Ayush Garg
PUID: 0033794803

Homework 1

CS 57800

1. Review Questions

1. (a) This forms a geometric distribution
P(first head at k+1 toss) = $\lambda(1-\lambda)^k$

   (b) Expected number of tosses to get first head in geometric distribution
E[number of tosses to get first head] = $\frac{1}{\lambda}$

2. (a) $\frac{\partial f}{\partial x} = 6x - y - 11$

   $\frac{\partial f}{\partial y} = 2y - x$

   (b) For getting the point that minimizes the function f, we equate the partial derivatives obtained in the first part to zero

$$6x - y - 11 = 0 \qquad 2y - x = 0$$
$$6x - y - 11 = 0 \qquad x = 2y$$

Plugging value of x from 2 in 1

$$11y = 11$$

$\implies$ y=1 and x=2

3. (a) i. For n=2

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \text{b} = 0$$

Let the values for $w_1 = 2$ and $w_2 = 1$ and b = 2, so the equation becomes:
$2x_1 + x_2 + 2 = 0$
This hyperplane is a line

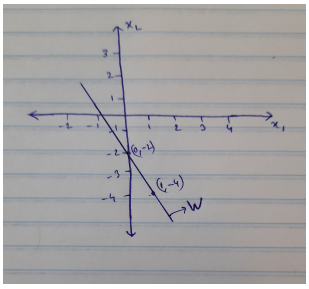

Figure 1: W is the hyperplane

ii. For n=3

$$\begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \text{b} = 0$$

Let the values for $w_1 = 1$ and $w_2 = 1$ and $w_3 = 1$ and b = 2, so the equation becomes:
$x_1 + x_2 + x_2 + 2 = 0$
This hyperplane passes through the points (0,0,-2), (0,-2,0), (-2,0,0) and (-1,-1,0).

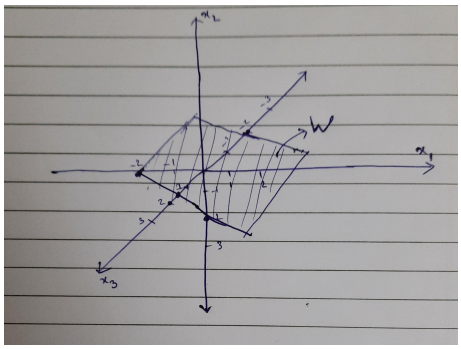

Figure 2: W is the hyperplane

(b) Distance between these two hyperplanes:
$\frac{|b_2 - b_1|}{||w||_2}$

2. Basic Concepts

1. (a) Training Set - It is the set of examples which are used during the learning process of a model and to fit the parameters.

   (b) Test Set - It is the set of examples which is used to assess the performance of a model. This is independent of the training set.

   (c) Validation Set - It is the set of examples which is used to fine tune the hyper-parameters of a model.

2. No, the validation set cannot be used as a test set. Validation set is used for fine tuning the hyper-parameters for the model. So it cannot be used as test set because the model requires it to fine tune the hyperparameters whereas we use the test set to check the accuracy of our best trained model.

3. Hypothesis h is said to be overfitting the training data if there is another hypothesis h', such that h has a smaller error than h' on the training data but h has larger error on the test data as compared to h'.

4. (a) False, because for a model to overfit the data, the training error should be less than testing error.

   (b) True, here f overfits the training data as training error is less than testing error. It fails to generalize on the test dataset.

3. Decision Trees

1. (a)

$$\begin{aligned}\text{Entropy of the target variable (Buy)} &= -\frac{4}{11}log_2(\frac{4}{11}) - \frac{7}{11}log_2(\frac{7}{11}) \\ &= -\frac{4}{11}(-1.46) - \frac{7}{11}(0.65) \\ &= 0.3636 * 1.46 + 0.6363 * 0.65 \\ &= 0.531 + 0.413 \\ &= 0.944\end{aligned}$$

   (b) The following attributes would be considered by the algorithm - Pages, Famous Author, Category and Cover Color. As Pages is a continuous variable, we will find a threshold value for which information gain is maximum. To get the threshold value, we first sort the values in Pages along with target labels and then check for information gain at the mean of two values where the target label is changed.

