1.背景技术和基本原理

机器学习(Machine Learning, ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径。本实验基于 Python，对给定数据进行分析，包括房价预测和薪资概率预测。所用的模型为线性回归和逻辑线性回归模型。

1.1 线性回归模型

线性回归(Linear Regression)是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。线性函数模型如下。

$$h_\theta(X)=\theta^T X \tag{1}$$

其中，$\theta=[\theta_0,\theta_1,...,\theta_{d-1}]^T\in R^{d\times 1}$ 是回归系数，$X=[1,x_1,x_2,...,x_{d-1}]^T\in R^{d\times 1}$ 是相关变量，其中变量个数为 d-1 个，θ_0 为该线性方程的截距。

已知一些数据 X，如何求里面的未知参数 θ，给出一个最优解 h_θ。一个线性矩阵方程，直接求解，很可能无法直接求解。现实生活中几乎没有唯一解的数据集。因此，需要退一步，将参数求解问题，转化为求最小误差问题，求出一个最接近的解，这就是一个松弛求解。其模型的好坏由以下损失函数判定。

$$J(\theta)=\sqrt{\frac{\sum_{i=1}^{n}(h_\theta(X^{(i)})-y^{(i)})^2}{n}} \tag{2}$$

其中，n 为测试样本数，$y^{(i)}$为测试样本的实际值，公式(2)是模型预测值与实际值的差的平方和。而学习的过程是通过迭代，求出 θ 的值。

$$\theta=\arg\min_{\theta} J(\theta) \tag{3}$$

当 X 是列满秩可以直接求解，否则用梯度下降法可以求得最小二乘解。

1.2 逻辑回归模型

逻辑回归(Logistic Regression)的模型是一个非线性模型，sigmoid 函数，又称逻辑回归函数。但是它本质上又是一个线性回归模型，因为除去 sigmoid 映射函数关系，其他的步骤，算法都是线性回归的。可以说，逻辑回归，都是以线性回归为理论支持的。只不过，线性模型，无法做到 sigmoid 的非线性形式，sigmoid 可以轻松处理 0/1 分类问题。其 sigmoid 函数如下。

$$f(t)=\frac{1}{1+e^{-t}} \tag{4}$$

把式(1)代入式(4)可得逻辑回归的输出值 $f(t)\in(0,1]$，公式如下。

$$f(J(\theta)) = \frac{1}{1+e^{-J(\theta)}} \tag{5}$$

逻辑回归的损失函数如下。

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} \log h_\theta(X^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(X^{(i)})) \right] \tag{6}$$

再次利用公式(3)可以学习出 θ 的值。

2.设计过程

2.1 线性回归模型预测房价

从所给的数据中，先进行数据分析。通过房间数量 x_1，起居室内部面积 x_2，房屋楼层 x_3 和房屋视觉效果 x_4 等调查的数据，去预测房价 $h_\theta(X)$。首先检查价格的分布情况，如图 1。

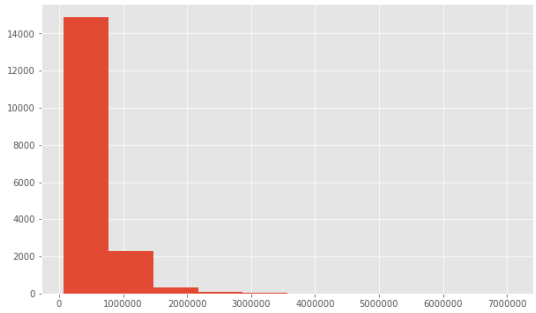

图 1 原始价格分布

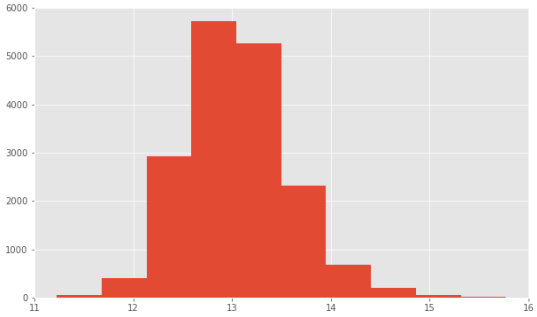

图 2 对数变换后价格分布

检查原始价格数据偏斜度(skewness)为 3.65010734496。对其做对数变换后的数据偏斜度为 0.440197120303。变换后的价格分布如图 2 所示。通过对数变换，可以改善数据的线性度。变换后数据偏斜度更接近于 0，分布为正太分布。最终预测的价格需要通过反变换，即取指数得到最终价格。

