Project 2 实验报告

曾正
软 71
2017011438

ABSTRACT

实验报告介绍了银行精准营销解决方案和青蛙叫声聚类分析的实现说明以及实验结果等。

Keywords

Classification, NBC, Clustering, K-Means

1. 银行精准营销解决方案

1.1 使用说明

环境依赖

```
pip install sklearn pandas numpy
```

运行方式

```
python classification.py
```

启动后会自动开始训练模型进行分类并输出分类结果。

1.2 分类器模型

- NaiveBayesClassifier（朴素贝叶斯分类器）：自行实现。

  分类器原理：有贝叶斯定理如下：

  $$P(A|B) = \frac{P(A|B)P(B)}{P(A)}$$

  由此我们可得到：

  $$P(Cla|Fea) = \frac{P(Fea|Cla)P(Cla)}{P(Fea)}$$

  利用上述公式我们就很容易求得目标

  $$P(Cla|Fea)$$

  ，即在给定特征集合（Features）的条件下，属于某一类别（Class）的概率。需要注意的是在贝叶斯定理中，有假设：特征集合（Features）中各特征相互独立。有如上假设便可得：

  $$P(Cla|Feas) = \frac{P(FeaA|Cla)P(Cla)*P(FeaB|Cla)P(Cla)}{P(Feas)}$$

  再根据中心极限定理，在样本数量足够大时，频率等于概率。

  以上便构成了实现朴素贝叶斯分类器的基本原理，基于上述原理很容易就可以实现朴素贝叶斯分类器。

- RandomForestClassifier（随机森林分类器）：调用 sklearn 库，使用默认参数。

- GradientBoostingClassifier(梯度提升树分类器)：调用 sklearn 库，使用默认参数。

1.3 特征组合选择

首先由于特征 ID 与客户信息完全无关，且数据集 y=1 集中在后面，会对分类结果造成较大负面影响，故将特其去除，其次注意到特征 Default 中大部分数据缺失，不具有参考价值，故也去除。之后分为以下两组特征组合：

- 从客户的个人信息入手进行分析。考虑客户的年龄、工作、受教育程度、婚姻等会影响客户财产状况及理财思维的因素，构成特征组

合：["age", "job", "marital", "education", "balance", "housing", "loan"]

- 从银行客服与用户的联系频率及效果入手进行分析。首先去除 Month、Day，这是由于缺乏年份，无法确定时间的早晚。剩余特征便可构成银行与客户沟通相关的特征组合：["contact", "duration", "campaign", "pdays", "previous", "poutcome"]

1.4 模型验证

由于提供的数据集全部为有标数据，因此采用了 K-折交叉验证的方法。考虑到数据集大小，此处采用的是 10 折交叉验证，即将全部数据集十等分，依次将每一份作为测试集，剩余 9/10 用作训练集通过比较结果从而验证模型的可靠性。考虑到前面提到的数据集特征 y 的有序性，故在读入数据时首先进行了乱序。

1.5 评价指标

评价指标采用分类正确率的大小进行比较，其范围为 $[0, 1]$。

1.6 分类效果差异及其原因

某一次训练的分类结果如下：

分类器 (组合 1)	RFC	GBC	NBC
1	0.872	0.879	0.669
2	0.871	0.883	0.662
3	0.877	0.889	0.665
4	0.863	0.870	0.655
5	0.872	0.887	0.651
6	0.863	0.872	0.665
7	0.870	0.884	0.635
8	0.866	0.882	0.659
9	0.875	0.887	0.655
10	0.867	0.881	0.667

分类器 (组合 2)	RFC	GBC	NBC
1	0.882	0.904	0.795
2	0.890	0.913	0.789
3	0.890	0.913	0.796
4	0.878	0.893	0.789
5	0.898	0.909	0.790
6	0.877	0.890	0.793
7	0.879	0.902	0.781
8	0.887	0.911	0.796
9	0.894	0.907	0.789
10	0.882	0.900	0.781

