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

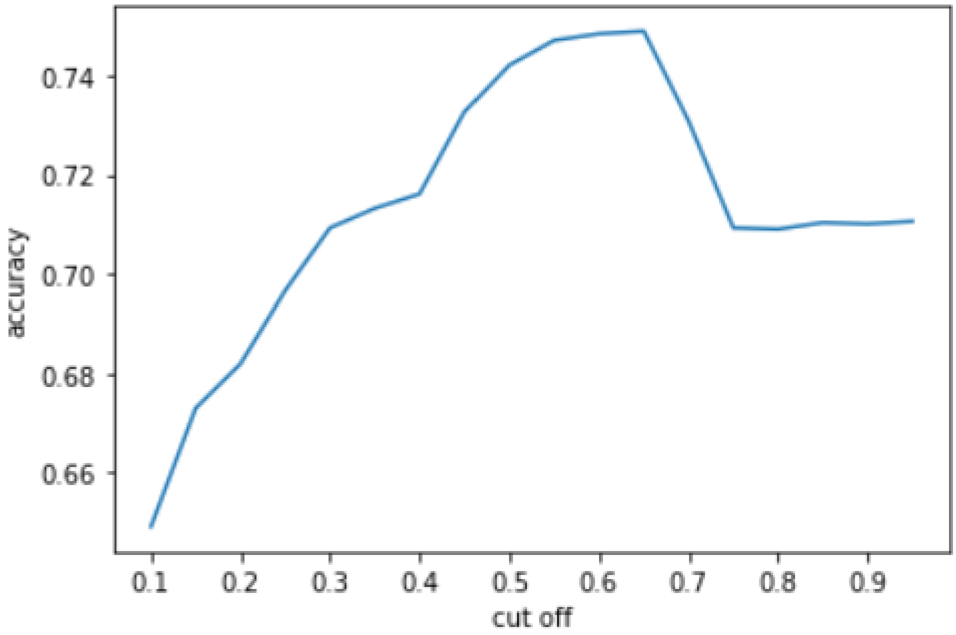
1. Naïve Bayes (40pts)

(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
$P(x_1 = 1)$	0.2839043178593148	0.4601869158878505
$P(x_2 = 1)$	0.26565984188120817	0.39869158878504674
$P(x_3 = 1)$	0.40097303871883233	0.4732710280373832
$P(x_4 = 1)$	0.44790188526251773	0.46317757009345795
$P(x_5 = 1)$	0.43361037907966754	0.45345794392523364
$P(x_6 = 1)$	0.66045003040746	0.5271028037383177
$P(x_7 = 1)$	0.6107845124670586	0.508411214953271
$P(x_8 = 1)$	0.5040543279951348	0.49700934579439254
$P(x_9 = 1)$	0.2350496655179404	0.6747663551401869
$P(x_{10} = 1)$	0.4606730184471924	0.5342056074766355

(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)

Initially, as the cut off increases, the accuracy also increases. And the peak is reached at about 0.6. So, as you can see, the optimal cut off is about 0.6 which is same with p(y='g' | x=1). After that, since number of class g is larger than number of class h, the accuracy persists around 72%.

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.

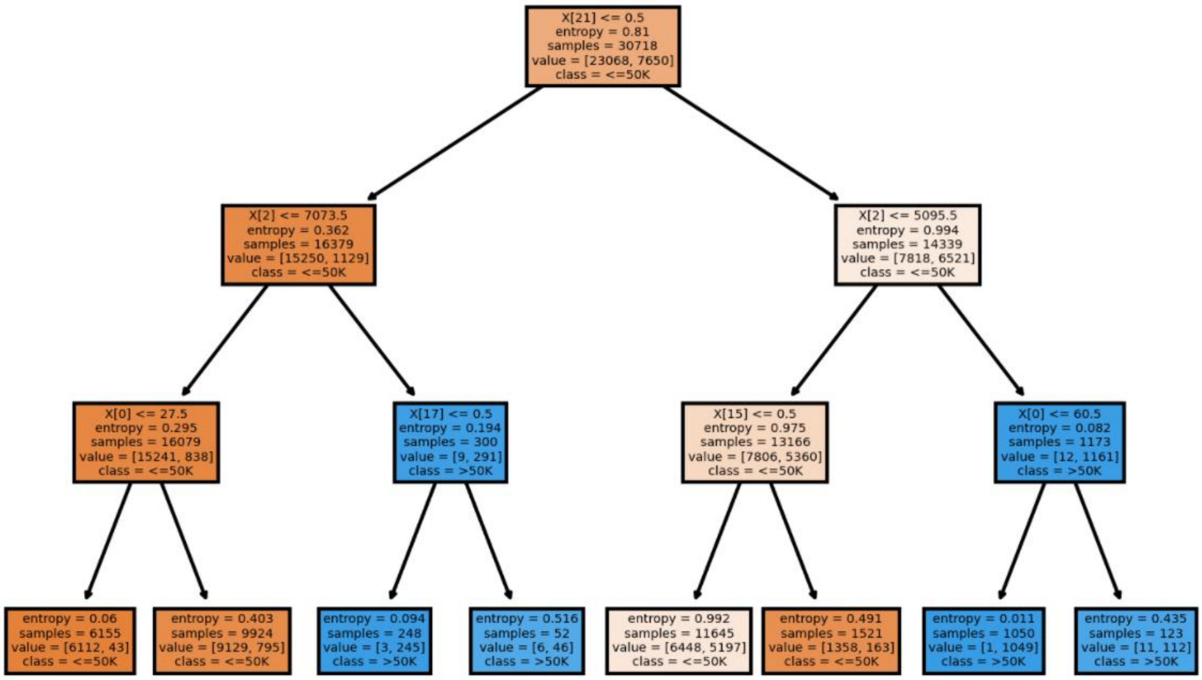
(1) Train a decision tree with the setting that max_depth=3, min_samples_split=100, min_samples_leaf=50 using entropy. Then, calculate overall accuracy, accuracy of class ">50K", and accuracy of class "<=50K". (5pts)

