Indian Institute of Technology, Kharagpur

Department of Computer Science and Engineering

Class Test 1, Autumn 2023

Machine Learning (CS60050)

Students:150 **Date**: 31-Aug-2023
Full marks: 26 **Time**: 12:00pm–1:00pm

Answer the questions in the spaces provided.

Roll Number		Section	
Name			

Question:	1	2	3	4	Total
Points:	6	8	8	4	26
Score:					

1. (6 points) (a) Given below table contains the values assigned to 6 instances of two variables **X** (input) and **Y** (output). Using Linear Regression, find the line that fits best through the given data points.

   Assume the functional form as $h_\theta(x) = \theta_0 + \theta_1 x$, where x is the input, θ_0, θ_1 are the parameters. Use mean squared error as the loss function. Starting from $\theta_0 = 0, \theta_1 = 1$, execute 4 iterations of gradient descent and write the final equation of the line. Note, in gradient descent, we use $\theta_i = \theta_i - \alpha \times \frac{\partial \text{loss}(\theta)}{\partial \theta_i}$ as the update rule. Use $\alpha = 1$. [6 points]

X	Y
2	21
4	27
6	29
8	64
10	86
12	92

$\theta_1^{new} = \theta_1^{old} - \alpha \times \frac{1}{n} \sum_{i=1}^{n} (\theta_0 + \theta_1 x - y)x$

$\theta_0^{new} = \theta_0^{old} - \alpha \times \frac{1}{n} \sum_{i=1}^{n} (\theta_0 + \theta_1 x - y)$

Calculation for the first iteration:

$\theta_1^{new} = \theta_1^{old} - 1 \times \frac{1}{6}[2(2-21) + 4(4-27) + 6(6-29) + 8(8-64) + 10(10-86) + 12(12-92)]$
$= 1 + 406 = 407$

$\theta_0^{new} = \theta_0^{old} - 1 \times \frac{1}{6}[(2-21) + (4-27) + (6-29) + (8-64) + (10-86) + (12-92)]$
$= 0 + 46.166 = 46.166$

After 1^{st} iteration: $\theta_0 = 46.166666666666664 \; and \; \theta_1 = 407.0$

After 2^{nd} iteration: $\theta_0 = -2795.8333333333335$ *and* $\theta_1 = -24140.833333333332$
After 3^{rd} iteration: $\theta_0 = 169038.99999999997$ *and* $\theta_1 = 1460440.5555555553$
After 4^{th} iteration: $\theta_0 = -10223030.72222222$ *and* $\theta_1 = -88322426.14814812$

2. (8 points) (a) Let the optimization objective for (modified) Logistic Regression is:

$$w^* = \arg\max_w \left(\sum_{i=1}^{n} y_i \text{sign}(w^T x_i)\right) \tag{1}$$

where w^* is the optimal w, $y_i \in \{+1, -1\}$ is the label that a point x_i can take and $w^T x_i$ is the distance of point x_i from the separator line/plane π. Assume, we have two candidate planes π_1 and π_2. Using the table given below, find the most optimal plane using the aforementioned optimization objective. **Note:** w is a unit vector ($||w|| = 1$). [5 points]

Points	Labels y_i	Distance from π_1	Distance from π_2
$x_1\,(1,1)$	$+1$	1	1
$x_2\,(2,1)$	$+1$	1	2
$x_3\,(3,1)$	$+1$	1	3
$x_4\,(-1,10)$	-1	10	-1
$x_5\,(1,-1)$	-1	-1	1
$x_6\,(2,-1)$	-1	-1	2
$x_7\,(3,-1)$	-1	-1	3

SOLUTION:
For plane π_1, summation will be:
$1*1+1*1+1*1+(-1)*1+(-1)*(-1)+(-1)*(-1)+(-1)*(-1) = 5$

For plane π_2, summation will be:
$1+1+1+(-1)*(-1)+1*(-1)+1*(-1)+1*(-1) = +1$

Hence, plane π_2 will be the optimal plane according to the objective mentioned in the question paper.

One can also then solve for w using different methods. No, need to show any extra calculations.

(b) In Figure 1, we show a dataset of 2-dimensional points that we wish to classify using Logistic Regression.

