Programming Assignment02

Submission guide

1. Write answer following questions in this file
2. Write your code using provided Jupyter notebook file
   - Do not use other packages that are not already imported in the script
   - You have to complete several functions under description
   - Please check **TODO**
   - After completing your code, run script and submit with the printed results for answering questions in this word file

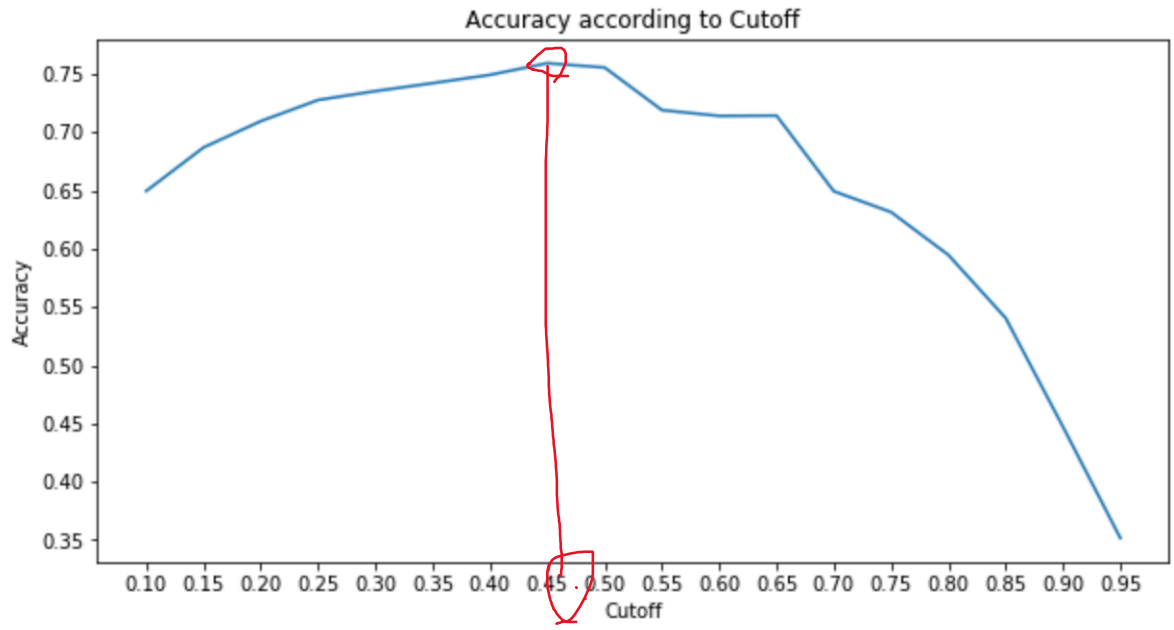
1. Naïve Bayes (40pts)

(1) Complete uploaded python code. (20pts)

(2) First, you have to binarize training set (trainX) of MAGIC Gamma Telescope data set. Each column is converted to binary variable based on the average value. If a value is greater than average, set a value as 1. Otherwise, set a value as 0. Then, using new binarized dataset, calculate $p_{ij} = P(x_j = 1|y_j = i)$ $(i = class, j = feature)$. (5pts)

	Class g	Class h
0	0.283948	0.460202
1	0.265707	0.398729
2	0.400993	0.473281
3	0.447912	0.463191
4	0.433624	0.453475
5	0.660418	0.527093
6	0.610762	0.508408
7	0.504054	0.497010
8	0.235103	0.674701
9	0.460681	0.534193

(3) Based on the calculated p_{ij}, calculate probability of class g for each test sample (testX) and calculate accuracy for testX with varying cutoff (**To binarize testX, use the mean of trainX**). Prior probabilities of classes are proportional to ratios of classes in training set. cutoff ∈{0.1,0.15,0.2,0.25,...,0.95}. Draw a line plot (x=cutoff, y=accuracy). (10pts)



(4) Explain why the shape of figure of Question 1-(3) looks like this. (5pts)

After cutoff is 0.5, the accuracy become lower and lower by changing cutoff.

