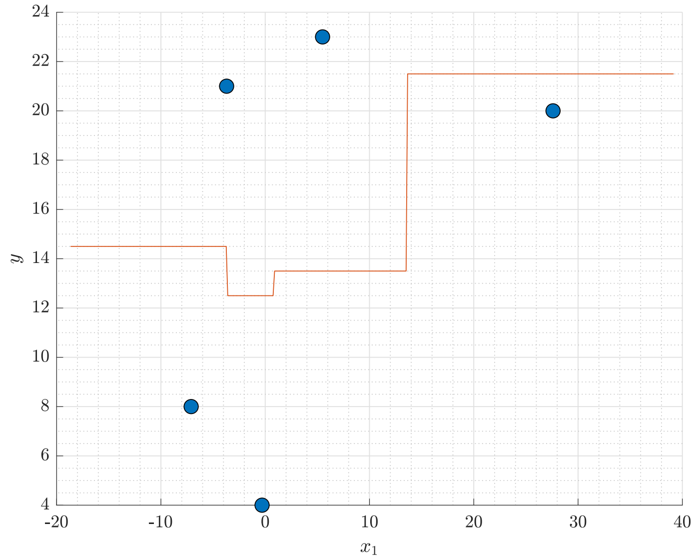

Figure 4: KNN regression model in which the red line is fitted to a small 1-dimensional dataset.

Question 16. Suppose a K-nearest neighbors regression model is fitted to a small 1-dimensional dataset with $N = 5$ observations. The predicted response is shown in Figure 4. How many neighbors (i.e. K) was used?

A. $K = 2$

B. $K = 4$

C. $K = 1$

D. $K = 3$

E. Don't know.

Solution 16.

The problem could be solved by using the definition of the KNN regression model and test various points, but it is much quicker solved using an intuitive argument. The KNN regression model consist of a series of steps, and the important information is where the discontinuities occur. If $K = 1$, the y-value has to pass through the training observations. On the other hand, if we consider the y-value at the right-most end of the x-axis, we note it is quite large consistent with it being computed using the two left-most observations, but not consisting with including the third observation from the right (or additional observations). Hence, we conclude that $K = 2$.

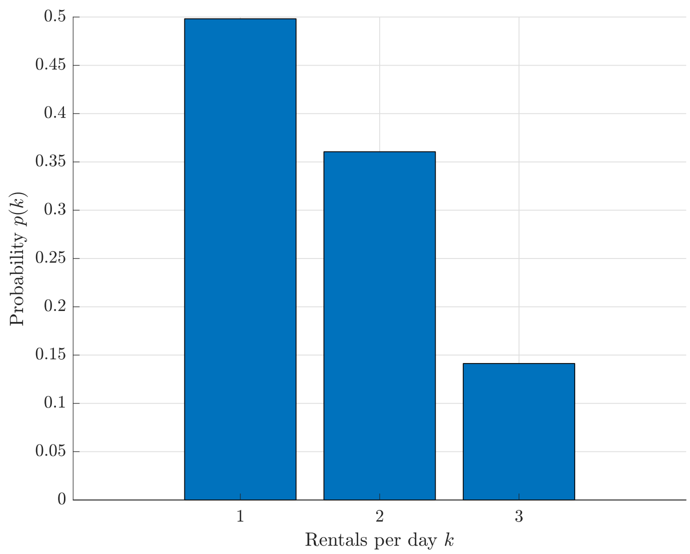

Figure 5: Probability $p(k)$ a citybike is rented exactly k times a day. The probability of $k \geq 4$ is negligible and can be ignored.

Question 17. The number of times a citybike is rented per day is an important factor in determining how often they should be replaced. Suppose the typical bike rentals per day is estimated from data, and the chance $p(k)$ a bike will be rented k times is shown in the discrete probability distribution shown in Figure 5. It is known that the mean of this distribution is 1.6, but what is the variance?

A. Variance is 3.4

B. Variance is 1.6

C. Variance is 0.2

D. Variance is 0.5

E. Don't know.

Solution 17.

