# 实验2《基于概率生成模型的二分类》

实验学时： 2 实验地点： 二综204 实验日期： 2019/10/16

1. **问题描述**

学会使用学习到的概率生成模型相关的知识，找出各类别最佳的高斯分布，从而达到通过输入测试，完成二分类任务，成功预测是否该用户年薪达到50k美元。

1. **设计简要描述**

**1．**数据集以及任务介绍

#### 1.1 任务

二分类任务：确定一个人是否年收入超过5万美元。

#### 1.2 数据集

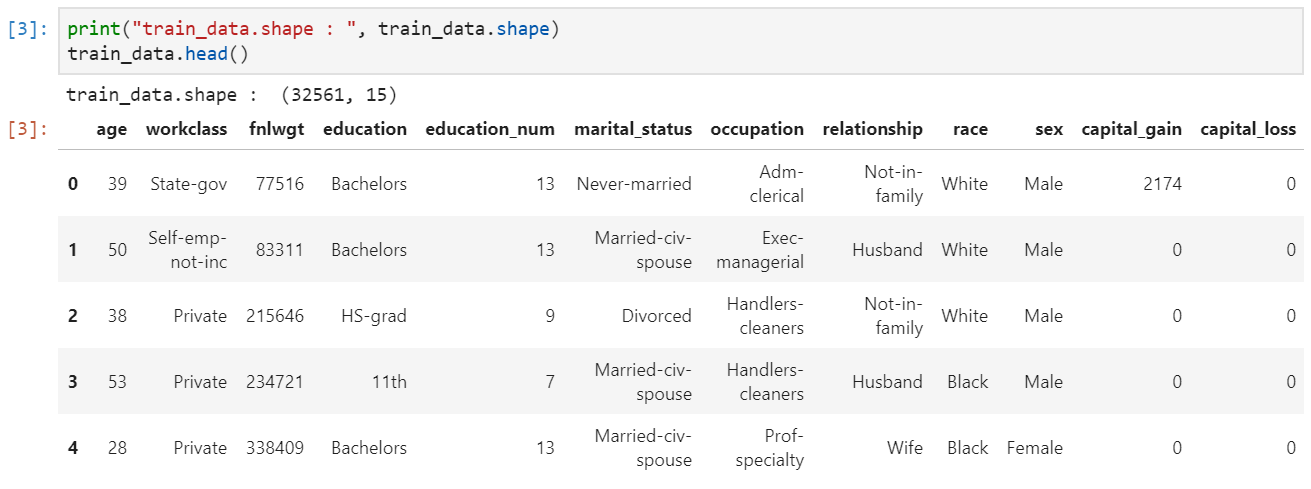
贝克尔(Barry Becker)从1994年的人口普查数据库中提取的数据。利用以下条件提取出一组相当清晰的记录:

((AGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)).

#### 1.3数据属性信息

总共是提供了两个数据集：train.csv，test.csv。并且一共是有以下属性：

age, workclass, fnlwgt, education, education num, marital-status, occupation relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, make over 50K a year or not（income）



#### 1.4 提供的特征属性格式

对于数据集：X\_train,Y\_train,X\_test：

1. 对于train.csv中离散的特征属性（例如work\_class,education...）——使用独热（onehot）编码来处理
2. 对于train.csv中连续的特征属性（例如age,capital\_gain...）——则是直接用数据集中的原值
3. 在X\_train, X\_test中 : 每行包含一个表示样本的106-维特性
4. Y\_train: label = 0 表示  “<= 50K” 、 label = 1表示“ >50K ”

