**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 |

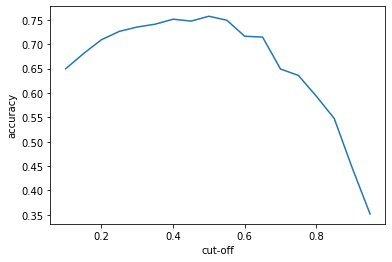
1. Naïve Bayes (40pts)

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

1. 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 (. (5pts)

|  |  |  |
| --- | --- | --- |
|  | Class g | Class h |
|  | 0.2839043178593148 | 0.4601869158878505 |
|  | 0.26565984188120817 | 0.39869158878504674 |
|  | 0.40097303871883233 | 0.4732710280373832 |
|  | 0.44790188526251773 | 0.46317757009345795 |
|  | 0.43361037907966754 | 0.45345794392523364 |
|  | 0.66045003040746 | 0.5271028037383177 |
|  | 0.6107845124670586 | 0.508411214953271 |
|  | 0.5040543279951348 | 0.49700934579439254 |
|  | 0.2350496655179404 | 0.6747663551401869 |
|  | 0.4606730184471924 | 0.5342056074766355 |

(3) Based on the calculated , 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)

**As P(y = ‘g’) is larger than 0.5, the graph tends to start with an accuracy of about 65%, and as cutoff value approach to center(0.50), Accuracy also increase at maximum which is similar with ordinal P(y = ‘g’ | x = 1).**

**After reaching to edge at cutoff 0.50, accuracy decrease significantly because of same reason for starting at 65%, which real P(y = ‘g’) is bigger than 0.5.**

**Due to those reasons, I see sum of accuracy when cutoff is 0.1 and 0.95 become 1.**

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)

**Because one parameter was chosen as decision tree, it have a skewed ratio.**

* **Reduce predictive power, and difficult to interpret.**

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

텍스트, 전자기기, 여러개이(가) 표시된 사진

자동 생성된 설명

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

**Martial-status\_Married-civ-spouse <= 50**

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

**age <= 60.5**

**accuracy = 0.7511340420374667**

(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)

**Gini’s case**

|  |  |  |
| --- | --- | --- |
| overall accuracy | accuracy of class “>50K” | accuracy of class “<=50K” |
| 0.8147991405690475 | 0.2733333333333333 | 0.9943644876018727 |

텍스트, 전자기기, 여러개이(가) 표시된 사진

자동 생성된 설명

For both trees, **Martial-status\_Married-civ-spouse** is a very important factor in distinguishing data at depth =0.

Decision tree based on gini factor is slightly more accurate performance than one based on entropy. Unlike decision tree with entropy chooses age and education(<High school) factors to determine depth=2, decision tree with gini chooses hours-per-week and capital-loss factors.

3. -means clustering (30pts)

This problem uses the data generated from 4 normal distributions for applying -means clustering.

k-means implemented in sci-kit learn can assign initial centroids through ‘init’. When init is set as by array ( = the number of clusters, = the number of features), each row is used as a centroid.

Ref: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

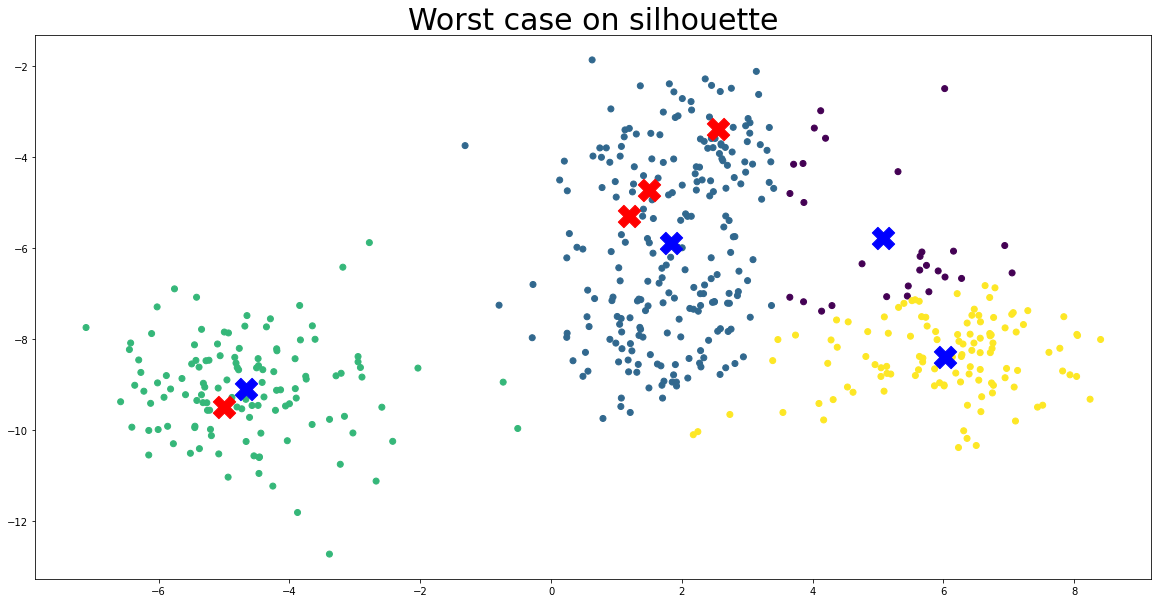
(1) Select randomly 4 samples from the given data set and use them as initial centroids. This procedure is repeated for 100 times. Then, calculate the average values of the silhouette coefficient and adjusted rand index values for 100 iteration. (5pts)

