Data Mining: Assignment 3

Luyifan 3110101132

June 19, 2014

1 K-Nearest Neighbor

1.1

Answer: see knn.m , and picture 1.a1,1.a2,1.a3

1.2

We have seen the effects of different choices of K. How can you choose a proper K when dealing with real-world data?

We can use cross validation, divide the training data into N folds,and N-1 folds is used as training data, and the other one fold is useding as test data. and then find the best K in chooses.

1.3

see getTrainData , and hack.m and picture 1.c1,1.c2,1.c3

2 Decision Tree and ID3

2.1

There are 27 balls, where one of them is lighter than the others. Please design a decision tree to find out which one is lighter with 3 trials. Prove that your decision tree is optimal in terms of number of trials.

2.2

Answer:First divide 27 balls into 3 parts, for each part there are 9 balls. compare two part whether is the same weight, if they are the same, the lighter ball is in the third part. if not , The lighter ball in the lighter part. then in the second trial, divide 9 balls into 3 part , and the same as before. and choose the part witch contains the lighter ball.And in the third trial,do the same as before , we can find the ligher ball.

Consider the scholarship evaluation problem: selecting scholarship recipients based on gender and GPA. Given the following training data: Draw the decision tree that would be learned by ID3 algorithm and annotate each non-leaf node in the tree with the information gain attained by the respective split.

Answer: First we calculate the Entropy , + has 200 , - has 250 , so

$$Entropy = -\frac{200}{450}log\frac{200}{450} - \frac{250}{450}log\frac{250}{450} = 0.298343$$

if we first use gender , we have 205 female and 105 + , and we have 245 male and 95 +

$$Entropy = -\frac{205}{450}(\frac{105}{205}log\frac{105}{205}+\frac{100}{205}log\frac{100}{205})-\frac{245}{450}(\frac{95}{245}log\frac{95}{245}+\frac{150}{245}log\frac{150}{245}) = 0.294962$$

$$\Delta Entropy = 0.003381$$

if we first use GPA , we have 235 High GPA and 185 + , and we have 215 Low GPA ans 15 +

$$Entropy = -\frac{235}{450}(\frac{185}{235}log\frac{185}{235}+\frac{50}{235}log\frac{50}{235})-\frac{215}{450}(\frac{15}{215}log\frac{15}{215}+\frac{200}{215}log\frac{200}{215}) = 0.169895$$

$$\Delta Entropy = 0.128448$$

so we choose GPA first and then Gender and in 235 High GPA , has 115 female and 95 + in it and 120 male and 90 + in it and in 215 Low GPA , has 90 female and 10 + in it and 125 male and 5 + in it

$$Entropy = -\frac{235}{450}(\frac{115}{235}(\frac{95}{115}log\frac{95}{115}+\frac{20}{115}log\frac{20}{115})+\frac{120}{235}(\frac{90}{120}log\frac{90}{120}+\frac{30}{120}log\frac{30}{120}))-$$

$$\frac{215}{450}(\frac{90}{215}(\frac{10}{90}log\frac{10}{90}+\frac{80}{90}log\frac{80}{90})+\frac{125}{215}(\frac{5}{125}log\frac{5}{125}+\frac{120}{125}log\frac{120}{125})) = 0.166964$$

$$\Delta Entropy = 0.00293$$

3 K-Means Clustering

3.1

Answer:see kmeanTest.m and kmeans.m and picture 3.a

3.2

You should observe the issue that the outcome of k-means algorithm is very sensitive to cluster centroids initialization form the above experiment. How can we get a stable result using k-means?

Answer: we can do it some times, and choose the result which have the smallest cost in total.

3.3

see Kmean_digit.m and picture 3.c1, 3.c2, 3.c3

3.4

What is the compress ratio if we set K to 64

see vq.m and picture 3.d1-8, 3.d1-16 , 3.d1-32 , 3.d1-64 , 3.d2-8, 3.d2-16 , 3.d2-32 , 3.d2-64 First we should store the 64 center point, eacch point 24bits , and each pixel should 64=2^6 use 6 bits, if the image is 640*480=307200, if not compress , it need 307200*24 bits , if it has compressed, only need 24*64+6*307200bits so the compress ratio

$$= 1 - \frac{307200 * 6 + 24 * 64}{24 * 307200} = 74.98\%$$

4 Spectral Clustering

4.1

see code spectral_exp1.m and spectral.m and knn_graph , and picture 4.a1 , 4.a2

4.2

see code spectral_exp2.m and picture 4.b and
Total times 100 , Spectral: Avg accracy : 0.873177 Avg mihtat :0.641648 , KTotal times 100 , Kmeans: Avg accracy : 0.520106 Avg mihtat :0.311197

4.3

see code spectral_exp3.m and answer.py build.py makeMatrix.py and I divide the renren friend into 5 group, by Spectral Clustering, we can see the answer.txt the first group is yunfeng the second group is my university friends the third group is about quanfuxia the fourth group is my friends in high school the fifth group is my friends in junior high school