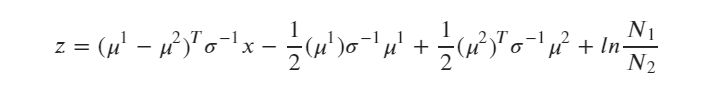
#### 1.5 提交的数据格式

结果输出到csv文件中：

第一行必须为id,label，从第二行开始为预测结果

每行分别为id以及预测的label，请以逗号分隔

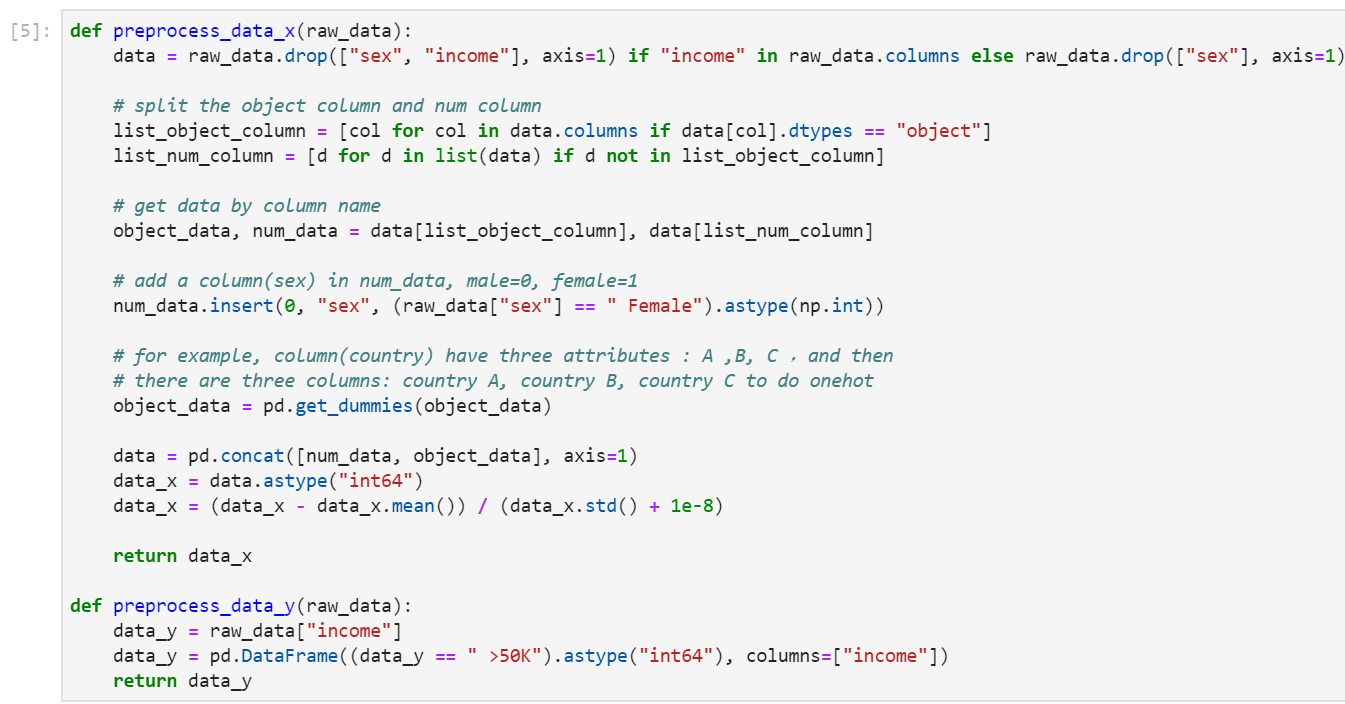
2.实现方法—概率生成模型



1. **程序清单**
2. **数据读取**

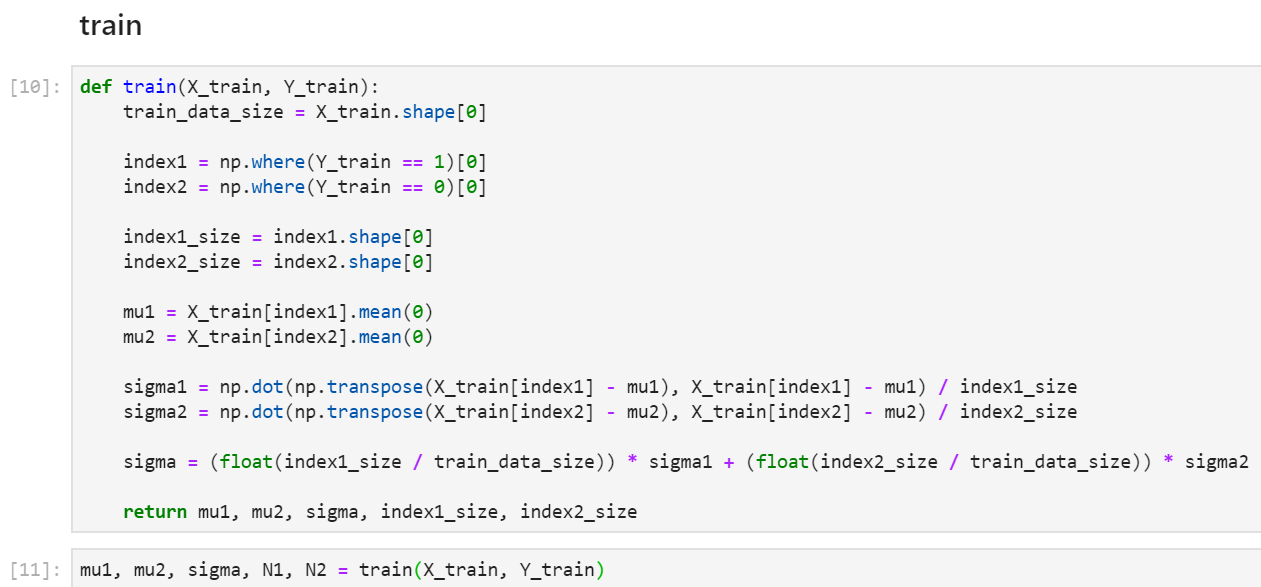
