Two hours

MOCK EXAM

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Foundations of Machine Learning

Examination date not specified

Time: Examination time not specified

Marking Scheme Included

Do not publish

Please answer ALL questions.

Each question is worth ten (10) marks.

The use of electronic calculators is permitted provided they are not programmable and do not store text.

---

1. In an election, voters can vote for one of two political parties: the Red Tomato Party and the Blue Berry Party. Votes can be made via postal mail or in-person.

   Surveys suggest that 35% of voters will use postal voting and that 55% of postal voters will choose the Blue Berry Party. In contrast, 52% of in-person voters chose the Red Tomato Party.

   a) What is the probability that any randomly chosen voter will choose the Red Tomato Party? **[2 MARKS]**

      A. 0.4955
      B. 0.4745
      C. 0.4381
      D. 0.4695

      A

      Solution: Let R be the event of a voter choosing the Red Tomato Party, and the I be the event that they made their vote in-person:

$$\begin{aligned} P(R) &= P(R|I)P(I) + P(\overline{R}|\overline{I})P(\overline{I}), \\ &= 0.52(1-0.35) + (1-0.55)(0.35), \\ &= 0.4955 \end{aligned}$$

   b) If a voter choose the Red Tomato Party, what is the probability that they voted by post? **[2 MARKS]**

      A. $P(\text{In-person}|\text{Red Tomato})P(\text{In-person})/P(\text{Red Tomato})$
      B. $1-P(\text{In-person}|\text{Red Tomato})P(\text{Red Tomato})/P(\text{In-person})$
      C. $1-P(\text{Red Tomato}|\text{In-person})P(\text{In-person})/P(\text{Red Tomato})$
      D. $P(\text{Red Tomato}|\text{In-person})P(\text{In-person})/P(\text{Red Tomato})$

      C

      Solution:

$$P(\overline{I}|R) = 1 - P(I|R) = 1 - \frac{P(R|I)P(I)}{P(R)}$$

   c) Calculate the probability that voter who choose the Red Tomato Party did so via postal vote? **[2 MARKS]**

      A. 0.50

B. 0.90

C. 0.53

D. 0.318

D

Solution:

$$\begin{aligned} P(\bar{I}|R) &= 1 - \frac{P(R|I)P(I)}{P(R)}, \\ &= 1 - 0.52 \times (1 - 0.35)/0.4955, \\ &= 0.318 \end{aligned}$$

d) Five voters were recruited independently but consecutively for a post-vote survey.

If X_i denotes the event that the i-th voter votes for the Red Tomato Party, what is the probability that the first two voters recruited will be Red Tomato Party voters and the last three Blue Berry Party voters? **[1 MARKS]**

A. $P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)$

B. $P(X_1)P(X_2)P(\overline{X}_3)P(\overline{X}_4)P(\overline{X}_5)$

C. $1 - P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)$

D. $P(\overline{X}_1)P(\overline{X}_2)P(X_3)P(X_4)P(X_5)$

B

Solution:

$$\begin{aligned} P(X_1 = R, X_2 = R, X_3 = B, X_4 = B, X_5 = B) &= P(X_1 = R)P(X_2 = R)P(X_3 = B)P(X_4 = B)P(X_5 = B). \\ &= P(X_1)P(X_2)P(\overline{X}_3)P(\overline{X}_4)P(\overline{X}_5). \end{aligned}$$

Compute the probability that the first two voters recruited will be Red Tomato Party voters and the last three Blue Berry Party voters? **[1 MARKS]**

A. 0.0320

B. 0.0313

C. 0.0315

D. 0.0310

C

Solution:

$$\begin{aligned} P(X_1 = R, X_2 = R, X_3 = B, X_4 = B, X_5 = B) &= P(X_1)P(X_2)P(\overline{X}_3)P(\overline{X}_4)P(\overline{X}_5), \\ &= 0.4955^2(1-0.4955)^3, \\ &= 0.03152614. \end{aligned}$$

e) You are informed that the first two voters recruited for the post-vote survey are Red Tomato Party voters.

What is the probability that the last three will be Blue Berry Party voters? **[1 MARKS]**

A. $P(\overline{X}_3, \overline{X}_4, \overline{X}_5 | X_1, X_2)$
B. $P(X_1, X_2 | X_3, X_4, X_5)$
C. $P(\overline{X}_3)P(\overline{X}_4)P(\overline{X}_5)$
D. $P(X_1, X_2)$

C

Solution: Because of independence:

$$P(X_3 = B, X_4 = B, X_5 = B | X_1 = R, X_2 = R) = P(X_3 = B)P(X_4 = B)P(X_5 = B),$$

since X_3, X_4, X_5 are independent of X_1, X_2.