The standard deviation of a discrete distribution is given by

$$\mathrm{Var}[K] = \sum_{k=1}^{3} p(k)(k-\mu)^2$$

and we are fortunate enough to be told that $\mu = 1.6$. The probabilities can be read from Figure 5 which gives us:

$$\mathrm{Var}[K] = 0.5(1-\mu)^2 + 0.36(2-\mu)^2 + 0.14(3-\mu)^2.$$

Upon insertion, we see the right answer is D.

Question 18. Which one of the following statements are true?

A. Regularization is not applicable to a logistic regression model.

B. When we apply Adaboost, the less errors a classifier makes in a given round of boosting, the more the weights will be increased for the wrongly classified observations.

C. When using McNemars test to determine if two classification models have different performance, one should apply two-level cross-validation (either hold out/K-fold or leave-one-out).

D. Let $\boldsymbol{x}_i$ be the i'th observation of a (non-standardized) dataset $\boldsymbol{X}$. Suppose we carry out a PCA analysis on $\boldsymbol{X}$ and we let $\boldsymbol{b}_i$ be the principal component coefficient vector (i.e., projection) corresponding to $\boldsymbol{x}_i$ when projected onto *all* the principal components. It is then true that $\|\boldsymbol{x}_i\| = \|\boldsymbol{b}_i\|$ (in the Euclidean norm).

E. Don't know.

Solution 18. The correct answer is B: Inspecting the adaboost algorithm, we see that as the number of error decrease, so does ϵ. When ϵ decreases, then $\alpha = \frac{1}{2}\log\frac{1-\epsilon}{\epsilon}$ will increase. Finally, note the wrongly classified observations are boosted proportional to e^{α_i}, so B is correct.

For the other options, note that regularization can easily be applied to logistic regression (same as linear regression). McNemars test only requires a set of prediction which can be obtained using normal 1-level cross-validation, and for the last problem, $\boldsymbol{x}_i$ does not have the mean-vector subtracted (whereas it is subtracted to compute $\boldsymbol{b}_i$) and so we should have no expectation the norms should be the same.

Question 19. Consider a regression problem where the goal is to predict a ratio variable y_i using the 1-dimensional input x_i. Suppose we wish to do this using a neural network with a single hidden layer (the hidden layer has a sigmoid activation function), no activation function (i.e. the identity activation function) for the output layer, and that we use the ordinary quadratic cost function suitable for regression. What is an appropriate cost function on a training set of size N (assuming all terms of the form $w^{(\cdot)}_{\cdot}$ are weights)?

A. $\sum_{i=1}^{N}\left(\frac{w_0^{(2)}}{1+e^{-y_i}} - \frac{w_1^{(2)}}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}}} - \frac{w_2^{(2)}}{1+e^{-w_{2,0}^{(1)}-x_i w_{2,1}^{(1)}}}\right)^2$

B. $\sum_{i=1}^{N}\left(w_0^{(2)} - \frac{w_1^{(2)}}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}-y_i w_{1,2}^{(1)}}} - \frac{w_2^{(2)}}{1+e^{-w_{2,0}^{(1)}-x_i w_{2,1}^{(1)}-y_i w_{2,3}^{(1)}}}\right)^2$

C. $\sum_{i=1}^{N}\left(y_i - w_0^{(2)} - \frac{w_1^{(2)}}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}}} - \frac{w_2^{(2)}}{1+e^{-w_{2,0}^{(1)}-x_i w_{2,1}^{(1)}}}\right)^2$

D. $\sum_{i=1}^{N}\left(y_i - w_0^{(2)} - \frac{w_1^{(2)}}{w_1^{(2)}-e^{w_{1,0}^{(1)}+x_i w_{1,1}^{(1)}}} - \frac{w_2^{(2)}}{w_2^{(2)}-e^{w_{2,0}^{(1)}+x_i w_{2,1}^{(1)}}}\right)^2$