****

1. **准备训练集，测试集数据**

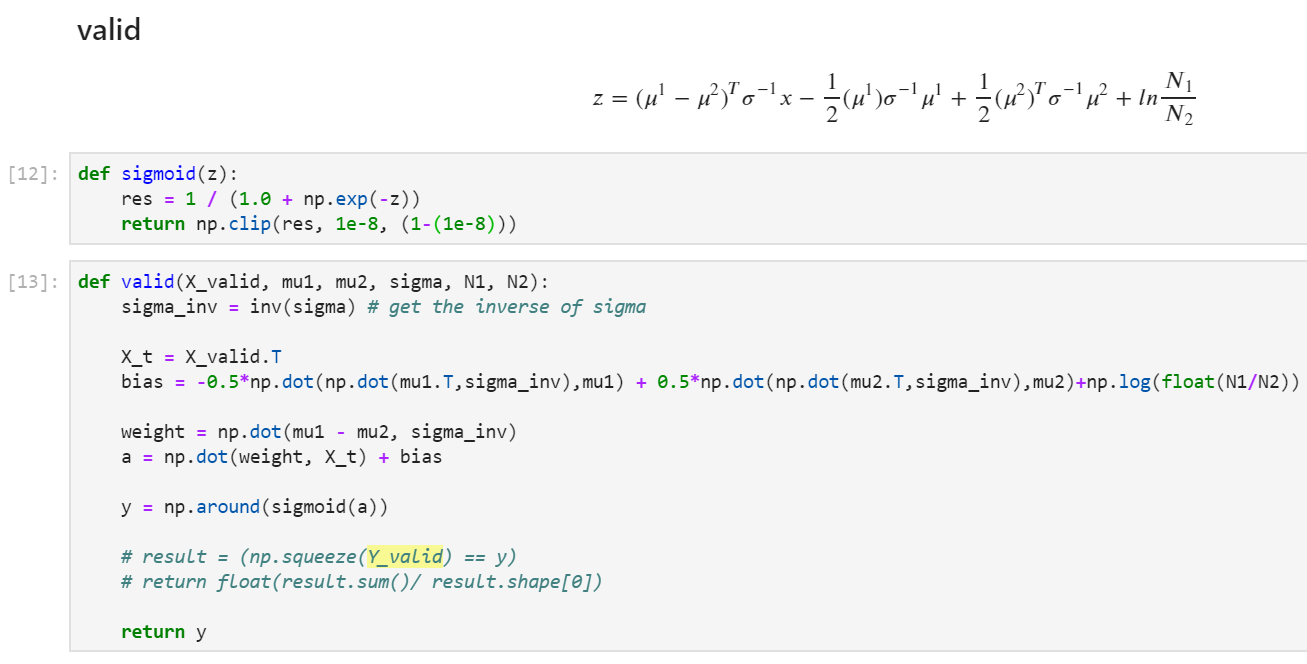
****

****

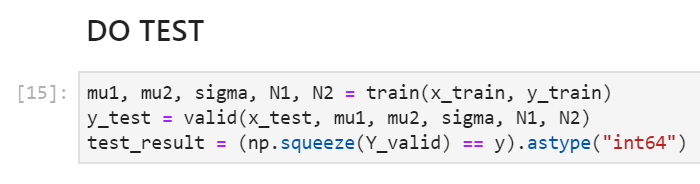
1. **训练**

****

1. **测试**

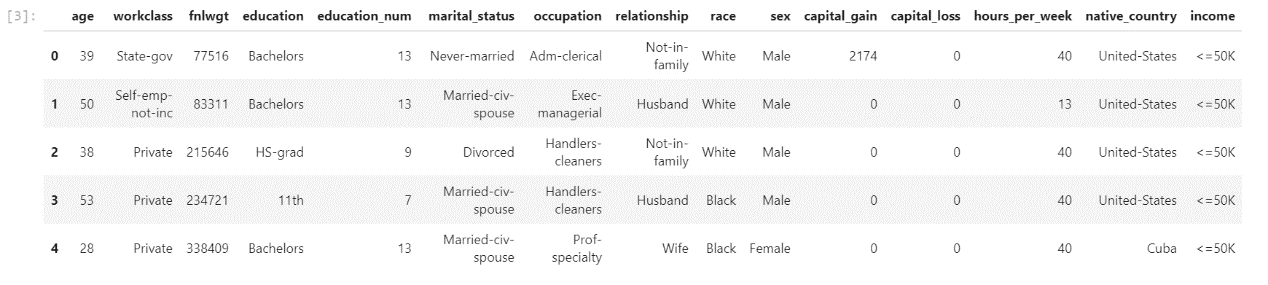
****

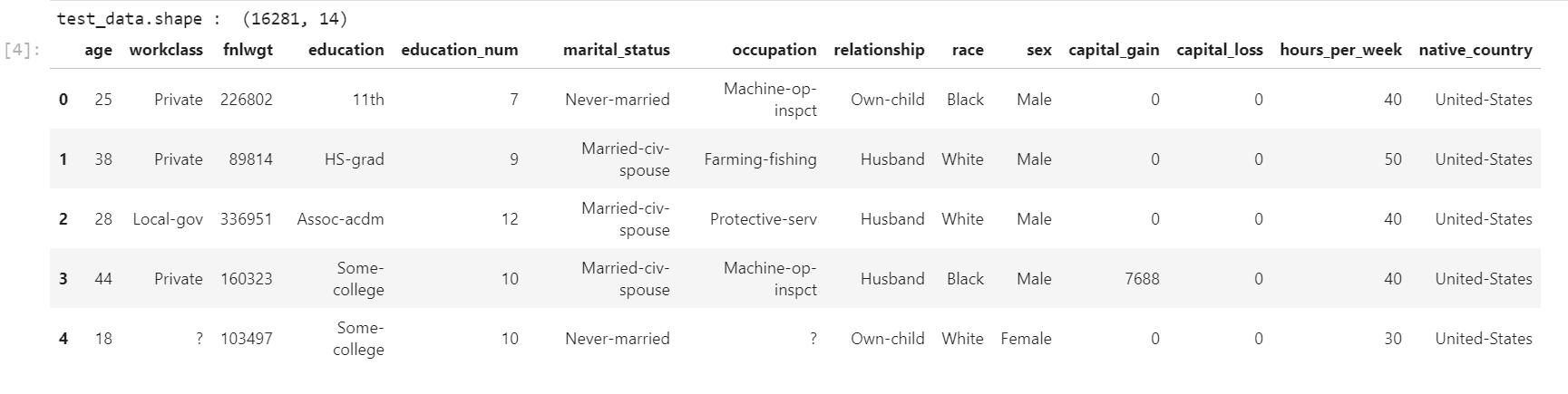
