Alexandria University
Faculty of Engineering

Introduction to Machine Learning
Assigned: Saturday, August 13, 2023
Due: Thursday, August 15,2023

# Sheet 2

Q1

Using logistic regression we can estimate the probability of a vector belonging to a class. The probability of vector (vec) can be computed as

Prob(vec)= exp(dot(vec,w))/(1+exp(dot(vec,w)))

1. What is the prob of vec=(-1,1,-1) if the weight vector w is (-ln(4),ln(2),-ln(3))? Show steps.
2. What can be a weight vector that makes the Prob. of vec=(-1,1,-1) reach 1 or very close to it? Tell why?
3. For which of these weight vectors are small changes between test instances likely to make large changes in classification? Which of these models do you think generalizes better and why?
    a. w1 = (10000, −2384092, 24249, 284924, −898)
    b. w2 = (1.213, −.123, 2.23, 3.4, −2)

Q2

Give the following data we need to find a linear separator between positive and negative examples. Every record has two numerical attributes a1, a2.

| ID | a1 | a2 | label |
|---|---|---|---|
| 1 | 1 | 1 | + |
| 2 | 1 | 2 | + |
| 3 | 2 | 2 | + |
| 4 | 4 | 1 | - |
| 5 | 4 | 2 | - |
| 6 | 3 | 3 | - |
| 7 | 3 | 2 | - |

1. Plot the data on graph paper and sketch a linear separator of your choice.
2. We will invent a new linear separator perpendicular to a line connecting any two samples from opposite classes.
    a. How many possible separators can be generated from the data? **Tricky**
    b. How many of the separators are perfect for the data? Which one is the best regarding the misclassification error?
    c. For the new samples (1,3) and (3,1) which class are you going to select given the best classifier, you chose in b.

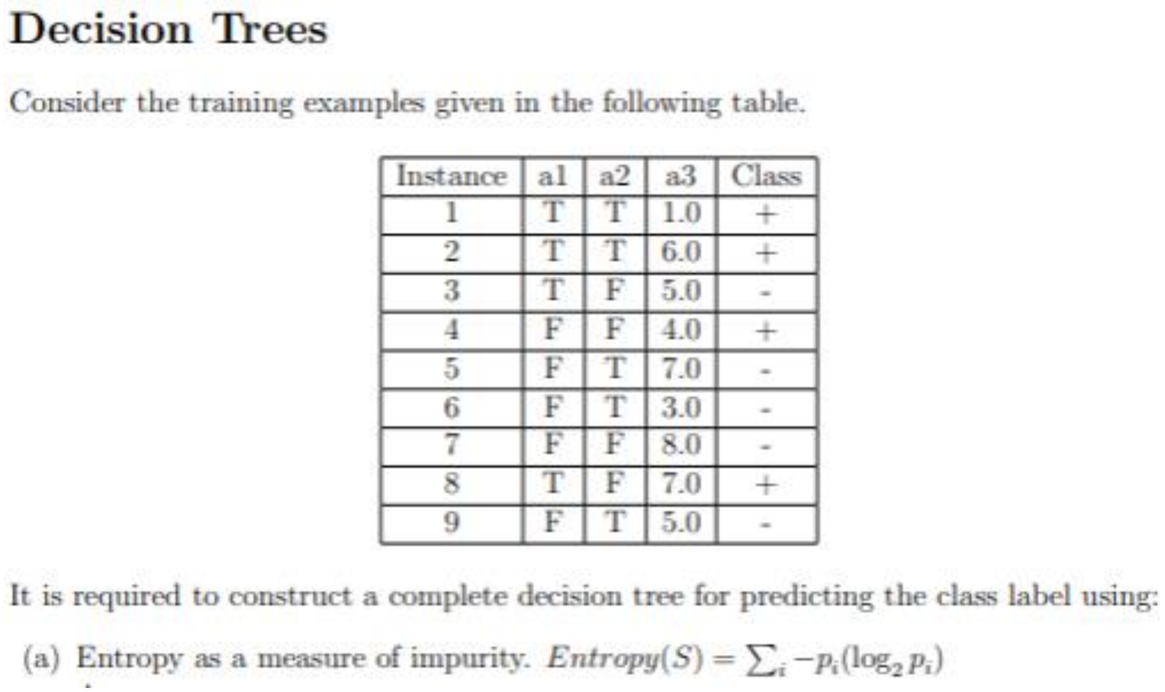
Q3

1. Consider a 1-layer neural net with three input units, 1 output unit, no hidden units and no bias terms. Suppose that the output unit uses a sigmoid activation function, i.e., **y = 1/(1 + e $^{-z}$ )**, where **z** is the total input to the unit. Let **y** be the computed output of the neural net. Let **d** be the desired output.
   Let **C = [− d log y − (1−d) log (1−y)]** be the cross-entropy error.

   1. Write down the equations for a single step of weight updates by gradient descent (based on a single data sample) and derive all the necessary derivatives.

   2. Simplify your answers and be sure to clearly identify all the variables you use. **Hint: use the chain rule and the following results: ∂y / ∂z = y(1 − y) and ∂ log u / ∂u = 1/ u**

Q4



## Decision Trees

Consider the training examples given in the following table.

| Instance | a1 | a2 | a3 | Class |
|---|---|---|---|---|
| 1 | T | T | 1.0 | + |
| 2 | T | T | 6.0 | + |
| 3 | T | F | 5.0 | - |
| 4 | F | F | 4.0 | + |
| 5 | F | T | 7.0 | - |
| 6 | F | T | 3.0 | - |
| 7 | F | F | 8.0 | - |
| 8 | T | F | 7.0 | + |
| 9 | F | T | 5.0 | - |

It is required to construct a complete decision tree for predicting the class label using:

(a) Entropy as a measure of impurity. *Entropy*$(S) = \sum_i -p_i(\log_2 p_i)$

Q5

For the following confusion matrix compute:

1. Accuracy, Error rate.

2. Precision and recall , F1- measure for every class , F1 = (2*Precision*Recall) / (Precision + Recall) Then compute the Average F1- Measure.

3. Discuss which measure (Accuracy or average F1) is better suited for this specific problem.

Predicted

| | | Amloki | Ata | Bilombo | Chalta | Orbori | Sapota |
|---|---|---|---|---|---|---|---|
| | Amloki | 33 | 0 | 5 | 0 | 0 | 0 |
| | Ata | 3 | 47 | 0 | 2 | 0 | 0 |
| Actual | Bilombo | 7 | 0 | 32 | 3 | 0 | 0 |
| | Chalta | 3 | 0 | 2 | 36 | 0 | 0 |
| | Orbori | 2 | 2 | 2 | 2 | 27 | 0 |
| | Sapota | 3 | 0 | 2 | 2 | 0 | 25 |

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Q6

Suppose binary-valued random variables X and Y have the following joint distribution:

| | $Y = 0$ | $Y = 1$ |
|---|---|---|
| $X = 0$ | $1/8$ | $3/8$ |
| $X = 1$ | $2/8$ | $2/8$ |

Determine the information gain IG(Y|X). You may write your answer as a sum of logarithms.