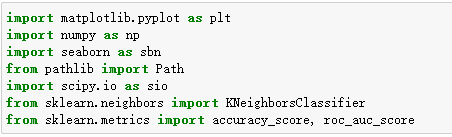
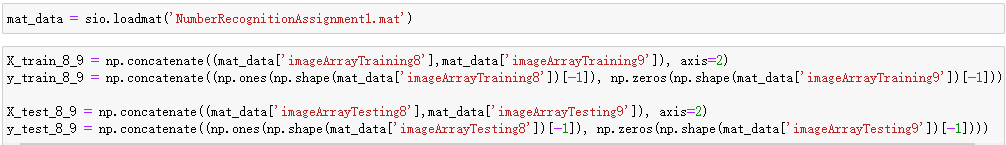
**Question 1**

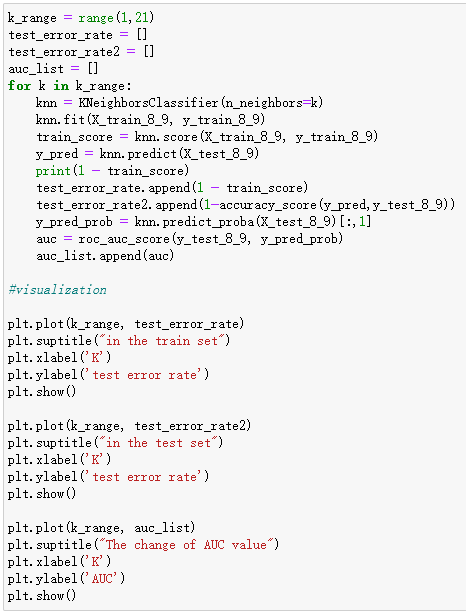
Import essential packages:



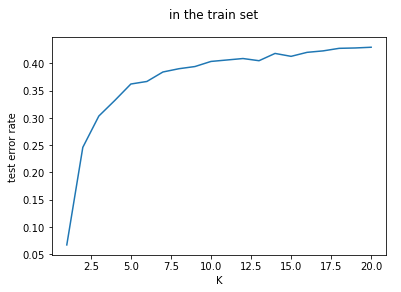
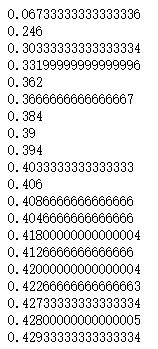
Import dataset and select training-set and testing-set:

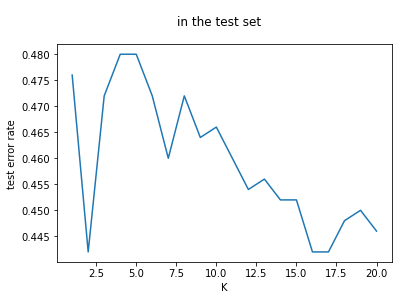


Knn（k=1,2,3,...,20）



result：





Question (a) Why does testing error rise at high values of K?

Answer：

Because in KNN algorithm, when K is higher, the model is becoming smoother, this means it will be more inclined to ignore noise and outliers, but in the mean time, it will ignore more value information. Therefore, when K is higher, model may miss the similarity easier, cause the test error increases.

Question (b) What is the error rate at the lowest K? Do you expect this to be a reliable performance estimate? Why?

Answer：

When k=1, the error rate is 0.067. when k is lowest, model when become too sensitive, leads to overfitting, therefore error rate is very low, but this doesn’t mean this is a reliable performance estimate, because when k is low, model is very sensitive and overfitting. In this case, it might performs good in training-set, but it might also performs very bad in new training-set. Thus, according to the chart, when k is in an appropriate range, error rate will reach the lowest, this is a reliable performance estimate.

**Question 2**

Dataset description:

Dataset is a classic one about adult income, this contains over 30000 samples, and 15 measurements.