接下来分析各个变量与变换后的房价的相关性，如表 1。

表 1 房价与变量的相关性

correlation	price	sqft_living	grade	sqft_above	…	long	date	id	zipcode
price	1.000	0.697	0.670	0.601	…	0.014	0.006	-0.020	-0.046

表 1 给出了与价格相关性较强的前四项和相关性较弱的后四项。从表 1 可以看出，date，id 和 zipcode 等与价格的相关性不高，所以可以考虑删除该特征。

考察部分相关变量与房价的散点图，如图 3 所示。

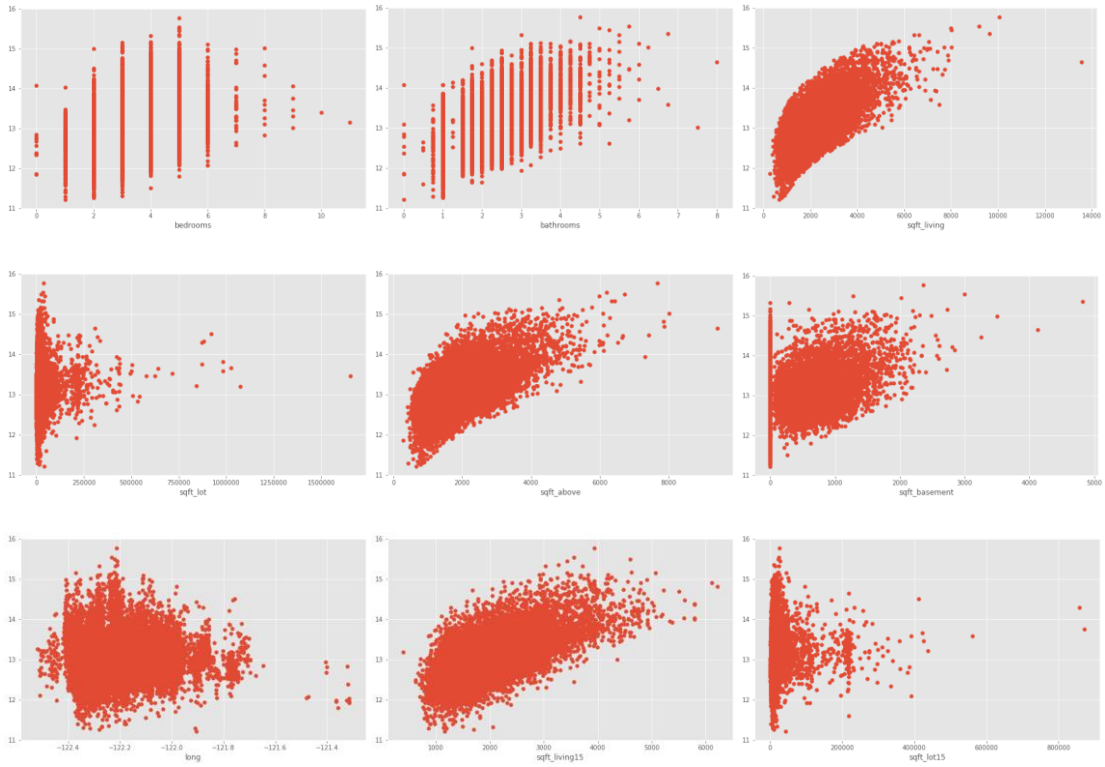

图 3 部分相关变量与房价散点图

从图 3 可以看出部分变量有部分散点较奇异，比较不集中在大部分数据中，这些数据很影响回归模型的拟合。比如 sqft_lot 在大于 50000 的时候有部分离散的点，这些数据可以删除掉，然后再进行线性回归拟合。

