Q2:

(a)Vehicle Silhouette Dataset:

We scale and pre-process the data using linear discriminant analysis from the previous assignment.

(b)

1. **Uncertainty Sampling:**

**Least Confident:**

Naïve-Bayes classifier is used for calculation of the class probabilities.

In the algorithm below, we first calculate class-wise probabilities of each row.

Then we find the maximum class probability in each row. The data is then arranged in an increasing order and the row which is the most difficult to classify i.e. the row with minimum value of maximum class probability is labelled and added to the labelled dataset.

library(naivebayes)

##Least Confidence

for(i in c(1:(nrow(vehicle\_dataset)\*percent))){

nb<-naive\_bayes(Class~.,labelledData.lda)

xtab<-table(predict(nb,unlabelledData.lda, type="class"),unlabelledData.lda$Class)

acc=sum(diag(xtab))/nrow(unlabelledData.lda)\*100

print(acc) a<-predict(nb,unlabelledData.lda, type="prob")

unlabelledData.lda<-cbind(unlabelledData.lda,"max"=apply(a, 1, max))

unlabelledData.lda<-unlabelledData.lda %>% arrange(max)

labelledData.lda<-rbind(labelledData.lda,unlabelledData.lda[1,1:4])

unlabelledData.lda<-unlabelledData.lda[2:nrow(unlabelledData.lda),1:4]

}

Initial accuracy of the Naïve-Bayes classifier is 72.14%

|  |  |
| --- | --- |
| **Additional points labelled** | **Accuracy** |
| 10% | 76.7% |
| 20% | 79.76% |
| 30% | 84.09% |
| 40% | 90.33% |

1. **Margin Sampling**

For margin sampling, after building a Naïve-Bayes classifier we find class probabilities and find the difference of the two maximum class probabilities for each data entry. We, then, select the data entry with the least difference between the two maximum class probabilities.

#MarginSampling

for(i in c(1:(nrow(vehicle\_dataset)\*percent))){

nb<-naive\_bayes(Class~.,labelledData.lda)

xtab<-table(predict(nb,unlabelledData.lda, type="class"),unlabelledData.lda$Class)

acc=sum(diag(xtab))/nrow(unlabelledData.lda)\*100

print(acc) a1<-predict(nb,unlabelledData.lda, type="prob")

maxes<-t(sapply(1:nrow(a1),function(i){

sort(a1[i,1:4],decreasing = TRUE)[1:2]

}))

diff<-maxes[,1]-maxes[,2]

unlabelledData.lda<-cbind(unlabelledData.lda,"diff"=diff)

unlabelledData.lda<-unlabelledData1.lda %>% arrange(diff)

labelledData1.lda<-rbind(labelledData1.lda,unlabelledData1.lda[1,1:4])

unlabelledData1.lda<-unlabelledData1.lda[2:nrow(unlabelledData1.lda),1:4]

}

Initial accuracy of the Naïve-Bayes classifier is 72.14%

|  |  |
| --- | --- |
| **Additional points labelled** | **Accuracy** |
| 10% | 76.99% |
| 20% | 80.78% |
| 30% | 84.68% |
| 40% | 90.33% |

1. **Entropy Sampling**

Entropy sampling takes into account all the class probabilities. Entropy is calculated as :

After calculating entropy row-wise, all the rows of data are sorted according to entropy values and the row with the highest entropy value is labelled and added to the labelled training set.

#Entropy Sampling

for(i in c(1:(nrow(vehicle\_dataset)\*percent))){

nb<-naive\_bayes(Class~.,labelledData.lda)

xtab<-table(predict(nb,unlabelledData.lda, type="class"),unlabelledData.lda$Class)

acc=sum(diag(xtab))/nrow(unlabelledData.lda)\*100

print(acc) a<-predict(nb,unlabelledData.lda, type="prob")

a.log<-log2(a)

unlabelledData.lda<-cbind(unlabelledData.lda,"entropy"=rowSums(-a.log \* a))

unlabelledData.lda<-unlabelledData.lda %>% arrange(entropy)

labelledData.lda<-rbind(labelledData.lda,unlabelledData.lda[nrow(unlabelledData.lda),1:4])

unlabelledData.lda<-unlabelledData.lda[1:(nrow(unlabelledData.lda)-1),1:4]

}

Initial accuracy of the Naïve-Bayes classifier is 72.14%

|  |  |
| --- | --- |
| **Additional points labelled** | **Accuracy** |
| 10% | 75.84% |
| 20% | 78.84% |
| 30% | 82.35% |
| 40% | 87.74% |