西南财经大学
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS
1925 - 2021
经世济民 孜孜以求

机器学习基础

朴素贝叶斯法

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大似然估计
- 贝叶斯决策
- 朴素贝叶斯算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大似然估计
- 贝叶斯决策
- 朴素贝叶斯算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

似然性和概率

- 似然性(likelihood)和概率都可以表示时间发生可能性的大小，但二者有很大区别
  - 似然性$L(\theta|x)$是从观测结果x出发，估计分布函数的参数θ的可能性大小
  - 概率$p(x|\theta)$是在已知参数θ的情况下，发生观测结果x可能性的大小

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

似然函数

- 定义似然函数为: $L(\theta|x) = p(x|\theta)$
  - 在 θ 已知, x 为变量的情况下, $p(x|\theta)$ 为概率, 表示通过已知的分布函数与参数, 随机生成出 x 的概率
  - 在 θ 为变量, x 已知的情况下, $p(x|\theta)$ 为似然函数, 它表示对于不同 θ, 出现 x 的概率是多少
    - 若对于两个参数 θ_1, θ_2, 有 $L(\theta_1|x) = p(x|\theta_1) > p(x|\theta_2) = L(\theta_2|x)$ 那么意味着 $\theta = \theta_1$ 生成的概率大于 $\theta = \theta_2$。因此, 观测数据为 x 时, θ_1 比 θ_2 更有可能为分布函数的参数

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

最大似然估计

- 对于给定的观测数据x，希望能从所有的可能参数$\theta_1, \theta_2, \cdots$中找出能最大概率生成观测数据的参数$\theta^*$作为估计结果，即

$$L(\theta^*|x) = p(x|\theta^*) \geq p(x|\theta) = L(\theta|x),\ \theta = \theta_1, \theta_2,$$

- 在求参数运算中，将θ看成是变量，计算以下优化问题

$$\theta^* = arg\max_{\theta} p(x|\theta)$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

- 通常一枚硬币是均匀的，正反面的概率都是0.5，假设H为投出正面，F为投出反面，p_H为投出正面的概率
  - 假设此时硬币正常，已知$p_H=0.5$。求三次试验结果为HHT的概率
  - 假设此时硬币不知道正常与否，p_H未知，但是已知经过三次试验结果为HHT，试估算出p_H

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

离散随机变量的最大似然估计

- 设 $X_1, X_2, \cdots, X_n$ 独立同分布，$x_1, x_2, \cdots, x_n$ 为相应的观测值，θ 为待估参数
- 分布函数随机取到 $x_1, x_2, \cdots, x_n$ 的概率为 $p(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$
- 构造似然函数 $L(\theta|x) = p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$
- 或者是对数似然函数

$$\ln L(\theta|x) = \sum_{i=1}^{n} \ln p(x_i|\theta)$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子

- 设 $X \sim b(1,p)$, $X_1, X_2, \cdots, X_n$ 是来自 X 的一个样本, $x_1, x_2, \cdots, x_n$ 为观测值。试求参数 p 的最大似然估计。

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大似然估计
- 贝叶斯决策
- 朴素贝叶斯算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

先验概率与后验概率

- 先验概率是在观测前就已经知道的概率分布
  - 根据以往的经验和分析
- 后验概率是某件事发生后对结果的估计
  - 后验是对先验“更加细致的刻画和更新”
- 后验比先验更有意义
  - 引入额外的观测信息，所以预测的准确度得到了加强

