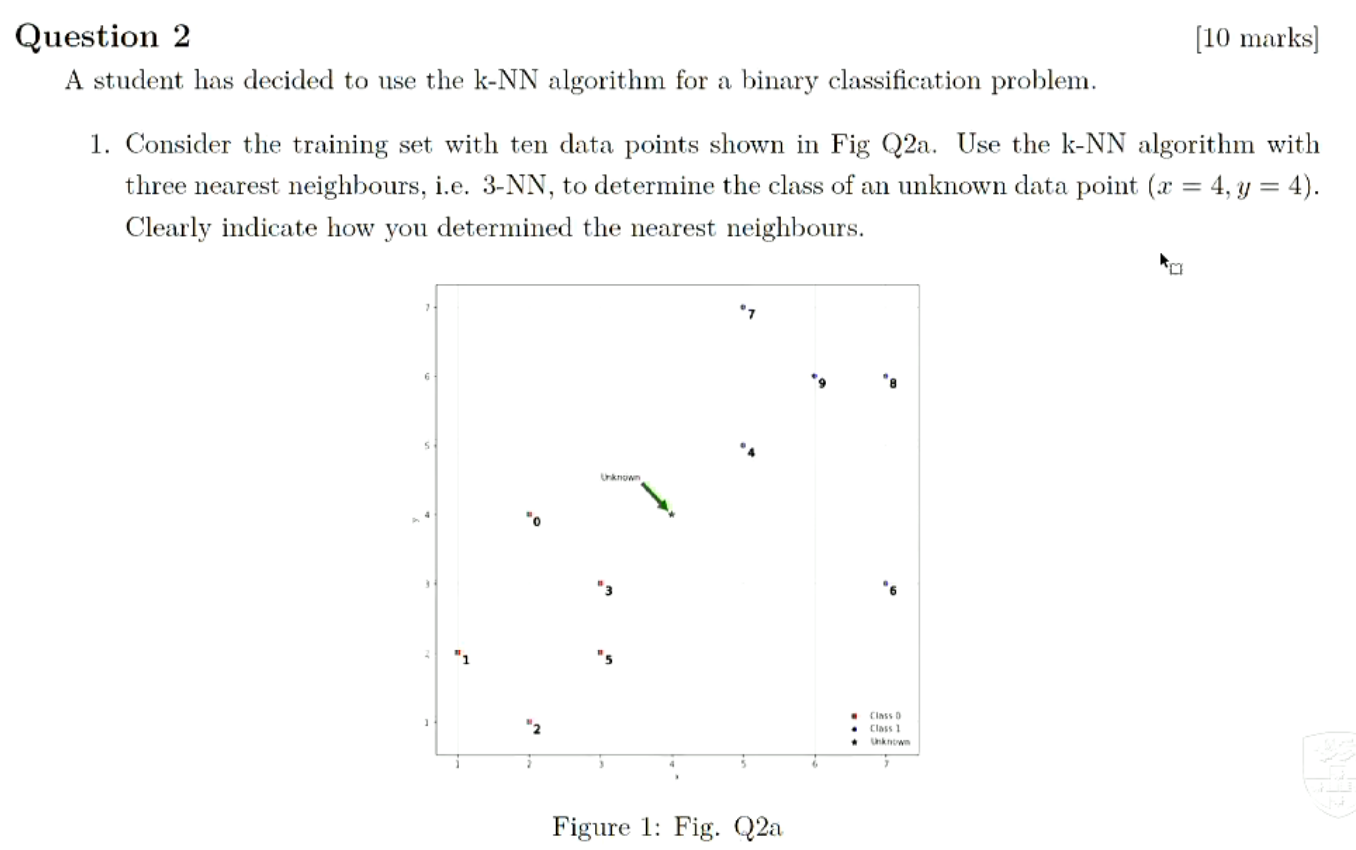
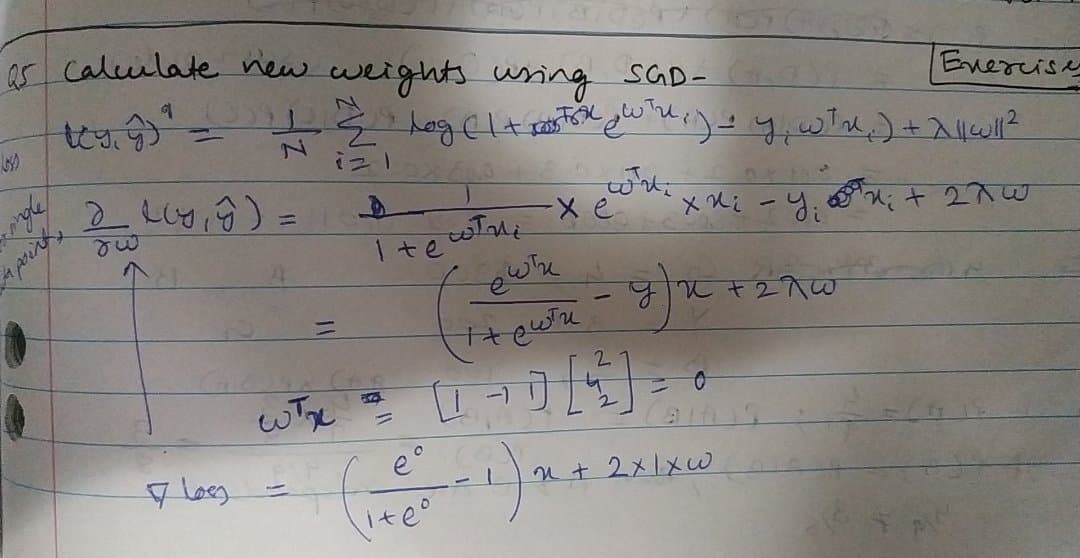
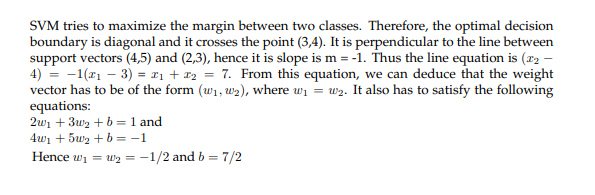
Sample Exam Questions: 2018-2019 
Question I 
[10 marks) 
Let A be a m * n matrix where m is 80 million and n is 4096. The eigen decomposition of ATA shows 
that 4086 eigen values are zero. The remaining ten eigen values (in descending order) are: 
2 '2 '2 •2 '2 •2 '2'2'2'2 
[10 , 9 , 8 , 7 , 5 , 4 , 3 , '2 , 1] 
I. What are the non-zero singular values of A? Show the calculation steps? 
2. We would like to create a representation that can capture all the information of A. How many 
dimensions at least this representation should have and why? Question 3 
[10 marks] 
A factory has three machines A, B, and C. From the history, we know that the defect rates of the throe 
machines are respectively 2%. and lu last year. the three machines respectivv•ly produced 2000, 
3000, and 5000 products. If two products are chosen at random from these total productsl and is found 
to be defective. what is the probability that it was produced by machine A? Question 4 
[10 marks) 
Given the dataset below, use the Naive Bayes classifier to predict whether the last borrower will default 
or not. Show your calculations as well as the final result. 
Id 
1 
2 
3 
4 
5 
7 
8 
9 
10 
Home Owner 
Yes 
No 
No 
Yes 
No 
No 
Yes 
No 
No 
Marital Status 
Single 
Nlarried 
Single 
Married 
Divorced 
Married 
Divorced 
Single 
Single 
Divorced 
Annual income 
High 
Medium 
High 
Medium 
Low 
High 
Low 
Medium 
Medium 
Low 
Default borrower 
No 
No 
No 
Yes 
No 
Yes 
No 
Yes Question 5 
[10 marks) 
We are going to optimise the regularised logistic regression by employing the stochastic gradient descent 
method (SCD). The objective. function is as follows: 
1 
The initial w: = [0, 0, 01, A = I, and the learning rate = 0.1. Given that the first randomly selected 
example is = [0, 0, = 0) and the second is (Tl = 
I) , what are the new weight 
values after two updates? 
Let f(x) = log.r, g(.r) = ex. Then, their first derivatives are as: f '(x) = and g/ (x) = ex. aasdasddadasdaa 

