Technical University of Denmark

Written examination: 28th May 2015, 9 AM - 1 PM. Page 1 of 12 pages.

Course name: Introduction to machine learning and data mining.

Course number: 02450.

Aids allowed: All aids permitted.

Exam duration: 4 hours.

Weighting: The individual questions are weighted equally.

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

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

Please write your name and student number clearly and hand in the present page (page 1) as your answer of the written test. Other pages will not be considered.

Answers:

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

Name: ______________________________

Student number: ____________________

HAND IN THIS PAGE ONLY

No.	Attribute description	Abbrev.
x_1	Student's sex (0: F, 1: M)	SEX
x_2	Student's age (15 to 22)	AGE
x_3	Mother's education (0: none, 1: 4th grade, 2: 5th to 9th grade, 3: secondary education or 4: higher education)	MEDU
x_4	Weekly study time in hours (1: 0 to 2, 2: 2 to 5, 3: 5 to 10, or 4: +10)	STUDYTIME
x_5	Number of past class failures (n if n from 1 to 3 else 4)	FAILURES
x_6	In romantic relationship (0: no, 1: yes)	ROMANTIC
x_7	Going out with friends (1: low to 5: high)	GOOUT
x_8	number of school absences (0 to 93 days)	ABSENCES
y	Final grade (0: low to 20: high)	GRADE

Table 1: Attributes of the *Student* dataset. The dataset includes 8 attributes ($x_1, \ldots, x_8$) of 395 students and their final grade. The purpose is to evaluate how various factors (such as being in a romantic relationship) affects the final grade.

Question 1. Consider the *Student*[1] dataset of table 1. Which of the following statements are true?

A. The most suitable way to apply logistic regression to predict if y is greater than 12 is to first remove the binary features from the dataset

B. The most suitable method to predict the ROMANTIC variable for females is using association mining

C. MEDU is ordinal discrete

D. ABSENCES is interval but not ratio

E. Don't know.

Solution 1.

A There are no good reasons to remove binary features in order to apply logistic regression

B Predicting ROMANTIC for females is a classification task, possibly only on the female subset of the data

C MEDU is indeed ordinal discrete

D ABSENCES is ratio

[1] Dataset obtained from http://archive.ics.uci.edu/ml/datasets/Student+Performance. Notice the dataset has been pre-processed for this exam.

Question 2. A principal component analysis is carried out on the *Student* dataset based on the attributes $x_1, \cdots, x_8$ found in table 1. The data is standardized by (i) substracting the mean and (ii) dividing each column by it's standard deviation to obtain the standardized matrix $\tilde{\boldsymbol{X}}$. A singular value decomposition is then carried out on the standardized matrix to obtain the decomposition $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \tilde{\boldsymbol{X}}$ where

$$\boldsymbol{S} = \begin{bmatrix} 25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 23 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 22 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 18 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 17 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 15 \end{bmatrix}$$

$$\boldsymbol{V} =$$

$$\begin{bmatrix} 0.1 & -0.6 & 0.2 & -0.1 & 0.3 & -0.5 & -0.1 & -0.5 \\ 0.5 & 0.2 & -0.0 & 0.1 & -0.1 & -0.7 & 0.0 & 0.4 \\ -0.3 & 0.0 & 0.7 & 0.1 & 0.1 & -0.0 & -0.5 & 0.4 \\ -0.3 & 0.6 & -0.1 & 0.4 & 0.1 & -0.3 & -0.4 & -0.5 \\ 0.6 & -0.1 & -0.2 & 0.0 & 0.1 & 0.4 & -0.7 & -0.0 \\ 0.3 & 0.4 & 0.3 & -0.3 & 0.7 & 0.1 & 0.2 & -0.1 \\ 0.3 & -0.1 & 0.3 & 0.8 & -0.0 & 0.2 & 0.3 & -0.1 \\ 0.3 & 0.3 & 0.5 & -0.3 & -0.6 & 0.1 & 0.0 & -0.4 \end{bmatrix}.$$

Notice the entries of the matrices have been rounded. Which one of the following statements is true?

A. The first four principal components accounts for more than 70% of the variance

B. An observation with a large projection onto the second principal component can be described as a *studious romantically involved female* (i.e. high STUDYTIME and ROMANTIC and a low value of SEX)

C. Since the variance is very similar for all 8 principal components this implies that any projection onto a single principal component will not be sufficient to predict the grade y.

D. The 3 principal components with the least variance account for less than 20% of the variance

E. Don't know.

Solution 2. Recall the variance of e.g. the first four components are

$$\text{var.} = \frac{\sum_{i=1}^{4} S_{ii}^2}{\sum_{j=1}^{8} S_{jj}^2}$$

Then the variance of the first four components is: 0.651 and the variance of the three last components: 0.246. We cannot say if any single principal component is sufficient to predict y by looking at the variances alone. This leaves option B which can be seen to be true by inspection.

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9
o_1	0.00	3.85	4.51	4.39	4.08	3.97	2.18	3.29	5.48
o_2	3.85	0.00	2.19	3.46	3.66	3.93	3.15	3.47	4.11
o_3	4.51	2.19	0.00	3.70	4.30	4.83	3.86	4.48	4.19
o_4	4.39	3.46	3.70	0.00	1.21	3.09	4.12	3.22	3.72
o_5	4.08	3.66	4.30	1.21	0.00	2.62	4.30	2.99	4.32
o_6	3.97	3.93	4.83	3.09	2.62	0.00	4.15	1.29	3.38
o_7	2.18	3.15	3.86	4.12	4.30	4.15	0.00	3.16	4.33
o_8	3.29	3.47	4.48	3.22	2.99	1.29	3.16	0.00	3.26
o_9	5.48	4.11	4.19	3.72	4.32	3.38	4.33	3.26	0.00

Table 2: The pairwise Euclidian distances, $d(o_i, o_i) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \sqrt{\sum_{k=1}^{M}(x_{ik} - x_{jk})^2}$ between 9 observations from the *Student* dataset (recall $M = 8$). Each observation o_i corresponds to a row of the student matrix $\boldsymbol{X}$ of table 1 (the matrix has been normalized). The colors indicate classes such that the blue observations $\{o_1, o_2, o_3\}$ belongs to class C_1 (low GRADE), the red observations $\{o_4, o_5, o_6\}$ belongs to class C_2 (medium GRADE) and the black observations $\{o_7, o_8, o_9\}$ belongs to class C_3 (high GRADE).

Question 3. Consider the distances in table 2. The class labels C_1, C_2, C_3 (corresponding to $\{o_1, o_2, o_3\}$, $\{o_4, o_5, o_6\}$ and $\{o_7, o_8, o_9\}$) will be predicted using a k-nearest neighbour classifier based on the distances in table 2. Suppose we use leave-one-out cross validation (i.e. the observation that is being predicted is left out) and a 1-nearest classifier (i.e. $k = 1$). What is the error rate for the $N = 9$ observations?

A. error rate $= \frac{3}{9}$

B. error rate $= \frac{4}{9}$

C. error rate $= \frac{5}{9}$

D. error rate $= \frac{6}{9}$

E. Don't know.

Solution 3. The true accuracy is 0.444444444444 or 4/9. This is easy to see by going through table 2 and notice the "wrongly" classified observations are o_1, o_6, o_7, o_8 (they are paired as (o_1, o_7), (o_6, o_8), (o_7, o_1) and (o_8, o_6)).