Compare when cutoff = 0.1 and when cutoff =0.9 , the accuracy (cutoff 0.9) is lower than (cutoff 0.1). -> this means the number of samples which is class 'h' is smaller than the number of samples which is class 'g'.

And the maximum Accuracy is almost 0.75 when cutoff is almost 0.45. this is the optimal cutoff in this model.

2. Decision Tree (30pts)

The aim of the given data set is to predict annual income of people based on the following factors.

- age: the age of an individual
- capital-gain: capital gains for an individual
- capital-loss: capital loss for an individual
- hours-per-week: the hours an individual has reported to work per week
- sex: 1 if male, 0 if female
- native-country: 1 if USA, 0 if others
- workclass_[#]: 1 if an individual belongs to workclass # otherwise 0 (eg. Workclass_Private is 1 if an individual works for private companies)
- education_[#]: 1 if an individual's education level is # otherwise 0(education level: Graduate > 4-year university > "<4-year university" > High school > "<High school" > Preschool)
- marital-status_[#] 1 if an individual's marital status is # otherwise 0 (Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces)
- occupation_[#]: 1 if an individual's occupation is # otherwise 0.
- race_[#]: 1 if an individual's race is #, otherwise 0

Target is 'income' (">50K" or "<=50K")

fnlwgt represents the number of people the census believes the entry represents, which is not used in training.

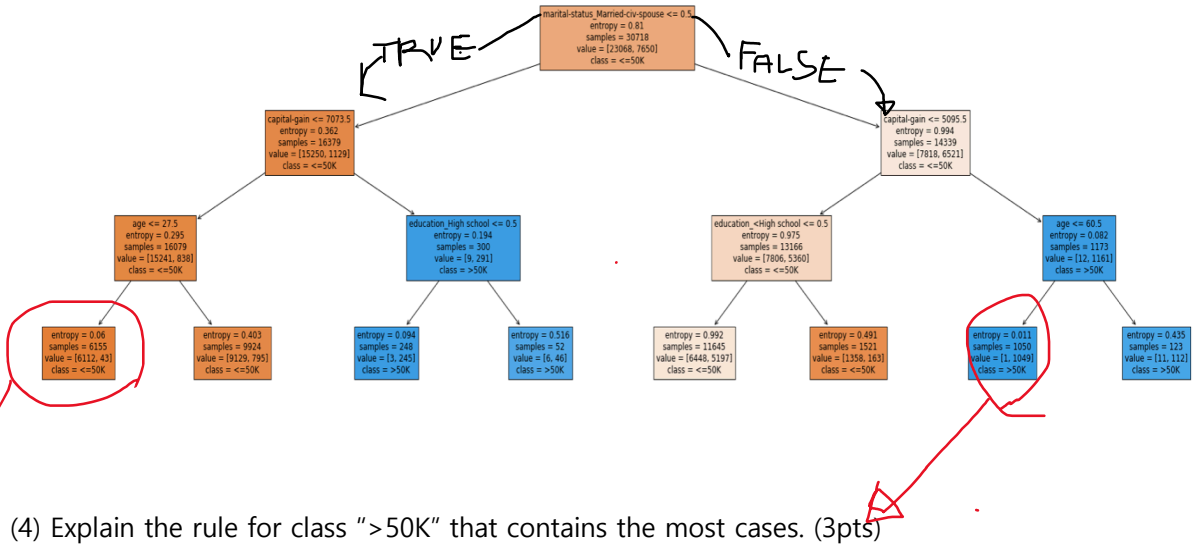
(1) Train a decision tree with the setting that max_depth=3, min_samples_split=100, min_samples_leaf=50 using entropy. Then, calculate overall accuracy, accuracy of class ">50K", and accuracy of class "<=50K". (5pts)

overall accuracy	accuracy of class ">50K"	accuracy of class "<=50K"
0.7975454131128329	0.18980392156862744	0.9990896479972255

(2) Based on the answer of Question 2-(1), describe the limitations of the trained decision tree model. (5pts)

Only the accuracy of a particular class is very high and the accuracy of another class is very low, which means that there is a high probability of making a right decision on a sample of the class <=50K, but the probability of making a right decision on a class of >50K is low.