**5.测试**

****

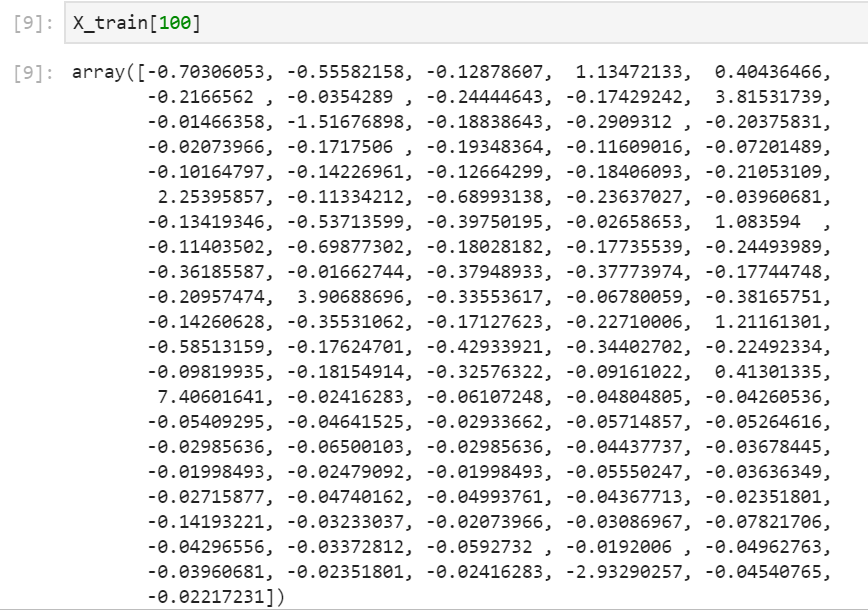
**6.保存测试结果**

****

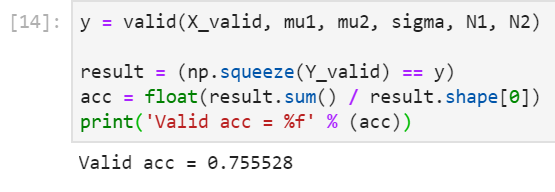
1. **结果分析**
2. **查看训练数据表表头部分**
3. **查看测试数据表表头部分**

****

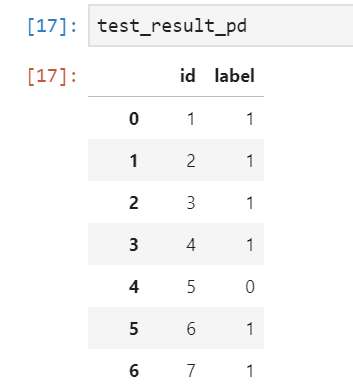
1. **分割后的训练集第100行数据**

****

1. **验证结果**

****

1. **测试集后训练结果**

****

1. **调试报告**
2. **所遇到的问题**

本次实验数据处理方面没有遇到太大的问题，使用的是pandas，操作非常简便。在写valid函数的时候突然卡住了，复习一遍PPT之后才勉强将公式理解，这或许是这次实验所遇到最大的困难吧

1. **经验和体会**

使用jupyter lab学会了很多快捷键之后，操作都不需要使用鼠标了，敲代码效率max。感觉机器学习的公式还是太难理解了，复杂公式的推导理解起来还是比较吃力，自己多花点时间去理解吧哎