Question 4. Consider the distances in table 2 and suppose we wish to apply mixture modelling and we use the normal density as the mixture distributions:

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \sigma) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \sigma) = (2\pi\sigma^2)^{-\frac{M}{2}} e^{-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}}$$

Suppose we wish to compute the density at o_9 based on a mixture model of $K = 8$ components, the parameters $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_8$ of each component is taken to be the position of the observations $o_1, \dots, o_8$ and the components are weighted equally. Suppose we set $\sigma = 5$, what is the probability density at the *last* observation o_9?

A. $p(o_9) = \frac{1}{8(\pi 50)^4}\left(e^{\frac{-5.48^2}{50}} + e^{\frac{-4.11^2}{50}} + e^{\frac{-4.19^2}{50}} + e^{\frac{-3.72^2}{50}} + e^{\frac{-4.32^2}{50}} + e^{\frac{-3.38^2}{50}} + e^{\frac{-4.33^2}{50}} + e^{\frac{-3.26^2}{50}}\right)$

B. $p(o_9) = \frac{1}{(\pi 50)^8}\left(e^{\frac{-5.48}{50}} + e^{\frac{-4.11}{50}} + e^{\frac{-4.19}{50}} + e^{\frac{-3.72}{50}} + e^{\frac{-4.32}{50}} + e^{\frac{-3.38}{50}} + e^{\frac{-4.33}{50}} + e^{\frac{-3.26}{50}}\right)$

C. $p(o_9) = \frac{1}{8(\pi 50)^4}\left(e^{\frac{-3.85^2}{50}} + e^{\frac{-4.51^2}{50}} + e^{\frac{-4.39^2}{50}} + e^{\frac{-4.08^2}{50}} + e^{\frac{-3.97^2}{50}} + e^{\frac{-2.18^2}{50}} + e^{\frac{-3.29^2}{50}} + e^{\frac{-5.48^2}{50}}\right)$

D. $p(o_9) = \frac{1}{(\pi 50)^4} \exp\left(\frac{-5.48}{50} + \frac{-4.11}{50} + \frac{-4.19}{50} + \frac{-3.72}{50} + \frac{-4.32}{50} + \frac{-3.38}{50} + \frac{-4.33}{50} + \frac{-3.26}{50}\right)$

E. Don't know.

Solution 4. Options B and D are not properly normalized by the number of mixture components (D is also of the wrong functional form). Option C uses the wrong distances, namely the distances from observation o_1 to the other elements o_i. Accordingly option A is the correct answer.

Question 5. We wish to compute the *average relative KNN density* (a.r.d) of observation o_1 of table 2 using the distances given in the table. Letting $d(\boldsymbol{x}, \boldsymbol{y})$ denote the Euclidian distance metric the a.r.d. is defined as

$$\begin{aligned} \text{density}(\boldsymbol{x}, K) &= \frac{1}{\frac{1}{K}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},K)} d(\boldsymbol{x}, \boldsymbol{y})} \\ \text{a.r.d}(\boldsymbol{x}, K) &= \frac{\text{density}(\boldsymbol{x}, K)}{\frac{1}{K}\sum_{\boldsymbol{z}\in N(\boldsymbol{x},K)} \text{density}(\boldsymbol{z}, K)}, \\ N(\boldsymbol{x}, K) &: \text{set of } K\text{-nearest neighbours of } \boldsymbol{x}. \end{aligned}$$

What is the a.r.d. of observation o_1 using $K = 2$ nearest neighbours?

A. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.868$

B. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.434$

C. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.569$

D. a.r.d$(\boldsymbol{x} = o_1, K = 2) \approx 0.502$

E. Don't know.

Solution 5. The nearest neighbour of o_1 is o_7, o_8 and the nearest neighbours of o_7 is o_1, o_2 and for o_8 it is

o_5, o_6. The densities are

$$\begin{aligned}\text{density}(o_1, K=2) &= 0.36563071298\\ \text{density}(o_7, K=2) &= 0.375234521576\\ \text{density}(o_8, K=2) &= 0.467289719626\end{aligned}$$

from which it follows

$$\begin{aligned}\text{a.r.d.}(o_1, K=2) &= \frac{0.36563071298}{\frac{1}{2}(0.375234521576 + 0.467289719626)}\\ &= 0.867941111008\end{aligned}$$

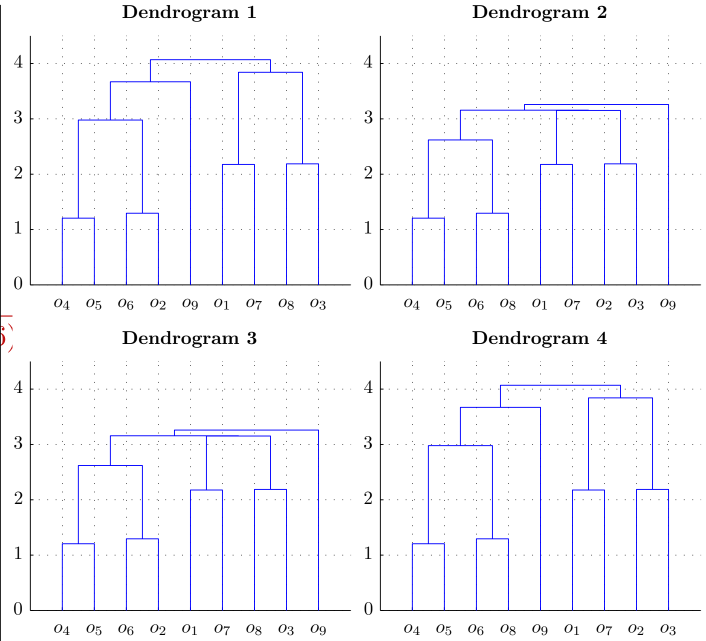

Figure 1: Proposed hierarchical clustering of the 9 observations considereded in table 2

Question 6. In table 2 is given the pairwise distances between 9 observations. A hierarchical clustering is applied to these nine observations using *minimum* linkage. Which of the dendrograms shown in fig. 1 corresponds to the clustering?

A. Dendrogram 1.

B. Dendrogram 2.

C. Dendrogram 3.

D. Dendrogram 4.

E. Don't know.

Solution 6. The true answer is B, dendrogram 2. Considering the distances between o_2 and o_6 allow us to rule out dendrograms 1 and 3. Then considering the minimum distance between observation o_9 and o_8 is 3.26 allow us to rule out dendrogram 4. This leave only option *B*.

Question 7.

In table 2 is given the pairwise euclidian distances between 9 observations of the *Student* dataset of table 1. Suppose we wish to train a Gaussian Mixture-model (GMM) using the EM-algorithm with $K = 2$ clusters. The cluster centers $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are initialized to coincide with the location of observations o_1 and o_4 respectively, the covariance matrices are initialized to be the unit matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}$ and the prior class-probability is selected to be $w_1 = 0.2$, $w_2 = 0.8$. The

EM algorithm is applied to compute the updated assignment of points to classes. Assuming the dimensionality of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ is $M = 8$, with what probability is point o_3 assigned to class C_1?