我算的0.9 -0.6 0.9

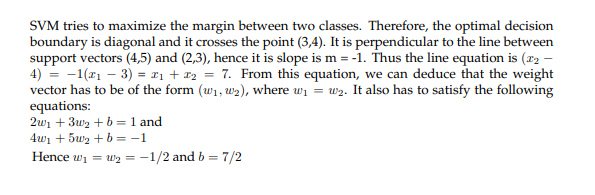
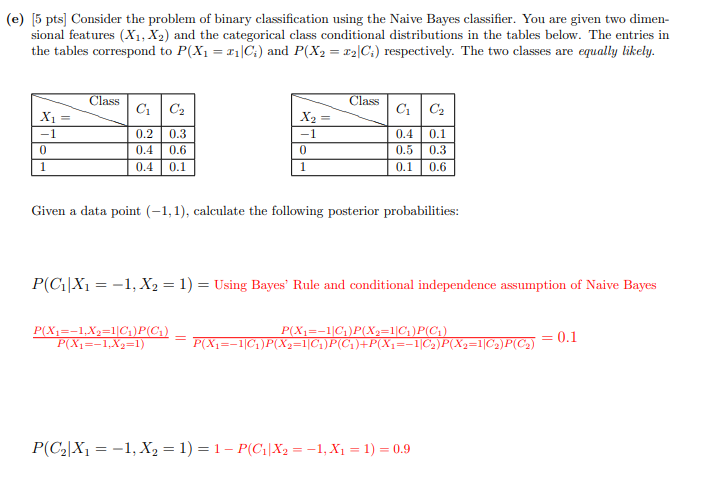
Question 6 
(20 marks) 
You have collected three datasets with two dimensional points for binary classification. Your ultimate 
goal is to classify the point (x; , x;) = (0, 2.5). You are provided with the following datasets: A, B, 
and C. 
For dataset A: The positive observations are (2, 2). (2, 3), and (3, 2). The negative observations are 
(-1, 1), (0, 0), (1, —1), and (-1, 0). 
Dataset B contains all the data from dataset A, but has one extra positive observation at (2, —2). 
For each of the above datasets, answer the following questions. 
Are the problems linearly separable in the original input space? You can use scatter plots to show 
I. 
these. 
If the problem is not linearly separable, provide an example of simple feature transformation 4b 
2. 
such that the resulting problem is linearly separable with the lowest number of features possible. 
Otherwise, use the two dimensional feature x2) 
3. Write down the equation for the decision boundaries in the original input space. 
4. Write down the coordinates of all support vector(s) for both classes based on your choice for the 
feature. 
Does belong to the positive or negative class? 
5. 

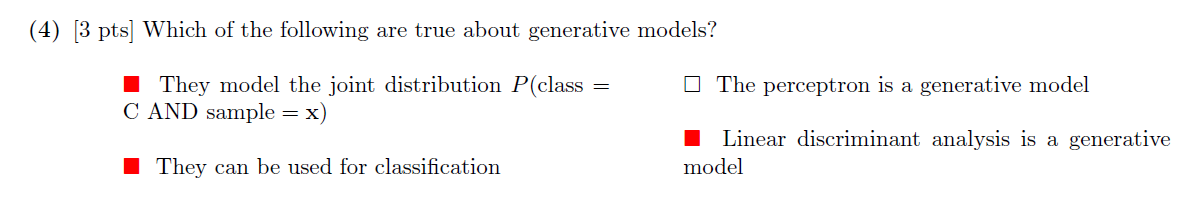
Question 7 
[10 marks] 
Consider a linear regression problem of estimating a non-linear function f with 30 training data points 
{ (Xi, "i . As shown in Fig Q8, three linear regressions were independently performed with polyno- 
mial features of polynomial orders I, 4 and 15. 
1. Identify which polynomial degrees (out of 1, 4 and 15) could correspond to models A, B and C. 
2. The models A, B, and C independently reported sum-of-squared-error (SSE) values between all 
training data points and corresponding estimates as 0.35, 6.78 and 0.15, respectively. Explain why 
model C has a lower SSE, although seems to have a spurious fit. 
30 
SSE = - g. 
Describe a procedure for this particular dataset to determine n suitable model complexity, i.e. ibe 
polynomial order. Question 8 
94 
92 
[10 marks] 
You are building a deep convolutional neural network for a classification problem on a large dataset of 
images. This network has three layers but as you add more elements in each layer, the performance start 
to degrade as indicated in the graph below. Describe two strategies to avoid this problem and explain 
why. 
95 
0.0 
0.2 
0.4 
0.8 
1.0 
of paratneters 

这道题 1.dropout 2.pooling两个策略对么

Question 9 
Given that the SVD of a matrix M EVT. 
[9 marks) 
If it is not, how 
I. 
2. 
3. 
Is it correct to say: "The matrix MT M can be decomposed as MT AI = V" EVT„. 
to make it become correct. 
Chopse a correct matrix to fill in the question mark: MT NI V 
Based on the above, what are eigenvectors and eigenvalues of MT M ? 