Compute the probability that the last three were Blue Berry Party voters? **[1 MARKS]**

A. 0.255
B. 0.122
C. 0.125
D. 0.128

D

Solution:

$$\begin{aligned} P(\overline{X}_3, \overline{X}_4, \overline{X}_5 | X_1 = R, X_2 = R) &= P(\overline{X}_3)P(\overline{X}_4)P(\overline{X}_5), \\ &= (1-0.4955)^3, \\ &= 0.1284055. \end{aligned}$$

2. i) A two-sided coin is flipped and a six-sided dice is rolled. What is the probability of obtaining a HEAD and a score less than 4? **[2 MARKS]**

A. $1/2$
B. $1/3$
C. $1/4$
D. $1/12$

C

Solution:

Relevant events $(H,1),(H,2),(H,3)$ out of 12 possibilities so probability is $3/12 = 1/4$.

ii) There are two routes that a commuter can use to travel by train from Manchester to London. The commuter would prefer to take a fast route but there is a 5% chance that the direct train will have a fault. Whilst the chance of a fault on the standard route is 1%. If the commuter randomly chooses a route with equal probability, what is the probability that they will be delayed on their journey to London? **[2 MARKS]**

A. 0.025
B. 0.005
C. 0.06
D. 0.03

D

Solution:

$$\begin{aligned} P(\text{Delay}) &= P(\text{Delay}|\text{Fast Route})P(\text{Fast Route}) + P(\text{Delay}|\text{Standard Route})P(\text{Standard Route}), \\ &= 0.05(0.5) + 0.01(0.5) = 0.03 \end{aligned}$$

iii) A discrete random variable Z has the following probability mass function:

z	2	4	6	8
$P(Z=z)$	0.8	0.1	0.05	0.05

Determine the expectation and variance of Z: **[2 MARKS]**

A. $(5, 6.67)$
B. $(5, 120)$
C. $(2.7, 9.8)$
D. $(2.7, 2.51)$

D

Solution:

$$\begin{aligned} E[Z] &= 0.8(2) + 0.1(4) + 0.05(6) + 0.05(8), \\ &= 2.7, \\ V[Z] &= 0.8(2^2) + 0.1(4^2) + 0.05(6^2) + 0.05(8^2) - 2.7^2, \\ &= 2.51 \end{aligned}$$

iv) Two six-sided dice are rolled and the total score is 9. What is the probability that one of the die rolls produced a 4?

[2 MARKS]

A. $1/2$
B. $1/4$
C. $1/8$
D. $1/18$

A

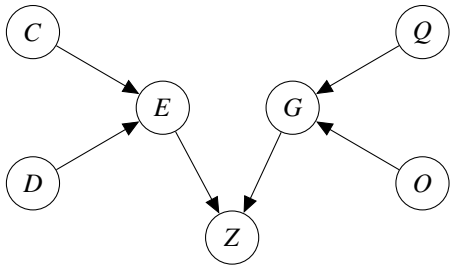
Solution:

Possible die roll combinations leading to 9: (3, 6), (6, 3), (4, 5), (5, 4). Fours only occur in 2 out of 4 possible combinations:

$$P(\text{At least one die is 4}|\text{Score is 9}) = 2/4 = 1/2$$

This is a *conditional* probability question so by conditioning on the total score of 9, we move from $6 \times 6 = 36$ combinations to 4 by conditioning on this event.

v) Given the following directed acyclic graph (DAG):



What is the probability $P(Z)$? **[2 MARKS]**

A. $\sum_{E,G} P(Z|E,G,C,D,O,Q)[\sum_{C,D} P(E|C,D)P(C)P(D)][\sum_{O,Q} P(G|O,Q)P(O)P(Q)]$
B. $\sum_{E,G} P(Z|E,G)[\sum_{C,D} P(E|C,D)P(C)P(D)][\sum_{O,Q} P(G|O,Q)P(O)P(Q)]$
C. $\sum_{E,G} P(Z,E,G)[\sum_{C,D} P(E|C,D)P(C)P(D)][\sum_{O,Q} P(G|O,Q)P(O)P(Q)]$
D. $\sum_{E,G} P(Z|E,G)[\sum_{C,D} P(E,C,D)P(C)P(D)][\sum_{O,Q} P(G,O,Q)P(O)P(Q)]$