(3) Draw the trained tree. (3pts)



(4) Explain the rule for class ">50K" that contains the most cases. (3pts)

In this case, there are three rules

mearital-status_Married-civ spuse > 0.5

-> Capital-gain > 5095.5

-> Age <= 60.5

(5) Explain the rule for class "<=50K" that contains the most cases with an accuracy of 0.7 or higher. (3pts)

In this case there are three rules,

mearital-status_Married-civ spuse <= 0.5

-> Capital-gain <= 7073.5ß

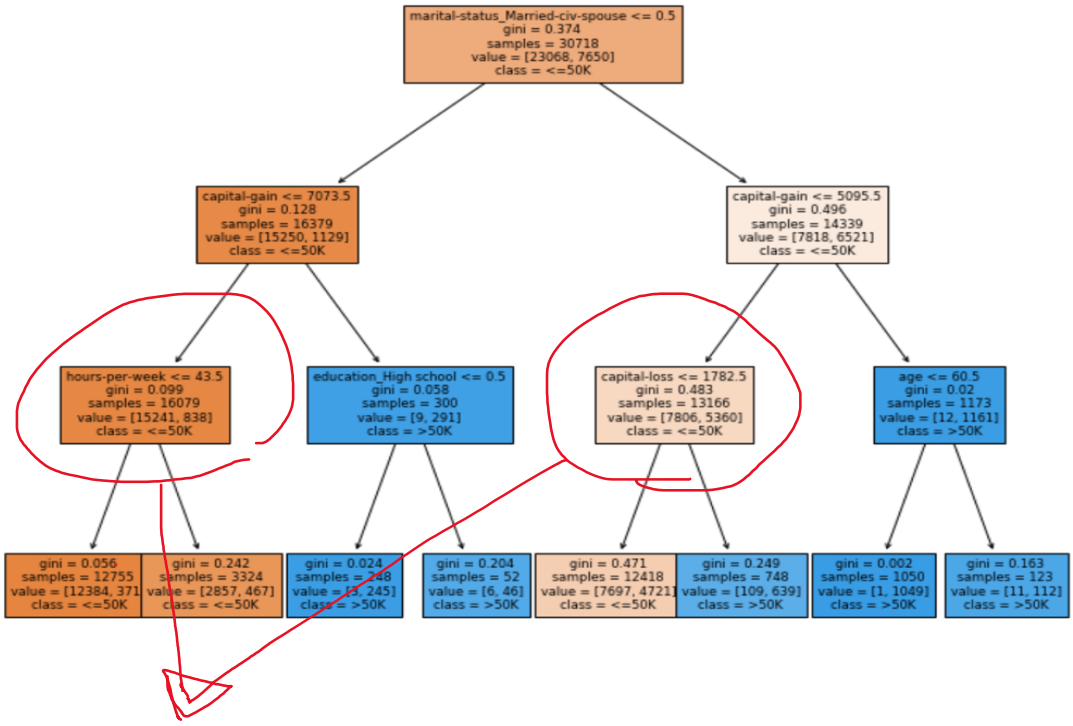
-> Age <= 27.5ß

(6) Train a new tree by changing a metric for finding split rules from entropy to gini impurity and compare two models in terms of the performance of the models and the generated rules (10pts)

```
When criterion = 'entropy' the overall accuracy is  0.7975454131128329
```

```
When criterion = 'gini impurity' the overall accuracy is  0.8147991405690475
```

Overall Accuracy is larger in criterion = gini impurity (more accuracy than entropy).