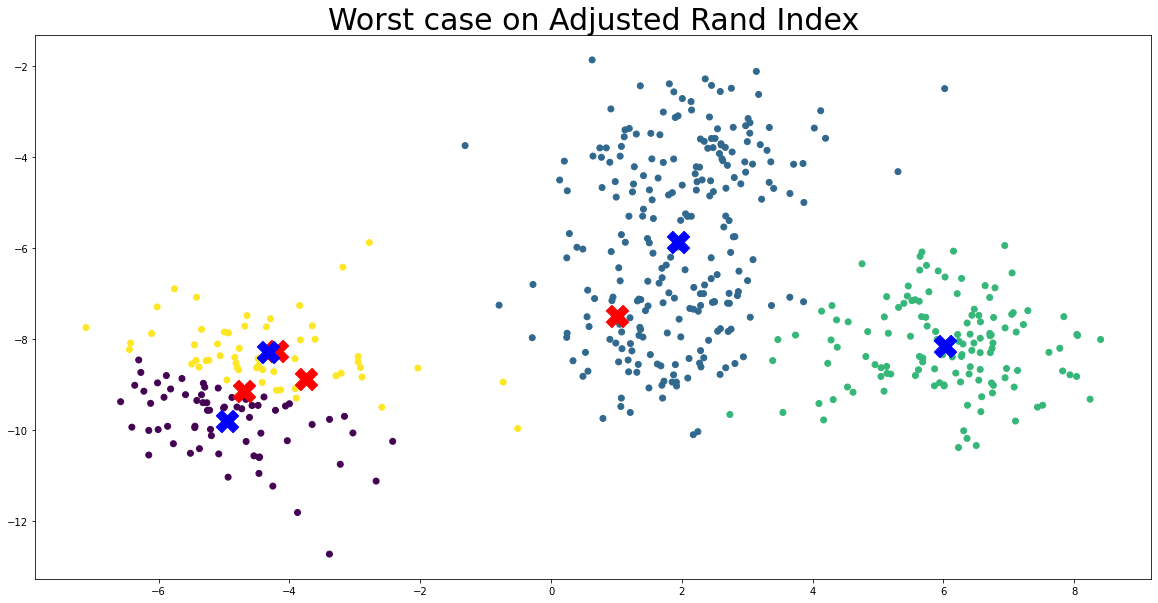
|  |  |
| --- | --- |
| silhouette coefficient | adjusted rand index |
| 0.5690794846567893 | 0.8825987448030784 |

(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)