B

3. Given the following data:

Item	1	2	3	4	5
x_1	-4	-3	1	2	2
x_2	-2	-1	1	1	2
Class	A	A	B	A	B

You are asked to develop a linear binary classification model that takes inputs (x_1, x_2) and predicts class A or B.

i) If the perceptron weights are $(w_1, w_2) = (1, 1)$ and bias $w_0 = 1$.
What is the form of the decision boundary? **[1 MARKS]**

A. $x_2 = -1 + x_1$.
B. $x_2 = 1 + x_1$.
C. $x_2 = 1 - x_1$.
D. $x_2 = -1 - x_1$.

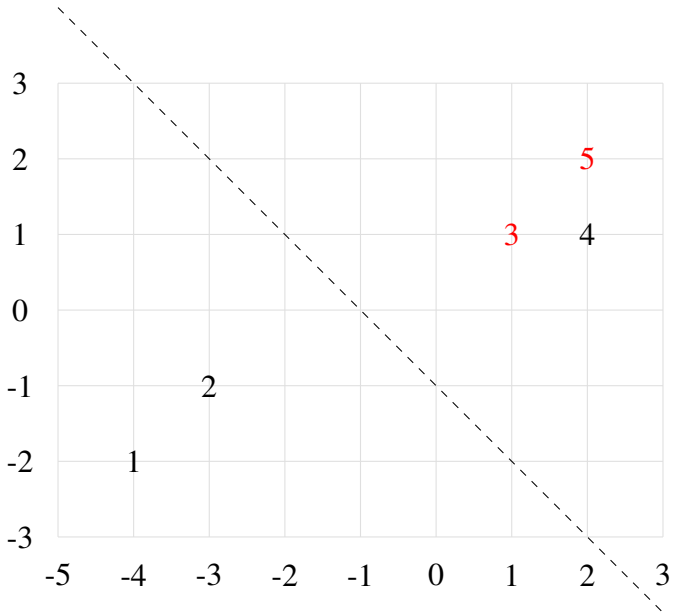
D

Solution:
Decision boundary is given by $1 + x_1 + x_2 = 0$ or $x_2 = -1 - x_1$.
Which data item would be misclassified? **[2 MARKS]**

A. 2.
B. 3.
C. 4.
D. 5.

C

Solution:
Data item 4 is misclassified assuming $A = -1$ and $B = +1$.



ii) Compute one iteration of a perceptron update. What are the updated perceptron weights (w_0, w_1, w_2)? State whether the algorithm has converged. **[3 MARKS]**

A. $(2,3,2)$. Not converged.
B. $(2,3,2)$. Converged.
C. $(0,-1,0)$. Converged.
D. $(0,-1,0)$. Not converged.

w1 = w1 - a*(y_pred - y_true) * x1
w0 = w0 - a*(y_pred - y_true)
w2 = w2 - a*(y_pred - y_true) * x2
can be judege by the rate of the change of w1 and w2

D

Solution:

Data item 4 is misclassified and should be used for the update.

The update is a misclassification of a true A ($y = -1$) item so we should update according to:

$$w_0 = w_0 - 1 = 1 - 1 = 0,$$
$$w_1 = w_1 - x_1 = 1 - (2) = -1,$$
$$w_2 = w_2 - x_2 = 1 - (1) = 0.$$

To check convergence, apply the updated weights to the data again:

Item	1	2	3	4	5
x_1	-4	-3	1	2	2
x_2	-2	-1	1	1	2
z	4	3	-1	-2	-2
y	1	1	-1	-1	-1

and use $A = -1$ and $B = 1$ as before. Algorithm has not converged as four items are misclassified.

iii) Which of the following weights (w_0, w_1, w_2) would lead to convergence? **[2 MARKS]**

A. $(-3,-3,7)$.
B. $(3,3,7)$.
C. $(-2,0,1)$.
D. $(-1,-1,1)$.

A

Solution: Compute $z = w_0 + w_1 x_1 + w_2 x_2$ for each set of weights and note that $w = (-3,-3,7)$ leads to $z = (-5,-1,1,-2,5)$ and the correct classification.

iv) An alternative classification model is built using logistic regression. It produces the following output which is missing one entry:

Item	1	2	3	4	5
x_1	-4	-3	1	2	2
x_2	-2	-1	1	1	2
P(Class A)	0.0009	0.007	0.73	X	0.953