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)

When performing the above decision tree prediction, parameters(max_depth, min_samples_split, min_samples_leaf) were set to avoid overfitting. However, from the accuracy of the two labels, considering that the number of features in the data is 46, it can be seen that the parameter setting is excessive. The max_depth 3 of tree is a number that is too small to reflect the large number of features and the differences between 2 label data.

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



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

The rules are

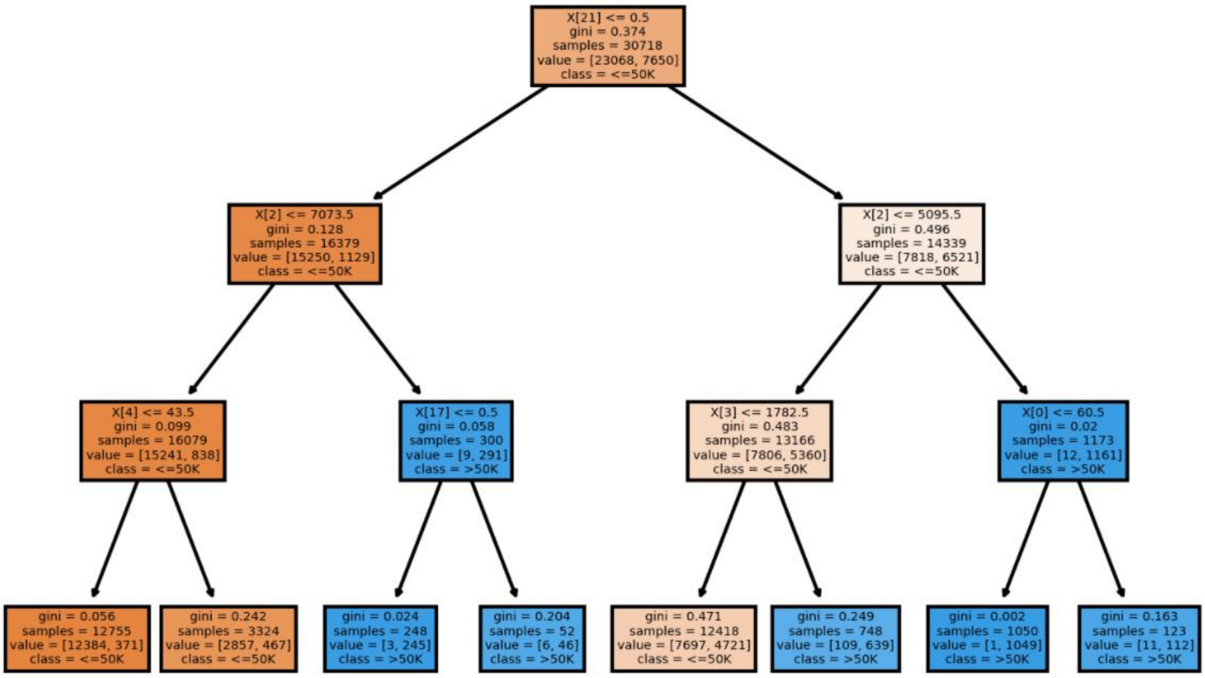
1. X[21] - Marital-status-Married-civ-spouse <= 0.5 – False
   A. The number of married people with a civilian spouse is greater than 0.5.
2. X[2] – captain-gain <= 5095.5 – False
   A. Capital gains for an individual are greater than 5095.5.
3. X[0] – Age <= 60.5 – True
   A. The age is less than or equal to 60.5.

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

1. X[21] - Marital-status-Married-civ-spouse <= 0.5 – True
   A. The number of married people with a civilian spouse is less than or equal to 0.5.
2. X[2] – capital-gain <= 7073.5 – True
   A. Capital gains for an individual are less than or equal to than 7073.5.
3. X[0] – Age <= 27.5 – False
   A. The age is greater than 27.5.