E. Don't know.

Solution 19.
For a general input x_i, the quadratic cost function will be of the form:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2$$

where $f(x)$ is the output of the neural network. If the hidden units activation is denoted z_1 and z_2 the output has the linear form

$$f(x) = w_0^{(2)} + w_1^{(2)} z_1 + w_2^{(2)} z_2$$

Finally, the output activation of the hidden units, for instance z_1, can be computed using the sigmoid activation function:

$$\sigma(w_{1,0}^{(1)} + x_i w_{1,1}^{(1)}) = \frac{1}{1+e^{-w_{1,0}^{(1)}-x_i w_{1,1}^{(1)}}}.$$

Comparing this information with the four expression we can rule out all options except C.

	x_1	x_5	y
Mean	12.9	4.1	11.5
Standard deviation	11.9	13.1	6.9

Table 4: Column-wise mean and standard deviation computed on the Bicycle rental dataset.

Question 20. Consider once again the bicycle rental dataset described in Table 1, but this time we will limit ourselves to just the features x_1 (Hour) and x_5 (Visibility) from the full dataset $\boldsymbol{X}$. The goal is still to predict the bike rental $y = y_r$, and to achieve this we will apply ridge-regression. Recall that ridge regression determines the constant offset w_0 and the two coefficients w_1 and w_2 of x_1 and x_5 respectively, by minimizing a cost function of the form:

$$\sum_{i=1}^{N}\left(y_i - w_0 - w_1\frac{X_{i,1}-\mu_1}{\sigma_1} - w_2\frac{X_{i,5}-\mu_5}{\sigma_5}\right)^2 + \lambda(w_1^2 + w_2^2).$$

In this expression, μ_k and σ_k are the mean and standard deviations of column k, and their values can be found in Table 4, along with the corresponding values for y. Assuming the regularization strength is $\lambda = 10.0$, which one of the following expressions will predict the value y for an input observation with $x_1 = 0$ and $x_5 = 1$?

A. $y = w_0 + 1.08w_1 + 0.39w_2$

B. $y = 0.14w_0 - 1.08w_1 - 0.24w_2$

C. $y = w_0 - 0.24w_2$

D. $y = 11.5 - 1.08w_1 - 0.24w_2$

E. Don't know.

Solution 20. To solve this problem, we should first apply the same transformation to the text-point $\boldsymbol{x}$ as is applied when training the model and then remember to add w_0 which is equal to the mean of $\boldsymbol{y}$. The prediction rule is

$$y = w_0 + w_1\frac{x_1-\mu_1}{\sigma_1} + w_2\frac{x_5-\mu_5}{\sigma_5}$$

The mean/standard deviations are given as the column-wise mean/standard deviations in Table 4. Inserting these values, as well as the values for x_1 and x_5, we see that answer D is correct.

Observation nr. i	$\boldsymbol{w}_1^\top \tilde{\boldsymbol{x}}_i$	$\boldsymbol{w}_2^\top \tilde{\boldsymbol{x}}_i$
1	0.03	-1.89
2	1.17	-0.89
3	1.15	-0.87
4	1.32	-0.71
5	-0.05	-1.9
6	0.64	-1.28
7	0.65	-1.27
8	1.25	-0.69

Table 5: Output of the linear transformation (prior to softmax normalization) of a multinomial regression model applied to the Bicycle rental dataset. The full dataset contains $N = 8760$ observations, but the table only contains the output for the first $i = 1, \ldots, 8$ observations.

Question 21. Consider the Bicycle rental dataset described in Table 1. Recall the dataset is comprised of $C = 3$ classes, and suppose we fit a multinomial regression model to predict the class label y_i given the $M = 8$-dimensional feature vector $\boldsymbol{x}_i$. This results in two weight-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ such that the class-label is predicted using the softmax activation as described in Section 15.3.3 in the lecture notes. Prior to softmax normalization, the output on the first 8 observations are shown in Table 5. According to the multinomial regression model, what is the probability observation $i = 1$ is assigned to the low demand class ($y = 1$)?