If the logistic regression model is of the form:

$$P(\text{Class A}) = p = \frac{1}{1+\exp(-[\beta_0+\beta_1 x_1+\beta_2 x_2])}$$

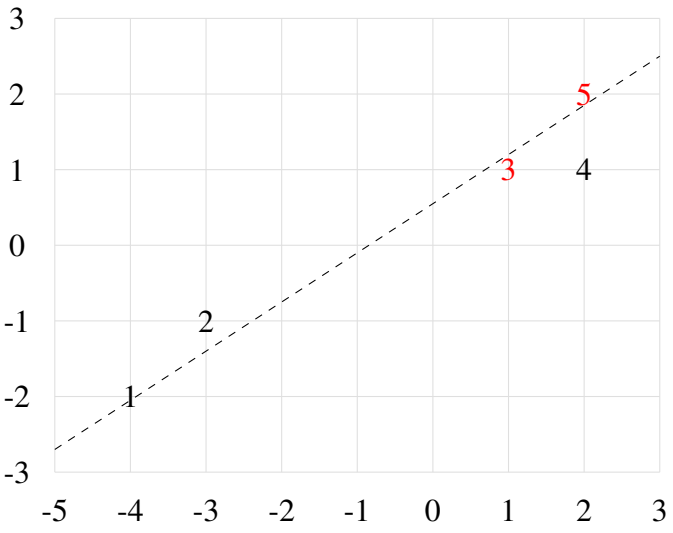
Which *one* of the following represents the probability of class A for data item 4? **[2 MARKS]**

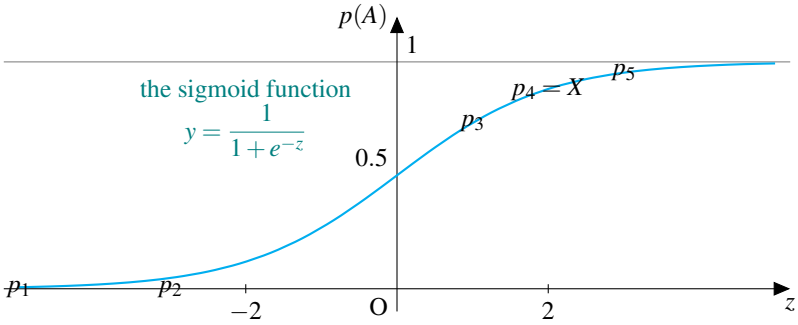
A. 0.98
B. 0.88
C. 0.12
D. 0.5

B

Solution:

As the probabilities are increasing from data item 1 to 5, this suggests that $z = \beta_0+\beta_1 x_1+\beta_2 x_2$ is increasing and that β_1 and β_2 are both positive-valued. Since the logistic function is monotonic, the value of X is likely to be between 0.73 and 0.953 and therefore only 0.88 fits.





4. A support vector machine classifier, *A*, is used to predict whether a person has at least one child (*child*), using the following features: (i) hours of sleep per night, (ii) years of education, and (iii) average cost of food per week. It is trained using data from 500 people. On a validation set with 350 people, it predicts *child* when the person has at least one *child*, in 150 cases. It predicts *child* when there is *no_child*, in 50 cases. It predicts *no_child* when there is *no_child* in 75 cases. Taking *child* to be the positive class:

   a) Which of the following confusion matrices is correct? **[2 MARKS]**

A.

	predicted:child	**predicted: no_child**
actual: child	75	50
actual: no_child	0	150

B.

	predicted:child	**predicted: no_child**
actual: child	150	0
actual: no_child	50	75

C.

	predicted:child	**predicted: no_child**
actual: child	150	50
actual: no_child	75	75

D.

	predicted:child	**predicted: no_child**
actual: child	150	75
actual: no_child	50	75