Hint: Notice the multivariate normal distribution for e.g. class C_1 becomes:

$$\mathcal{N}(o_i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = \frac{1}{(2\pi)^{\frac{M}{2}}} \exp\left(\frac{-d(o_i, o_1)^2}{2}\right)$$

A. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0089$

B. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0347$

C. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0003$

D. $p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \approx 0.0013$

E. Don't know.

Solution 7. The probability can be computed by using Bayes rule

$$\begin{aligned} & p(z_3 = C_1 | o_3, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) \\ & \quad = \frac{\mathcal{N}(o_3; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) w_1}{\mathcal{N}(o_3; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) w_1 + \mathcal{N}(o_3; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) w_2} \\ & \quad = \frac{w_1 e^{\frac{-d(o_3, o_1)^2}{2}}}{w_1 e^{\frac{-d(o_3, o_1)^2}{2}} + w_2 e^{\frac{-d(o_3, o_4)^2}{2}}} \\ & \quad = 0.00891253249006 \end{aligned}$$

Question 8. In table 2 is given the pairwise euclidian distances between 9 observations from the *Student* dataset of table 1. Suppose the Euclidian norm of observations o_1 and o_2 is:

$$\|o_1\| = \sqrt{\sum_{k=1}^{M} x_{1k}^2} = 2.99, \quad \|o_2\| = \sqrt{\sum_{k=1}^{M} x_{2k}^2} = 2.26$$

What can be concluded about the extended Jaccard similarity of these two observations? (Hint: recall for vectors $\boldsymbol{x}$, $\boldsymbol{y}$ that $\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\boldsymbol{x}^\top \boldsymbol{y}$)

A. $\text{EJ}(o_1, o_2) \approx -0.0523$

B. $\textbf{EJ}(o_1, o_2) \approx -0.0268$

C. $\text{EJ}(o_1, o_2) \approx -0.0261$

D. $\text{EJ}(o_1, o_2) \approx -0.1052$

E. Don't know.

Solution 8. Notice the inner product can be recovered as

$$o_1^\top o_2 = \frac{\|o_1\|_2^2 + \|o_2\|_2^2 - d(o_1, o_2)^2}{2} = -0.3874$$

and the definition of the extended Jaccard similarity is

$$\text{EJ}(o_1, o_2) = \frac{o_1^\top o_2}{\|o_1\|_2^2 + \|o_2\|_2^2 - o_1^\top o_2}$$

GRADE	less than 10	from 10 to 12	more than 12
SEX=F	55	44	88
SEX=M	75	59	74

Table 3: Number of students of the two sexes in the three classes of GRADE. For instance, there are 44 females with a grade between 10 and 12

Question 9.

Consider the *Student* dataset of table 1. Suppose the variable GRADE is divided into three classes depending on whether GRADE is (i) < 10 (ii) between $10-12$ (iii) > 12, thereby creating a 3-class classification problem. Suppose we attempt to train a decision tree and we initially consider a split on the variable SEX. If the number of students in the three classes of either sex is as listed in table 3, what is the *impurity gain* Δ of the split if the *Gini* impurity measure is used?

A. $\Delta \approx 0.00329$

B. $\Delta \approx 0.65548$

C. $\Delta \approx 0.64415$

D. $\Delta \approx 0.00497$

E. Don't know.

Solution 9. The relevant definitions can be found in section 4.3 of Tan et.al. We first need the frequencies for all students as well as for the males and females. Letting C_1, C_2 and C_3 denote the low, medium and high classes:

$$\begin{aligned}\text{Female: } & p(C_1|F) = 55/187, p(C_2|F) = 44/187, p(C_3|F) = 88/187\\ \text{Male: } & p(C_1|M) = 75/208, p(C_2|M) = 59/208, p(C_3|M) = 74/208\\ \text{All: } & p(C_1|A) = 130/395, p(C_2|A) = 103/395, p(C_3|A) = 162/395\end{aligned}$$

From this we can compute the impurity: $I(x) = 1 - \sum_i p(i|x)^2$

$$\begin{aligned}\text{Female: } & I(F) = 0.636678200692\\ \text{Male: } & I(M) = 0.662953032544\\ \text{All: } & I(A) = 0.655484697965.\end{aligned}$$

Then combining these we have

$$\begin{aligned}\Delta &= I(A) - (187/395)I(F) - (208/395)I(M)\\ &= 0.00497063644952.\end{aligned}$$

Question 10.

Consider the $N = 9$ students from table 2 and assume the data has been processed to the 9×6 binary matrix described in table 4. Suppose we only consider the first two features f_1, f_2 and train a Naïve-Bayes classifier to distinguish between class C_1, C_2 and C_3 based on only these two features. If an observation has $f_1 = 1, f_2 = 0$, what is the probability the observation belongs to class C_3 according to the Naive-Bayes classifier?

A. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.143$

B. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.133$

C. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.375$

D. $p_{NB}(C_3|f_1 = 1, f_2 = 0) = 0.125$

E. Don't know.

Solution 10. True answer is: 0.125. This can be found by computing the per-class probabilities

$$\begin{aligned}p(f_1 = 1|C_1) = 1,\ & p(f_2 = 0|C_1) = 1/3\\ p(f_1 = 1|C_2) = 2/3,\ & p(f_2 = 0|C_2) = 2/3\\ p(f_1 = 1|C_3) = 1/3,\ & p(f_2 = 0|C_3) = 1/3\end{aligned}$$

The class label priors are the same $p(C_i) = \frac{1}{3}$ and so the Naive-Bayes estimate is

$$p_{NB}(C_3|f_1 = 1, f_2 = 0) = \frac{p(f_1 = 1|C_3)p(f_2 = 0|C_3)p(C_3)}{\sum_{i=1}^{3} p(f_1 = 1|C_i)p(f_2 = 0|C_i)p(C_i)} = \frac{1}{8}$$

	f_1	f_2	f_3	f_4	f_5	f_6
o_1	1	0	1	0	1	0
o_2	1	1	1	1	1	0
o_3	1	1	0	0	1	0
o_4	1	1	1	0	0	1
o_5	1	0	1	0	0	1
o_6	0	0	1	1	0	1
o_7	1	1	1	1	1	1
o_8	0	0	1	1	1	1
o_9	0	1	0	1	0	1

Table 4: Processed version of the $N = 9$ observations of table 2. For each student we record 6 features corresponding to $f_1 : SEX$, $f_2 : AGE$, $f_3 : MEDU$, $f_4 : STUDYTIME$, $f_5 : GOOUT$ and $f_6 : ABSENCES$. The features are binarized by thresholding at the mean value. The categories still reflect GRADE, i.e. the blue category C_1 (o_1, o_2, o_3) is low GRADE, the red category C_2 (o_4, o_5, o_6) is medium GRADE and the black category C_3 (o_7, o_8, o_9) is high GRADE.

Question 11. Suppose we consider the binary matrix in table 4 as a market-basket problem consisting of $N = 9$ "transactions" $o_1, \dots, o_9$ and $M = 6$ "items" $f_1, \dots, f_6$. Which of the following options represents all itemsets with support greater than 0.4?