Measurements：

'age': age

'workclass'：employment status

'fnlwgt'：the weight of each respondent in the total population

'education': education background

'education-num'：education years

'marital-status'：marital status

'occupation'：occupation

'relationship'：relationship

'race'：race

'sex'：sex

'capital-gain'：capital gain

'capital-loss'：capital loss

'hours-per-week'：hours per week

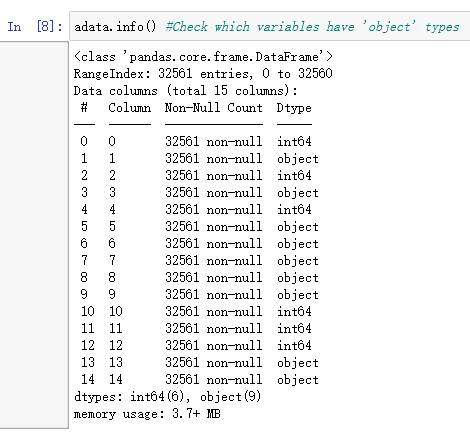
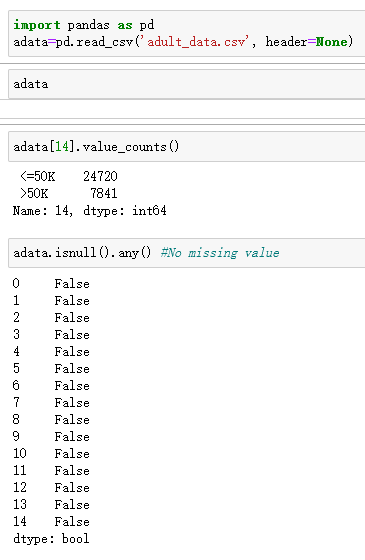
'native-country'：native country

'income'：income

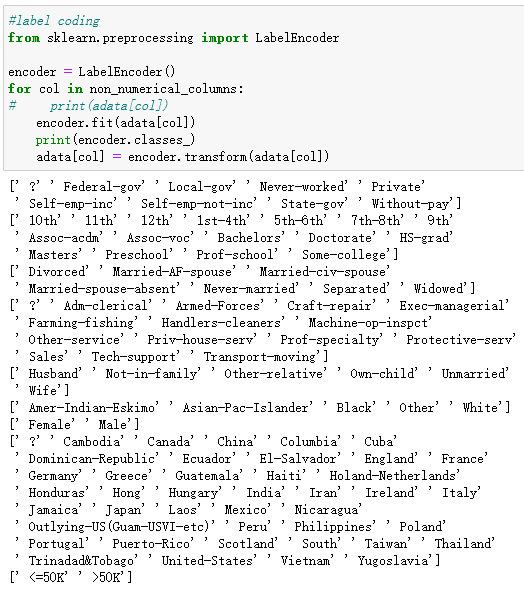
The variable (income) includes more than 50k and less than 50k, the sample number of more than 50k is 7841, and less than 50k is 24720.

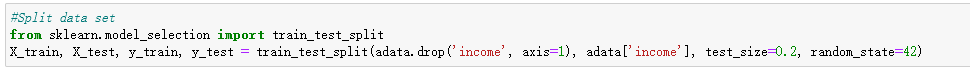
I used the logistic regression algorithm model to calculate the AUC value of each variable, the code is showing below,

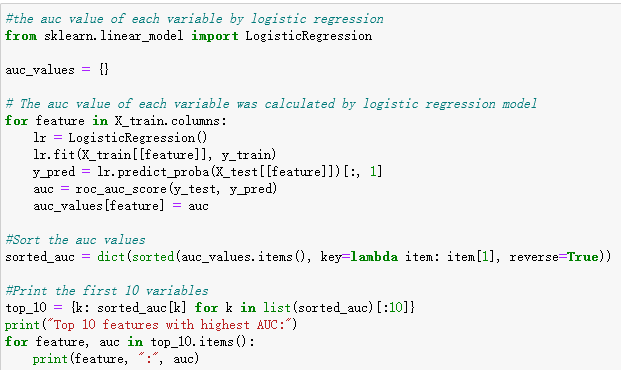
-Data import and preprocessing.