D

   b) What is the sensitivity of classifier *A*? **[2 MARKS]**

A. 0.60
B. 0.50
C. 0.67
D. 0.75

C

   c) What is the specificity of classifier *A*? **[2 MARKS]**

A. 0.60
B. 0.50
C. 0.67
D. 0.75

A

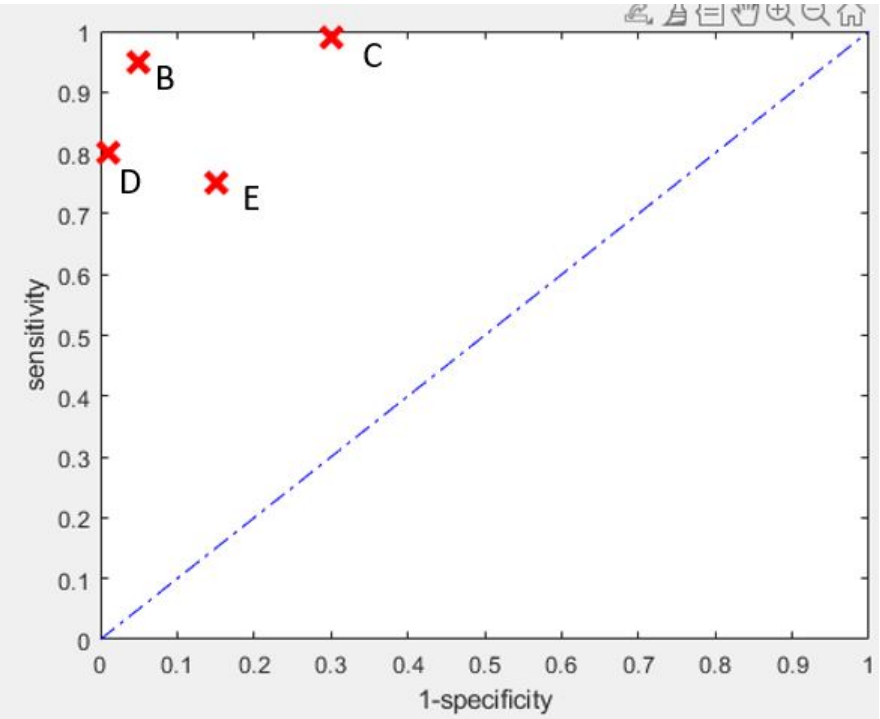
d) What is the precision of classifier A? **[2 MARKS]**

A. 0.60
B. 0.50
C. 0.67
D. 0.75

D

Four different machine algorithms (C-H) are trained on the same data. The resulting points are depicted in an ROC plot below.

Suppose we wish to use the classifier in a screening test to predict the presence of a disease, in which we correctly the detect the disease 99% of the time, if the disease is present. Which classifier would be most useful in this case?? **[1 MARK]**



A. B
B. C
C. D
D. E

B

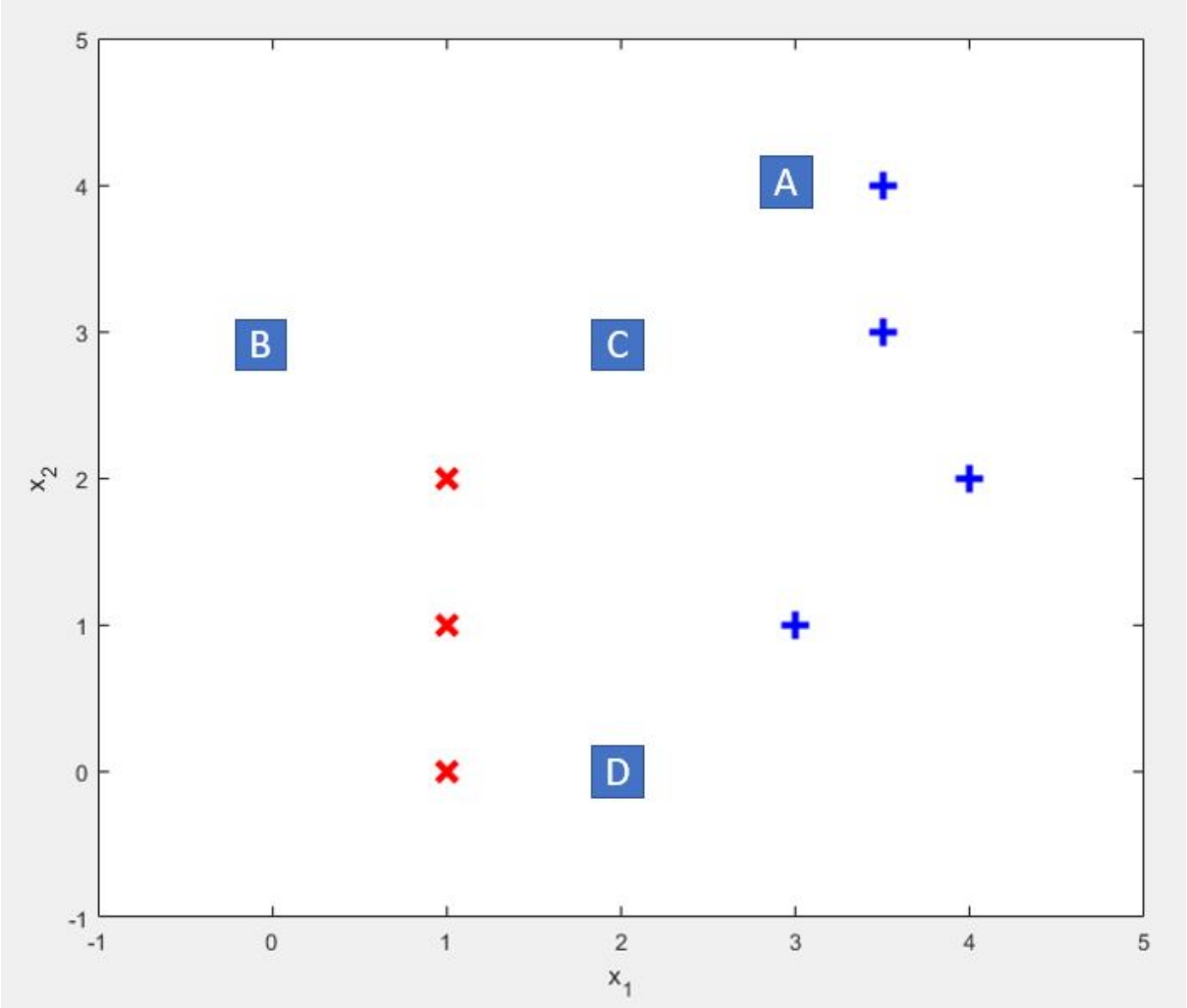
e) Which classifier would we select if the sensitivity and specificity are of equal importance? **[1 MARK]**

