**机器学习第一次作业项目报告**

1. **实验要求**

从<http://archive.ics.uci.edu/ml/> 中选择一组你感兴趣的分类问题的数据集，采用两个分类模型进⾏分类，并比较分类效果和性能。

1. **数据的选择及分析**

Adult Data Set（<http://archive.ics.uci.edu/ml/datasets/Adult>）

数据集作为我们的测试数据。该数据集包含了以下14个特征，通过这14个特征预测其收入>50K或<=50k。其中存在缺失值，通过“？”来表示。其中，连续型特征6个，离散型特征8个。

**age:** continuous.

**workcass:** Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov,  State-gov, Without-pay, Never-worked.

**fnlwgt:** continuous.

**education:** Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

**education-num:** continuous.

**marital-status:** Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

**occupation:** Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

**relationship:** Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

**race:** White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

**sex:** Female, Male.

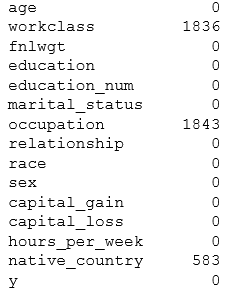
**capital-gain:** continuous.

**capital-loss:** continuous.

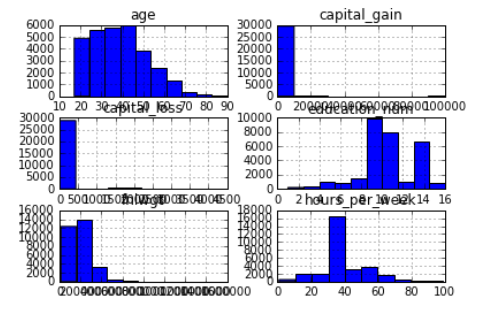
**hours-per-week:** continuous.

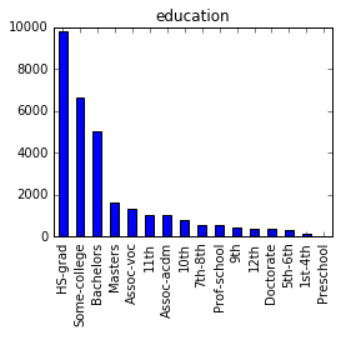
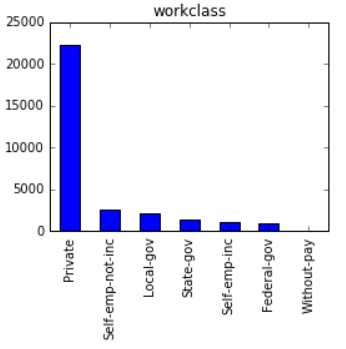
**native-country:** United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

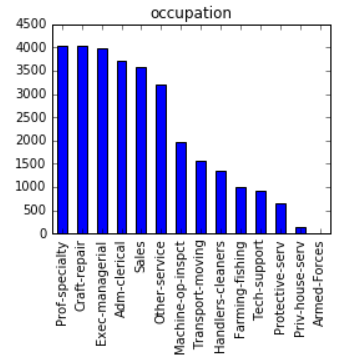
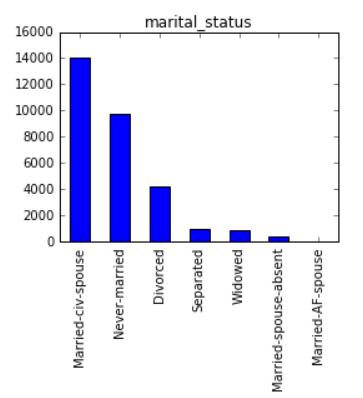
数据分析：训练数据共有32561组，我们发现其特征的缺失情况如下。通过观测，我们发现缺失值所占的比例并不大，因而我们直接去除了所有缺失值。所得到的的训练数据30162组。