   (c)

$$\begin{aligned}\text{Entropy for Famous Author} &= \frac{7}{11}(-\frac{5}{7}log_2(\frac{5}{7}) - \frac{2}{7}log_2(\frac{2}{7})) + \frac{4}{11}(-\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4})) \\ &= \frac{7}{11}(0.3467 + 0.516) + \frac{4}{11}(1) \\ &= 0.549 + 0.3636 \\ &= 0.912\end{aligned}$$

$$\begin{aligned}\text{Entropy for Category} &= \frac{5}{11}(-\frac{4}{5}log_2(\frac{4}{5}) - \frac{1}{5}log_2(\frac{1}{5})) + \frac{6}{11}(-\frac{3}{6}log_2(\frac{3}{6}) - \frac{3}{6}log_2(\frac{3}{6})) + 0 \\ &= \frac{5}{11}(0.2575 + 0.4644) + \frac{6}{11}(1) \\ &= 0.3281 + 0.5454 \\ &= 0.8735\end{aligned}$$

$$\begin{aligned}\text{Entropy for Cover Color} &= \frac{9}{11}(-\frac{6}{9}log_2(\frac{6}{9}) - \frac{3}{9}log_2(\frac{3}{9})) + \frac{2}{11}(-\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2})) \\ &= \frac{9}{11}(0.39 + 0.528) + \frac{2}{11}(1) \\ &= 0.7513 + 0.1818 \\ &= 0.933\end{aligned}$$

For the continuous attribute like Pages, we first sort the values along with their corresponding labels and then check for the split on the middle value of the of rows where there is flip in the target label (Buy).
In the sorted order, the flip comes in between 45 and 50, 72 and 100, 100 and 120, 150 and 200, 200 and 300, 300 and 1000

i. Split at 47.5

$$\begin{aligned}\text{Entropy for Pages} &= \frac{1}{11}(-\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1})) + \frac{10}{11}(-\frac{3}{10}log_2(\frac{3}{10}) - \frac{7}{10}log_2(\frac{7}{10})) \\ &= \frac{1}{11}(0) + \frac{10}{11}(0.881) \\ &= 0 + 0.800 \\ &= 0.800\end{aligned}$$

ii. Split at 86

$$\begin{aligned}\text{Entropy for Pages} &= \frac{3}{11}(-\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3})) + \frac{8}{11}(-\frac{5}{8}log_2(\frac{5}{8}) - \frac{3}{8}log_2(\frac{3}{8})) \\ &= \frac{3}{11}(0.9182) + \frac{8}{11}(0.9543) \\ &= 0.250 + 0.694 \\ &= 0.944\end{aligned}$$

iii. Split at 110

$$\begin{aligned}\text{Entropy for Pages} &= \frac{4}{11}(-\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4})) + \frac{7}{11}(-\frac{5}{7}log_2(\frac{5}{7}) - \frac{2}{7}log_2(\frac{2}{7}))\\ &= \frac{4}{11}(1) + \frac{7}{11}(0.862)\\ &= 0.3636 + 0.548\\ &= 0.912\end{aligned}$$

iv. Split at 175

$$\begin{aligned}\text{Entropy for Pages} &= \frac{7}{11}(-\frac{5}{7}log_2(\frac{5}{7}) - \frac{2}{7}log_2(\frac{2}{7})) + \frac{4}{11}(-\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{4}))\\ &= \frac{7}{11}(0.862) + \frac{4}{11}(1)\\ &= 0.548 + 0.3636\\ &= 0.912\end{aligned}$$

v. Split at 250

$$\begin{aligned}\text{Entropy for Pages} &= \frac{3}{11}(-\frac{2}{3}log_2(\frac{2}{3}) - \frac{1}{3}log_2(\frac{1}{3})) + \frac{8}{11}(-\frac{5}{8}log_2(\frac{5}{8}) - \frac{3}{8}log_2(\frac{3}{8}))\\ &= \frac{3}{11}(0.9182) + \frac{8}{11}(0.9543)\\ &= 0.250 + 0.694\\ &= 0.944\end{aligned}$$