A. $\{f_1\}$, $\{f_3\}$, $\{f_6\}$

B. $\{f_1\}$, $\{f_2\}$, $\{f_3\}$, $\{f_1, f_3\}$, $\{f_4\}$, $\{f_5\}$, $\{f_6\}$, $\{f_3, f_6\}$

C. $\{f_1\}$, $\{f_2\}$, $\{f_1, f_2\}$, $\{f_3\}$, $\{f_1, f_3\}$, $\{f_4\}$, $\{f_3, f_4\}$, $\{f_5\}$, $\{f_1, f_5\}$, $\{f_3, f_5\}$, $\{f_6\}$, $\{f_3, f_6\}$, $\{f_4, f_6\}$

D. $\{f_1\}$, $\{f_2\}$, $\{f_1, f_2\}$, $\{f_3\}$, $\{f_1, f_3\}$, $\{f_2, f_3\}$, $\{f_1, f_2, f_3\}$, $\{f_4\}$, $\{f_2, f_4\}$, $\{f_3, f_4\}$, $\{f_5\}$, $\{f_1, f_5\}$, $\{f_2, f_5\}$, $\{f_1, f_2, f_5\}$, $\{f_3, f_5\}$, $\{f_1, f_3, f_5\}$, $\{f_4, f_5\}$, $\{f_3, f_4, f_5\}$, $\{f_6\}$, $\{f_1, f_6\}$, $\{f_2, f_6\}$, $\{f_3, f_6\}$, $\{f_1, f_3, f_6\}$, $\{f_4, f_6\}$, $\{f_3, f_4, f_6\}$

E. Don't know.

Solution 11. Recall by chapter 6.1 of Tan et al. the support count is the number of "transactions" containing a given set of items. The problem is then to find all subsets of items that occur in at least 4 of the 9 transactions. These are easily seen to be those in option C and no other.

Question 12. We consider again the $N = 9$ students from table 4 as 6-dimensional binary vectors. Which one of the following statements is true regarding the Jaccard/cosine similarity and the simple matching coefficient?

A. $\mathrm{SMC}(o_2, o_6) > \mathrm{J}(o_2, o_7)$

B. $\mathrm{SMC}(o_2, o_6) > \mathrm{COS}(o_2, o_6)$

C. $\mathbf{COS}(o_2, o_7) > \mathbf{J}(o_2, o_7)$

D. $\mathrm{COS}(o_2, o_6) > \mathrm{COS}(o_2, o_7)$

E. Don't know.

Solution 12. It is easily verified only option C is correct by plugging in the following values:

$$\begin{aligned} \mathrm{SMC}(o_2, o_6) &= 0.333333333333 \\ \mathrm{J}(o_2, o_7) &= 0.833333333333 \\ \mathrm{COS}(o_2, o_6) &= 0.516397779494 \\ \mathrm{COS}(o_2, o_7) &= 0.912870929175 \end{aligned}$$

Question 13. Suppose we consider the binary matrix of table 4 as a market-basket problem consisting of $N = 9$ "transactions" $o_1, \dots, o_9$ and $M = 6$ "items" $f_1, \dots, f_6$. What is the *lift* of the rule $\{f_2, f_4, f_6\} \to \{f_1, f_5\}$ if the lift for a rule $A \to B$ is defined as

$$\mathrm{Lift}(A \to B) = \frac{\mathrm{Confidence}(A \to B)}{\mathrm{Support}(B)}$$

A. $\mathbf{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{9}{8}$

B. $\mathrm{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{8}{9}$

C. $\mathrm{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{6}{9}$

D. $\mathrm{Lift}(\{f_2, f_4, f_6\} \to \{f_1, f_5\}) = \frac{5}{9}$

E. Don't know.

Solution 13. the lift is 1.125. Recall the confidence is defined as (see chapter 6.1 of Tan et al)

$$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

All items f_1, f_2, f_4, f_5, f_6 occur only in 1 transaction, the items f_2, f_4, f_6 only in 2 transactions and finally f_1, f_5 only in 4 transactions. The lift is then:

$$\mathrm{Lift}(A \to B) = \frac{\frac{1/9}{2/9}}{4/9} = \frac{9}{8}$$

Feature(s)	A_{train}	A_{test}
None	0.737	0.574
x_1	0.569	0.526
x_2	0.644	0.563
x_3	0.572	0.52
x_4	0.678	0.581
x_1, x_2	0.625	0.589
x_1, x_3	0.805	0.645
x_1, x_4	0.709	0.607
x_2, x_3	0.535	0.5
x_2, x_4	0.634	0.608
x_3, x_4	0.609	0.499
x_1, x_2, x_3	0.738	0.623
x_1, x_2, x_4	0.738	0.614
x_1, x_3, x_4	0.763	0.596
x_2, x_3, x_4	0.547	0.525
x_1, x_2, x_3, x_4	0.579	0.552

Table 5: The *accuracy* on a training set A_{train} and test set A_{test} of linear regression models (predictions are made by thresholding at 0.5) trained on different subsets of features of the Students dataset of table 1

Question 14. Consider the Students dataset of table 1 and suppose GRADE has been binarized to only take two values and we only consider the first four features x_1, x_2, x_3, x_4. Suppose we wish to examine which subset of these features gives the **highest** accuracy on a test set. In table 5 is shown how different combinations of features give rise to different values of the accuracy on a training and test set for a classifier. Which one of the following statements is true?

A. Forward and backward selection will select the same number of features.

B. Forward selection will select a model with higher accuracy on the test set than backward selection.

C. Backward selection will select *more* features than forward selection.

D. Backward selection will select *less* features than forward selection.

E. Don't know.

Solution 14. Firstly notice the column with the training set accuracy can be disregarded. Forward selection then first selects x_4, then x_2, x_4 and finally x_1, x_2, x_4 and then terminates with an accuracy on the test set of 0.614. Backward selection will first select x_1, x_2, x_3, then x_1, x_3 and terminate with an accuracy of 0.645 on the test set. Accordingly only option D is correct.

Question 15. Consider the attributes ROMANTIC, GOOUT, and GRADE of table 1. Assume each attribute has been binarized by thresholding at the median value giving the 3 binary attributes:

- RO = yes, no: *In a romantic relationship or not*
- GO = yes, no: *Going out in the evening or not*
- GR = high, low: *Has a high or low grade*

and there are thus $2^3 = 8$ possible outcomes. Suppose we are given the following information

$$\begin{aligned} p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}) &= 0.36 \\ p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{y}) &= 0.39 \\ p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{n}) &= 0.47 \\ p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{n}) &= 0.48 \\ \text{and} \qquad p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{y}) &= 0.23 \\ p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{y}) &= 0.46 \\ p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{n}) &= 0.11 \\ p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{n}) &= 0.21 \end{aligned}$$

What is then the probability that a student goes out and has a romantic relationship given the student attains high grades?

A. $p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.45$

B. $p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.32$

C. $p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \approx 0.18$

D. $p(\mathbf{RO}{=}\mathbf{y},\mathbf{GO}{=}\mathbf{y}|\mathbf{GR}{=}\mathbf{h}) \approx 0.2$

E. Don't know.

Solution 15. The problem can be solved by applying Bayes theorem:

$$\begin{aligned} p(\mathrm{GR}{=}\mathrm{h}) = {} & p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y})p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{y}) \\ & + p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{n})p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{n}) \\ & + p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{y})p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{y}) \\ & + p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{n},\mathrm{GO}{=}\mathrm{n})p(\mathrm{RO}=\mathrm{n},\mathrm{GO}=\mathrm{n}) \end{aligned}$$

$$\begin{aligned} & p(\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y}|\mathrm{GR}{=}\mathrm{h}) \\ & = \frac{p(\mathrm{GR}{=}\mathrm{h}|\mathrm{RO}{=}\mathrm{y},\mathrm{GO}{=}\mathrm{y})p(\mathrm{RO}=\mathrm{y},\mathrm{GO}=\mathrm{y})}{p(\mathrm{GR}{=}\mathrm{h})} \\ & \approx 0.2 \end{aligned}$$

