DS 250 Exam 1

Name:

Roll Number:

Instructions:

a. Read all questions carefully before you start to answer them. In case of any doubt, you can write your assumptions before answering the question.

b. This is a CLOSED BOOK exam. You can use calculators etc. Only a single sheet of formulae is allowed.

c. Use of INTERNET or Google Searches etc. is NOT ALLOWED.

d. Discussion with any other person is NOT ALLOWED. Any unfair means used will result in D or F grade and will be reported to the competent authority.

f. You will get 90 minutes to solve all the questions. We won't be giving extra time. Please submit whatever you have done at the end of the allotted time. Anyone found writing after completion time will get -20 marks.

g. PLEASE ALWAYS PROVIDE A JUSTIFICATION FOR YOUR ANSWERS. CORRECT SOLUTIONS WITHOUT ANY JUSTIFICATION WILL NOT GET FULL CREDITS. THEY MAY ALSO BE CONSIDERED AS POTENTIAL UNFAIR MEANS.

h. Make sure to write your name and roll number on the top of your sheet(s) when submitting.

i. Please write the answers in the EXAM PAPER itself. The justifications and calculations can be in the answer sheet. We will look at answer sheet only if needed.

Q1: Suppose we have three prototype images for Apple, Banana and Orange given as: x_a, x_b, x_o and a test image x_t. This question explores the connection between Cosine similarity C(x, y) and Euclidean distance, E(x, y) between a pair of images x and y.

Each of the image is an N-dimensional vector and $||x_a||_2 = ||x_b||_2 = ||x_o||_2 = ||x_t||_2 = N$

i. (4 marks) Relate the Euclidean distance of the test image with a prototype to its cosine similarity to the same prototype. Derive the formula:

$E(x, x_t) =$

ii. (3 marks) Plot E(x, x_t) vs C (x, x_t)

$iii.(1\ marks)\ If\ C(x_a, x_t) = 0.5$, then E $(x_a, x_t) =$

iv. (2 marks) If cosine similarities of x_t with $x_a, x_b, and\ x_o$are 0.5, 0.45 and 0.6 respectively, which of the statement(s) are true:

a. x_a will have the minimum Euclidean distance to x_t

b. x_o will have the minimum Euclidean distance to x_t

c. We can't conclude anything about minimum Euclidean distance to a prototype as it depends on the value of N

d. x_t must be classified as an orange.

Q2: σ is the sigmoid function. Plot the following functions in the box provided here:

i. $\sigma(2 - x)$

(2 marks)

ii. $ReLU\ (\sigma(x) - 0.5)$

(3 marks)

iii. (10 marks) Now, build a neural network which takes input x and can approximate the following curve.

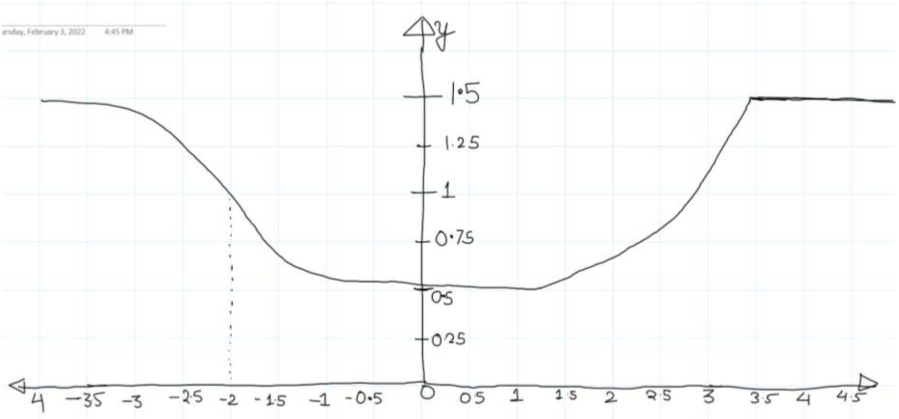


Figure 1: Neural Network

F(x) corresponding to the Neural Network is:

Briefly justify that it satisfies the following properties:

