项目

Creating Customer Segments

此部分属于 Machine Learning Engineer Nanodegree Program

项目审阅

注释

与大家分享你取得的成绩！

Meets Specifications

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。

很棒的回答，和数据集的统计特征进行了很好的比较。做得很棒，我们建议选择 percentile 而非 mean，因为在未知数据分布的情况下使用均值作为比较对象是比较危险的——因为不清楚概率分布，所以用 percentile、median 这样的统计特征会相对好一点。更多你可以参考描述统计学、或者数据的统计量特征。

✓

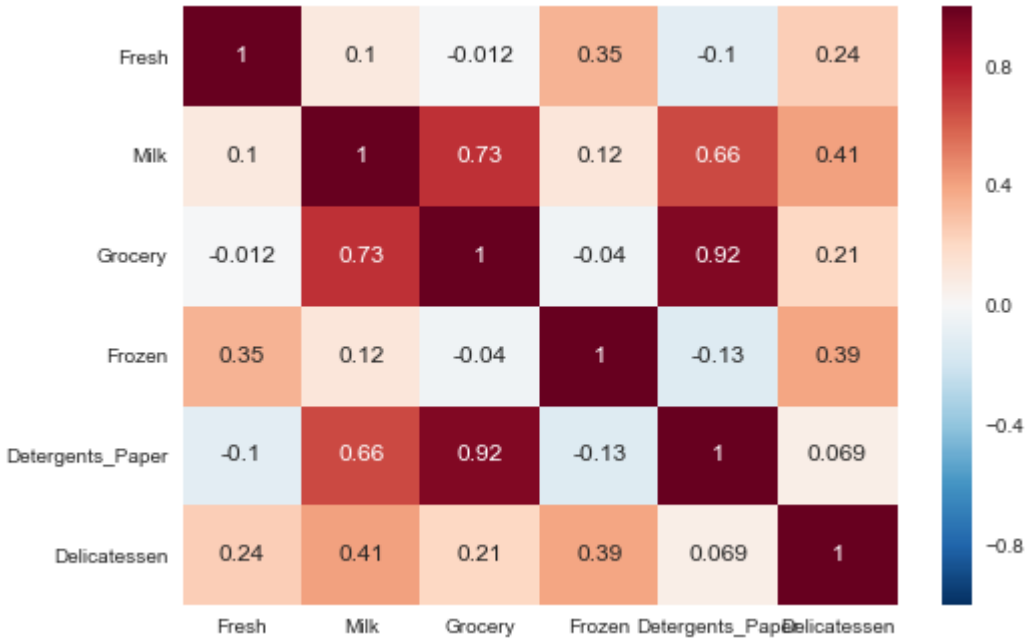
准确报告被删除属性的预测分数，合理解释被删除属性是否具有相关性。

✓

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。

这里补充一种判断属性关联度的方法,希望你可以尝试一下：

```
import seaborn as sns
sns.heatmap(data.corr(), annot=True)
```



数据处理

✓

数据和样本的特征缩放已在代码中正确实施。

✓

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。

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PCA的理解，你可以参考这篇文章，以及下面我的一些解读。
  - 在这边，我们使用了主成分分析法，将原来的6个特征通过数学变换，变换为了另外6个特征。对方差的计算，是为了让我们能够选择方差较大的特征以保留它们。每个新特征，实际上都是由原来的特征通过某种带权重的组合得到的，权重就是图中柱状图柱高度。考虑权重的绝对值，权重绝对值越大，说明权重对应的原特征对这个新特征带来的影响越大，反之亦反。对不同的feature，若权重值为同号，则说明他们有正相关性；异号则说明它们是负相关性。A和B有正相关性可以理解为，买更多的A意味着有很大可能买更多的B；负相关性意味着买更多的A意味着有很大可能买更少的B。