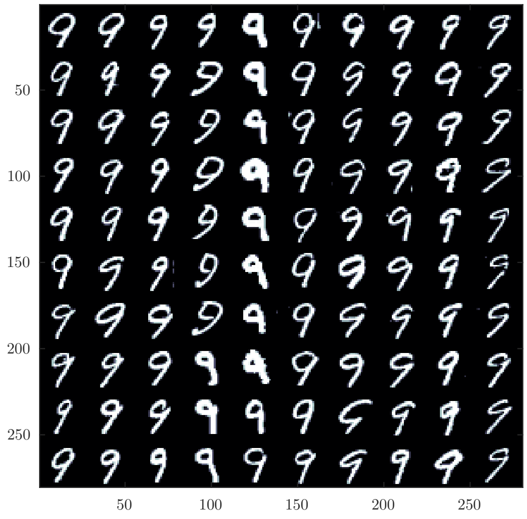

Figure 2: 100 images of the number 9 in a 10 × 10 grid

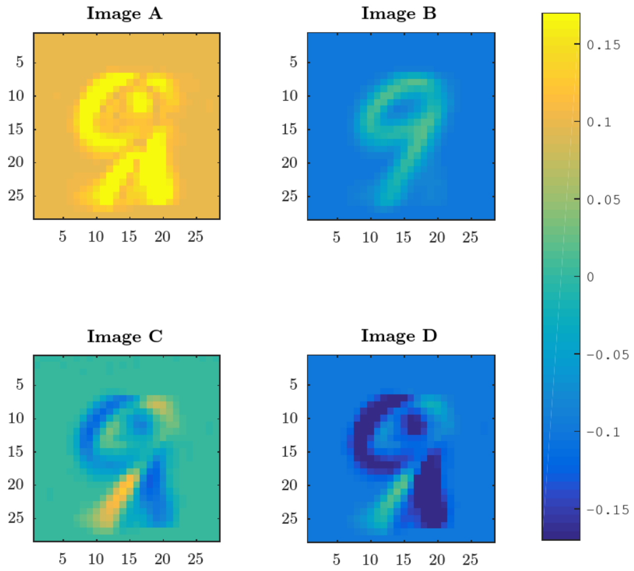

Figure 3: Four candidates for the first principal component of the image dataset of fig. 2. Notice the colorbar indicating coordinate magnitude

Question 16. In fig. 2 is shown a dataset[2] consisting of 100 images of the number 9 each of size 28×28 pixels. Each image is considered as a vector of $28^2 = 784$ coordinates and a principal component analysis is carried out as usual on the dataset. Which one of the four images, Image A, Image B, Image C or Image D in fig. 3 represents the first principal component reshapen as a 28×28 image and plotted using colors to indicate coordinate magnitude? (Hint: Use the colorbar)

A. Image A

B. Image B

C. Image C

D. Image D

E. Don't know.

Solution 16. Notice the instances of the number 9 only vary in pixel intensity in the middle region of the picture (or put in another way, most pixels along the boundary of the image remains zero). We can expect the variance along these boundary pixels to be zero, and this only leaves image C. Alternatively one can observe the large number of non-zero values implies only image C can be normalized.

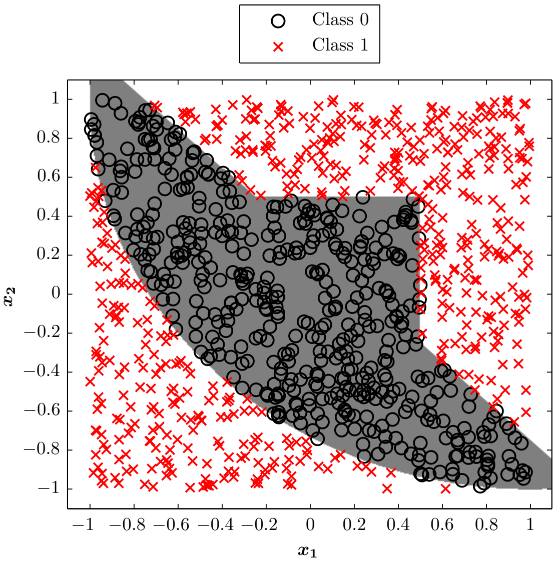

Figure 4: Two-class classification problem

Question 17. Suppose we wish to solve the two-class classification problem in fig. 4 using a classification tree of the form given in fig. 5. What rules, acting on the coordinates $\boldsymbol{x} = (x_1, x_2)$, should be assigned to the three internal nodes A, B and C of the tree to give rise to the indicated decision boundary?

A. $A: \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2,\ \ B: \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 < 2.25$
$C: \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$

B. $A: \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2.25,\ \ B: \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$
$C: \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2$

C. $A: \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2.25,\ \ B: \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2$
$C: \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$

D. $A: \left\|\boldsymbol{x} - \begin{bmatrix}1\\1\end{bmatrix}\right\|_2 > 2,\ \ B: \|\boldsymbol{x}\|_\infty \geq \frac{1}{2}$
$C: \left\|\boldsymbol{x} - \begin{bmatrix}-1\\-1\end{bmatrix}\right\|_1 > 2.25$

E. Don't know.

Solution 17. First consider the point $(0.4, 0.4)$ which should belong to the dark class 0. For this point $\|[0.4, 0.4] - [-1, -1]\|_1 = 1.4 + 1.4 = 2.8$ and so in option B and C it will be classified incorrectly leaving option A and D.

For option A, consider the point $(1, 1)$. Obviously node A will evaluate to false but node B will evaluate to $\|[1, 1] - [-1, -1]\|_1 = 2 + 2 = 4$ and so this node will evaluate to false and the point will incorrectly register as belonging to the dark class 0.

[2] The numbers are a subset of the MNIST dataset obtained from http://yann.lecun.com/exdb/mnist/

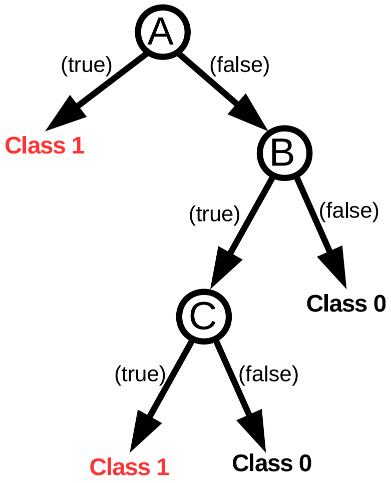

Figure 5: Decision tree with 3 nodes A, B and C

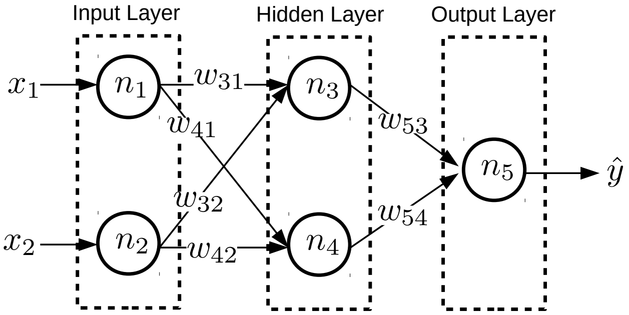

Figure 6: Simple neural network of 6 weights

Question 18. Consider the feedforward neural network shown in fig. 6. The network has no bias weights. Suppose the weights of the neural network after training are

$$\begin{aligned} w_{31} &= 0.05, & w_{41} &= 0, & w_{32} &= 0.1, \\ w_{42} &= -0.05, & w_{53} &= 0.1, & w_{54} &= -10 \end{aligned}$$

and the activation function of all five neurons n_1, n_2, n_3, n_4 and n_5 is the rectified linear unit

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \frac{1}{10}x & \text{otherwise.} \end{cases}$$

Suppose the network is evaluated on input $x_1 = 0.5$, $x_2 = 1$, what is the output?