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

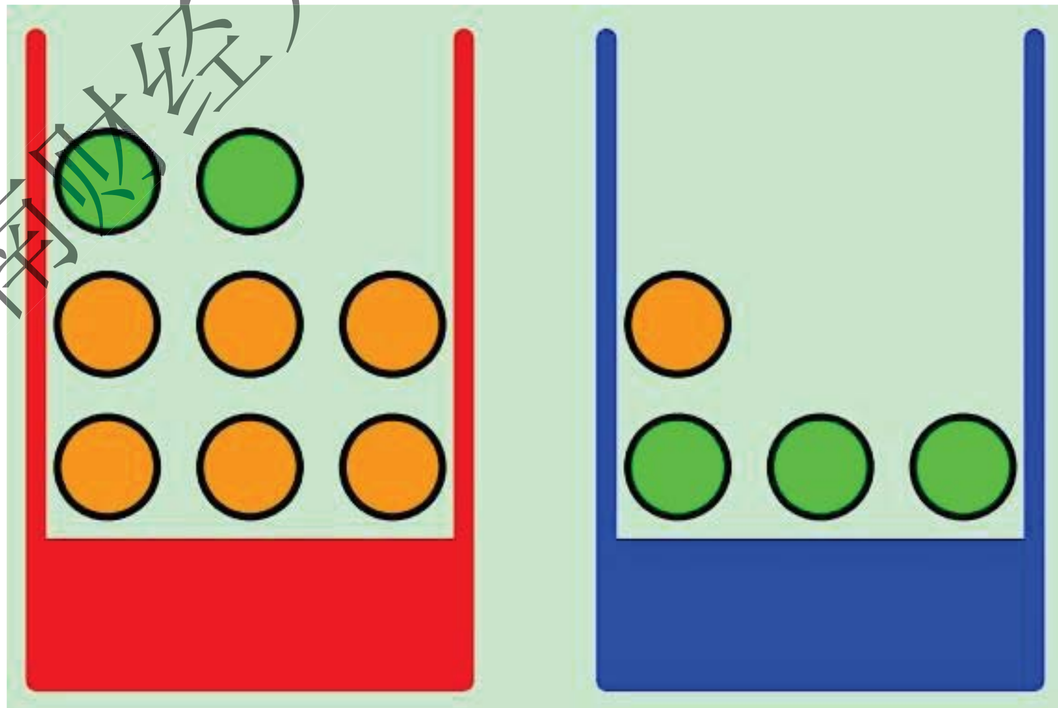
例子

- 先验概率 $P(Box=blue)=0.6$

  $P(Box=red)=0.4$
- 已知随机从一个盒子中挑出了橙色小球
- 后验概率

$$P(Box=blue|Ball=o)=\frac{1}{3}$$

$$P(Box=red|Ball=o)=\frac{2}{3}$$



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

基本思想

- 通过训练数据集学习联合概率分布 $P(X,Y)$
  - 先验概率分布 $P(Y=c_k)$， $k=1,2,\cdots,K$
  - 条件概率分布

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k),\quad k=1,2,\cdots,K$$

- 对给定输入 x， 利用贝叶斯公式计算后验概率

$$P(Y=c_k|X=x)=\frac{P(X=x,Y=c_k)}{P(X=x)}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

贝叶斯决策

- 类别决策为后验概率最大的一类

$$y = f(x) = arg\max_{c_k} P(Y = c_k | X = x)$$

- 等价形式（二分类问题）
  - 后验概率 $P(c_i|x) = \max_{j=1,2} P(c_j|x)$
  - 贝叶斯公式 $P(c_i|x) = \max_{j=1,2} P(x|c_j)P(c_j)$
  - 似然比 $$l(x) = \frac{P(x|c_1)}{P(x|c_2)} > \frac{P(c_2)}{P(c_1)}$$
  - 对数似然比 $$h(x) = -\ln(l(x)) > \ln(P(c_1)) - \ln(P(c_2))$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大似然估计
- 贝叶斯决策
- 朴素贝叶斯算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

条件独立性假设

- 所有先验概率和条件概率由极大似然估计得到
  - 当样本量足够多时，概率近似等于频率
- 条件概率 $P(X=x|Y=c_k)$ 需要的样本数随特征数呈指数增加
  - 维度灾难：多数条件概率不会有任何样本发生
- 条件独立性假设：分类的特征在类确定的条件下都是条件独立

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

条件独立性假设

- 条件独立性假设（朴素贝叶斯）

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$$

$$=\prod_{i=1}^{n}P(X^{(i)}=x^{(i)}|Y=c_k)$$

- 能避免条件概率为0，但会牺牲分类准确率

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

贝叶斯公式

- 后验概率的计算根据贝叶斯公式

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{P(X=x)}$$

$$=\frac{P(Y=c_k)\prod_{i=1}^{n}P(X^{(i)}=x^{(i)}|Y=c_k)}{P(X=x)},\quad k=1,2,\cdots,K$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

朴素贝叶斯分类器

- 根据后验概率最大为依据，朴素贝叶斯分类器可表示为

$$y = f(x) = arg\max_{c_k} \frac{P(Y = c_k)\prod_{i=1}^{n} P(X^{(i)} = x^{(i)} | Y = c_k)}{P(X = x)}$$

- 注意到对不同的 c_k，分母 $P(X = x)$ 都相同，所以

$$\boxed{y = f(x) = arg\max_{c_k} P(Y = c_k)\prod_{i} P(X^{(i)} = x^{(i)} | Y = c_k)}$$

- 后验概率最大等价于期望风险最小

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

参数估计

- 频率替代概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N},\quad k=1,2,\cdots,K$$

$$P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

贝叶斯估计

- 极大似然估计可能会出现所要估计概率为0
- 设 $\lambda > 0$：

$$P_\lambda(X^{(j)} = a_{jl} | Y = c_k) = \frac{\sum_{i=1}^{N} I(x_i^{(j)} = a_{jl}, y_i = c_k) + \lambda}{\sum_{i=1}^{N} I(y_i = c_k) + S_j \lambda} > 0$$

- 当 $\lambda = 0$ 时称为拉普拉斯平滑

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大似然估计
- 贝叶斯决策
- 朴素贝叶斯算法
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

总结

- 利用贝叶斯公式计算后验概率
  - 后验概率最大等价于期望风险最小
- 朴素贝叶斯：条件独立性假设
- 参数估计：极大似然估计
- 拉普拉斯平滑保证条件概率大于0

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法优缺点

- 优点
  - 算法逻辑简单，结果解释性强，易于实现
  - 分类过程中时空开销小
  - 对缺失数据不敏感
- 缺点
  - 条件独立性假设难成立，特征之间相关性大，分类效果不好
  - 难以获得先验概率等计算要件，需要通过统计方法进行估计，存在一定误差