一、数据清洗

部分数据（workclass、occupation、country）包含缺省值。

如果用其他变量做预测模型来算出缺失变量，由于workclass、occupation、country此类数据本身特点，如果其他变量和缺失变量无关，则预测的结果无意义；如果预测结果相当准确，则又说明这个变量是没必要加入建模的。

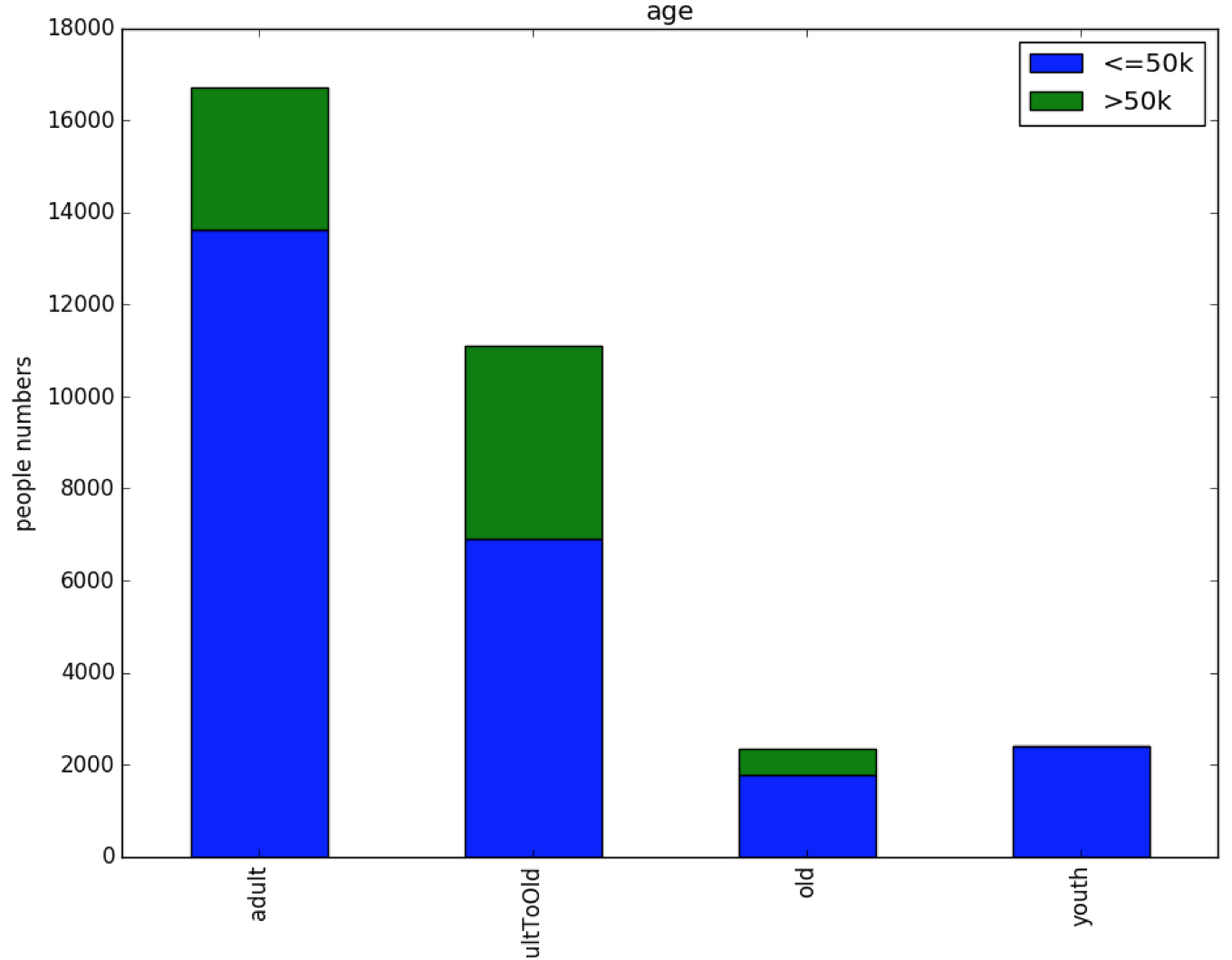
因此本文选择把变量映射到高维空间。比如workclass，有State-gov、Self-emp-not-inc、缺失等情况，则映射成多个变量：是否State-gov、是否Self-emp-not-inc、是否缺失等。这样做的好处是完整保留了原始数据的全部信息、不用考虑缺失值、不用考虑线性不可分之类的问题。

