Day3 学习报告

**今日练习内容：多元线性回归**

多元线性回归是尝试通过用一个线性方程组来适配观测数据，这个线性方程组是在两个或两个以上的特征和响应之间建立一个关系。

多元线性回归的实现步骤与之前的简单线性回归相似，只是在评价部分有所不同。

#数据预处理

·导入数据集

[[165349.2 136897.8 471784.1 'New York']

[162597.7 151377.59 443898.53 'California']

[153441.51 101145.55 407934.54 'Florida']

[144372.41 118671.85 383199.62 'New York']

[142107.34 91391.77 366168.42 'Florida']

[131876.9 99814.71 362861.36 'New York']

[134615.46 147198.87 127716.82 'California']

[130298.13 145530.06 323876.68 'Florida']

[120542.52 148718.95 311613.29 'New York']

[123334.88 108679.17 304981.62 'California']]

[192261.83 191792.06 191050.39 182901.99 166187.94 156991.12 156122.51

155752.6 152211.77 149759.96 146121.95 144259.4 141585.52 134307.35

132602.65 129917.04 126992.93 125370.37 124266.9 122776.86 118474.03

111313.02 110352.25 108733.99 108552.04 107404.34 105733.54 105008.31

103282.38 101004.64 99937.59 97483.56 97427.84 96778.92 96712.8

96479.51 90708.19 89949.14 81229.06 81005.76 78239.91 77798.83

71498.49 69758.98 65200.33 64926.08 49490.75 42559.73 35673.41

14681.4 ]

·将类别数据数字化

labelencoder:

[[165349.2 136897.8 471784.1 2]

[162597.7 151377.59 443898.53 0]

[153441.51 101145.55 407934.54 1]

[144372.41 118671.85 383199.62 2]

[142107.34 91391.77 366168.42 1]

[131876.9 99814.71 362861.36 2]

[134615.46 147198.87 127716.82 0]

[130298.13 145530.06 323876.68 1]

[120542.52 148718.95 311613.29 2]

[123334.88 108679.17 304981.62 0]]

onehot:

[[0.0000000e+00 0.0000000e+00 1.0000000e+00 1.6534920e+05 1.3689780e+05

4.7178410e+05]

[1.0000000e+00 0.0000000e+00 0.0000000e+00 1.6259770e+05 1.5137759e+05

4.4389853e+05]

[0.0000000e+00 1.0000000e+00 0.0000000e+00 1.5344151e+05 1.0114555e+05

4.0793454e+05]

[0.0000000e+00 0.0000000e+00 1.0000000e+00 1.4437241e+05 1.1867185e+05

3.8319962e+05]

[0.0000000e+00 1.0000000e+00 0.0000000e+00 1.4210734e+05 9.1391770e+04

3.6616842e+05]

[0.0000000e+00 0.0000000e+00 1.0000000e+00 1.3187690e+05 9.9814710e+04

3.6286136e+05]

[1.0000000e+00 0.0000000e+00 0.0000000e+00 1.3461546e+05 1.4719887e+05

1.2771682e+05]

[0.0000000e+00 1.0000000e+00 0.0000000e+00 1.3029813e+05 1.4553006e+05

3.2387668e+05]

[0.0000000e+00 0.0000000e+00 1.0000000e+00 1.2054252e+05 1.4871895e+05

3.1161329e+05]

[1.0000000e+00 0.0000000e+00 0.0000000e+00 1.2333488e+05 1.0867917e+05

3.0498162e+05]]

#躲避虚拟变量陷阱

#拆分数据集为训练集和测试集

#在训练集上训练多元线性回归模型

·创建一个LinearRegression的对象regressor，接着，用LinearRression类的fit()方法，用对象regressor在数据集上进行训练。

#在测试集上预测结果

输出保存在Y\_pred中的结果：

[103015.20159796 132582.27760815 132447.73845175 71976.09851258

178537.48221056 116161.24230166 67851.69209676 98791.73374687

113969.43533013 167921.06569551]

**注意和补充：**

·当变量过多时，会造成模型的精确度降低，比如存在一些与结果无关的变量或者一些对其他变量造成很大程度影响的变量。

·虚拟变量陷阱是指两个或者两个以上变量之间高度相关的情形。解决虚拟变量陷阱：

X = X[: , 1:]