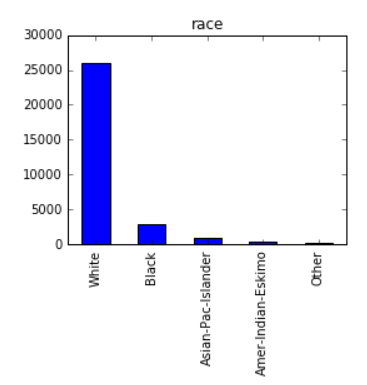
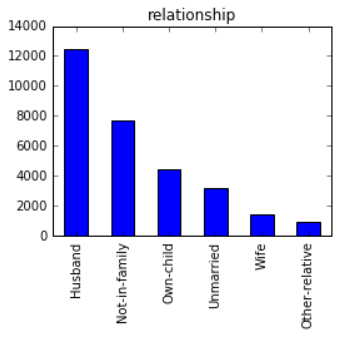


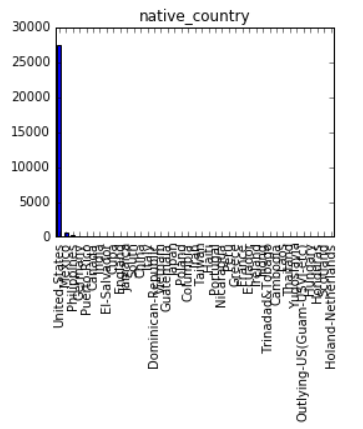
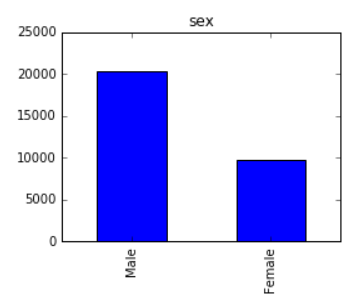
各特征取值的分布情况如下：