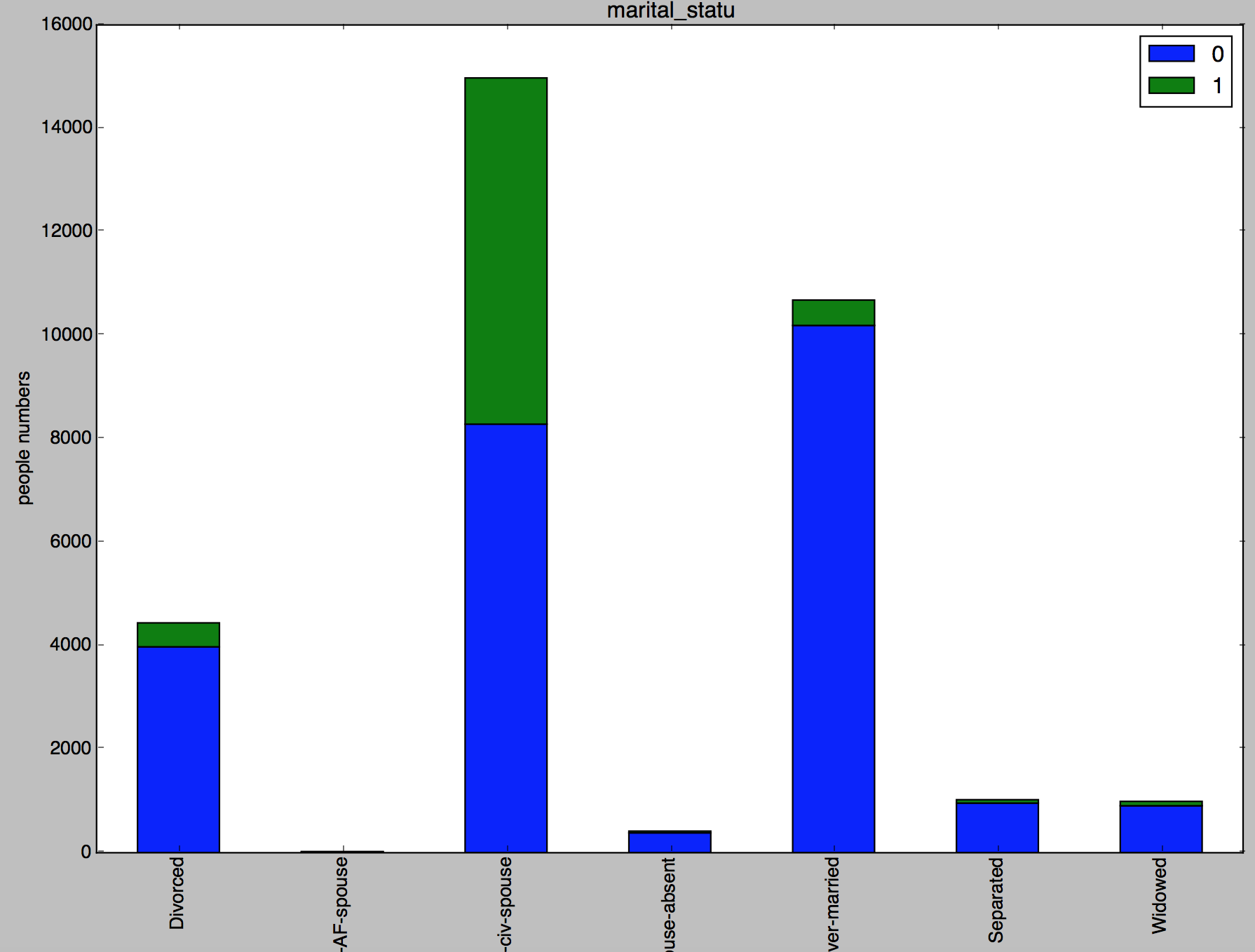
二、数据预处理

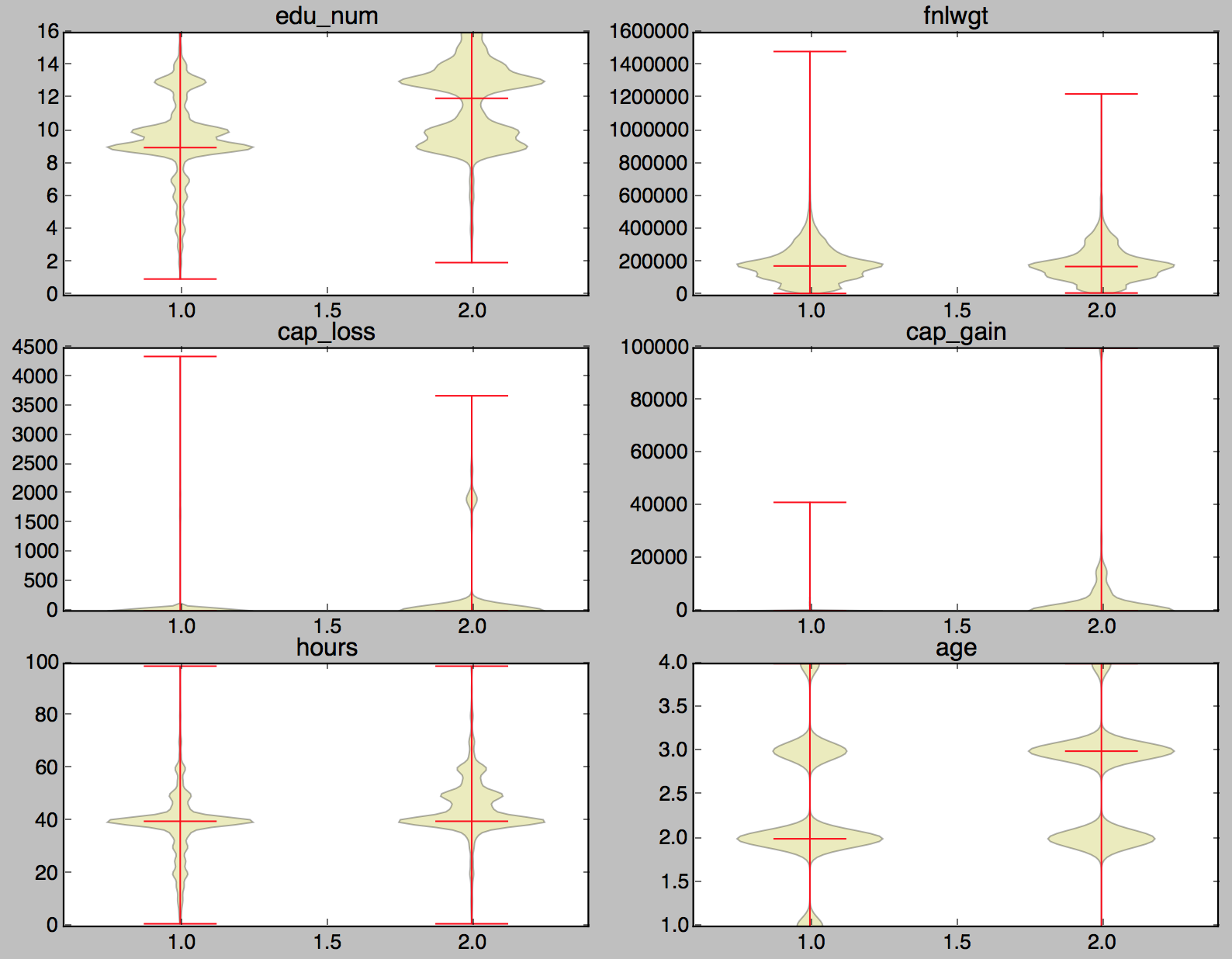
为了方便的数据分析，我们需要先将非数值型数据转化为数值型数据。一共有两种处理方式：

1. Category Indexing. 将每个属性的不同取值直接修改为数值型数据（0,1,2,…,numCategories-1）
2. One-Hot Encoding. 将每个属性的不同取值传化成numCategories个二元特征向量，特征之间互斥。

在本文数据集中，包含有序变量（如education：Preschool - Doctorate）,也包含离散型变量（如relationship：Wife, Husband, Own-child等等）。对有序变量做第一种处理（Category Indexing），对离散型变量做第二类处理（One-Hot Encoding）。