A. $\hat{y} = 0.5125$

B. $\hat{y} = 0.05125$

C. $\hat{y} = -0.00375$

D. $\hat{y} = 0.0625$

E. Don't know.

Solution 18. To compute the output, initialize $n_1 = f(0.5) = 0.5, n_2 = f(1) = 1$. Then we can compute:

$$n_3 = f(n_1 * 0.05 + n_2 * 0.1) = f(1/8) = 1/8$$
$$n_4 = f(n_1 * 0 + n_2 * (-0.05)) = f(-1/20) = -1/200$$

Then for the output of the neural network we have

$$\hat{y} = n_5 = f(n_3 * 0.1 + n_4 * (-10)) = f(1/16) = 0.0625.$$

Question 19. Consider the classification problem given in fig. 4. Suppose the problem is solved using the following four classifiers

(1NN) A 1-nearest neighbour classifier

(TREE) A decision tree

(LREG) Logistic regression

(NNET) An artificial neural network with four hidden units

All classifiers are using only the two attributes x_1, x_2, corresponding to the position of each observation, as well as the class label. Which of the descriptions (1NN),(TREE),(LREG),(NNET) matches the boundariers of the four plots (Plot A, B, C and D) indicated in fig. 7?

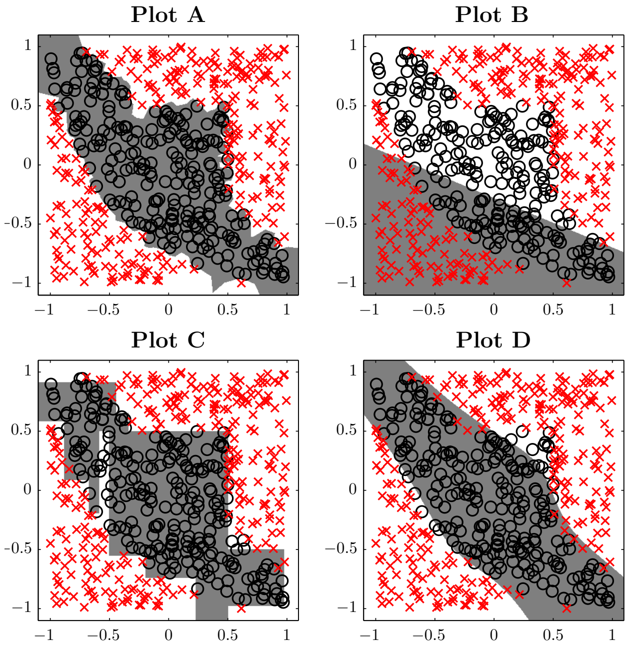

Figure 7: Four classifiers applied to a two-class classification problem

A. Plot A is 1NN, Plot B is LREG, Plot C is NNET, Plot D is TREE.

B. Plot A is 1NN, Plot B is LREG, Plot C is TREE, Plot D is NNET.

C. Plot A is 1NN, Plot B is NNET, Plot C is TREE, Plot D is LREG.

D. Plot A is LREG, Plot B is 1NN, Plot C is TREE, Plot D is NNET.

E. Don't know.

Solution 19. Plot A is a 1NN classifier (notice all points are correctly classified), *B* is the only classifier with a linear boundary and must be logistic regression and *C* has the "boxes" characteristic for a decision tree.

Question 20. Suppose the 2D dataset shown in fig. 8 was generated from a Gaussian mixture-model (GMM) with three components. Which of the following is the most likely equation of the density of the mixture model?

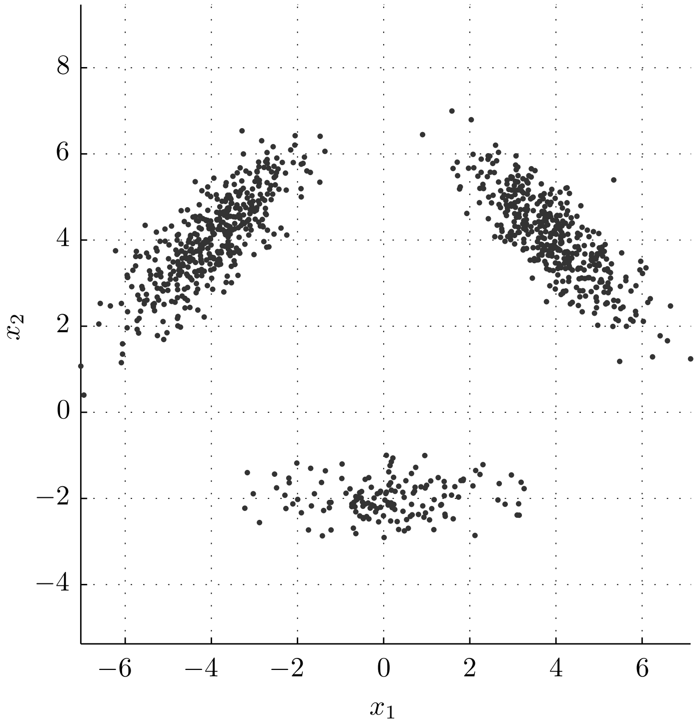

Figure 8: Scatter plot of observations

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.0 & -0.8 \\ -0.8 & 1.0 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.8 & 0.0 \\ 0.0 & 0.2 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix},$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -4 \\ 4 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}, \boldsymbol{\mu}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix},$$

A. The density is:

$$p(\boldsymbol{x}) = 0.15\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_3)$$

B. The density is:

$$p(\boldsymbol{x}) = 0.15\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1) + 0.425\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_3)$$

C. The density is:

$$p(\boldsymbol{x}) = 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_1) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$$

D. The density is:

$$p(\boldsymbol{x}) = 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_2) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1) + 0.33\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_3)$$

E. Don't know.

Solution 20. Options C and D can be ruled out because the densities cannot be weighted equally in the true mixture distribution. Then simply recall a covariance matrix with negative off-diagonal elements (such as $\boldsymbol{\Sigma}_1$) corresponds to a density slanted in the top-left to the bottom-right direction.

Question 21.

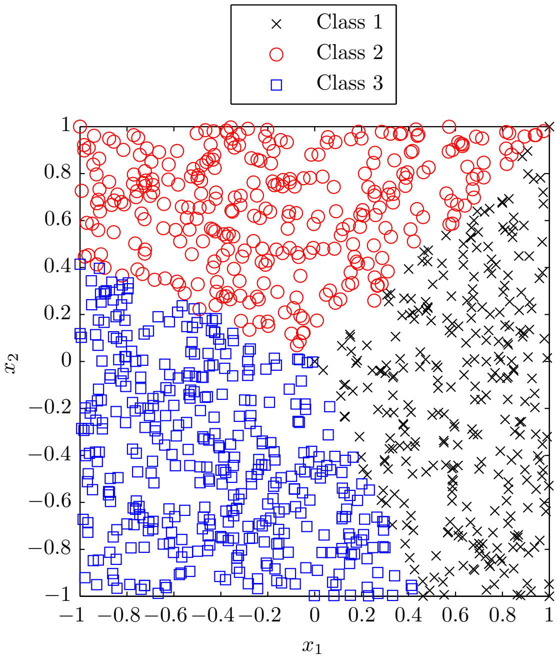

Figure 9: Observations labelled with the most probable class

Consider a multinomial regression classifier for a three-class problem where for each point $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$ we compute the class-probability by first computing the intermediate values

$$y_1 = \boldsymbol{w}_1^\top \boldsymbol{x}, \quad y_2 = \boldsymbol{w}_2^\top \boldsymbol{x}, \quad y_3 = \boldsymbol{w}_3^\top \boldsymbol{x}$$

and then combine these to the per-class probability

$$P(\hat{y} = k) = \frac{e^{y_k}}{e^{y_1} + e^{y_2} + e^{y_3}}$$

A dataset of $N = 1000$ points where each point is labelled according to the maximum class-probability is shown in fig. 9. Which setting of the weights was used? (Hint: consider points (x_1, x_2) where either x_1 or x_2 is zero)