A. B
B. C
C. D
D. E

A

5. This question uses the training data in the table below. There are 4 test data points: $A : (x_1 = 3, x_2 = 4)$, $B : (x_1 = 0, x_2 = 3)$, $C : (x_1 = 2, x_2 = 3)$ and $D : (x_1 = 2, x_2 = 1)$. The points are also plotted in the figure below.

x_1	1	1	1	3	3.5	3.5	4
x_2	0	1	2	1	3	4	2
class	×	×	×	+	+	+	+



a) Which of the following statements are true about the k-nearest neighbours (k-NN) method? (select 2) **[2 MARKS]**

A. k-NN is a probabilistic classification method
B. k-NN requires more computation on test data, rather than during model training
C. Large values of k are likely to lead to an over-fitted model
D. In k-NN, an optimal value for k can be chosen using cross validation
E. the k-NN decision boundary is linear
F. the k-NN decision boundary is smoother with smaller value of k

B

Solution: B and D

b) By applying k-NN to the dataset, with $k = 1$ and using a Euclidean distance metric, which test data points belong to class $(\times)$? **[2 MARKS]**

A. A
B. B
C. C
D. D

Solution: B, C, D

Decision boundary is given by $1 + x_1 + x_2 = 0$ or $x_2 = -1 - x_1$.

c) By applying k-NN to the dataset, with $k = 3$ and using a Euclidean distance metric, which test data points belong to class $(\times)$? **[2 MARKS]**

A. A
B. B
C. C
D. D

D

Solution: B,D

d) For n features, the Manhattan distance between two points p and q, is defined as:

$$d_1(p,q) = \sum_{i-1}^{n} |p_i - q_i|$$

What is the Manhattan distance between points A and C? **[2 MARKS]**

A. 2
B. $\sqrt{2}$
C. 1
D. $-\sqrt{2}$

A

e) Using $k = 1$ and a Manhattan distance metric which data points are classified as $(\times)$? **[2 MARKS]**

A. A
B. B
C. C
D. D

Solution: B,D

6. This question uses the training data in the table below.

Hours Studied	Mock Exam	Sex	Degree	Grade
50	Pass	Male	Maths	Pass
120	Pass	Female	Comp Sci	Pass
55	Pass	Male	Maths	Pass
50	Pass	Female	Maths	Pass
60	Fail	Male	Comp Sci	Pass
45	Fail	Female	Physics	Pass
25	Fail	Male	Comp Sci	Fail
50	Fail	Female	Physics	Fail
7	Fail	Male	Physics	Fail
42	Fail	Female	Maths	Fail
0	Fail	Male	Physics	Fail
35	Pass	Female	Physics	Fail

a) Calculate the entropy (in bits) of *Grade* given that $Sex = Female$ AND $Mock\_Exam = Pass$ **[2 MARKS]**

