Homework 1

Nishant Agarwal

September 17, 2015

1.Decision Trees

1.(a) IF x_1 is TRUE then YES
IF x_1 is FALSE then
IF x_2 is FALSE then NO
ELSE IF x_2 is TRUE then
IF x_3 is TRUE the YES
ELSE IF x_3 is FALSE then NO

1.(b) IF x_1 is TRUE then
IF x_2 is TRUE then NO
IF x_2 is FALSE then YES
ELSE IF x_1 is FALSE then
IF x_2 is TRUE then YES
IF x_2 is FALSE then NO

1.(c)) IF x_1 is TRUE then
IF x_2 is TRUE then YES
ELSE IF x_2 is FALSE then
IF x_3 is TRUE the YES
ELSE IF x_3 is FALSE then NO
IF x_1 is FALSE then
IF x_2 is FALSE then NO
ELSE IF x_2 is TRUE then
IF x_3 is TRUE the YES
ELSE IF x_3 is FALSE then NO

2.(a) The possible number of functions to map the four feature to Boolean functions is $2^{2.2.3.4} = 2^{48}$. The number of functions consistent with the given training dataset is 9.

2.(b)The ratio of positive labels is 5/9 and ratio of negative labels is 4/9

$$\begin{aligned} Entropy(S) = H(S) &= -p_+ \log_2 p_+ + p_- \log_2 p_- \\ &= -(5/9.log_2(5/9) + 4/9.log_2(4/9)) \\ &= 0.991 \end{aligned}$$

2.(c) The formula for information gain is :-

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} |S_v|/|S| Entropy(S_v)$$

Using the Entropy Formula from 2(b)
Friday:NO
$p_+ = 4/6, p_- = 2/6$
$H_N = -(4/6 \log_2(4/6) + 2/6 \log_2(2/6)$
$H_N = 0.918$

Friday:Yes $p_+ = 1/3, p_- = 2/3$
$H_N = -(1/3 \log_2(1/3) + 2/3 \log_2(2/3)$
$H_N = 0.918$

Expected Entropy = 6/9.0.918 + 3/9.0.918 = 0.918

Information Gain = 0.991 − 0.918 = 0.073 (0.991 is entropy of labels calculated in 2b)

Similarly for all attributes of each feature vector we get:
Hungr:Yes
Entropy = 0.722

Hungry:No
Entropy = 0.811

Expected Entropy = 0.762

Information Gain = 0.991 - 0.762 = 0.229

Patrons:Some
Entropy = 0

Patrons:Full
Entropy= 0.811

Patrons:None
Entropy = 0

Expected Entropy = 0.360

Information Gain = 0.991 - 0.360 = 0.631

Type:French
Entropy = 1

Type:Thai
Entropy = 0.918

Type:Chinese
Entropy = 0.918

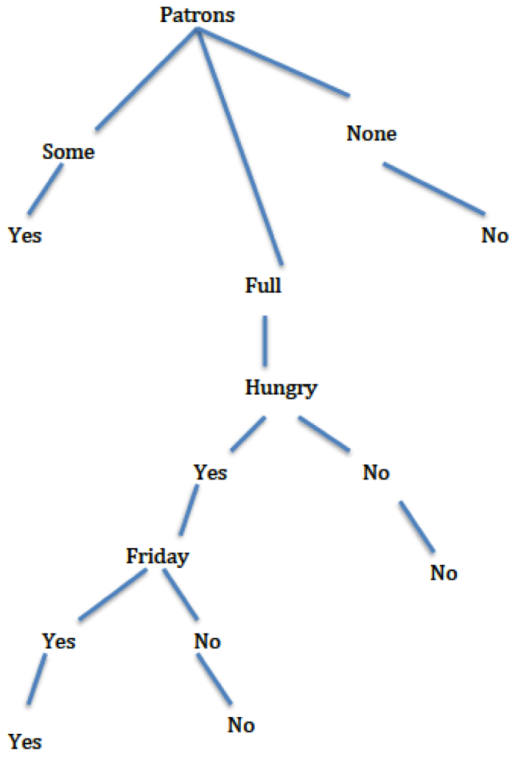
Type:Italian
Entropy = 0

Expected Entropy = 0.834

Information Gain = 0.991-0.834 = 0.157

2.(d) As per the information gain calculated above in 2(c) the highest gain is of Patrons. Therefore Patrons will be the root of the ID3.

2(e)



2(f) Predicted labels:

$$i)Yes$$
$$ii)No$$
$$iii)Yes$$

Hence the accuracy is $2/3 = 0.6667$ or 66.67%

3(a) Misclasification(S) $= 1 - \max_i p_i$
i) Hence the
$InformationGain = Misclassification(S) - \sum_{v \in A} |S_v|/|S| Misclassification(S_v)$

ii) Using the Misclassification Formula from 3(a) (i)

Missclassification of labels $= 1 - \max 4/9, 5/9 = 1 - 5/9 = 4/9$
Friday:NO
$p_+ = 4/6, p_- = 2/6$
$Misclassification(No) = 1 - \max(4/6, 2/6) = 1 - 4/6 = 2/6$

Friday:Yes $p_+ = 1/3, p_- = 2/3$
$Misclassification(Yes) = 1 - \max 1/3, 2/3 = 1 - 2/3 = 1/3$

Expected Misclassification $= 6/9.2/6 + 3/9.1/3 = 1/3$

Information Gain $= 4/9 - 1/3 = 1/9 = 0.11$

Similarly for all attributes of each feature vector we get:

Information Gain for Hungry = 4/9 - 2/9 = 0.222

Information Gain for Patrons = 4/9 - 1/9 = 3/9 = 0.333

Information Gain for Type = 4/9 - 1/3 = 0.111

The root is Patrons

3(b) $Gini(s) = \sum_i p_i(1 - p_i)$
Using the above equation we get:

Gini(labels) $= 4/9.5/9 + 5/9.4/9 = 40/81$

Friday:NO
$p_+ = 4/6, p_- = 2/6$
$Gini(No) = 4/6.2/6 + 2/6.4/6 = 16/36$

Friday:Yes $Gini(Yes) = 4/9$

Expected Gini = 72/162

Information Gain $= 40/81 - 72/162 = 0.049$

Similarly for all attributes of each feature vector we get:

Information Gain for Hungry = 0.149

Information Gain for Patrons= 0.327

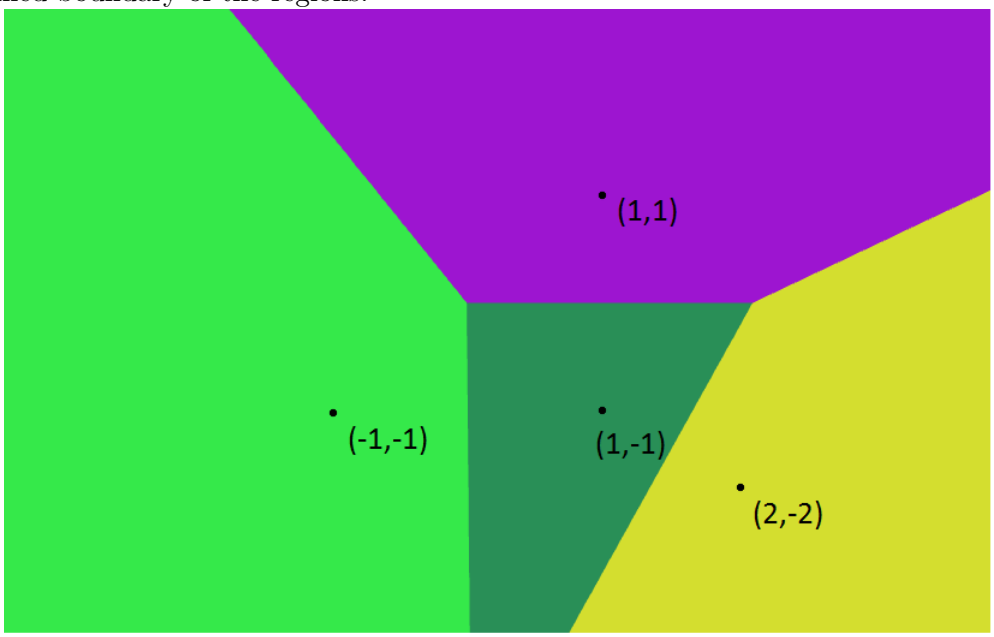
Information Gain for Type= 0.086

The root is Patrons

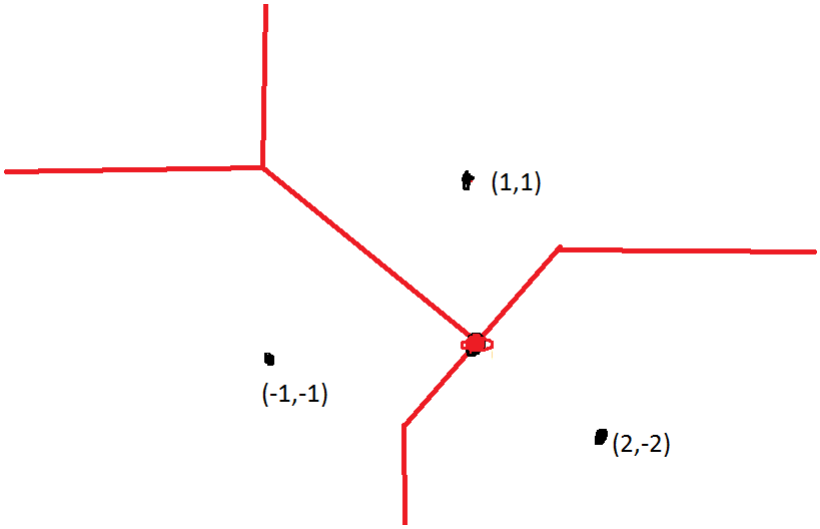
2.Neighbors

1. Euclidean Voronoi Map

In this using the euclidean distance we passed a perpendicular bisector through it each euclidean line. The intersection points of each bisector formed boundary of the regions.



2. Manhattan Distance Voronoi Using the manhattan distance formula we found regions from where the points were equidistant to regions in doin so for the three points common regions got absorbed as they fell being close to



that point.

3.taking the points from the above question we can use it to form our training set.

the point (2,0) is closer to c in manhattan distance but closer to A in the euclidean distance.

3.Experiment

1. The decison tree structure is like a tree generalized for multiple links and values for links. The structure stores if it is a leaf and also on which column the tree has split. The formation of the tree is part of ID3 algorithm where the subtree is added to the parent node. The tree node stores label data in case of leaf.

2. The accuracy of my decision tree on the tic tac toe data is 85.7%

3. In KNN i have used comparison of characters to calculate distance. If the character is same distance is 0 else 1. This is done against columns of testdata with training data. Each row of test data against all of the training data.

4. The average accuracy of each K value is as follows:

K value:1 80.315%
K value:2 82.021%
K value:3 89.1076%
K value:4 88.7139%
K value:5 91.8635%
Final K value is : 5 with avg accuracy 91.8635%
The Final Accuracy using K = 5 is 92.8571%