Intro. to Machine Learning

Chpt 6. Bayes Classification

Lecturer: MaoQiang Xie
College of Software, Nankai University

Outline

1. 贝叶斯公式
2. 朴素贝叶斯分类器 (Naïve Bayes Classifier)
3. 贝叶斯信念网 (Bayes Belief Network)
4. 贝叶斯与语言模型(Language Model)
5. 极大似然估计(Maximum Likelihood Estimation, MLE)
6. EM（期望最大化）算法

基于概率模型的分类器（回顾）

- 定义
  - X：图像集合$\{x_i\}$=

    {  }

  - Y:识别结果集合$\{c_i\}$={0，1，9，M，z}
- 任务:估计以下的概率
  - P(y=0 | x= ),
  - P(y=1 | x= ),
  - P(y=9 | x= ),
  - P(y=M | x= ),
  - P(y=z | x= )

使用最大后验概率作为分类的判别依据

$$h^*(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c\,|\,\boldsymbol{x})$$

1. 贝叶斯公式（1763年）

- 由条件概率公式 $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- 可得贝叶斯公式 $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$

- 全概率公式： $P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$

从预测角度看估计后验概率 $P(c \mid \boldsymbol{x})$

$$P(c \mid \boldsymbol{x}) = \frac{P(c, \boldsymbol{x})}{P(\boldsymbol{x})} = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}$$

$P(\boldsymbol{x})$：样本 $\boldsymbol{x}$ 在样本空间中出现的概率

$P(\boldsymbol{x}|c)$：代表样本x相对于类别 c 的类条件概率（class-conditional probability），或称为似然（likelyhood）

$P(c)$ ：$\mathcal{Y}$ 中各 $c \in \mathcal{Y}$ 的先验概率

2. 朴素贝叶斯分类器(Naïve Bayes Classifier)

- 在估计$P(\boldsymbol{x}|c)$时很难估计$\boldsymbol{x}$所有维联合发生的概率，原因在于很难用频率估计概率，因此，通过假设“各维度条件独立”，将联合概率变成各维度概率连乘。

$$P(c \mid \boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

朴素贝叶斯分类器的判决函数

- 由于对于给定 $\boldsymbol{x}$，其 $P(\boldsymbol{x})$ 对于所有 $c \in \mathcal{Y}$ 是一样的。
- 朴素贝叶斯的判决函数简化为：

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

类先验概率、类条件概率的估计

$P(c)$：使用样本空间中各类样本所占比重来估计。根据大数定理，当训练集包含充足的独立同分布样本时，可使用高频率来估计概率。

$$P(\boldsymbol{x}|c) = P(x_1, x_2, \cdots, x_d|c) = \prod_{i=1}^{d} P(x_i|c)$$

直接使用样本出现频率来估计对关于 所有特征的联合概率将会遇到严重困难——很多种取值未在训练集中出现。（“未被观测到” v.s. “出现概率为0”）

假设样本 d 维特征都是二值的。联合随机事件数 2^d，独立的事件数 $2d$

判决函数中的各概率的估计

$$h_{nb}(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\max}\, P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

- 类先验概率：

$$P(c) = \frac{|D_c|}{|D|}$$

- 类条件概率：

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

(离散属性)

判决函数中的各概率的估计

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} P(x_i \mid c)$$

- 类先验概率： $$P(c) = \frac{|D_c|}{|D|}$$
- 类条件概率： $$P(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}}\exp\left(-\frac{(x_i-\mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$
  (连续属性)
- 第c类样本在第i个属性上取值的 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$

参数化方法估计类条件概率

– 形式上简单

– 估计的准确性严重依赖于“假设是否成立”

– 实际应用中需要融合任务本身的经验知识，用先验知识引导类条件概率的分布假设。

估计$P(\boldsymbol{x}|c)$ 和$P(c)$时碰到的问题

- 训练样本不充足，导致概率估计为零。
- 可以进行“平滑”smoothing。比如拉普拉斯修正（Laplacian Correction）

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} ,$$

N: 数据集中的类别数

$$\hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i} .$$

N_i: 第 i 维特征的可能取值数

P153

总结：朴素Bayes分类器的训练算法

1. 数据预处理（特征的离散化）
2. 估计每个类别的类先验概率 $P(c)$
3. 估计每个类别条件下各特征下每种取值出现的概率 $P(\boldsymbol{x}_i|c)$

总结：朴素Bayes分类器的预测算法

1. 数据预处理(按照训练的预处理方法进行)
2. 利用预计算好的$P(c)$和$P(\boldsymbol{x}_i|c)$，根据贝叶斯公式为每个类别 c 计算

$$P(c \mid \boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

3. 选择最大后验概率对应的类别作为结果输出。

从最小化风险的角度来看NB

训练Naïve Bayes其实也是寻找判别模型 $h: \mathcal{X} \mapsto \mathcal{Y}$ 以最小化总体风险

$$R(h) = \mathbb{E}_{\boldsymbol{x}}[R(h(\boldsymbol{x})|\boldsymbol{x})]$$

基于 $P(c_i|\boldsymbol{x})$ ，可获得奖样本 $\boldsymbol{x}$ 分类为 c_i 所产生的条件风险（或损失）

$$R(c_i|\boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\boldsymbol{x})$$