A. 0.01

B. 0.07

C. 0.26

D. 0.47

E. Don't know.

Solution 21. Solving this problem is simply a matter of using the definition of the multinomial regression model. The probability can be computed as

$$p(y = 1|\boldsymbol{x}_1) = \frac{e^{\boldsymbol{w}_1^\top \boldsymbol{x}_1}}{e^{\boldsymbol{w}_1^\top \boldsymbol{x}_1} + e^{\boldsymbol{w}_2^\top \boldsymbol{x}_1} + 1}$$

Inserting the number from Table 5 we see that answer D is correct.

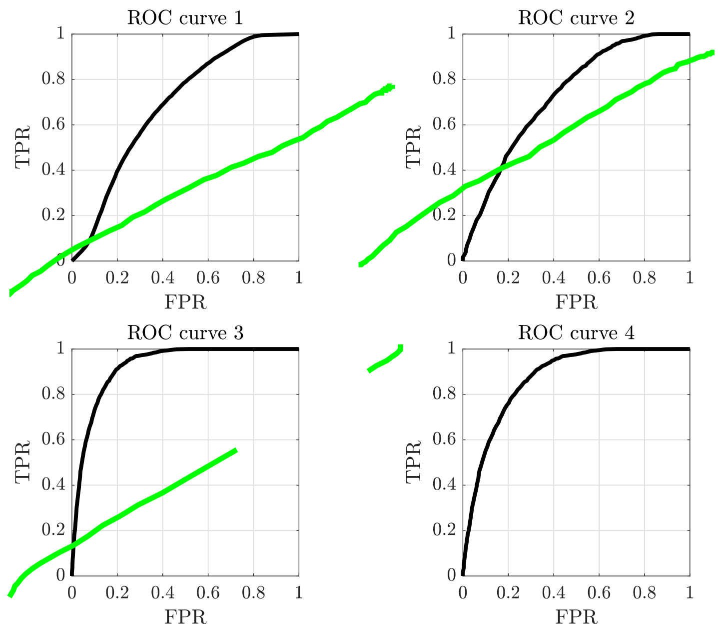

Figure 6: Candidate ROC curves for the classifier.

Question 22. We wish to predict whether an observation from the Bicycle rental dataset (see Table 1) belongs to the low demand class (or not). To accomplish this, we fit a logistic regression model to the dataset, and for each observation $\boldsymbol{x}_i, y_i$ obtain a class-probability prediction $\hat{y}_i \in [0, 1]$. We threshold the class-probability at different values θ thereby obtaining, for each value of θ, the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). These are plotted as functions of θ in Figure 7. Which of the receiver-operator characteristic (ROC) plots shown in Figure 6 corresponds to these graphs?

A. ROC curve 1

B. ROC curve 2

C. ROC curve 3

D. ROC curve 4

E. Don't know.

Solution 22. From the TP curve (left-most value) we get that the total number of positive-class observations are $P = 1652$ and from the TN curve (right-most value) we get $N = 2728$. The simplest approach is to compute a point on the ROC curve. Most values will do, however we choose the point corresponding to $\theta = 0.5$, at which the number of false positives is $FP = 570$ and true positives is $TP = 1285$. We