Question 10 
14 marks) 
Give a vector a = (1, 3, 4) and other vectors = (4, 3, = (0.4, 10, 50), and = (1, 4, 10). 
1. Report the minimum distance of a to x2, and and the nearest neighbor of a. 
2. Find which of makes the smallest angle with a and report that angle. Question 11 
[12 marks] 
Alice and Bob both need to buy a bicycle. The bike store has a stock of 4 green, 3 yellow and 2 red 
bikes. Alice randomly picks one of the bikes aud buy it. Immediately after, Bob does the same. The 
sale price of the green, yellow and red bikes are $300, $200 and $100, respectively. 
Let be the event that Alice bought a green bike, and B be the event that Bob bought a green bike. 
I. NVhat is P(A)? What is P(AIB)? Are A and B independent events? .Justify your answer. 
2. NVhat is the probability that Alice and Bob bought bicycles of different colors? 
3. NVhat is the probability that at least one of them bought a green bike? 
4. Given that Bob bought a green bike, what is the expected value of the amount of money spent by 
Alice? Question 12 
In Linear Regression given the following cost function 
2 
[9 marks] 
(1) 
where ho (xo)) = 9Tx(i), feature vector xo) € Rd of the i-th sample, and there are n data samples. We 
usually use the gradient descent to learn the minimum value of the cost function: O — OVJ(O). 
I. 
2. 
3. 
\Vhat is the name of the cost function above? 
Show step-by-step the gradient descent update for this cost function. 
How will the gradient update change if we add the regularization term? Question 13 
[10 marks) 
Support vector machines learn a decision boundary leading to the largest margin from both classes. You 
are training SVNI on a tiny dataset with 4 points shown in Figure [4 This dataset consists of two 
examples with class label 
—1 (denoted with triangles), and two examples with class label +1 (denoted 
with plus). Class -1 
Class +1 Figure 2: A tiny dataset for SVM 
Find the weight vector w and bias b. What is the equation corresponding to the decision boundary? 

Question 14 
[10 marks] 
Consider the problem of binary classification using the Naive Bayes classifier. You are given two di- 
mensional features (Xl, X2) and the categorical class conditional distributions in the tables below. The 
entries in the tables correspond to P(XI = ICi) and P(X2 = X21Ci) respectively. The two classes are 
equally likely. 
Class 
0.3 
0.2 
0.5 
Class 
-1 
1 
0.1 
0.6 
0.3 
-1 
o 
1 
Given a data point (I, I), calculate the following posterior probabilities: 
P(C21X1 = 1) 
0.5 
0.3 
0.2 
0.6 
0.3 
0.1 
P(CIIXI 1,X2 = 
I) and  The following questions are multiple choices questions. Please check ALL CORRECT CHOICES 
and circle your answers. Note that every question should have at least one right answer. 
Question 15 
Which of the following are true about generative models? 
(5 marks) 
A. They model the joint distribution 
P(class = C AND sample = x). 
C. The Perceptron is a generative model. 
B. They can be used for classification. 
D. Linear discriminant analysis is a generative 
model. 

Question 16 
[5 marks) 
Suppose we train a hard-margin linear SVM on n > 100 data points in R2, yielding a hyperplane 
with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the 
maximum possible number of support vectors for the new hyperplane (assuming the n + 1 points are 
linearly separable)? 
B. 
3 

Question 17 
[5 marks) 
Suppose we are given data comprising points of several different classes. Each class has a different 
probability distribution from which the sample points are drawn. We do not have the class labels. We 
use k-means clustering to try to guess the classes. Which of the following circumstances would undermine 
its effectiveness? 
A. Each class has the same mean. 
B. Choose k = n,the number of sample points. 
C. Some of the classes aren't normally distributedD. The variance of each distribution is small in all 
directions. 

