UDACITY

项目

Creating Customer Segments

此部分属于 Machine Learning Engineer Nanodegree Program

项目审阅

代码审阅

注释

与大家分享你取得的成绩！

Meets Specifications

恭喜你通过了这个项目！
你的修改非常棒，取得了很大的进步！
加油！继续后面的学习吧～

数据研究

已选取三个数据样本，提出建立表达式并给出合理解释。

做得不错，你选择了三个客户，并且根据统计数据进行了初步分析。

建议：

我注意到你将客户采购的数量与平均值进行了比较，这是一个不错的思路。不过在后面的图表中你会发现这个数据集不是正态分布的，分布是向右倾斜的（大部份数据点位于该图的左侧），因此平均值会比中位数（50%）大很多，从上面的**数据集的统计描述**中也能看得出来。因而你可以尝试使用中位数作为比较的标准。参考上一位reviewer给出的代码，你可以将mean改为median，或者直接添加一项median比较看看。

准确报告被删除属性的预测分数，合理解释被删除属性是否具有相关性。

这一次你很好地理解了题目的要求，你通过计算得出了Detergents_Paper是一个能够很大程度上被其它特征解释的特征这样的结论。在之后的可视化过程中你就能发现是哪一个特征与它相关了。

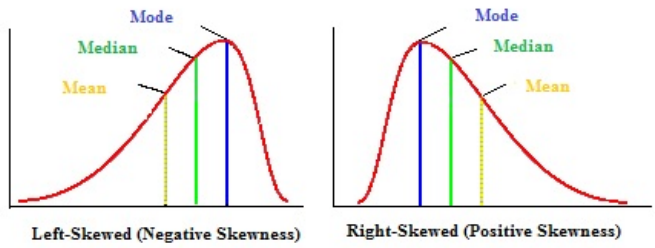
学生找出具有关联的属性并将其与预测属性相比较，随后深入讨论这些属性的数据分布模式。

非常好，你通过观察散布矩阵得出的结论验证了你上一个问题中的观点并识别出有相关性的特征。

提示：

你能够从散点图上得到这些信息：

- 分布是向右倾斜的（大部份数据点位于图的左侧，是正偏态分布，这一点你可以从上一位reviewer给你的图像中得出来，你也可以参照下图再次理解它）；
- 大部分的特征都含有一些异常值（即你提到的距离中值很远的值）；
- 中位数低于平均值。



数据处理

数据和样本的特征缩放已在代码中正确实施。

你在特征缩放上做得不错！从可视化图表中可以看出现在的数据基本呈正态分布了。

学生找出极端的异常值，讨论是否删除这些异常值，并说明删除各数据点的理由。

非常好，你准确地找出了数据中的异常值并移除了它们，这对你后面的数据分析非常有利。
你统计异常值的方法非常棒！

属性转换

准确报告主要成分分析数据的二个维度与四个维度的总方差。将前四个维度合理解释为对消费者支出的表达。

对二维缩放数据及样本数据的主要成分分析已在代码中正确实施。

成功地对数据进行了降维。

聚类

高斯混合模型和K-均值算法已进行详细比较。学生选择的算法符合算法和数据的特点。

非常好，你抓住了K-Means聚类算法和高斯混合模型最重要的不同点。

提示：

对于高斯混合模型 和 k-均值算法的主要区别就在于是否为软聚类方法。

准确报告多个轮廓分数，根据报告的最佳分数选择最佳集群数量。已给出的集群视觉化将根据已选的聚类算法生成最佳的集群数量。

很好！通过准确计算出不同聚类数目的分数来确定最佳聚类数目是非常稳妥和可靠的办法。

提示：

这里的计算量不大，你可以尝试更多的聚类数目。

根据数据集的统计描述提出每个客户细分所代表的类型。对集群中心的逆变换和反比例级联已在代码中正确实施。

非常好，你成功地通过施加反向的转换恢复了客户的花费，并且根据数据集的统计描述分析了每个客户细分所代表的类型。

客户细分正确识别样本数据点，讨论各样本数据点的预测集群。

结论

提出了某些功能改进方法，可以改进从 A/B 测试获取结果的功能。

你的实验设计的不错！

提示：

当我们选择实验组的用户时尽量从每个聚类的中心选择，这些用户更能够代表聚类的特征。

学生讨论了聚类数据如何可以通过监督学习预测新的属性。

聚类数据可以通过监督学习预测新的属性。这是特征工程的一个重要方法。

客户细分与客户通道数据进行对比，对通道数据识别客户细分的问题进行讨论，包括该表达是否符合早期结果。

模型对于两边的数据点表现得不错，但是中间重合的部分不能完美划分，不过由于你选择的是高丝混合模型，对于你感兴趣的点你可以查看它们属于每个类别的概率。

下载项目

返回 PATH

给这次审阅打分

学员 FAQ