We will use this simple linear logistic regression model:

$$P(y=1|x,w) = g(w_0 + w_1 x_1 + w_2 x_2) = \frac{1}{1+\exp^{-(w_0+w_1x_1+w_2x_2)}} \tag{2}$$

Notice that the training data can be separated with zero training error with a linear separator. Consider training regularized linear logistic regression models where we try to maximize

$$\sum_{i=1}^{n} P(y_i|x, w_0, w_1, w_2) - C w_j^2, \tag{3}$$

where C is very large, and w_j can be any one of the weights at a time (not all the weights together). In other words, only one of the parameters is regularized in each case. Given the

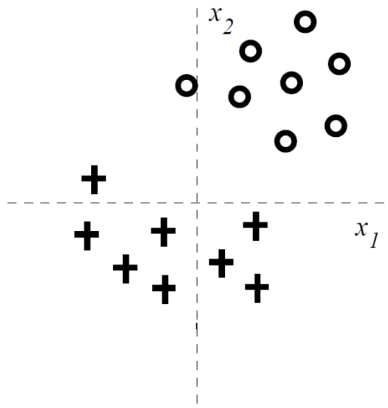


Figure 1: The 2-dimensional labeled training set, where '+' corresponds to class $y = 0$ and '∘' corresponds to class $y = 1$.

training data in Figure 1, how does the training error change with regularization of each parameter w_j? Say how training error increases or decreases for each of the three cases (when w_1 is regularized, w_2 is regularized, w_3 is regularized). Give brief justification. [3 points]

SOLUTION:

regularizing w_0 :

Training error increases. When we regularize w_0, then the boundary will eventually go through the origin (bias term set to zero). For very large C, the training error increases as there is no good linear vertical separator of the training data.

regularizing w_1 :

Training error remains the same. When we regularize w_1, the resulting boundary can rely less and less on the value of x_1 and therefore becomes more horizontal and the training data can be separated with zero training error with a horizontal linear separator.

regularizing w_2 :

Training error increases. When we regularize w_2, the resulting boundary can rely less and less on the value of x_2 and therefore becomes more vertical. For very large C, the training error increases as there is no good linear vertical separator of the training data.

3. Consider the following training instances described in terms of two numerical attributes A1 and A2.

A1	A2	Class
4	1	1
8	6	1
5	4	0
7	3	1
1	2	0
9	5	1
6	6	0

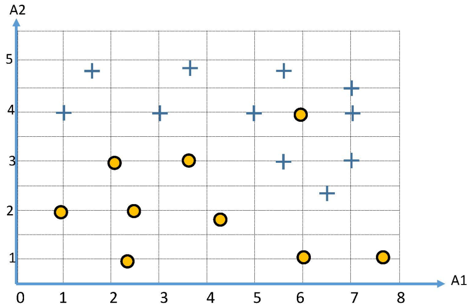
(a) (2 points) What is the entropy at the root of the tree?

(b) (6 points) Use Information Gain to decide the split at the root. Find the best information gain for each of the attributes. Clearly state the final attribute split value,

entropy at the leaf nodes and the information gain.
You may use the following approximate values is required:

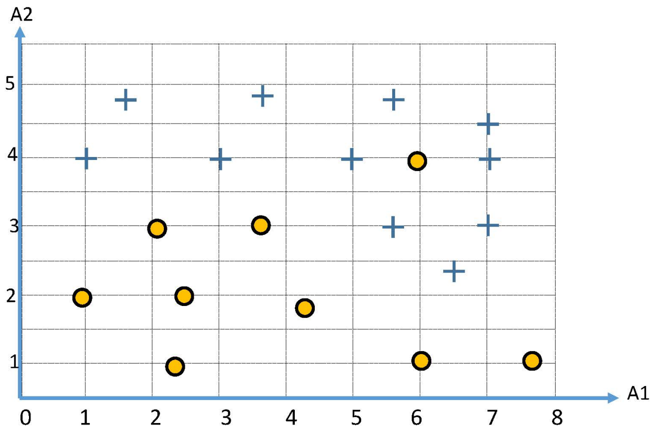
$$\log_2 2 = 1, \log_2 3 = 1.585, \log_2 5 = 2.322, \log_2 7 = 2.807$$

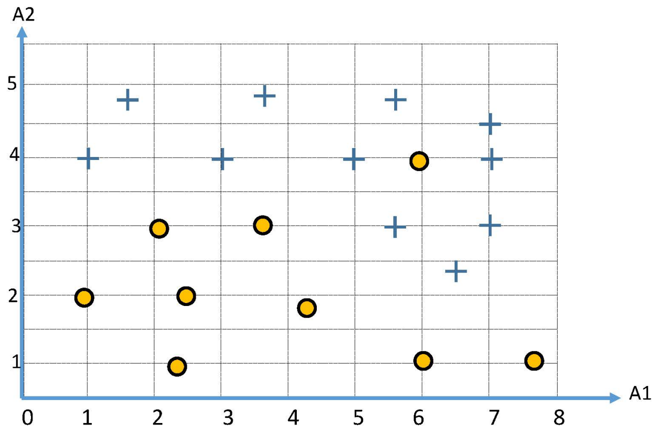
4. (4 points) Consider the following training instances corresponding to a 2-class classification problem in terms of two numerical attributes.



(a) Draw the decision boundaries of a possible decision tree that may overfit the data. You may either draw the decision tree or show the decision boundaries in the diagram below.

(b) Draw the decision boundaries of a possible decision tree that is likely to underfit the data.





[Extra space for rough work]