A. $\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\boldsymbol{w}_3 = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}$

B. $\boldsymbol{w}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, $\boldsymbol{w}_3 = \begin{bmatrix} 0.8 \\ 0.8 \end{bmatrix}$

C. $\boldsymbol{w}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\boldsymbol{w}_3 = \begin{bmatrix} 0.3 \\ 0.3 \end{bmatrix}$

D. $\boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $\boldsymbol{w}_3 = \begin{bmatrix} -0.8 \\ -0.8 \end{bmatrix}$

E. Don't know.

Solution 21. Consider for instance $x_1 = -1$ and $x_2 = 0$. Then we have for the four settings of weights:

$$\begin{aligned} A &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & -0.3 \end{bmatrix} \\ B &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -0.8 \end{bmatrix} \\ C &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -0.3 \end{bmatrix} \\ D &: \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0.8 \end{bmatrix} \end{aligned}$$

Next, since the multinomial regression function preserves order we need only consider the maximal value. Accordingly A is classified to class 2, B and C to class 1 and only D correctly to class 3.

Question 22. Consider a two-dimensional data set consisting of $N = 8$ observations shown in fig. 10. The dataset contains three classes indicated by the black crosses (class 1), red circles (class 2) and blue squares (class 3). In the figure, the decision boundaries for four K-nearest neighbor classifiers (KNN) are shown. Which of the plots correspond to the $K = 3$ nearest-neighbour classifier assuming ties are broken by assigning to the *nearest* neighbour's class?

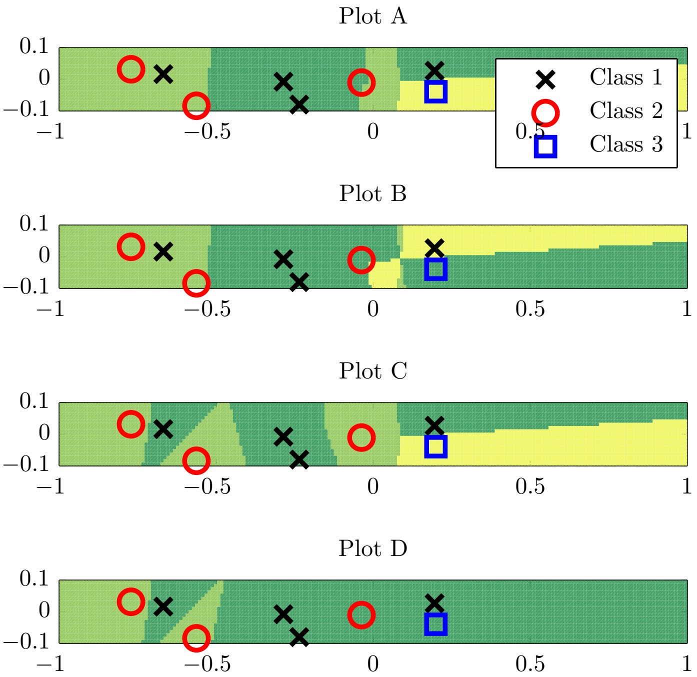

Figure 10: Decision boundaries for four KNN classifiers.

A. Plot A

B. Plot B

C. Plot C

D. Plot D

E. Don't know.

Solution 22. Lets focus on the blue square (Class 3). The three nearest neighbours at the blue square represents all three classes and (since ties are broken by assigning to the nearest class) the blue square should belong to it's own class. Evidently this rules out D (only two classes) and B (why should the blue class extend so far to the left?).

Consider then plot C. The left-most black cross and red circle has the same 3-nearest neighbours consisting of one black cross and two red circles, accordingly they should be in the same class. This rules out option C leaving only A.

X	1	3	4	6	7	8	13	15

Table 6: Simple 1-dimensional dataset comprised of $N = 8$ observations.

Question 23. Consider the 1-dimensional data set comprised of $N = 8$ observations shown in table 6. Which one of the following clusterings corresponds to converged state of a K-means algorithm using standard Euclidian distances?

A. {1, 3, 4}, {6, 7, 8}, {13, 15}

B. {1}, {3, 4, 6}, {7, 8}, {13, 15}

C. {1, 3, 4}, {6, 7}, {8, 13, 15}

D. {1, 3, 4}, {6, 7}, {8, 13}, {15}

E. Don't know.

Solution 23. The problem can be solved by explicit calculation, however it is easier solved by drawing the points on a paper and ruling out the clusterings that look the most "odd". For instance:

- For B notice the second cluster has mean 4.33 and the third has mean 7.5. Accordingly $x = 6$ is in the wrong cluster.
- For partition C notice the two last clusters have mean 6.5 and 12 and so $x = 8$ is in the wrong cluster.
- For options D the two last clusters has mean 10.5 and 15 and accordingly $x = 13$ is in the wrong cluster.

It is easy to check the first option has converged.

Question 24. Consider a dataset comprised of two classes as shown in fig. 11. For each observations i, there is an associated value y_i, $i = 1, \ldots, n$, and the curves indicate the density of each class. The two classes are composed of a comparable number of observations. I.e. most of the black observations (the *negative* class) has y-values between -2 and 2 and most of the red observations (the *positive* class) has y values between 2 and 10.

By thresholding at different levels θ, i.e. assign each observation i to class 0 (the predicted negative class) if $y_i \leq \theta$ and otherwise to class 1 (the predicted positive class), one obtains different values of the TP and FN (true positives and false negatives) which in turn allow us to compute the TPR (true positive rate) curve. Which of the true positive rate (TPR) curves A,B,C or D shown in fig. 12 corresponds to the two-class problem of fig. 11?

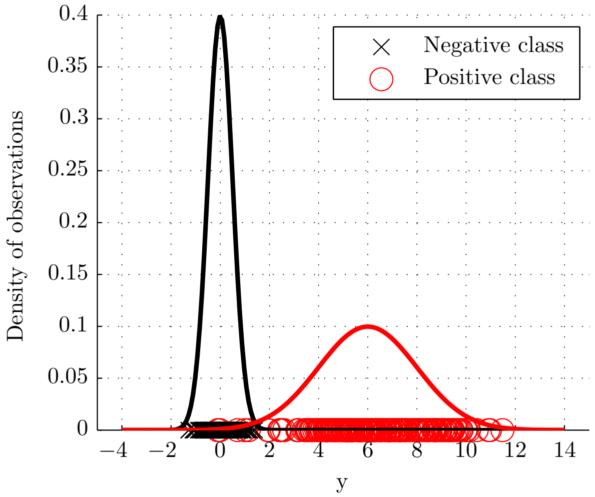

Figure 11: A simple two-class problem. Colors indicate the two classes and the curves indicates the density of each class. Each "point" is simply the value of y_i for observation i.