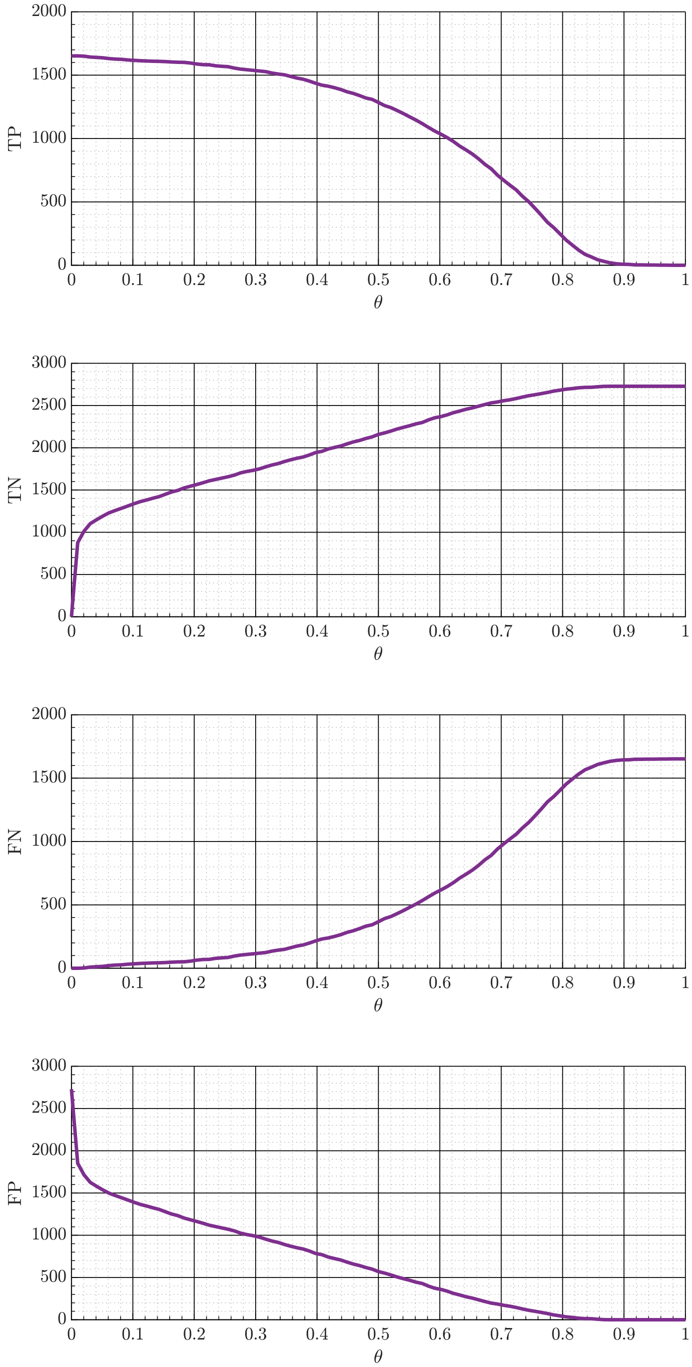

Figure 7: TP, TN, FN, and FP as functions of threshold value θ.

therefore see that the following point must lie on the ROC curve:

$$(\text{fpr}, \text{tpr}) = (\frac{FP}{N}, \frac{TP}{P}) = (0.21, 0.78)$$

this rules out all options except D.

Question 23. Which one of the following statements is true?

A. Suppose hold-out cross-validated backward selection is applied to select which features to include in a linear regression model. Each time a new model is selected by backward selection, the training error for that model will be smaller than (or equal to) the training error in the previous step.

B. Consider how Bagging and Boosting makes predictions in a binary classification task. Recall that both bagging and boosting train multiple classifiers on the same dataset. The only difference between the methods is how the training sets used to train the classifiers is sampled from the full dataset. Both sample the datasets with replacement, but bagging sample them uniformly, whereas Adaboost sample them according to weights which are iteratively updated.

C. In terms of a bias-variance trade-off, a logistic regression model with a well-tuned regularization parameter has a negligible bias but a fairly high variance.

D. When comparing two classifiers, leave-one-out cross-validation is a suitable cross-validation method to use in conjunction with McNemars test.

E. Don't know.

Solution 23. The correct answer is D: McNemars test can be used for any cross-validation procedure as long as the methods are tested on the same sets of observations. This is guaranteed for leave-one-out cross-validation. A is wrong because backward selection will remove features, and is thereby guaranteed to increase the training error as the models become less and less expressive. AdaBoost uses a weighted combination of classifiers and therefore B is wrong. Finally, regardless of regularization parameter, logistic regression is a fairly inflexible (and therefore biased) model type.

Question 24. Let $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In Figure 8 is given 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Each observation is colored and marked in terms of which cluster it came from in the Gaussian Mixture model.