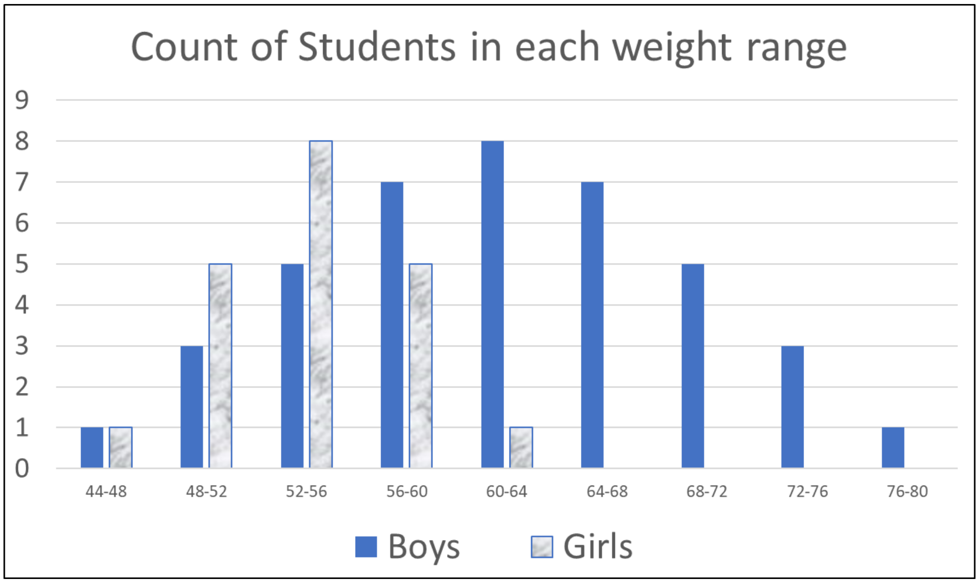
a. y=1.5 is an asymptote as x -> — ∞

b. x=-2 is an inflection point

c. Minimum value of y is 0.5

d. Maximum value of y is 1.5

3: The weights of 60 students (40 boys and 20 girls) in a CS class were measured and the following histogram plot (bar chart) was made.



The std. dev. (σ) was calculated to be 8 kg. for Boys and 4 kg. for Girls.

Write formulae for p.d.f. of weight of Girls (X_g)and Boys (X_b)

i. (2 marks) P.D.F.: $p_b(X_b$= w) =

ii. (2 marks) P.D.F.: $p_g(X_g$= w) =

iii. (3 marks) The KL divergence between the two distributions,

KL ($p_b || p_g$)=

iv. (10 marks) I have the weights of two students from the same class. Using Bayesian Analysis, predict the probability of their being boy or girl.

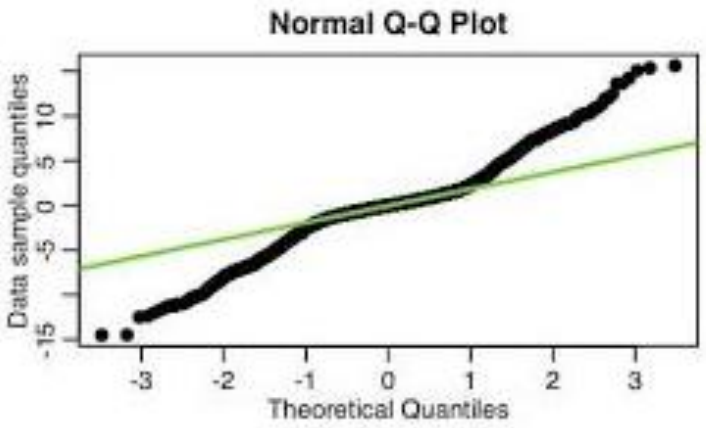
Weight of Student (kg), w	p(S =Boy|w)	p (S=Girl|w)
S_1= 58 kg		
S_2 = 50 kg		

Hints for calculations:

X	y
0	0.398942
0.5	0.352065
1	0.241971
1.5	0.129518
2	0.053991

Where y=$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

v. (3 marks) Weights of another group of 40 boys from EE were measured and observed to have the same mean and variance as boys from CS. A QQ plot was made w.r.t. normal distribution from CS. The QQ plot is shown below. What does this say about the histogram of weights of EE boys as compared to the histogram of weights of CS boys?



4. (25 marks) Engineers from American Motors (AMC) are working on a new car (AMX). They are analyzing how the mileage of this car matches up with their other cars and the competition. It is common knowledge that the mileage of a car is dependent on its Engine's displacement or size of the engine in CC and the Weight, W. They come up with two approaches.

a. First Approach: Using linear regression

They develop an equation to predict mileage as miles per gallon (mpg) using the following regression model where CC is measured in units of 100cc and W in tonne.

$$P = w_o + w_1 CC + w_2 W$$

Current value of w_0 is 30 and w_1 is -1 and w_2 is -1.