为了更好的利用年龄数据，本文将具体的年龄划分成了年龄段（如age<20则为’youth’）。如图2.1所示，年龄在40-60岁的人最有可能赚到50K以上。

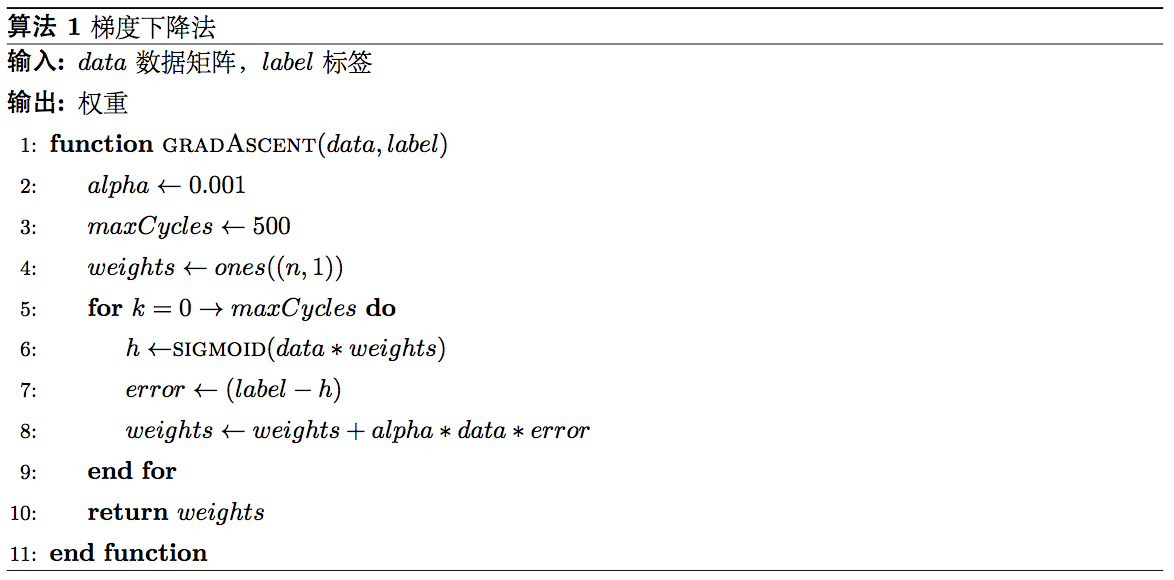
如图2.2所示，married civilian spouse 更有可能赚到50k以上。

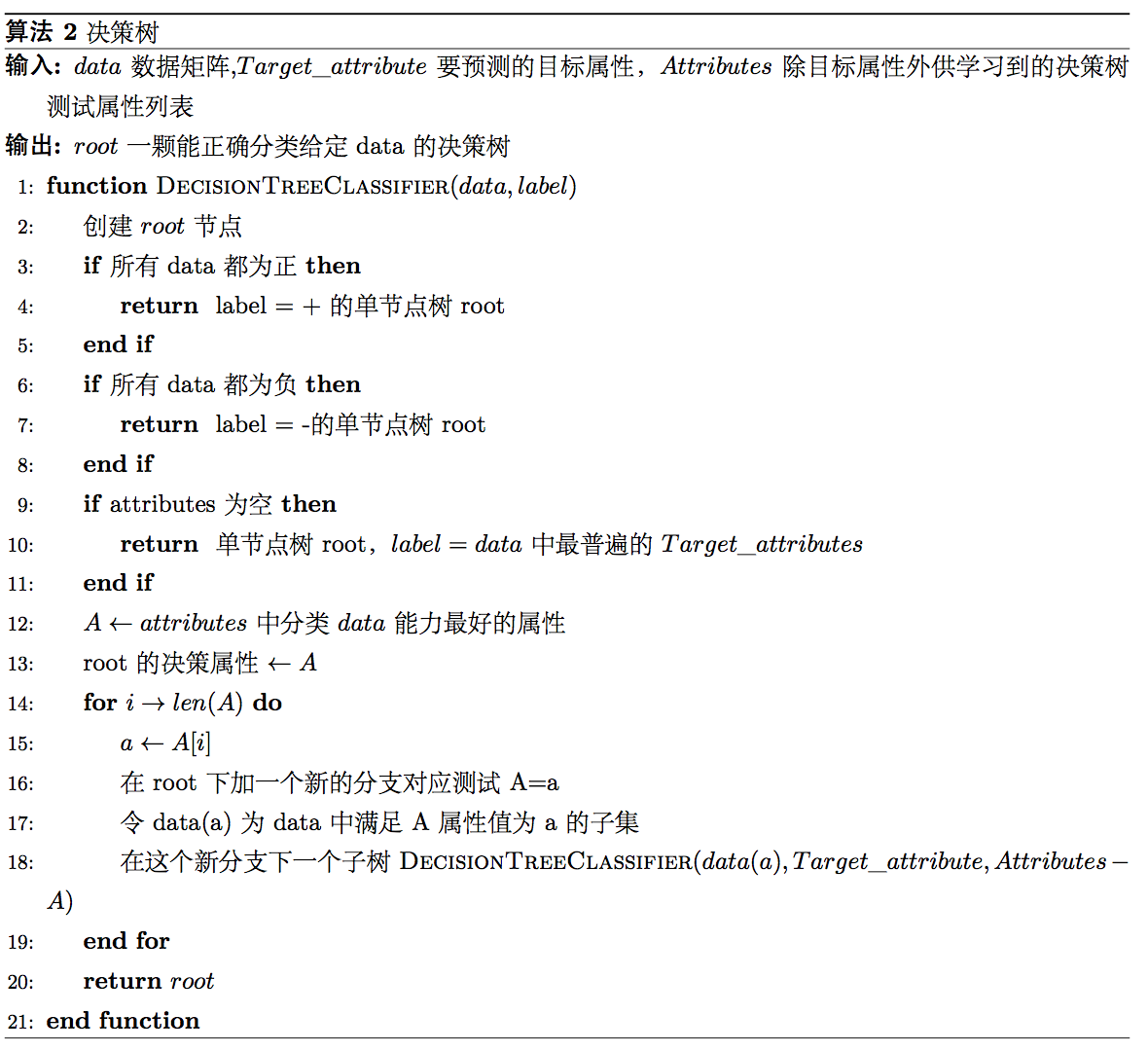
如图2.3所示，hours per week、fnlwgt在两类收入下的数据分布差异不大，对分类可能没有明显效果。