✓

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。

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。

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。

你还可以尝试更多的聚类数。

同时，你会发现，随着轮廓分数的增大，轮廓分数又会递增乃至达到一个局部极大值？这并不是一个好现象，因为当我们选择轮廓分数作为我们的评测基准时，我们希望这是一个凸函数（即只有一个峰），这样我们可以很顺利地对这个函数最优化。但是这个函数出现了多个峰，也就是除了2个聚类时对应的最大值以外，还有其他聚类情况对应的极大值。你可以搜一搜相关的信息，自己探索一下相关的问题，并可以写在回答中~

你可以参考这个页面。

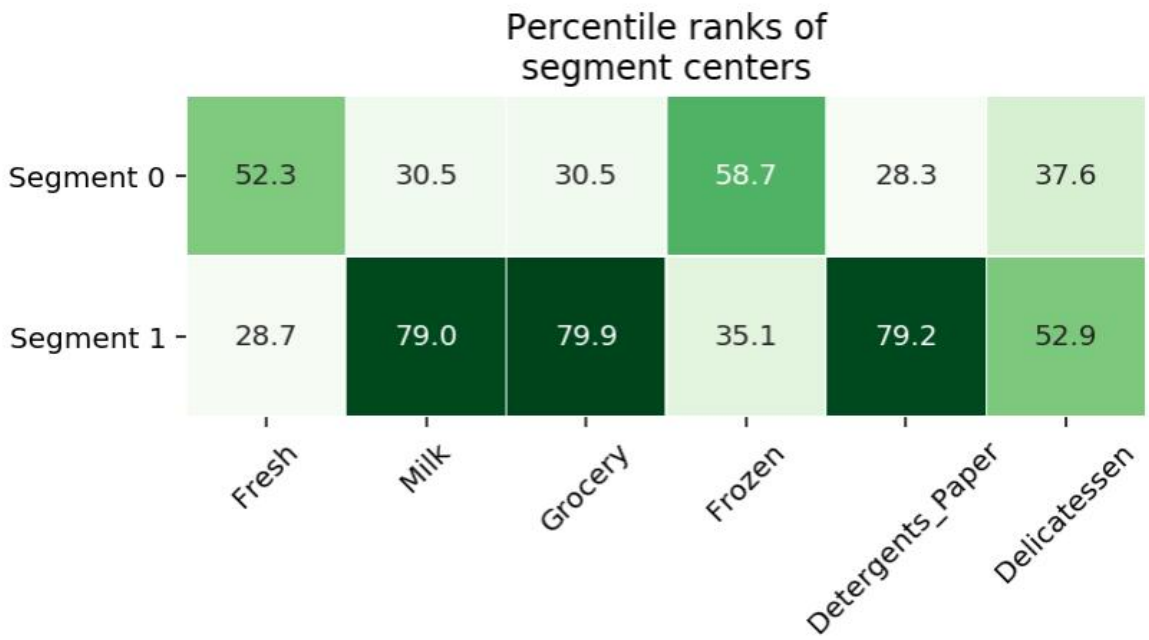
根据数据集的统计描述提出每个客户细分所代表的类型。对集群中心的逆变换和反比例级联已在代码中正确实施。

不错的分析，注意到你分析了他们的`percentile rank`，这里提供一种可视化方法，以得到更为详实的结果。

```
import seaborn as sns
import matplotlib.pyplot as plt
# add the true centers as rows to our original data
newdata = data.append(true_centers)

# show the percentiles of the centers
ctr_pcts = 100. * newdata.rank(axis=0, pct=True).loc[['Segment 0', 'Segmen
t 1']].round(decimals=3)
print ctr_pcts

# visualize percentiles with heatmap
sns.heatmap(ctr_pcts, annot=True, cmap='Greens', fmt='.1f', linewidth=.1,
 square=True, cbar=False)
plt.xticks(rotation=45, ha='center')
plt.yticks(rotation=0)
plt.title('Percentile ranks of\nsegment centers');
```



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。

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。

- 关于特征工程的进一步了解，你可以参考这个知乎回答来了解更多相关的信息：特征工程到底是什么?

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。

下载项目