Question 1 [10pts]

Let A be a $m \times n$ matrix where m is 80 million and n is 4096. The eigen decomposition of $A^\top A$ shows that 4086 eigen values are zero. The remaining ten eigen values (in descending order) are:

$$[10^2, 9^2, 8^2, 7^2, 6^2, 5^2, 4^2, 3^2, 2^2, 1^2].$$

What are the non-zero singular values of A? Show the calculation steps.

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?

The non-zero singular values of A are sqrt of eigan values. i.e. [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]
There should at least have 10 dimensions to represent and capture all the information of A.
This is because apart from the 10 eigenvalues above, the rest are 0 indicating they yield variance/information.

Question 2 [10pts]

A student has decided to use the k-NN algorithm for a binary classification problem.

a. Consider the training set with six data points shown in Fig Q2.a. Use the k-NN algorithm with three nearest neighbours, i.e. 3-NN, to determine the class of an unknown data point $(x_1 = 4, x_2 = 6)$. Clearly indicate how you determined the nearest neighbours.

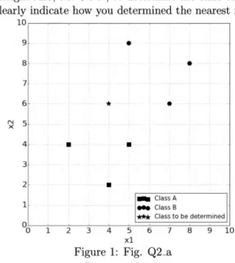


Figure 1: Fig. Q2.a

(2-4)^2 + (4-6)^2 =
(4-4)^2 + (2-6)^2 =
(5-4)^2 + (4-6)^2 =
(7-4)^2 + (6-6)^2 =
(5-4)^2 + (9-6)^2 =
(8-4)^2 + (8-6)^2 =

b. A friend has suggested using the naïve Bayes classifier. Using a test set, the student has found that "the sum of true positive and true negative" values for the two classifiers are almost the same. Considering the aforementioned finding and the ROC curves given in Fig Q2.b, deduce if the naïve Bayes classifier based model outperforms the k-NN based model.

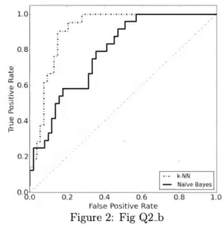


Figure 2: Fig Q2.b

No, the ROC curve fo k-NN indicates it has more true positives than Naive Bayes for any given false positive up up until 0.5. K-nn reaches its max true positive rate much quicker and has a smaller flash positive rate.

kNN curve hugs left corner more. Shows it has more predictive power.

Question 3 [10pts]

A factory has three machines A, B, and C. From the history, we know that the defect rates of the three machines are respectively 1%, 1%, and 3%. In last year, the three machines respectively produced 2000, 3000, and 5000 products. If a product is chosen at random from these total products and is found to be defective, what is the probability that it was produced by machine A?

A:20, B:30, C:150
Defective Total = 200
Using Bayes theorem
P(A|Defective) = P(D|A)*P(A)/P(D)
= 0.01 * (2000/10000) / (200/10000)
= 0.1

Question 4 [10pts]

Given the dataset below, use the Naïve Bayes classifier to predict whether the last borrower will default or not. Show your calculations as well as the final result.

No = 6
Yes = 3

Id	Home owner	Marital status	Annual income	Default borrower
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	High	No
4	Yes	Married	Medium	No
5	No	Divorced	Low	Yes
6	No	Married	High	No
7	Yes	Divorced	Low	Yes
8	No	Single	Medium	No
9	No	Single	Medium	Yes
10	Yes	Divorced	Low	?

P(Home Owner = "yes"| Default borrow = "No") = 2/9
P(Martial Status = "Divorced"| Default borrow = "No") = 0/9
P(Annual Income = "Low"| Default borrow = "No") = 0/9

P(Home Owner = "yes"| Default borrow = "yes") = 1/3
P(Martial Status = "Divorced"| Default borrow = "yes") = 2/3
P(Annual Income = "Low"| Default borrow = "yes") = 2/3
1/3 * 2/3 * 2/3 = 4/27
4/27 > 0. Therefore, will default.

Question 5 [10pts]

We are going to optimise the regularised logistic regression by employing the stochastic gradient descent method (SGD). The objective function is as follows:

$$\frac{1}{N}\sum_{i=1}^{N}(\log(1+e^{\mathbf{w}^\top \mathbf{x}_i}) - y_i \mathbf{w}^\top \mathbf{x}_i) + \lambda\|\mathbf{w}\|^2.$$

Currently $\mathbf{w}^\top = [1, -1, 1]$, $\lambda = 1$, and the learning rate $\eta = 0.1$. Given that the randomly selected example is $(\mathbf{x}^\top = [2, 4, 2], y = 1)$, what are the new weight values in the next step?

(Let $f(x) = \log x$, $g(x) = e^x$. Then, their first derivatives are as: $f'(x) = \frac{1}{x}$ and $g'(x) = e^x$.)

wTx = 2 - 4 + 2 = 0

Deferentiate: (1/(1 + e^0)) * xe^(wTx) - yx + 2w
=1/2 * x - x + 2w = [1, -4, 1]T
Wnew = Wold + 0.1 * [1, -4, 1]T

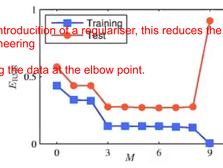
Question 6 [10pts]

In a regression problem, the error on the test set will generally decrease as we get more training data since the model will be better estimated. However, as shown in the figure below, as the complexity of model grows the error on the test set increases even though the error on the training set goes to zero. Explain why. Provide two different ways to avoid this phenomenon.

This phenomenon is called overfitting. This is when the model can only predict data that is very similar to the training data, in other words it has been over trained and become too dependent on the training data.

The first way to avoid this is the introduction of a regulariser, this reduces the weight that one feature might have if its too domineering

The second way is to stop training the data at the elbow point.



Question 7 [20pts]

You have collected three datasets with two dimensional points for binary classification. Your ultimate goal is to classify the point $x^* = (x_1^*, x_2^*) = (0, 2.5)$. You are provided with the following datasets: A, B, and C.

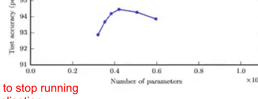
- For dataset A: The positive observations are $(2, 2)$, $(2, 3)$, and $(3, 2)$. The negative observations are $(-1, 1)$, $(0, 0)$, $(1, -1)$, and $(-1, 0)$.
- Dataset B contains all the data from dataset A, but has one extra positive observation at $(2, -2)$.
- Dataset C contains all the data from dataset B, but has another extra positive observation at $(-2, -2)$.

For each of the above datasets, answer the following questions.

1. Are the problems linearly separable in the original input space? You can use scatter plots to show these. A = yes, B = yes, C = no
2. If the problem is not linearly separable, provide an example of simple feature transformation ϕ such that the resulting problem is linearly separable *with the lowest number of features possible*. Otherwise, use the two dimensional feature $\phi(x_1, x_2) = (x_1, x_2)$.
   C: phi(x1, x2) = x1^2, x2^2
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.
5. Does x^* belong to the positive or negative class?

Question 8 [10pts]

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.



Implement a callback to stop running
Introduce batch normalisation
Introduce a bias/regulaiser
Introduce drop out