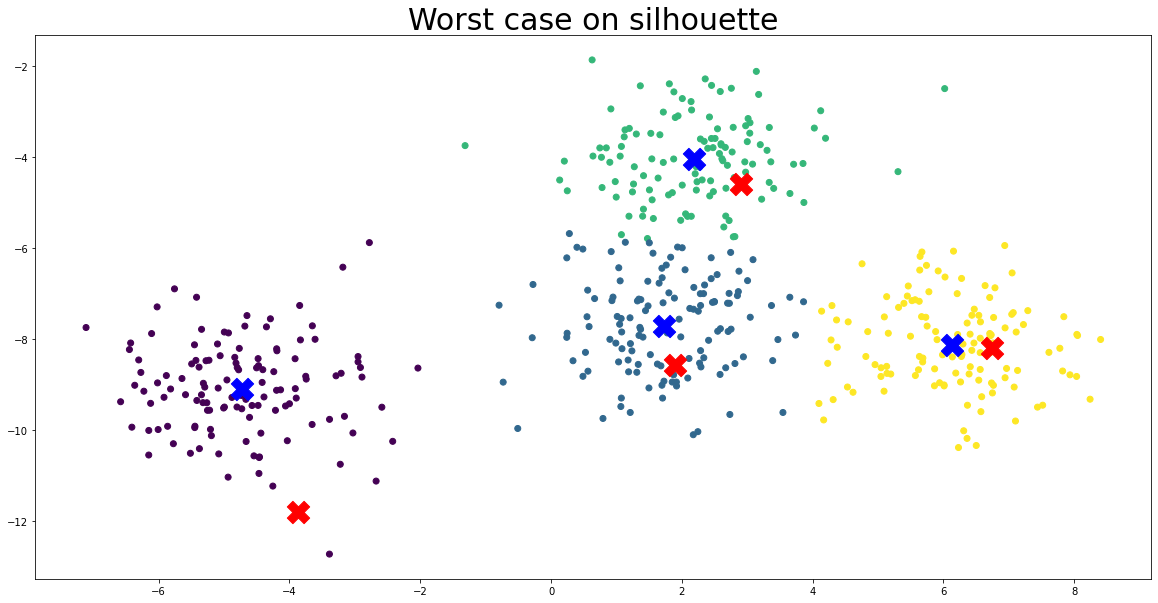
|  |  |
| --- | --- |
| silhouette coefficient | adjusted rand index |
| 0.592853012984585 | 0.9404523904828056 |

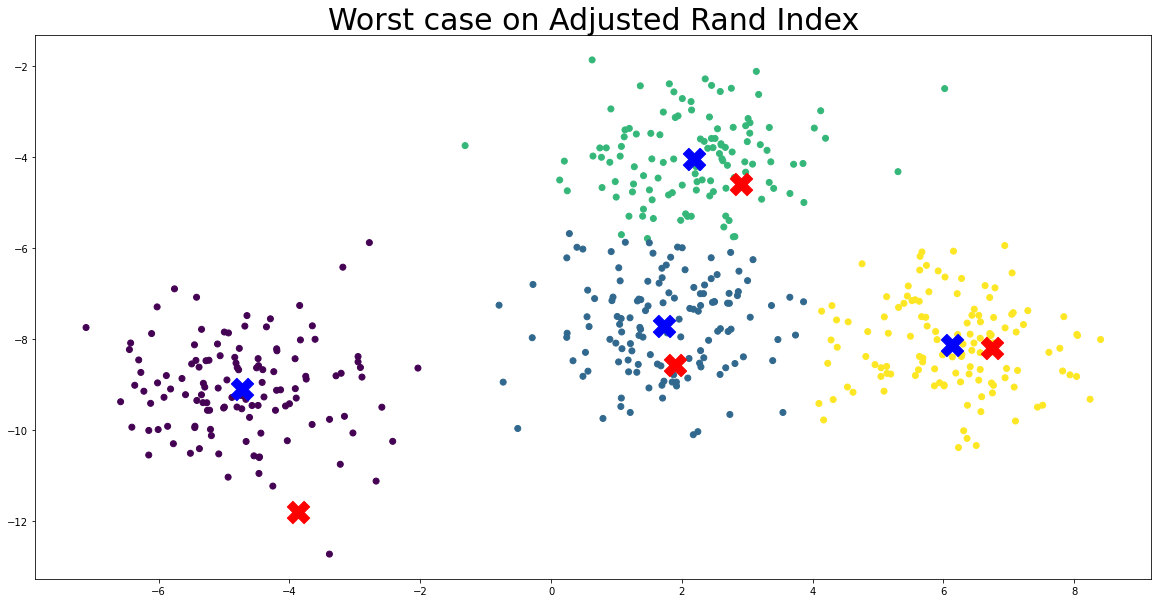
(3) Draw scatter plots for the given data with initial centroids and final centroids for the worst cases among 100 trials in Question 3-(1) in terms of silhouette coefficient and adjusted rand index, respectively. The initial centroids should be marked as red ‘X’ and the final centroids should be marked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





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





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

**By comparing these two different results, we can know that the early centering has a large impact on the overall result. Also, it is more accurate and efficient to choose random centroids from distinct groups than completely random. Therefore, choosing a random centroid for each clusters may make better results than completely random.**