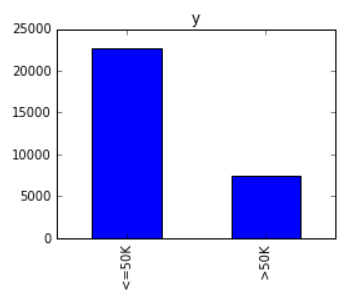










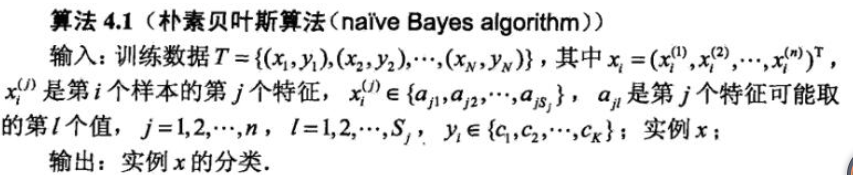


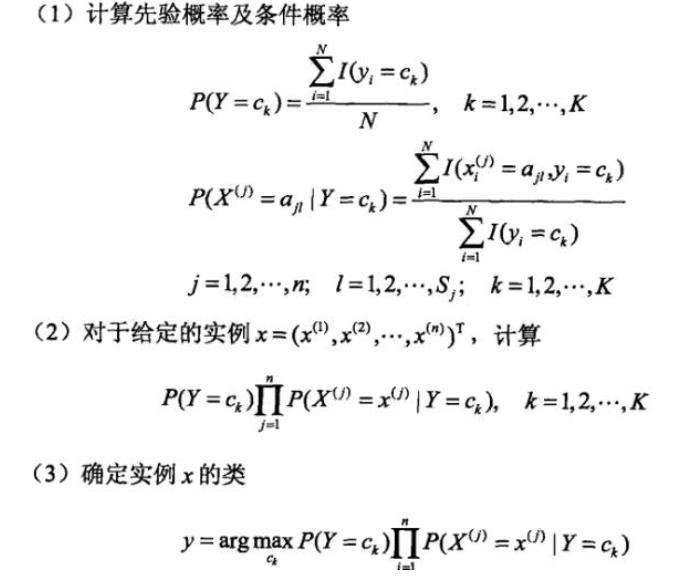
对于测试数据集我们也进行了同样的分析和处理，得到了15060组测试数据。

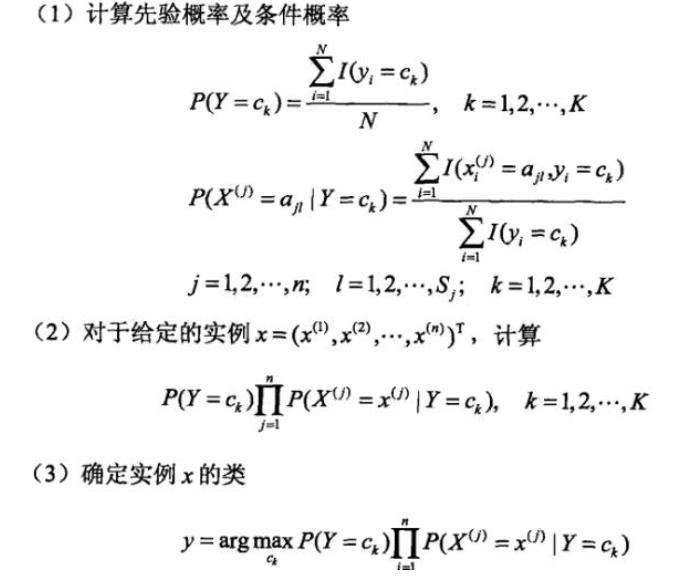
1. **实验内容**

1．首先，我们通过调⽤程序包，使用RandomForestClassifier和DecisionTreeClassifier两个分类器通过对训练数据进行学习生成分类器，利用生成的分类器对测试数据进行了预测并与原始结果进行了对比，从而评价分类器的性能。特别地，对于离散型特征，我们通过one\_hot\_encoding (dat, tdat, categorical\_features)方法，将每个特征的取值都看作是一个单独的特征。

1. 其次，我们编程实现了朴素贝叶斯算法。朴素贝叶斯分类的正式定义如下：





由于朴素贝叶斯算法要求所有特征的取值都为离散型，因此，首先要解决的问题就是连续数据的离线化。这里主要用了panda库中的一个函数cut,。data\_of\_feature表示某一个特征的数据集，是一列数据，而bin\_num就是一个数字，表示将个列连续数据分为多少个离散组。利用这两个参数panda.cut可以自动离散化数据，并且进行合适的分割。 值得一提的是，这里需要将训练数据和测试数据放在一起离散化，因为如果分开离散化，训练数据、测试数据的分割点不同。

1. **实验结果**

1）正确率（accuracy）

　　正确率是我们最常见的评价指标，accuracy = （TP+TN）/(P+N)，这个很容易理解，就是被分对的样本数除以所有的样本数，通常来说，正确率越高，分类器越好；

　　2）精度（precision）

　　精度是精确性的度量，表示被分为正例的示例中实际为正例的比例，precision=TP/（TP+FP）；

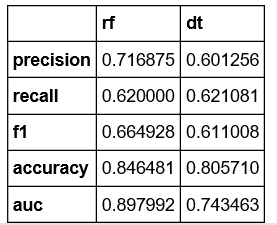
　　3）召回率（recall）

召回率是覆盖面的度量，度量有多个正例被分为正例，recall=TP/(TP+FN)=TP/P=sensitive，可以看到召回率与灵敏度是一样的。

4）F1-score

F1-score是accurancy 和recall两者的综合。F1-score 越高，说明分类模型越稳健。

随机森林和决策树



朴素贝叶斯

Accuracy: 82.137%

Precision: 60.663%

Recall: 77.649%

F1-score: 0.681

从结果可以看出，三者的正确率都达到了80%以上。随机森林算法的accurancy最高，说明其对负样本的区分能力最强。朴素贝叶斯算法的Recall值最大，说明其对正样本的识别能力最强。从F1只可以看出，朴素贝叶斯和随机森林的分类性能都比较稳健，而决策树稍差一些。Precision反映了正类中被正确预测为正类的比例，其中随机森林表现最好。