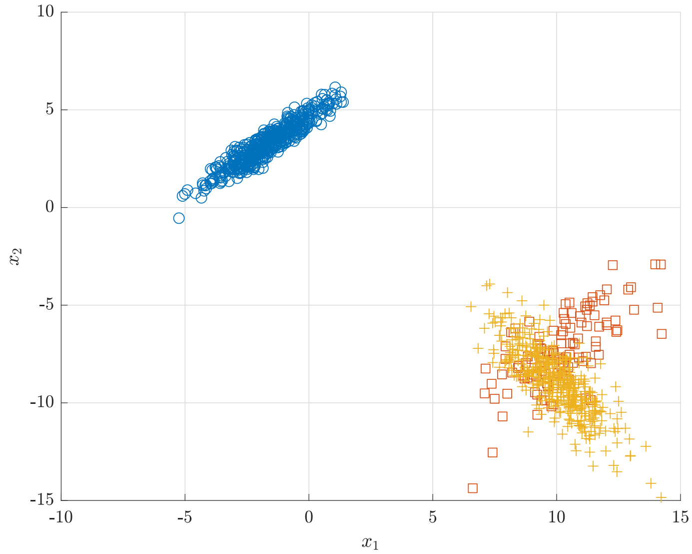

Figure 8: 1000 observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

Which one of the following GMM densities was used to generate the data?

A.

$$p(\boldsymbol{x}) = \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right) + \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right)$$

B.

$$p(\boldsymbol{x}) = \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right)$$

C.

$$p(\boldsymbol{x}) = \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right)$$

D.

$$p(\boldsymbol{x}) = \frac{1}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}-1.5\\3.4\end{bmatrix}, \begin{bmatrix}2.4 & 1.6\\1.6 & 3.0\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}10.1\\-7.2\end{bmatrix}, \begin{bmatrix}1.6 & -1.7\\-1.7 & 2.9\end{bmatrix}\right) + \frac{5}{11}\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix}9.9\\-8.8\end{bmatrix}, \begin{bmatrix}1.6 & 1.3\\1.3 & 1.2\end{bmatrix}\right)$$

E. Don't know.

Solution 24.

D The three components in the candidate GMM densities can be matched to the colored observations by their mean values. Then, by considering the basic properties of the covariance matrices, we can easily rule out all options except A. Alternatively, in Figure 9 is shown the densities for densities corresponding to option B (upper left), C (upper right) and D (bottom center).

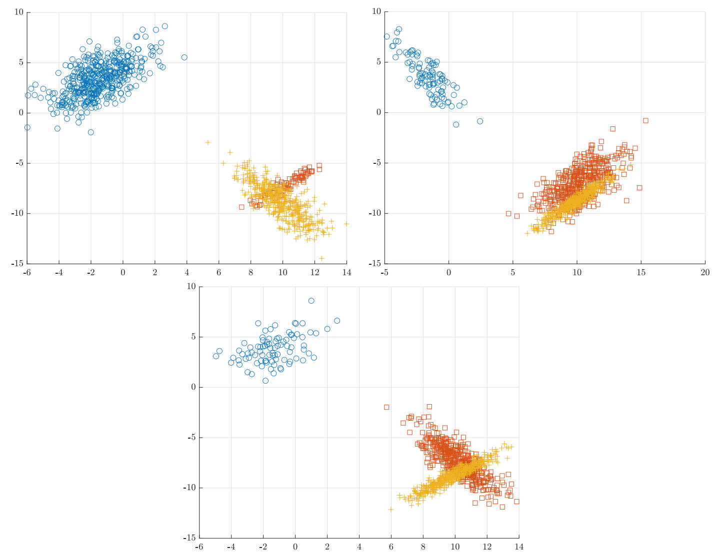

Figure 9: GMM mixtures corresponding to alternative options.

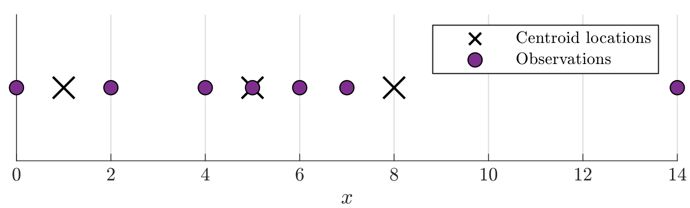

Figure 10: A small 1-dimensional dataset and initial values of centroids.