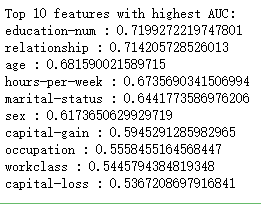




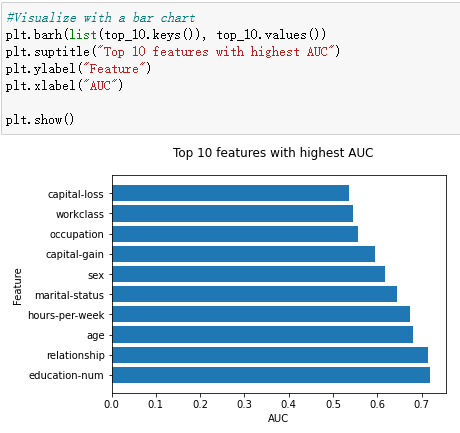




Result：



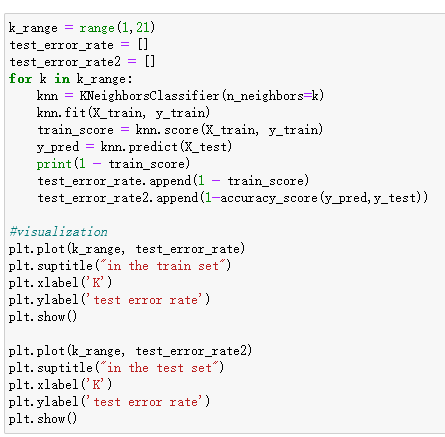
visualization：



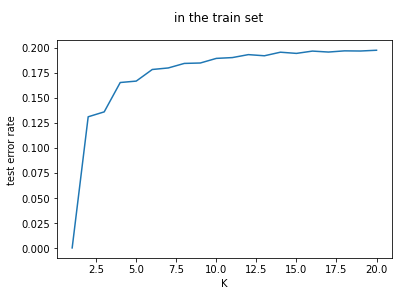
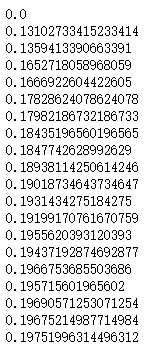
The top ten feature names are years of education, relationship, age, hours per week, marital status, sex, capital gain, occupation, working class and capital loss.

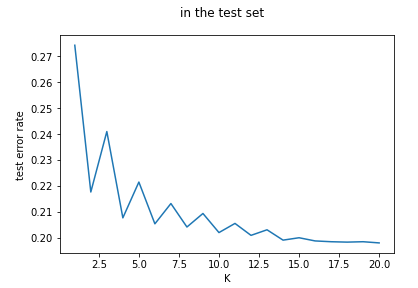
**Question 3**

Classify this dataset with the model of the first question and observe the error value.



结果如下：





In these two different datasets, the compactness between the data and the difference between the data points are different, however, the changing of the k-error rate relationship diagram in the training-set is the same, the reason of this was being answered in question 1. The difference is the k error rate relationship chart on the testing-set, the general rule is the opposite of the training-set: when the k is lower, the error is higher, and when k is higher the error is lower. (but because the data among with these two different datasets has their own distribution characteristics, they are showing different changing characteristics.) this is because when k is higher, KNN algorithm will have less impact on neighbors, model will become smoother. When k is lower, neighbors will have bigger impact on model, model will become more complex. For training-set, if model is too complex, will cause overfitting and has bad performance in the training-set. But for testing-set, if model is too simple, will cause underfitting, and bad performance on the testing-set.

So when k is higher, the error rate for training-set will increase, but decreasing for the testing-set, because the model is simpler and smoother, it can fix the underfitting problem in the training-set, also can fix the overfitting problem in the testing-set.