A. Figure A

B. Figure B

C. Figure C

D. Figure D

E. Don't know.

Solution 24. Recall the true positive rate is defined as

$$\text{TPR} = \frac{\sharp\text{true positives}}{\sharp\text{total positives}}$$

Accordingly we only need to consider the positive class (red circles). Put in a different way, for a given threshold θ we have:

$$\text{TPR} = \frac{\sharp\text{red circles to the right of } \theta}{\sharp\text{red circles}}$$

so if we for instance select $\theta = 2$, the true positive rate should be about 0.95, ruling out all but option C. Alternatively the other options can be ruled out since they reflect changes in the area where we only have black squares (the negative class).

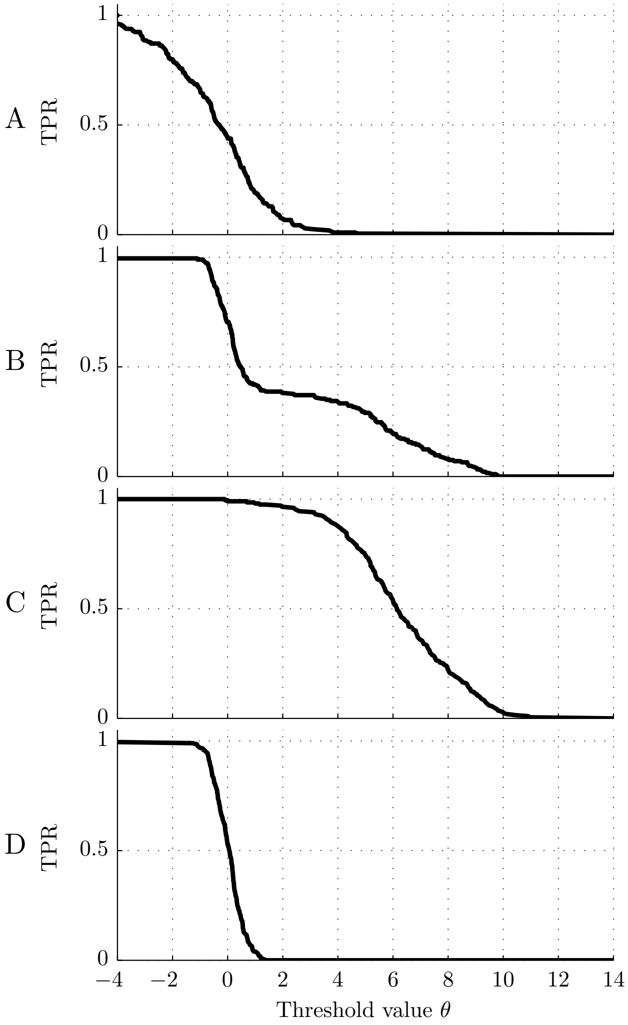

Figure 12: Proposed TPR (true positive rate) curves

Question 25. Suppose for a given problem the true positive rate (TPR) as a function of the threshold θ is as shown in fig. 12(D). Suppose we consider predictions made at a threshold of $\theta = 0$, and suppose we are told that the number of true positives at $\theta = 0$ is $TP = 113$ and the TPR at this threshold is $TPR = 0.55$, what is the (approximate) total number of observations in the positive class?

A. Actual positives: 252

B. Actual positives: 205

C. Actual positives: 159

D. Actual positives: 420

E. Don't know.

Solution 25. This is easily calculated by observing

$$\text{TPR} = \frac{\text{TP}}{\sharp\text{observations in the positive class}}$$

so

$$\sharp\text{positives} = \frac{113}{0.55} \approx 205.45$$

Question 26. Suppose a neural network classifier is applied to a small binary classification problem of only $N = 4$ observations shown in table 7. We attempt to improve the performance by applying AdaBoost (the version in the lecture notes, chapter 15). AdaBoost works by first sampling a new dataset D_1 with replacement, then training a classifier C_1 on D_1 and proceeding with the subsequent steps of the AdaBoost algorithm.

i	x_i	y_i	$C_1(x_i)$
1	50	1	1
2	22	1	0
3	20	0	1
4	76	0	0

Table 7: True values y_i and predictions $C_1(x_i) = \hat{y}_i$ for a neural network classifier C_1 trained on the (subsampled) dataset D_1 (see text) in an AdaBoost iteration.

Suppose in the first iteration of the AdaBoost algorithm a dataset D_1 is selected and the classifier C_1 trained on D_1. The predictions of the classifier C_1 is given in table 7. If AdaBoost is applied for $k = 1$

rounds of boosting what is the resulting (approximate) value for the weights $\boldsymbol{w}$?

A. $\boldsymbol{w} = \begin{bmatrix} 0.072 & 0.428 & 0.428 & 0.072 \end{bmatrix}$

B. $\boldsymbol{w} = \begin{bmatrix} 0.019 & 0.481 & 0.481 & 0.019 \end{bmatrix}$

C. $\boldsymbol{w} = \begin{bmatrix} 0.250 & 0.250 & 0.250 & 0.250 \end{bmatrix}$

D. $\boldsymbol{w} = \begin{bmatrix} 0.130 & 0.370 & 0.370 & 0.130 \end{bmatrix}$

E. Don't know.

Solution 26. The classifier C_1 classifies observation 2 and 3 incorrectly. From the lecture notes we have for a classifier C_i that

$$\varepsilon_i = \left[\sum_{j=1}^{N} w_j I\left(C_i(\boldsymbol{x}_j) \neq y_j\right) \right]$$

$$\alpha_i = \frac{1}{2} \log\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

and accordingly $\varepsilon_1 = \frac{1}{4} \times 2 = \frac{1}{2}$. This gives

$$\alpha_1 = \frac{1}{2} \log \frac{1-\frac{1}{2}}{\frac{1}{2}} = \frac{1}{2} \log 1$$

and so for $\boldsymbol{w}$ we get

$$\boldsymbol{w} \propto \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

and normalizing:

$$\boldsymbol{w} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$$

accordingly option C is correct.

Question 27. Consider a neural network model applied to a dataset of $N = 1000$ observations. Suppose we wish to select both the optimal number of hidden units of the network as well as estimate the generalization error. To simplify the problem, we only consider 4 possible number of hidden units

$$n_{\text{hidden}} = 2, 4, 6, 8.$$

We opt for a two-level cross-validation strategy in which we use an inner loop of K_2-fold cross-validation to estimate the optimal number of units and an outer loop of K_1 fold cross-validation to estimate the generalization error. That is, for each of the K_1 outer folds, the dataset is divided into a validation set $D_{\text{validation}}$ and the remainder D_{CV} is used for the K_2-cross-validation to select the optimal number of neurons for this outer fold. Then, having estimated the optimal number of neurons for this outer fold, we train a new model on D_{CV} and use it to predict the values in $D_{\text{validation}}$ in order to estimate the generalization error. The full generalization error is obtained as the average of the K_1 outer folds.

Suppose we select $K_1 = 5$ and $K_2 = 10$. How many times in total must we *train* a neural network model?

A. 205

B. 200

C. 210

D. 55

E. Don't know.

Solution 27. This can easily be obtained noting for each of the K_1 outer folds we must both (i) train K_2 models on the $L = 4$ different settings of the number of hidden units (ii) train a single new model to estimate the generalization error for this fold. Accoringly the number of trained models is

$$K_1(K_2 L + 1) = 5(10 \cdot 4 + 1) = 205$$