Question 25. Consider a small dataset comprised of $N = 7$ one-dimensional observations shown as the filled circles in Figure 10.

Suppose a k-means algorithm is applied to the dataset with $K = 3$ and using Euclidean distances. We will assume the location of the centroids are initialized to the values indicated by the crosses in Figure 10. After initialization, the k-means algorithm is evaluated for one step, comprised of assigning observations to centroids and updating the location of the centroids. After the first step, what will be the new location of the centroids?

A. $\mu_1 = 1$, $\mu_2 = \frac{11}{2}$, and $\mu_3 = 14$.

B. $\mu_1 = 4$, $\mu_2 = 6$, and $\mu_3 = 7$.

C. $\mu_1 = 1$, $\mu_2 = 5$, and $\mu_3 = \frac{21}{2}$.

D. $\mu_1 = 2$, $\mu_2 = \frac{11}{2}$, and $\mu_3 = \frac{21}{2}$.

E. Don't know.

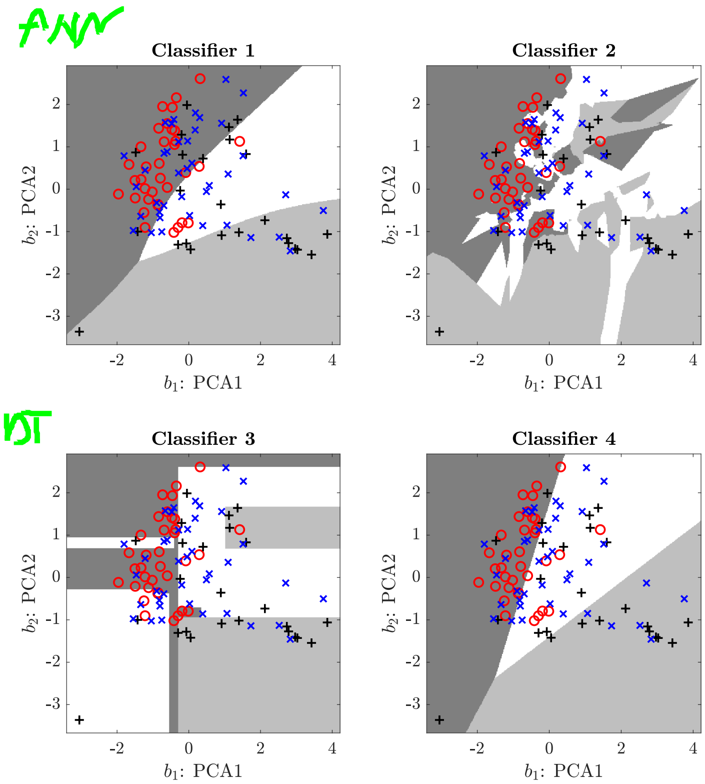

Figure 11: Decision boundaries for four different classifiers trained on the Bicycle rental dataset when projected onto the first two principal components.

Solution 25. The location of the observations and centroids is first read from Figure 10. When this is done, each observation is assigned to the nearest centroid. The observations are thereby partitioned into the three clusters $\{0, 2\}$, $\{4, 5, 6\}$, $\{7, 14\}$.

The new location of the centroids are simply the mean of the observation in each of these three sets. Doing this, we see C is the correct answer.

Question 26. We will consider a subset of the Bicycle rental dataset (described in Table 1) after it has been projected onto the first two principal components b_1 and b_2 given in Equation (1), thereby giving rise to a smaller two-dimensional dataset.

We will consider the following four classifiers:

MREG: Multinomial regression

ANN: Artificial neural network with 5 hidden units

CT: Classification tree with regular axis-aligned splits ($b_i < c$)

KNN: K-nearest neighbours with $K = 3$

Suppose the classifiers are trained on the two-dimensional dataset and the decision boundary for each of the four classifiers is given in Figure 11. Which one of the following statements is correct?

A. Classifier 1 corresponds to ANN,
Classifier 2 corresponds to KNN,
Classifier 3 corresponds to CT,
Classifier 4 corresponds to MREG.