vi. Split at 675

$$\begin{aligned}\text{Entropy for Pages} &= \frac{1}{11}(-\frac{0}{1}log_2(\frac{0}{1}) - \frac{1}{1}log_2(\frac{1}{1})) + \frac{10}{11}(-\frac{3}{10}log_2(\frac{3}{10}) - \frac{7}{10}log_2(\frac{7}{10}))\\ &= \frac{1}{11}(0) + \frac{10}{11}(0.881)\\ &= 0 + 0.800\\ &= 0.800\end{aligned}$$

So, all these calculations show that we should split the data on the Pages attribute as that would give us the highest information gain (max difference with entropy of the target variable). In that, we can either split on 47.5 or 675 as both give the same information gain.

(d) To deal with the missing values in the dataset, we can use any of these techniques:

i. fill the missing values by the mode value for that attribute (mainly categorical attribute)
ii. fill the missing values by the mean or median value for that attribute (mainly continuous attribute)
iii. drop the rows which have missing attributes

2. (a) Code files along with dataset uploaded through turnin.

(b) Trends :
The validation accuracy decreases somewhat at depth 4 and then increases till depth 7. Afterwards, the validation accuracy remains almost constant. The test accuracy keeps on fluctuating a bit in the starting and after one point (depth =7) decreases and then after few more depths becomes almost constant.
MaxDepth :

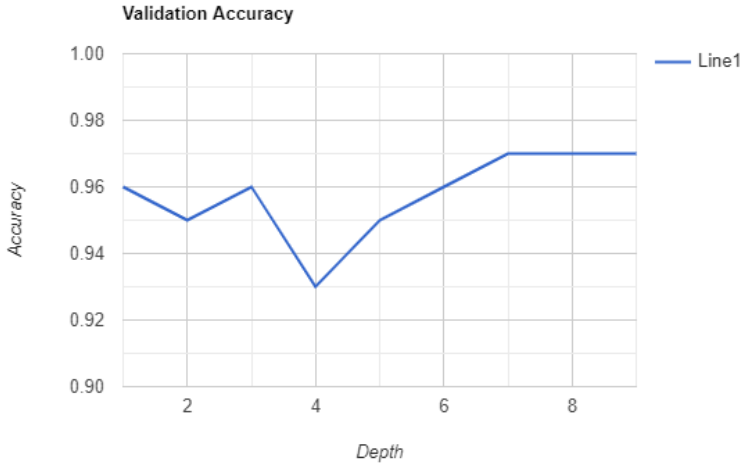

Figure 3: Validation Accuracy

Based on the graphs shown, I have chosen the maximum depth at 7 because afterwards the validation accuracy become almost constant. It is not a good choice to take increase it more than 7 because the model then starts to overfit the data.

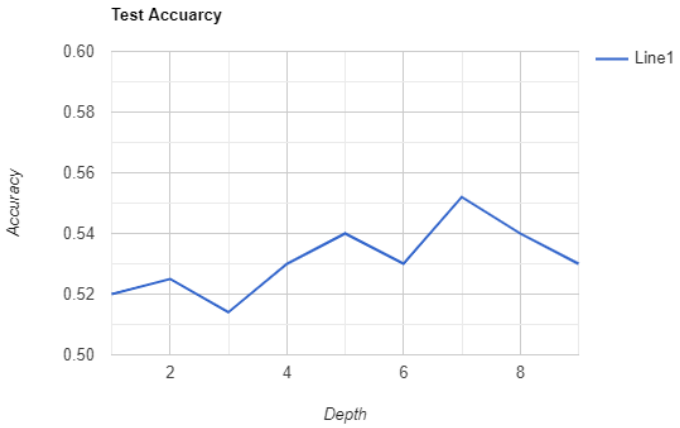

Figure 4: Test Accuracy