This two split rules are changed compared with criterion = entropy

3. k-means clustering (30pts)

This problem uses the data generated from 4 normal distributions for applying k-means clustering. k-means implemented in sci-kit learn can assign initial centroids through 'init'. When init is set as c by p array (c = the number of clusters, p = the number of features), each row is used as a centroid. Ref: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

(1) Select randomly 4 samples from the given data set and use them as initial centroids. This procedure is repeated for 100 times. Then, calculate the average values of the silhouette coefficient and adjusted rand index values for 100 iteration. (5pts)

```
--------- Average Value of Silhouette Coefficient for 100 times repeated---------
                          0.5695060889509516

--------- Average Value of Adjusted Rand Index for 100 times repeated---------
                          0.8827533178355585
```

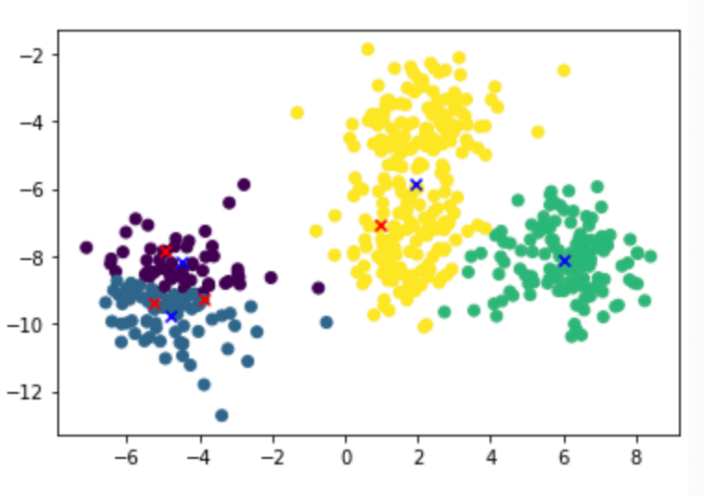
(2) Select randomly one sample from each normal distribution and use them as initial centroids. This procedure is repeated for 100 times. Then, calculate the average values of the silhouette coefficient and adjusted rand index values for 100 iteration. (5pts)

```
--------- Average Value of Silhouette Coefficient for 100 times repeated---------
                          0.5928509526139325

--------- Average Value of Adjusted Rand Index for 100 times repeated---------
                          0.9405029233193548
```

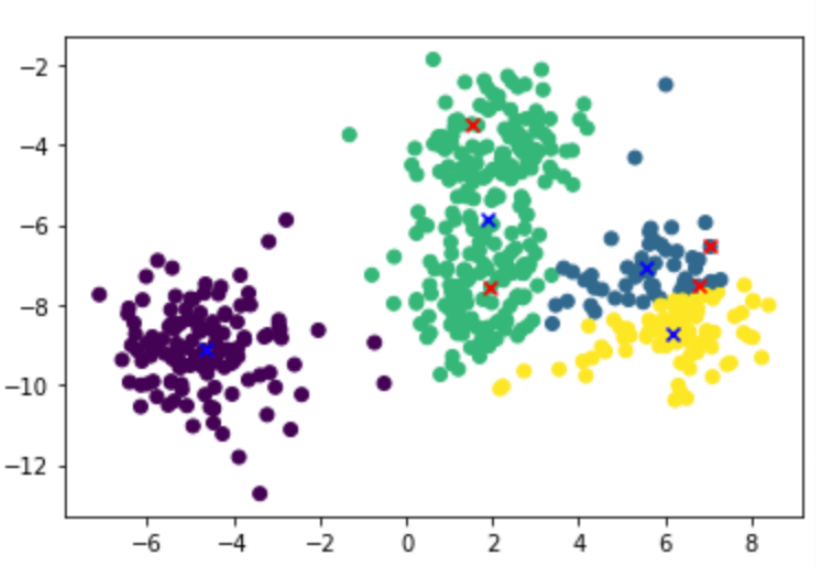
(3) Draw scatter plots for the given data with initial centroids and final centroids for the worst cases among 100 trials in Question 3-(1) in terms of silhouette coefficient and adjusted rand index, respectively. The initial centroids should be marked as red 'X' and the final centroids should be marked as blue 'X'. (5pts)

