Naive Bayes

Mia Feng

2018 年 4 月 19 日

1 概述

Naive Bayes：监督学习，生成式模型。用于分类，朴素指的是各特征条件独立 [**?**]。用于垃圾邮件分类等。

求解目标：分类，以后验概率最大时对应的类别作为预测分类结果。

$$y = \arg\max_{c_k} P\big(Y = c_k\big) \prod_{j=1}^{n} P\big(X_j = x_j | Y = c_k\big) \tag{1}$$

求解思路：最大化后验概率，或者说省略分母不看后最大化似然, 取后验概率最大或者似然值最大对应的类标作为预测类标。Concretely, 分别计算各类别出现概率 $P\big(Y = c_k\big), k = 1, 2, \cdots, m$；分别计算各类别下对应特征出现的概率 $P\big(X = x_j | Y = c_k\big), j = 1, 2, \cdots, n, k = 1, 2, \cdots, m$；按公式 (1) 预测类标。求解方法：MAP 或者最大化似然。

1.1 推导

推导 取 I 为示性函数。a_l 表示 X 的第 l 个特征。样本有 n 个。类标 m 个，特征 s 个。

$$P\big(Y = c_k\big) = \frac{\sum_{i=1}^{n} I\big(y_i = c_k\big)}{n}, k = 1, 2, \cdots, m \tag{2}$$

$$P\big(X_j = a_{jl} | Y = c_k\big) = \frac{\sum_{i=1}^{n} I\big(x_i^j = a_{jl}, y_i = c_k\big)}{\sum_{i=1}^{n} I\big(y_i = c_k\big)} \tag{3}$$

其中，$i = 1, 2, \cdots, n$,$l = 1, 2, \cdots, s$,$k = 1, 2, \cdots, m$

改进 为了避免分母为 0，进行了拉普拉斯平滑，即在分母上加了类数目。

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

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

   对于每一个类 j，更新 cluster centroids：

$$\mu_j := \frac{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\left\{c^{(i)} = j\right\}} \tag{5}$$

   }

3 Implementation

聚类测试：数据在 data.csv

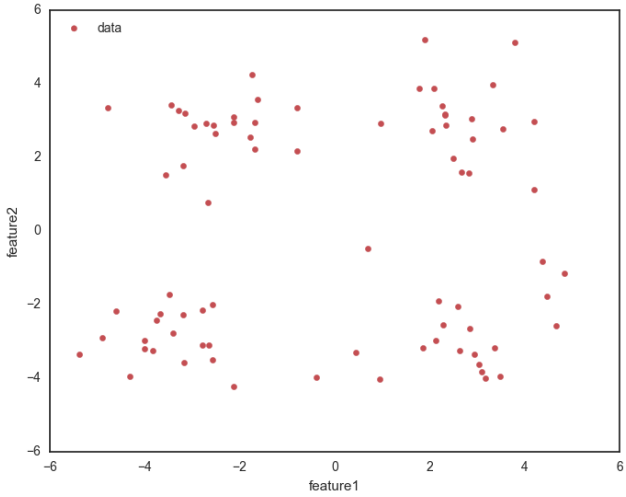

图 1: 训练数据

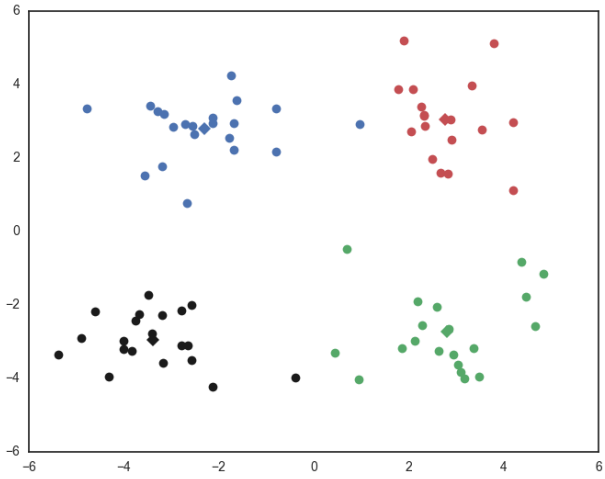

图 2: kmeans 运行结果，iter=1，k=4。菱形标记聚类中心，点标记数据

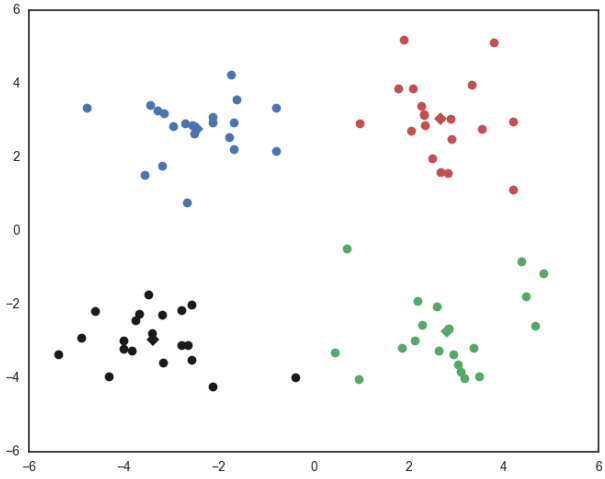

图 3: kmeans 运行结果，iter=2，k=4。菱形标记聚类中心，点标记数据

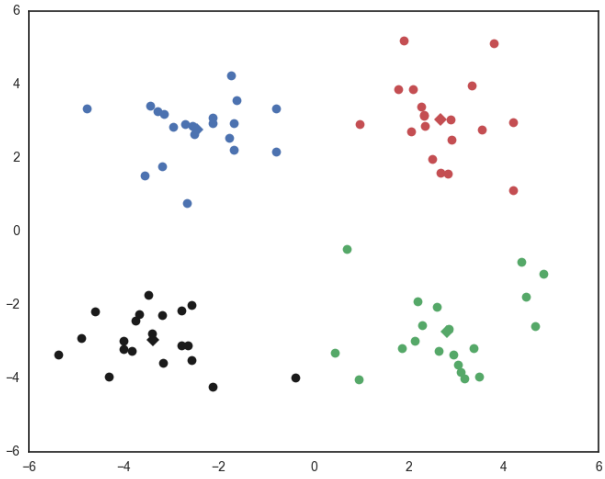

图 4: kmeans 运行结果，iter=3，k=4。菱形标记聚类中心，点标记数据