UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

7 SPECIFICATIONS REQUIRE CHANGES

同学你好，作为第一次提交，这是一份质量不错的作业。你的思路与方法基本是正确的，只是有一些问题的讨论还需要进一步量化和加强。继续加油！

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。

这是个不错的开始，你的推论也很合理。但是为了让讨论更加准确，你最好可以将样本数据与数据的统计信息做一下比较。

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。

说的没错。R^2测量的是真实值中的方差 (v = var(y_true - y_pred))有多少被预测值(u= var (y_true - y_true))所解释。当预测效果特别的坏，其方差会远大于真实方差，此时R^2 (=1-u/v) 会远小于0.

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。

"部分指标间有一定程度的相关性":
这里最好具体说说，都有哪几个指标存在相关性?

数据处理

数据和样本的特征缩放已在代码中正确实施。

做的很好。
另外一种数据缩放的办法是Box-Cox转化：

```
from scipy.stats import boxcox
x_boxcox, _ = boxcox(x)
```

参考（英文）：http://scipy.github.io/devdocs/generated/scipy.stats.boxcox.html

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。

你的思路是正确的，但是其实还有一个点也在多于一个特征中被当作异常值，你能找出它来吗?

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。

你的计算是正确的，但是这里最好能基于消费模式更加详细的解释一下前四个特征的所代表的类型。比如说，第一个维度在detergents_paper (0.75), milk (0.4), and grocery (0.44)上有很高的比重，因此它可能代表零售业的客户。另外，也需要注意到有一些维度存在负相关的情况。

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。

你很好的使用了PCA来降维。此时你也可以再次观察数据在二维下的分布：

```
pd.scatter_matrix(reduced_data, alpha = 0.3, figsize = (10,6), diagonal = 'kde');
```

或许你会发现，第一维的分布与之前的milk, grocery, 和detergents_paper很接近。我们在之前观察到，这几个特征在第一个维度有很高的权重。

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。

确实，K-means算法简单明了，并且聚类效果很好，尤其适合运用于大型数据集。

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。

做的很好。你也可以考虑设定k-means的random_state, 以让结果可重复。

如果你持续增大聚类数目，或许轮廓系数会反弹至接近1。但是很多的聚类并没有实际的意义。鉴于这个项目的数据量很少，我们也不希望有太多的聚类数目。因此，聚类数目为2是很合理的。

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

同之前所讲，这里最好能把讨论进一步量化一下。比如将每个产品类型的花费与平均数做一下比较。

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。

这个问题实际有两个部分需要回答。可以分成下面两步：
第一步：比较样本点与类的中心点，样本更有可能属于哪一类?
第二步：运行聚类算法，结果是否与第一步的预测一致?

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。

你的结论是没错的，但是这里也需要讨论一下如何设计A/B test。我们需要回答下面几个问题：

- 客户的类是如何定义的?
- 我们是否需要对每一类客户分别进行A/B测试?

- 进行A/B测试的时候，参考组与实验组应该如何选择?

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。

确实，一个常见的办法就是将聚类的结果作为标签，从而将问题转化为分类问题。

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。

对于那些在两个类的分界上的点，很难说它们确切的类属。这时候，或许ＧＭＭ会更加有效，因为它会给出数据属于两类的概率。

RESUBMIT

DOWNLOAD PROJECT

Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

Watch Video (3:01)

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review