A. 1
B. 0.98
C. 0.92
D. 0.76

C

b) We wish to create an ID3 decision tree to predict *Grade*, which feature would be selected first? **[2 MARKS]**

A. Hours Studied
B. Mock Exam
C. Sex
D. Degree

A

c) Calculate the entropy for the first feature selected (the answer to the previous question) **[2 MARKS]**

A. 0
B. 0.66
C. 0.65

D. 0.33

D

d) We train the decision tree again, but this time set the minimum number of samples in each leaf node equal to 12. Calculate the training error in this case.

[2 MARKS]

A. $\frac{0}{12} = 0$
B. $\frac{12}{12} = 1$
C. $\frac{6}{12} = 0.5$
D. Not enough information to calculate

C

e) Which of the following statements are true about decision tree (post-)pruning? **[2 MARKS]**

A. The main purpose of pruning is to reduce bias in the classifier
B. The training error tends to decrease as a decision tree is pruned
C. Decision tree pruning is a way of minimising the entropy to maximise the information gain
D. The main purpose of pruning is to reduce overfitting to the training set

D

7. a) Below is a Python function related to a machine learning method. The input, X, is an array of training data. What does it do? **[2 MARKS]**

```
1  def my_function(X):
2      no_examples = X.shape[0]
3      idx = randint(0, no_examples, size =
           no_examples)
4      X_out = X[idx,:]
5      return X_out
```

A. Calculates the distance between two data points
B. Normalises a training data set
C. Calculates one step of gradient descent
D. Creates a bootstrapped sample training set

D

b) Below is another Python function related to a machine learning method. The input, X, is an array of training data. What does it do? **[2 MARKS]**

```
1  def my_function2(X):
2      import numpy as np
3      mu = np.mean(X,0)
4      sigma = np.std(X,0)
5      X_out = (X-mu)/sigma
6      return X_out
```

A. Calculates the distance between two data points
B. Normalises a training data set
C. Calculates one step of gradient descent
D. Creates a training set from bootstrapped samples

B

c) We incorporate my_function2 from 7.b) into a script that trains and tests a logistic regression classifier, below. Which of the following options for using the function is correct? **[2 MARKS]**

```
1  import numpy as np
2  from sklearn import datasets
3  from sklearn.model_selection import
       train_test_split
```

```
4                   from sklearn.linear_model import
                       LogisticRegression
5
6                   X = iris.data
7                   Y = iris.target
8                   train_X, test_X, train_y, test_y =
                       train_test_split(X, Y, test_size =
                       0.33)
9                   logreg = LogisticRegression()
10                  logreg.fit(train_X, train_y)
11                  y_pred = logreg.predict(test_X)
```

A. *X = my_function2(X)* should be inserted between lines 7 and 8
B. *train_X = my_function2(train_X)* should be inserted between lines 8 and 9
C. *train_X = my_function2(train_X)* should be inserted between lines 10 and 11
D. *X = my_function2(X)* should be inserted between lines 8 and 9

B

d) There is one further error in this code snippet. Which of the following best describes the error? **[2 MARKS]**

A. The training set has not been normalised
B. The test set is too small a percentage of the total data set
C. The test set has not been normalised
D. The test labels should be used to generate y_pred

C

e) For a logistic regression, what is the impact, of normalising the training data? **[2 MARKS]**

A. No impact at all
B. The algorithm may converge more quickly, but the final result (weights and bias) is the same
C. The final result (weights and bias) will be different
D. You should not normalise the data for logistic regression

B

8. a) What is the purpose of the "Kernel Trick"? **[1 MARK]**

A. To transform the problem from nonlinear to linear
B. To transform the problem from regression to classification
C. To transform the data from non-linearly separable to linearly separable
D. To transform the problem from supervised to unsupervised learning