将数据处理为数值型数据后，将每个属性／每列数据进行标准化（去除均值和方差缩放），见公式2.1。得到的结果是，对于每个属性/每列来说，所有数据都聚集在0附近，方差为1。

三.算法简介

本文使用的部分分类算法伪代码如下。

1、LogisticRegression见算法1

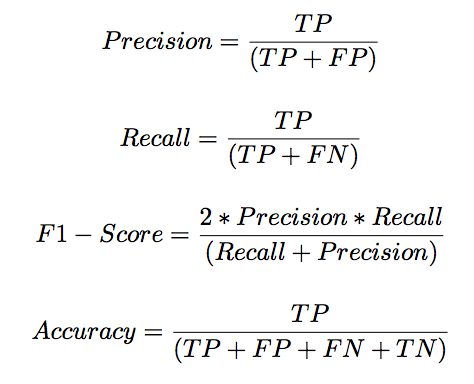
2、DecisionTreeClassifier见算法2

四.精度分析与比较

在此部分比较LogisticRegression、AdaBoostClassifier、GaussianNB、DecisionTreeClassifier、RandomForestClassifier几种分类器的精确度、召回率和F值。其标准定义如下：

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Truth | |
| Class=“<=50K” | Class=“>50K” |
| Decision | Class=“<=50K” | TP | FP |
| Class=“>50K” | FN | TN |

其中准确率、精确度、召回率和F值见图1.4.1。



本文使用sklearn实现，分类结果如下：

|  | precision | recall | f1-score | accuracy |
| --- | --- | --- | --- | --- |
| LogisticRegression | 0.83 | 0.84 | 0.83 | 0.84 |
| AdaBoostClassifier | 0.85 | 0.86 | 0.85 | 0.86 |
| GaussianNB | 0.58 | 0.76 | 0.66 | 0.76 |
| DecisionTreeClassifier | 0.84 | 0.85 | 0.84 | 0.85 |
| RandomForestClassifier | 0.84 | 0.80 | 0.74 | 0.78 |

1、普通的准确度无法全面地衡量算法的效果，从Precision、Recall、和 F1-Score等方面来更好地把握算法的运行结果。

2、如果训练集很小，那么高偏差/低方差分类器（如朴素贝叶斯分类器）要优于低偏差/高方差分类器（如k近邻分类器），因为后者容易过拟合。然而，随着训练集的增大，低偏差/高方差分类器将开始胜出（它们具有较低的渐近误差），因为高偏差分类器不足以提供准确的模型。

3、综合来看，AdaBoostClassifier效果更好。

4、实验使用python实现，在运行前先运行代码目录下的脚本安装环境。