CSE 537 - Artificial Intelligence

Report: Project 4

Khan Mostafa 109365509 **Abhijit Betigeri** 109229784

{khan.mostafa, abhijit.betigeri}@stonybrook.edu
Department of Computer Science
Stony Brook University

Decision Tree, Naïve Bayes, K-means

Q1. Implementing the Decision Tree Algorithm

Methodology Used: We use the ID3 to build a decision tree from the fixed set of data instances given. The resulting tree is used to classify the future samples – since classes created by ID3 are inductive, here we use the training set given, and classes created by ID3 are used for all future instances . The leaf nodes of the decision tree contains class name & non-leaf node contains the decision made. The decision node here is the attribute test with each branch being the possible value of the attribute. ID3 used information gain logic to see which attribute goes into the decision node. We use the property of information gain for attribute selection describing how well the given attributes separates the training examples into targeted classes. The one with the highest information is selected. To define gain we use the concept of entropy – signifying the amount of information in an attribute.

Growing the tree can be limited up to a number of examples. We stop branching when the size of examples decrease a passed threshold.

Execution Details
We have reporting the results till the threshold of 50.

Command: `python learning.py -q1`

Threshold	accuracy
0	0.920973
1	0.920973
2	0.920973
3	0.942249
4	0.942249
5	0.942249
6	0.936170
7	0.936170
8	0.936170
9	0.957447
10	0.957447
11	0.957447

12	0.957447
13	0.957447
14	0.957447
15	0.957447
16	0.957447
17	0.957447
18	0.957447
19	0.957447
20	0.957447
21	0.957447
22	0.957447
23	0.957447
24	0.957447

25	0.957447
26	0.957447
27	0.957447
28	0.957447
29	0.957447
30	0.957447
31	0.957447
32	0.957447
33	0.957447
34	0.957447
35	0.957447
36	0.957447
37	0.957447

38	0.957447
39	0.957447
40	0.957447
41	0.957447
42	0.957447
43	0.957447
44	0.957447
45	0.957447
46	0.957447
47	0.957447
48	0.957447
49	0.957447

Q2. Implementing the Naïve Bayes Algorithm

Methodology Used: In Naïve Bayes classifier an assumption that the presence of a particular feature of the class is unrelated to the presence of any other feature. Using the Bayesian interpretation there is a linking of the degree of belief in a proposition before and after accounting for evidence.

We use a technique called Laplace smoothing for the parameter estimation which accounts for the unobserved event. The technique is more robust and will not fail completely when the data that has never been observed in the training shows up.

Execution Details

With Laplace smoothing: 0.930091:
`python learning.py -q2`
Accuracy: 0.930091

Baseline (no smoothing) 0.884498:
`python learning.py -q2`
Accuracy: 0.884498

Tie break: republican

Q3. Implementing the K-means Algorithm

Methodology Used: This unsupervised learning algorithm is used in classifying the data set through a certain number of clusters fixed a priori.

The steps used are:

1. We place |P| points into the space represented by the objects that are being clustered. Use the initial set of the centroids already given.
2. Assign each point to a cluster that minimizes distance to centroid.
3. After assigning all points to some cluster, recalculate each of the k centroids.
4. Repeat steps 2 and 3 until centroids converge, i.e. no longer move or change. This helps in separation of the objects into groups from which the metric to be minimized is calculated.

Execution Details
Command 1: `python learning.py -q3.1 -q3.2`

`centroids = [(30, 30), (150, 30), (90, 130)])`
Question 3.1 centroid: (32,82)
Question 3.1 centroid: (108,23)
Question 3.1 centroid: (126,125)

Question 3.2 centroid: (47,88)
Question 3.2 centroid: (129,40)
Question 3.2 centroid: (125,100)

centroids = [(30, 60), (150, 60), (90, 130)]
Question 3.1 centroid: (32,82)
Question 3.1 centroid: (108,23)
Question 3.1 centroid: (126,125)

Question 3.2 centroid: (48,47)
Question 3.2 centroid: (128,80)
Question 3.2 centroid: (48,103)