大数据分析实验报告——聚类

苏致成 201250104

注明：

本文档采用markdown书写，因此，转为word部分公式可能出现显示异常等问题。

使用方法

SimpleKMeans

聚类属于无监督学习，KMeans聚类是最基础常用的聚类算法，其基本思想是，通过迭代寻找k个聚簇的一种划分方案，使得聚类结果对应的损失函数最小。其中，损失函数可以定义为各个样本距离所属聚簇中心点的误差平方和。即：

$$J(c,\mu)=\sum_{i=1}^{M}||x_i-\mu_{c_i}||^2$$

其中， x_i 代表第 i 个样本，c_i 是 x_i 所属的聚簇，μ_{c_i} 代表聚簇对应的中心点，M 代表样本总数。

核心部分即为先固定中心点，调整样本所属类别来减少 J ;再固定每个样本类别，调整中心点继续减小 J 。两个过程交替循环， J 单调递减直到最小值，中心点和样本划分的类别同时收敛。

算法流程

1. 数据预处理。包括标准化、异常点过滤等。
2. 随机选取k个中心，记为 $\mu_1^{(0)},\mu_2^{(0)},\dots,\mu_k^{(0)}$ 。
3. 定义损失函数：$J(c,\mu)=min\sum_{i=1}^{M}||x_i-\mu_{c_i}||^2$ 。
4. 令 $t=0,1,2...$ 为迭代次数，重复如下过程直至 J 收敛：
    - 对于每一个样本 x_i ，将其分配到距离最近的中心，即：

$$c_i^t<-argmin_k||x_i-\mu_k^t||^2$$

- 对于每一个类中心 k ,重新计算该类的中心。即：

$$\mu_k^{(t+1)}<-argmin_\mu\sum_{i:c_i^t=k}^{b}||x_i-\mu||^2$$

EM

EM算法是一种迭代优化策略，它的每一轮迭代过程都分为两步，一个是期望步（E步），另一个是极大步（M步）。基本思想是根据观测值，估计出模型参数的值，然后再根据上一步估计出的参数值估计缺失数据的值，然后再根据估计出的缺失数据加上观测值重新再对参数值进行估计，反复迭代，直至收敛。

缺点：对初始值敏感，聚类结果随不同的初始值波动较大。

算法流程

1. 输入观察的数据 $x = (x_1, x_2, \ldots, x_n)$ ，联合分布 $p(x, z; \theta)$,条件分布 $p(z|z, \theta)$,最大迭代次数 J 。
2. 随机初始化模型参数 θ 的初值 θ_0 。
3. 令 $j = 1, 2, \ldots J$ 开始 EM 算法迭代步骤：
    - E步骤：计算联合分布的条件概率期望：

$$Q_i(z_i) = p(z_i|x_i, \theta_j)$$

$$l(\theta, \theta_j) = \sum_{i=1}^{n} \sum_{z_i} Q_i(z_i) log \frac{p(x_i, z_i; \theta)}{Q_i(z_i)}$$

    - M步骤：极大化 $l(\theta, \theta_j)$,得到 θ_{j+1} ：

$$\theta_{j+1} = argmax l(\theta, \theta_j)$$

    - 若 θ_{j+1} 已经收敛，则算法结束。否则继续进行E步骤和M步骤进行迭代。

数据集处理思路

SimpleKMeans

1. 缺失值处理。将含有缺失值的条目均去除。
2. 读入 $iris.arff$，调用 $SimpleKMeans$,并且将其聚类的数量设置为3（注：在KMeans中聚类的数目k的值一般比较难以确定，此处确定为3的部分原因是已知其结果的可靠性）。
3. 输出结果。

EM

1. 缺失值处理：观察数据集可知，并无缺失数据值。
2. 读入 $iris.arff$，调用 $setClassIndex$ 将最后一列属性不予纳入聚类考虑。
3. 设置最大迭代次数为100，并且聚簇的数目是3。
4. 打印出对数似然度量等值。

实验结果

SimpleKMeans

观察到KMeans方法正好分为3种类别的鸢尾属植物。

可以观察到诸如花瓣宽度方面， $virginica$ 明显比 $setosa$ 高等指标信息。

SimpleKMeans_

```
Number of iterations: 3
Within cluster sum of squared errors: 7.817456892309574

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor
Cluster 2: 6.9,3.1,5.1,2.3,Iris-virginica

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data              0              1              2
                          (150.0)         (50.0)         (50.0)         (50.0)
================================================================================
sepallength                5.8433          5.936          5.006          6.588
sepalwidth                  3.054           2.77          3.418          2.974
petallength                3.7587           4.26          1.464          5.552
petalwidth                 1.1987          1.326          0.244          2.026
class                 Iris-setosa Iris-versicolor     Iris-setosa  Iris-virginica
```

EM

从下图可知每个类的分类聚簇后各类的数量、各项指标等信息。

可以观察到诸如花瓣宽度方面， $virginica$ 明显比 $setosa$ 高等指标信息。

```
                Cluster
Attribute            0       1       2
                  (0.41)  (0.33)  (0.25)
=====================================
sepallength
  mean           5.9275   5.006  6.8085
  std. dev.      0.4817  0.3489  0.5339

sepalwidth
  mean           2.7503   3.418  3.0709
  std. dev.      0.2956  0.3772  0.2867

petallength
  mean           4.4057   1.464  5.7233
  std. dev.      0.5254  0.1718  0.4991

petalwidth
  mean           1.4131   0.244  2.1055
  std. dev.      0.2627  0.1061  0.2456

Clustered Instances

0       64 ( 43%)
1       50 ( 33%)
2       36 ( 24%)
```

并且根据下图可知，聚类过程很好地将三个子类进行了划分，并且错误率（9.3333%）较低。

```
Log likelihood: -2.055


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 50  0  0 | Iris-versicolor
 14  0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :    14.0      9.3333 %

logLikelihood: -2.274316661821451
```