从上述数据我们可以得到两个较为明显的特征：

- 分类器分类效果：GBC > RFC > NBC
- 特征组合分类效果：组合 2 > 组合 1，其中 RFC 和 GBC 收到影响较小，对 NBC 影响很大。

原因分析：

GBC 和 RFC 的分类效果接近，均达到 0.88 左右，相比之下 NBC 的表现就要差了不少，组合 2 有 0.79 左右正确率，而组合 1 时仅有 0.65 左右的正确率。关于这一点根据前面提到的朴素贝叶斯分类器的基本原理我们很容易找到原因，贝叶斯定理的前提条件是，特征组合的各个特征之间需要相互独立。而在我们所选择的无论组合一还是组合二都很容易发现它们彼此之间并非是完全独立的，例如组合二中的 Campaign（在本次活动中，与该客户交流过的次数）就和 Duration（最后一次联系的交流时长）可能存在一些关系，即两者之间并不独立。基于这一点就很容易理解为何 NBC 的分类效果较差。

对于特征二我们可以从之前选择特征组合的出发点进行考虑。组合一是客户的个人特征，几乎与银行无关，而组合二则是客户与银行、与该产品都相关的重要信息。此外，加上各特征间的独立性增强，因此在由组合一到组合二时，NBC 的正确率得到了很大的提升。相比之下，GBC 和 RFC 由于正确率已经达到了其自身分类能力的阈值故正确率较难得到很大的增长。

2. 青蛙叫声聚类分析

2.1 使用说明

环境依赖

```
pip install matplotlib
```

运行方式

```
python clustering.py
```

启动后会自动开始训练模型进行聚类并展示可视化聚类结果。

2.2 分类器模型

- K-Means：自行实现。

  K-Means 聚类算法的基本思想很简单，就是将全部数据划分为 k 个簇（Cluster），使得簇内的各点间尽可能紧密连接在一起，而簇间的距离尽可能地大。

  基于上述基本思想我们首先将全部数据划分为 $C_1, C_2, ..., C_k$，给个簇的质心分别为 $\mu_1, \mu_2, ..., \mu_k$，其表达式为

  $$\mu_i = \frac{\sum\limits_{x \in C_i} x}{|C_i|}$$

  利用上式，我们可以得到平方误差为

  $$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$$

  这样我们的目标就变成了求解 E 的最小值，但是这是一个 NP-hard，故采用启发式的迭代过程。

  首先随机选取 k 个簇的质心，计算所有点到该质心距离，比较距离取较小者即可得到首次迭代后的簇划分。之后便可求得各簇的质心，之后重复上述过程直到簇不再改变。

  按照上述原理我们便可以实现 K-Means 聚类模型。

- Birch：调用 sklearn 库，$n\_cluster = k$。

2.3 特征组合选择

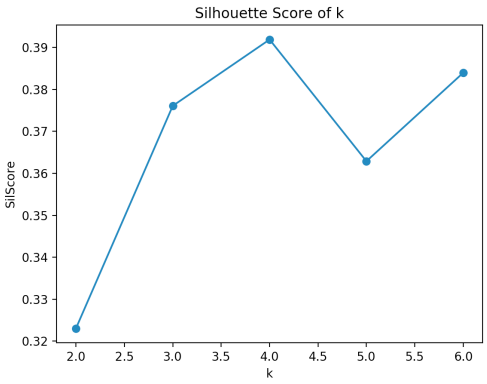
特征组合的选取采用了两种方式：

- 手动选取：MFCC 的各个特征量间并没有明显的主次之分，可以考虑随机选取。此处是选择最后 5 个特征量。
- 使用 PCA 降维，减少特征量。

2.4 距离度量及超参数选择

- 距离度量：采用欧几里得距离。
- 超参数选择：K-Means 和 Birch 的超参数均为 k，这里 k 的确定是通过计算 $k = [2, 6]$ 时得到的轮廓系数大小进行比较，选取轮廓系数较大的 k 作为超参数。但注意到此次聚类是对青蛙所属的科（Family）进行聚类比较，共四科，故 $k = 4$