最后，对训练数据进行测试，把训练数据随机分成两部分，1/3 用来测试，2/3 用来训练线性回归模型。训练的结果如图 4 所示。

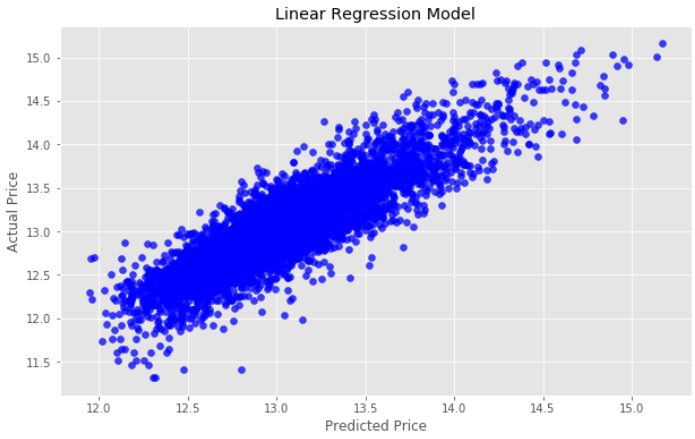

图 4 用训练的数据集进行测试

最终算出来的损失函数值为 0.0602484829117，比较小，符号大致的需求。对其线性模型加一个正则化项，另 alpha 为 0.01,0.1,1,10,100 发现并不能提升多少性能。所以最终确定出来的模型不加正则化项。

2.2 逻辑回归模型预测薪资

读取所给训练数据集和测试数据集，发现两份数据的特征变量除了有数字特征外，还有文字特征。数字特征的身份号“id”和身份背景“fnlwgt”与逻辑回归模型相关性不是很大，选择删掉。查看非数字特征选项，如表 2 所示。

表 2 部分训练和测试数据集统计

数据集	train.csv			test.csv	
相关变量	sex	native_country	income	sex	native_country
count	26561	26561	26561	6000	6000
unique	2	42	2	2	41
top	Male	United-States	<=50K	Male	United-States
freq	17802	23786	20132	3988	5384

从表 2 中发现，sex 和 income 只有两类，所以我们可以先处理他们为二值化数据，比如 sex 为 1 则表示 Male，否则不是。再看两类数据集的国籍，发现训练数据集有 42 类国籍，而测试数据集有 41 类国籍，但有一个共同的特点就是美国国籍占多数，所以，采取二值化处理为美国国籍为 1，非美国国籍为 0。

对于其他非数字特征，我们采用 one-hot encoding 可以将分类数据转化为数值特征。如图 5 所示。

Index	race
0	White
1	White
2	Black
3	Other
4	Black

Index	White	Black	Other
0	1		
1	1		
2		1	
3			1
4		1	

图 5 one-hot encoding

从图 5 知道，当一个变量有 3 个类别的时候，通过 one-hot encoding 会增加多 3 列变量特征。所以上诉国籍变量较多，用该方法会增加数字特征的维度，增加运算量。

最后，还是对训练数据进行测试，把训练数据随机分成两部分，1/3 用来测试，2/3 用来训练逻辑回归模型。通过不断调整特征，最终确定抛弃 id，fnlwgt 和 education_num，训练的结果损失函数的值为 0.00436842449597。

3.实验结果

利用所给的训练数据训练模型，用 house_train.csv 训练线性回归模型预测 house_test.csv 的房价并保存成 house_result.csv 和用 income_train.csv 训练逻辑回归模型预测 income_test.csv 并保存 income_result.csv。实验结果如下图 6 所示。

id	price
1222029077	137169.7
7983100150	218319.2
7575600610	373401
7399000360	404068.4
2822079012	279070.2
1492800296	370894.4
3521069142	432210.1

(a)线性回归预测房价

id	income_prob
30146	0.134750798
30772	0.031854785
19298	0.867035277
6064	0.831096537
11436	0.033676488
12984	0.021831052
26915	0.094380149

(b)逻辑回归预测薪资

图 6 部分预测结果

本实验结果总共有 4 个 py 文件：house_analysis.py，predict_house_price.py，income_analysis.py 和 predict_employee_income.py。2 个 csv 文件：house_result.csv 和 income_result.csv。

4.总结与分析

最初在预测房价的时候，我只用了一部分特征，结果预测的房价有负数的，这肯定是不可以的，所以我用了很多特征，但效果还是不行，通过查找文档，发现需要把价格取对数，效果才能明显提高。至于去除异常点，效果也只是提高一点点而已。而在预测薪资的时候，发现虽然没有空值，但又部分数据是以“?”的形式存在，把它替换为出现频率最高的词汇时，发现效果也是一般，所以不替换它，而是作为一类去训练逻辑回归模型。