However, they feel that using gradient descent they can improve this model.

car name	$\boldsymbol{CC_i}$	$\boldsymbol{W_i}$	**Actual Mpg $(\boldsymbol{M_i})$**	**Predicted $\boldsymbol{P_i}$**	**Error $(\boldsymbol{E_i = P_i - M_i})$**	$\boldsymbol{\frac{\partial L_i}{\partial w_1}}$	$\boldsymbol{\frac{\partial L_i}{\partial w_2}}$	$\boldsymbol{P_i'}$	$\boldsymbol{E_i'}$
ford pinto	1	2	25	27	2				
amc gremlin	2.3	2.6	19	25.1	6.1				
amc hornet sportabout (sw)	2.6	3	18	24.4	6.4				
ford torino 500	2.5	3.3	19	24.2	5.2				
ford galaxie 500	3.5	4.2	14	22.3	8.3				

The above table already calculated the value of Predictions and Current Error.

We will use the following Loss function: $L = \sum_i L_i = \frac{1}{2}\sum_i E_i^2$=$\frac{1}{2}\sum_i (P_i - M_i)^2$

(10 marks) Please complete the above table where you have to evaluate the gradients at each training point, and update the predictions, P' and calculate new errors, E'. Use learning rate $\eta = 0.01$

Note: One place after decimal is fine for calculations.

(2 marks) The updated Regression Model is:

(3 marks) Better or Worse: AMX has a CC = 2.3 and W = 3.3 and its mileage is 19 m.p.g. Using the updated Regression Model, answer if it is better or worse than other cars and its competition.

b. Second Approach: Nearest neighbors can also be used to solve the problem. They use K=3 nearest neighbors to AMX (CC = 2.3 and W = 3.3) in terms of Euclidean distance and perform a weighted average to estimate the expected MPG.

$$MPG = \frac{\sum_{i \in KNN} w_i M_i}{\sum_{i \in KNN} w_i}$$

Weight $w_i = \frac{1}{1+d(x,y)}$; where d(x, y) is the Euclidean distance between x and y.

(6 marks) Calculate the weights of neighbors and the products

Index	Neighbor Weight	Neighbor MPG	Product
1			
2			
3			
Total		n.a.	

(2 marks) Compute the estimated MPG as a weighted average for the CC and W of AMX.

(2 marks) Using the KNN approach, is AMX (mileage = 19 m.p.g.) better or worse than other cars/competition?

Q5. (20 marks) In this question, we will be analyzing the reasons for heart disease and developing a decision tree for predicting the probability of having a disease.

Gender: gender of the individual using the following format: 1 = male; 0 = female

(CP) Chest-pain type: type of chest-pain experienced by the individual using the following format:
1 = typical angina; 2 = atypical angina; 0 = asymptomatic

(CA) Number of major vessels (0–1) colored by flourosopy: displays the value as integer or float.

Diagnosis of heart disease: Displays whether the individual is suffering from heart disease or not.

Gender	CP	CA	No Disease	Disease
0	0	0	9	15
0	1	0	0	11
0	2	0	0	26
1	0	0	22	19
1	1	0	3	23
1	2	0	7	24
0	0	1	1	1
0	1	1	1	3
0	2	1	1	8
1	0	1	31	1
1	1	1	3	1
1	2	1	6	5

(2.5 marks) Entropy of the dataset is:

Hint: Entropy = $\mathrm{H}(\boldsymbol{S}) = -\sum_{c \in C} p_c \ \log_2 p_c$

(2.5 marks) Calculate the information gain if we split the dataset by the following rule: Gender<=0.5

Hint: $IG = H(S) - \frac{|S_1|}{|S|} H(S_1) - \frac{|S_2|}{|S|} H(S_2)$

(2.5 marks) Calculate the information gain if we split the dataset by the following rule: CA<=0.5

(10 marks) Which of the above two rules should we use for the root-node of the decision tree? Assuming that this is the best choice, can you build the decision tree with 2 levels of internal nodes?

(2.5 marks) A male patient has typical angina chest pain. What is his probability of having a heart disease?