(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)

overall accuracy	accuracy of class ">50K"	accuracy of class "<=50K"
0.8147991405690475	0.2733333333333333	0.9943644876018727



Based on decision tree by gini, accuracy performance has improved a little bit. And, For the class of "<= 50K", the split rule in the internal node just before the leaf node has been changed.

Entropy	Gini
X[0] - age<=27.5	X[4] - hours-per-week <43.5
X[15] - education_High_school <=0.5	X[3] - capital-loss<=1782.5

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

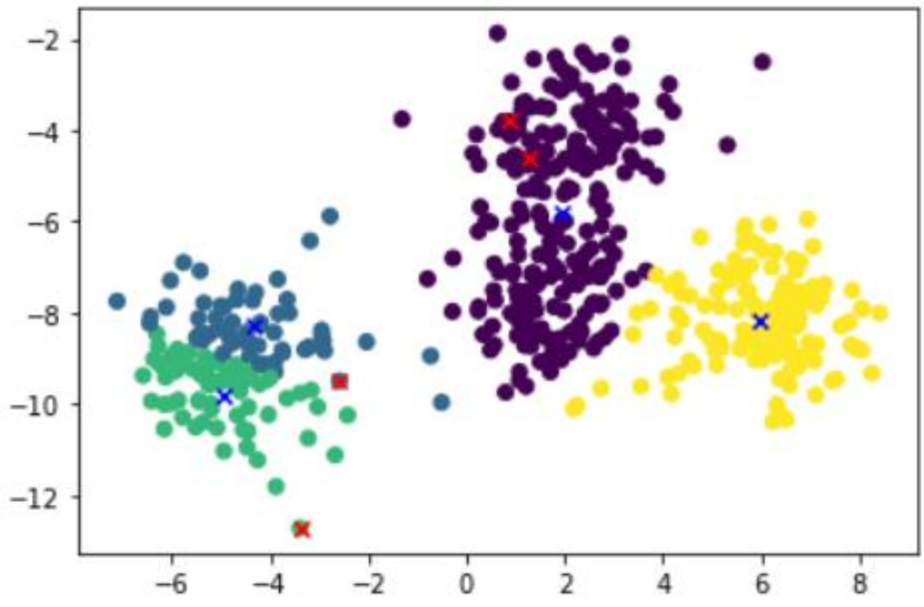
(1) Select randomly 4 samples from the given data set and use them as initial centroids. This procedure is repeated for 100 times. Then, calculate the average values of the silhouette coefficient and adjusted rand index values for 100 iteration. (5pts)

silhouette coefficient	adjusted rand index
0.5594989879790816	0.8600967238659396

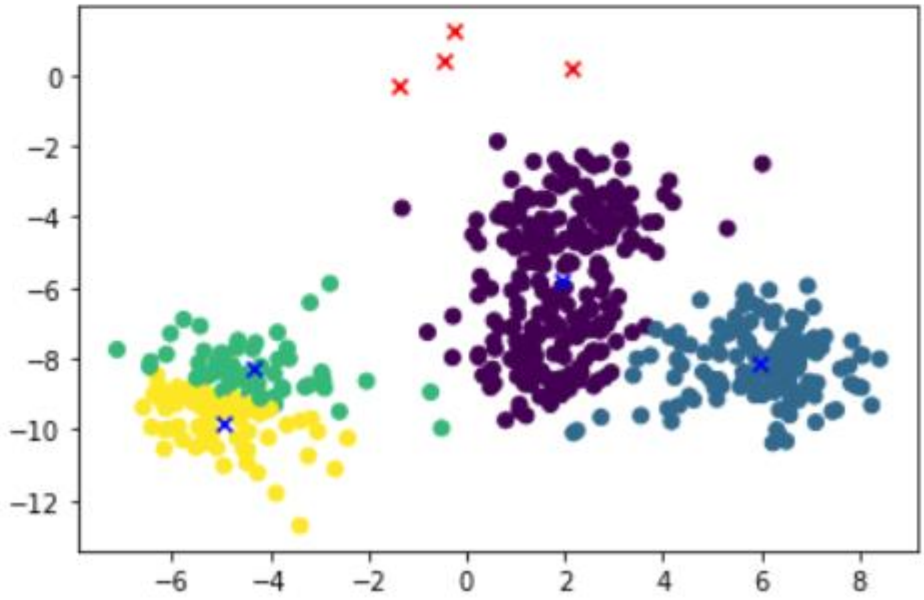
(2) Select randomly one sample from each normal distribution and use them as initial centroids. This procedure is repeated for 100 times. Then, calculate the average values of the silhouette coefficient and adjusted rand index values for 100 iteration. (5pts)

silhouette coefficient	adjusted rand index
0.5130167351517236	0.7429527051510538

(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 the other two results, we can find initial centroids have a great influence on entire modeling. And It's more accurate to get random one initial centroid from each group.