<In Terms of ARI values , I select worst case>



Blue X represent initial centroids
Red X represent initial centroids

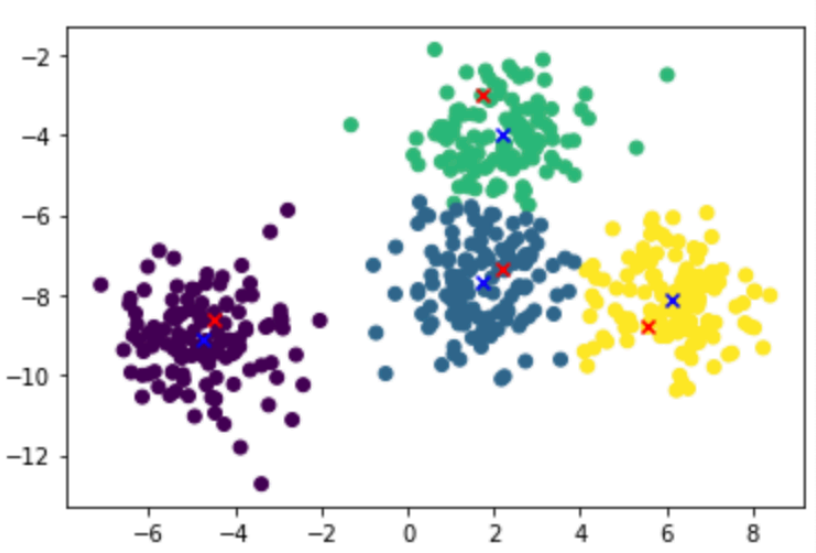
<In Terms of Silhouette values , I select worst case>



Blue X represent initial centroids
Red X represent initial centroids

(4) Draw scatter plots for the worst case of Question 3-(2) in the same way as in Question 3-(3). (5pts)

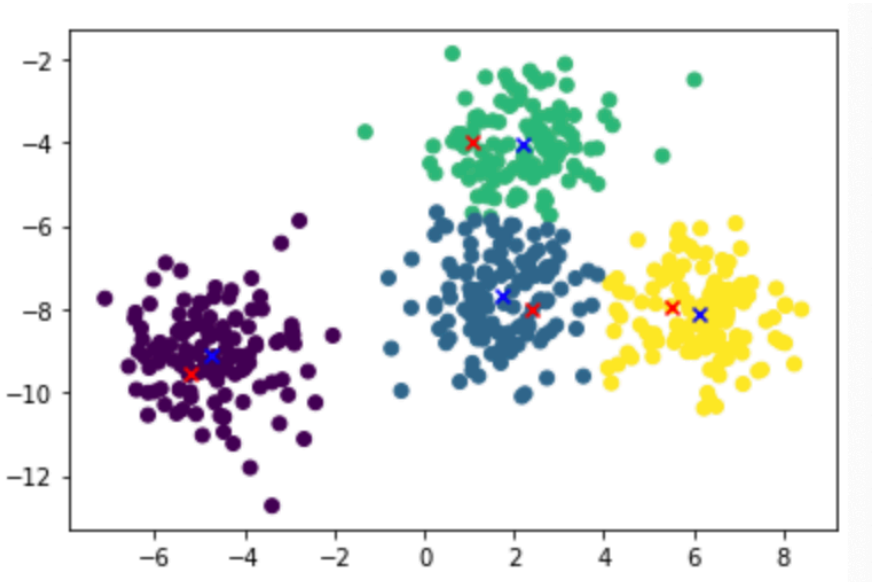
<In Terms of ARI values , I select worst case>



Blue X represent initial centroids

Red X represent initial centroids

<In Terms of Silhouette values , I select worst case>



Blue X represent initial centroids

Red X represent initial centroids

(5) Based on the different results from 100 trials for each case, compare two different methods to determine initial centroids. (10pts)

According to Answers of 3-(1) and 3-(2), in 3-(1) initial centroids is randomly choosed by all samples in 3-(2) centroid is selected randomly by each Normal Distribution.
Both average ARI and Sihouette values in 100 trials in 3-(1) are larger than 3-(2)

In terms of the ARI mean, it is generally more optimal model when it is 3-(2).
In terms of the Siouette mean, generally model is more concentrated on 3-(2)

In conclusion, it means that the Centroid decision method chosen in the 3-(2) model is better.