

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 [2]。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。

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

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 $P(Y = c_k), k = 1, 2, \dots, m$; 分别计算各类别下对应特征出现的概率 $P(X = x_j | Y = c_k), j = 1, 2, \dots, n, k = 1, 2, \dots, m$; 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。

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

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

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

2 算法实现

见 CS229[1]

1. 随机初始化 cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$

2. 迭代直至收敛 {

对于每一个样例 i , 计算类标

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

对于每一个类 j , 更新 cluster centroids:

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

}

3 Implementation

聚类测试: 数据在 data.csv

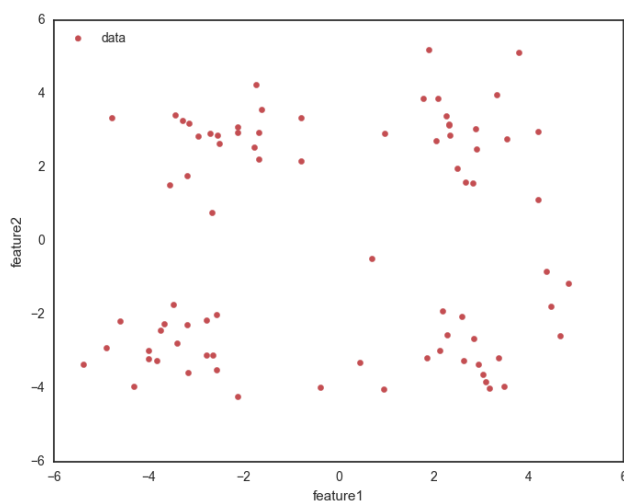


图 1: 训练数据

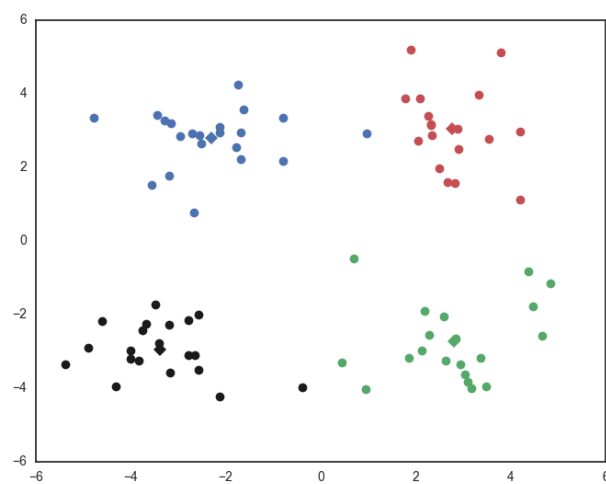


图 2: kmeans 运行结果, $\text{iter}=1$, $k=4$ 。菱形标记聚类中心, 点标记数据

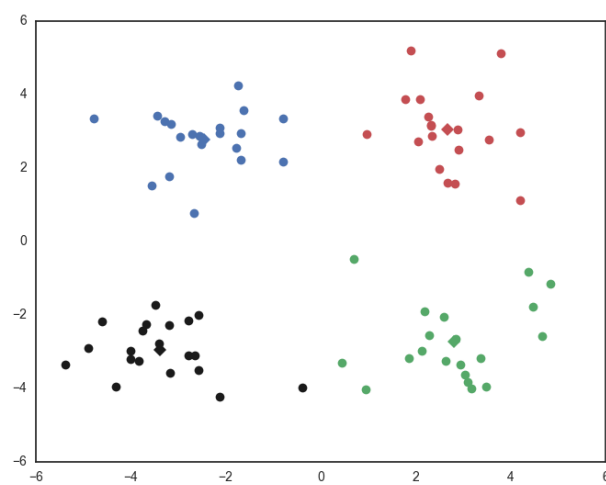


图 3: kmeans 运行结果, $\text{iter}=2$, $k=4$ 。菱形标记聚类中心, 点标记数据

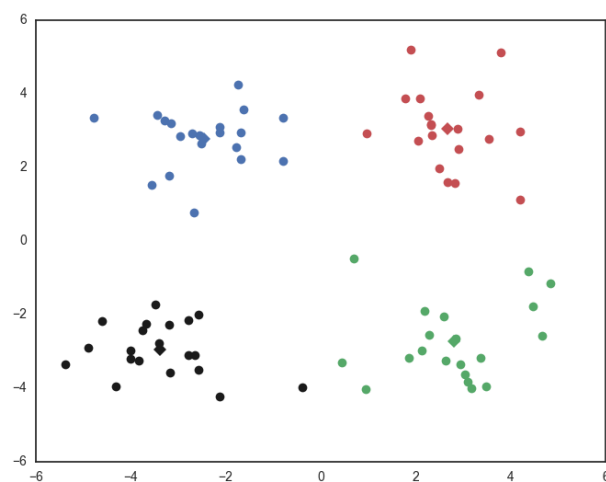


图 4: kmeans 运行结果, iter=3, $k=4$ 。菱形标记聚类中心, 点标记数据

参考文献

- [1] Andrew Ng. cs229-notes2. <http://cs229.stanford.edu/notes/cs229-notes2.pdf>, 2012. Stanford-CS229, 2012.
- [2] 李航. 统计学习方法. 清华大学出版社, 2012.