Figure 1: $k - SilScore$



次数选择 $k = 4$

2.5 评价指标

- Calinski-Harabaz Index（CH 指标）：

  $$CH(k) = \frac{trB(k)/(k-1)}{trW(k)/(n-k)}$$

  故 CH 指标越大簇自身越紧密，簇与簇之间越分散，聚类结果越好。

- Silhouette Coefficient（轮廓系数）：

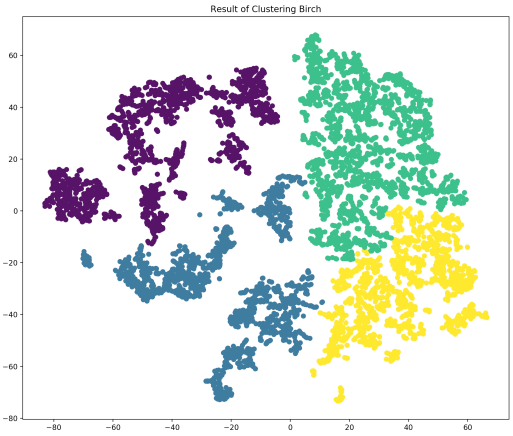
  $$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

  故轮廓系数越接近 1 聚类效果越好，越接近 -1 结果越差

2.6 聚类效果差异及其原因

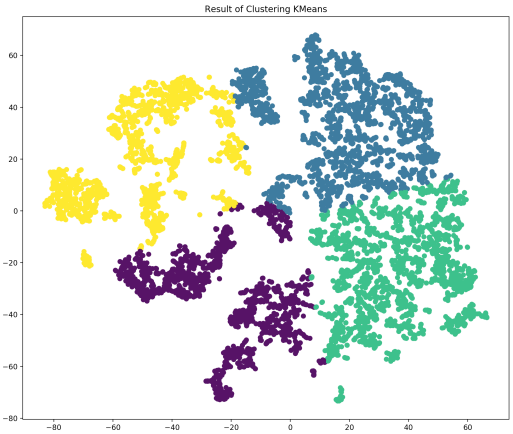
- Birch，特征组合 1。CH 指标：6540.1，轮廓系数：0.389

Figure 2: $Birch-1$



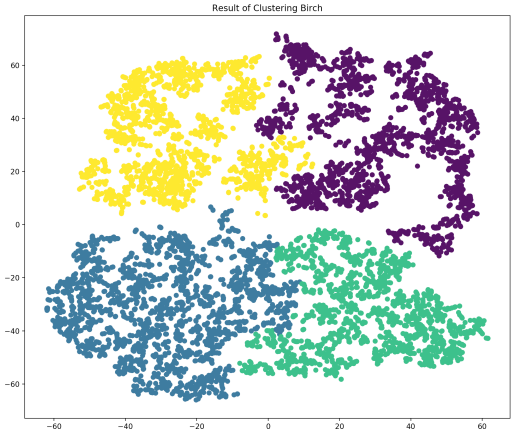
- K-Means，特征组合 1。CH 指标：7377.3，轮廓系数：0.414

Figure 3: $KMeans-1$



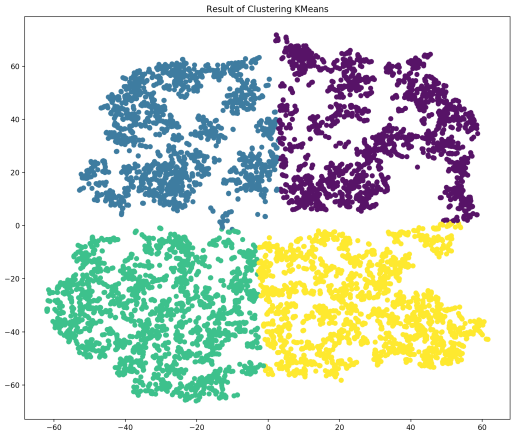
- Birch，特征组合 2。CH 指标：722.4，轮廓系数：0.416

Figure 4: $Birch-2$



- K-Means，特征组合 2。CH 指标：8317.9，轮廓系数：0.441

Figure 5: $KMeans-2$



原因分析：

使用特征组合 2 时聚类效果有一定提升，这是由于特征组合 2 是由 PCA 进行降维之后再去训练模型，提升了训练集的可靠性。

此外 KMeans 的聚类效果优于 Birch，这是由于 Birch 更适用于 k 值较大，即簇类较多的情况，此处使用的 $k=4$，故聚类效果较差。