CSAL4243

Introduction to Machine Learning

Mid Term

1. Categorize the following as regression or classification problem. Give reason.
   a. Car speed from camera image installed on top of the camera.

   Ans: Speed is a real number, hence regression.

   b. Predicting who is going to win next cricket world cup from players and teams performance.

   Ans: Multiclass classification since the output is one of the team participating in the tournament, which is a categorical value.

   c. To determine whether a transaction of Rs. 10,000,456 is a fraud.

   Ans: Fraud or not fraud is a binary classification problem.

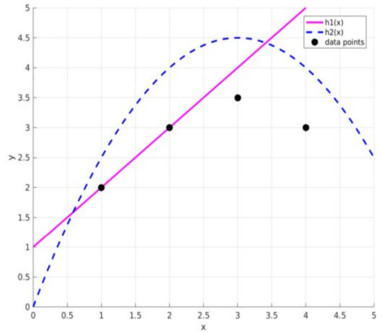
   d. Predicting company's profit at the end of the fiscal year based on performance in first two quarters.

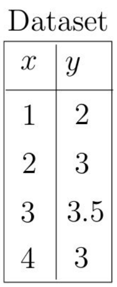
   Ans: Regression since profit is a real value.

   e. To determine whether a product review of 500 words is positive or negative based on the content of the review.

   Ans: Positive or negative review is a binary classification problem.

2. We were given a dataset as shown in table below. We used linear regression to fit two models h1(x) and h2(x) as shown below in magenta and blue respectively. The parameters of the model h1(x) are [1, 1] and for model h2(x) they are [0, 3, -0.5] while the second feature/attribute in h2(x) is x^2. Which model best fit the dataset. Justify your answer.





Dataset

x	y
1	2
2	3
3	3.5
4	3

Ans: Best fit is determined using the cost function. The model giving the lowest cost is the best fit. Cost for linear regression is given by

$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h(x_i) - y_i)^2$ **where** $h(x) = \theta^T X$

For dataset m = 4 since there are four examples given.

Model 1:

$\theta = [\theta_0 , \theta_1] = [1 , 1]$

$h(x) = \theta_0 + \theta_1 x = 1 + x$

x	y	$h(x) = 1 + x$	$(h(x) - y)^2$
1	2	2	0
2	3	3	0
3	3.5	4	0.25
4	3	5	4

Cost of model 1 = $J_1(\theta) = \frac{1}{2*4}(0 + 0 + 0.25 + 4) = 0.53$

Model 2:

$\theta = [\theta_0 , \theta_1 , \theta_2] = [0 , 3, -0.5]$
$h(x) = \theta_0 + \theta_1 x + \theta_2 x^2 = 3x - 0.5x^2$

x	y	$h(x) = 3x - 0.5x^2$	$(h(x) - y)^2$
1	2	2.5	0.25
2	3	4	1
3	3.5	4.5	1
4	3	4	1

Cost of model 2 = $J_2(\theta) = \frac{1}{2*4}(0.25 + 1 + 1 + 1) = 0.4$

Hence model 2 is a better fit than model 1.

3. Consider you are given a dataset for all the players participated in Pakistan Super League (PSL) and the category that they were assigned. A player is given a category from the list {Platinum, Diamond, Gold, Silver, Emerging, Supplementary} based on his previous record. You need to train a machine learning model that could predict player category given his data. Please answer the following:
    a. The problem is a classification or a regression problem, why ?

    Ans: Classification since output is a categorical value.

    b. What algorithm you would like to use and why?

    Ans: Any classification algorithm would work as long as its accuracy is high enough. I would go with neural network though.

c. How long will you run the algorithm?

Ans: Run it for as long as the cost is reducing at each step and stop when cost become stable. Its also called algorithm convergence.

d. How would you determine whether your model is trained properly or not?

Ans: We find the accuracy of our model on both test and training set. If both are high and there is not too much difference between the two i.e. no underfitting or overfitting then the model is trained properly.

4. What is the accuracy of below given neural network on the provided test dataset. Use approximate values i.e. 0.99 ~ 1 and 0.01 ~ 0.

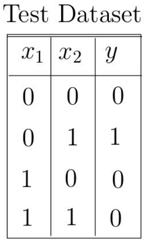
To find the accuracy of the neural network, first we find predicted output h(x). For neural network:

$a^1 = X = [x_0\ x_1\ x_2]$

$z = \theta^T a$

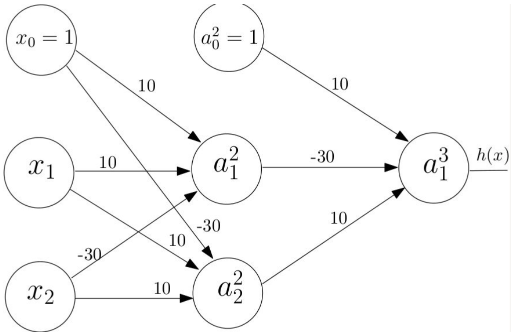
$a = g(z)$

$$h(x) = \begin{cases} 1 & \text{if } a > 0.5 \\ 0 & \text{if } a \le 0.5 \end{cases}$$



Test Dataset

x_1	x_2	y
0	0	0
0	1	1
1	0	0
1	1	0



x_1	x_2	**y**	$a_1{}^2$	$a_2{}^2$	$a_1{}^3$	**h(x)**
0	**0**	**0**	**1**	**0**	**0**	**0**
0	**1**	**1**	**0**	**0**	**1**	**1**
1	**0**	**0**	**1**	**0**	**0**	**0**
1	**1**	**0**	**0**	**0**	**1**	**1**

Accuracy = $\frac{\textit{number of examples predicted correctly}}{\textit{total number of examples}}$ **=** $\frac{3}{4}$ **= 0.75 or 75%**

5. **Bonus Question:** State a problem specific to Pakistan (be creative) that can be solved using machine learning. Mention the data (features and output) and how machine learning can be used to solve it.