B. Classifier 1 corresponds to **CT**,
Classifier 2 corresponds to **MREG**,
Classifier 3 corresponds to **KNN**,
Classifier 4 corresponds to **ANN**.

C. Classifier 1 corresponds to **MREG**,
Classifier 2 corresponds to **CT**,
Classifier 3 corresponds to **KNN**,
Classifier 4 corresponds to **ANN**.

D. Classifier 1 corresponds to **KNN**,
Classifier 2 corresponds to **ANN**,
Classifier 3 corresponds to **CT**,
Classifier 4 corresponds to **MREG**.

E. Don't know.

Solution 26. To solve this problem, we have to use our intuition about what the typical decision boundaries for the different methods look like:

- A KNN method will have decision boundaries dictated by the nearest neighbors. That is, points (x, y) where the nearest K neighbors are in one class must be in the same class and therefore the boundaries will be fairly complex and respect the data distribution well.
- A decision tree has axis aligned splits, therefore the boundaries must be vertical or horizontal
- A multivariate regression model must have linear boundaries
- An artificial neural network with few hidden units can have some non-linearity, but otherwise have boundaries of limited complexity and consisting of relatively simple shapes

It is easy to see this rules out all but option A.

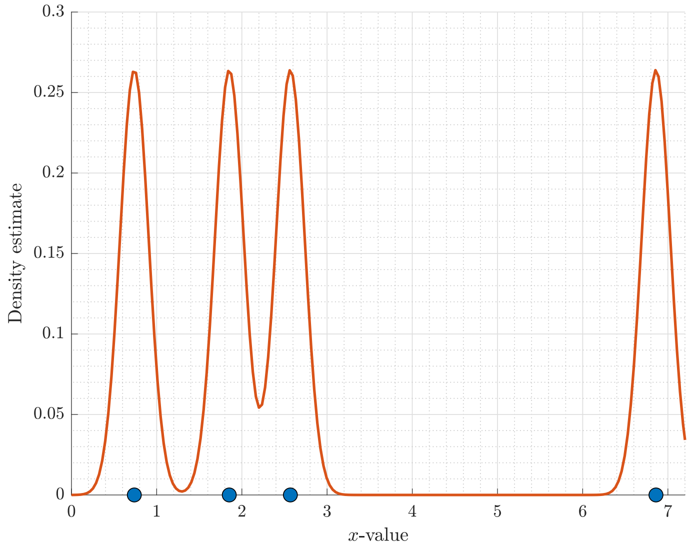

Figure 12: Plot of the density function of a kernel density estimator applied to a 1-dimensional dataset using a Gaussian kernel with kernel width $\lambda = 0.168$. Only a subset of the dataset, indicated by the circles, is shown.

Question 27. A small 1-dimensional dataset of N observations, along with the kernel density estimate, is shown in Figure 12. The kernel is the usual Gaussian kernel with kernel width $\lambda = 0.168$ (i.e., the individual Gaussian components in the KDE have variance $\sigma^2 = \lambda^2$). Note the x-axis has been truncated so not all observations are shown. How many observations were in the dataset?

A. $N = 9$

B. $N = 6$

C. $N = 21$

D. $N = 17$

E. Don't know.

Solution 27. In Figure 12 we see the density $p(x)$ of the KDE. Recall the general formula for a KDE:

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{N}(x|x_i,\sigma^2).$$

Since the kernel width $\lambda = \sigma$ is fairly small relative to the component distance, the density at an x-value corresponding to the peak of one of the components, i.e. x_i, will only be determined by the density of that component (and none of the other ones). In other words we get:

$$\begin{aligned} p(x_i) &\approx \frac{1}{N}\mathcal{N}(x_i|x_i,\sigma^2) \\ &= \frac{1}{N}\frac{1}{\sqrt{2\pi}\sigma}. \end{aligned}$$

We can read off $p(x_i) \approx 0.26$ from the figure and solve to find $N = \frac{1}{\sqrt{2\pi}\sigma p(x)}$. By doing so we see that A is correct.