# KNN

Mia Feng

2018 年 4 月 21 日

## 1 概述

KNN：有监督，生成式模型，竞争学习算法，懒惰学习。用于聚类，即将没有类标的数据集分组为一些异质性的子群（cohesive "clusters"）[**?** ]。算法对初始值 $k$ 敏感。

懒惰学习是指直到需要预测时算法才建立模型。它很懒，因为它只在最后一刻才开始工作。优点是只包含了与未知数据相关的数据，称之为局部模型。缺点是，在大型训练数据集中会重复相同或相似的搜索过程，带来昂贵的计算开销。

求解目标：聚类，每个样本点被标以类标。

$$c^{(i)} = \{j | j \in [1, 2, \cdots, k]\} \tag{1}$$

求解思路：最小化误差平方和, 取距离样本点最近的类标作为样本的 label。Concretely, 通过迭代寻找 $k$ 个聚类，使得这 $k$ 个聚类的均值所代表相应各类样本时所得的总体误差最小。记 $c$ 为类标，$\mu$ 为 cluster 中心，代价函数为 [**?** ]：

$$\boldsymbol{J}(c, \mu) = \sum_{i=1}^{k} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2 \tag{2}$$

求解方法：误差平方和最小化。

### 1.1 推导

**聚类中心** cluster centroids，是 cluster 内样本点的均值。有 $k$ 个，在初始化时候指定。更新时，聚类中心是求取当前 cluster 的均值。对于类 $j$，

其类中心 $\mu_j$（对于聚类中心的当前猜值）为

$$\mu_j := \frac{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\}} \tag{3}$$

**类标**

$$c^{(i)} := \arg\min_j \left\|x^{(i)} - \mu_j\right\|^2 \tag{4}$$

# 2 算法实现

见 CS229[**?** ]

1. 随机初始化 cluster centroids $\mu_1, \mu_2, \cdots, \mu_k \in \mathbb{R}^n$
2. 迭代直至收敛 {

   对于每一个样例 $i$，计算类标

   $$c^{(i)} := \arg\min_j \left\|x^{(i)} - \mu_j\right\|^2 \tag{5}$$

   对于每一个类 $j$，更新 cluster centroids：

   $$\mu_j := \frac{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\}} \tag{6}$$

   }

# 3 Implementation

聚类测试：数据在 data.csv