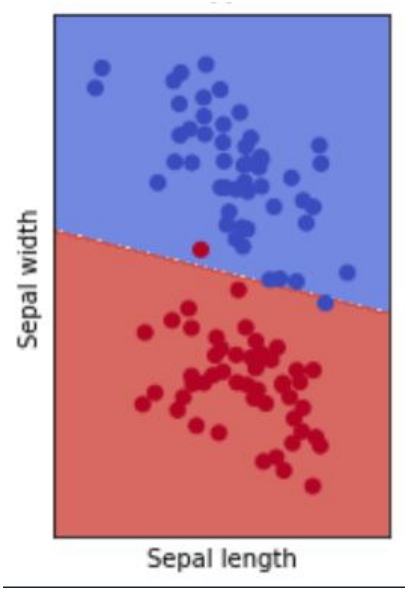
C

b) What is the purpose of "slack variables"? **[1 MARK]**

A. To provide slack in the maximum margin
B. To allow data points to be on the wrong side of the decision boundary
C. To ensure that data points are on the correct side of the decision boundary
D. To trade off between maximising the margin and minimising misclassification

B

c) Training data for a classification problem, as well as the SVM decision boundary, are shown in the figure below. Which set of hyperparameters generated the decision boundary? C is the regularisation parameter. **[2 MARKS]**:

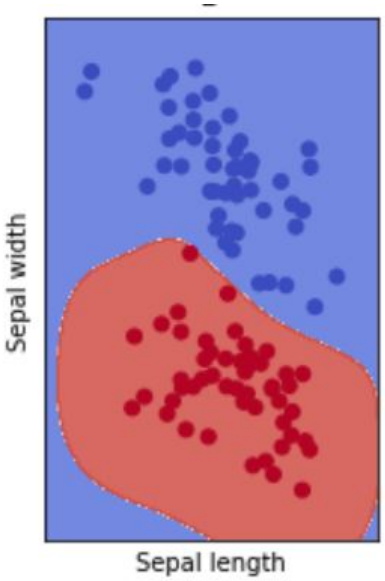


it allows data point on the wrong side here, so C is small

A. kernel = linear, C = 0.01
B. kernel = linear, C = 1000
C. kernel = polynomial, degree = 3, C = 1
D. kernel = rbf, gamma = 0.7, C = 1

A

d) Training data for a classification problem, as well as the SVM decision boundary, are shown in the figure below. Which set of hyperparameters generated the decision boundary? C is the regularisation parameter. **[2 MARKS]**



the boundary is really straight and unsmooth.

A. kernel = linear, C = 0.01

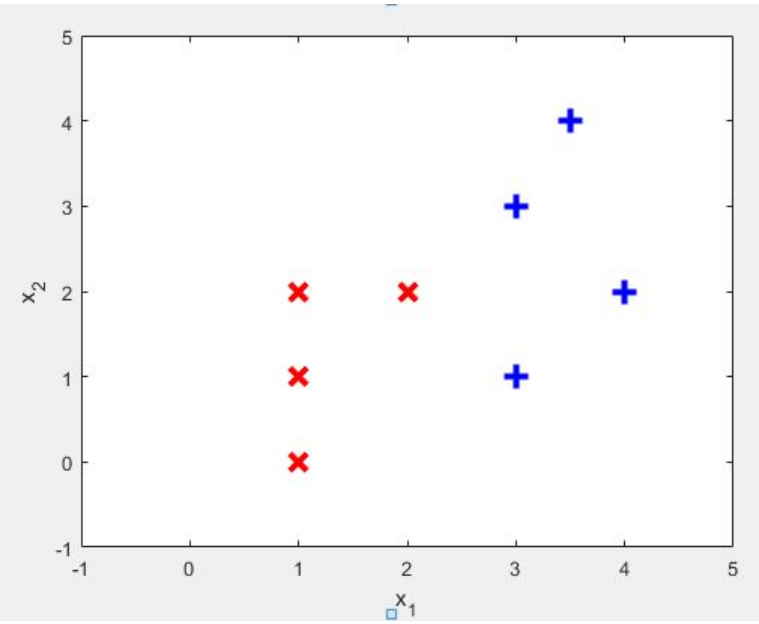
B. kernel = linear, C = 1000

C. kernel = polynomial, degree = 3, C = 1

D. kernel = rbf, gamma = 0.7, C = 1

D

e) For the data set plotted below, suppose we classify using a hard margin linear SVM. Which of the following equations describes the decision boundary?: **[2 MARKS]**



A. $x_2 - 2.5 = 0$
B. $x_1 - 2.5 = 0$
C. $x_1 + x_2 = 0$
D. $-x_1 - 3 = 0$

B

f) For the same data set, which coordinates are the support vectors? **[2 MARKS]**

A. (2,2), (3,1)
B. (2,2), (3,1), (3,3)
C. (2,2), (3,3)
D. (1,1), (2,1), (3,2)

B