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{otherwise}. \end{cases}$$

贝叶斯最优分类器与贝叶斯风险

Bayes Decision Rule：为最小化总体风险，只需在每个样本上选择使条件风险 $R(c|\boldsymbol{x})$ 最小的类别标记，即

$$h^*(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\min}\, R(c\,|\boldsymbol{x})$$

*h**：Bayes Optimal Classifier，对应的总体风险 $R(h^*)$ 被称为 Bayes Risk。 $1 - R(h^*)$ 是通过机器学习所能产生的模型精度的理论上限。

插播：贝叶斯分类器是生成式模型

- 生成式模型（Generative Models）
  - 需要对生成样本的分布进行建模(如下述3个分布)，然后估计 $P(c\,|\boldsymbol{x})$

$$P(\boldsymbol{x}, c) \quad P(\boldsymbol{x}|c) \quad P(c)$$

- 判决式模型（Discriminative Models）
  - 直接建模 $P(y=c|\boldsymbol{x})$（即直接求解判决模型(参数)）

Outline

1. 贝叶斯理论
2. 朴素贝叶斯分类器 (Naïve Bayes Classifier)
3. 贝叶斯信念网 (Bayes Belief Network)
4. 贝叶斯与信息检索语言模型(Language Model)
5. 极大似然估计(Maximum Likelihood Estimation, MLE)
6. EM（期望最大化）算法

3. Bayes Belief Network

- 使用条件概率表（Conditional Probability Table, CPT）来描述属性的联合概率分布。

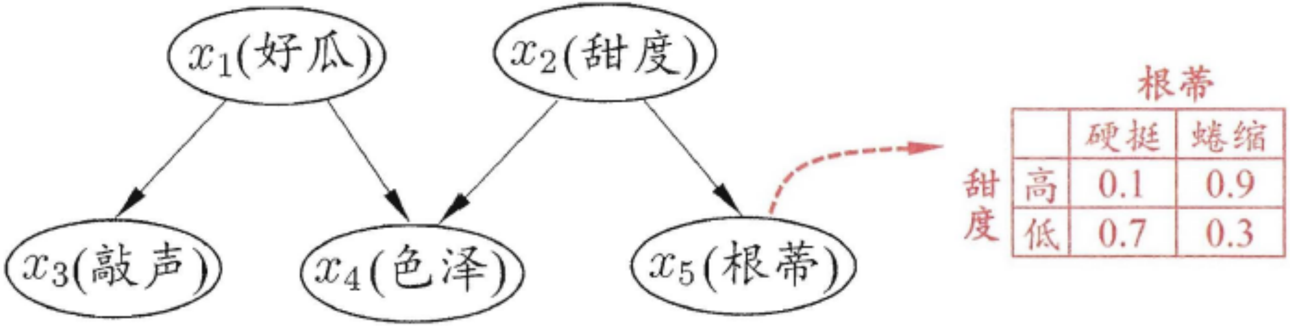


图 7.2 西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3|x_1)P(x_4|x_1, x_2)P(x_5|x_1)$$

周志华《机器学习》7.5贝叶斯网

贝叶斯信念网的特点

- 多维变量中联合概率估计时属性组合爆炸问题
- 样本稀疏导致概率估计值为0的问题
- 朴素贝叶斯分类器属性条件独立性假设不满足的问题
  - 只需各类条件概率排序正确、无需精准概率值即可
  - 若属性间依赖对所有类别影响相同，或依赖关系的影响能够相互抵消，则假设不符合对性能影响不大
  - 在信息检索领域中，应用效果很好

Outline

1. 贝叶斯理论
2. 朴素贝叶斯分类器 (Naïve Bayes Classifier)
3. 贝叶斯信念网 (Bayes Belief Network)
4. 贝叶斯与信息检索语言模型(Language Model)
5. 极大似然估计(Maximum Likelihood Estimation, MLE)
6. EM（期望最大化）算法

4. 语言模型(Language Model)

- 语言模型的核心是估计句子(Sentence)被生成的概率

$$
\begin{aligned}
P(S) &= P(w_1, w_2, w_3, \cdots, w_k) \\
&= P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)\cdots P(w_k|w_{k-1}, w_{k-2}\cdots w_1)
\end{aligned}
$$

其中，句子S由单词w_i组合而成。

- 条件概率估计对样本量的需求巨大，例如，在估计下述概率得0可能性很大

$$P(w_k|w_{k-1}, w_{k-2}\cdots w_1)$$

N-gram for Language Model

Markov 假设:下一个词出现的概率仅依赖前面一个或几个词

- 1-gram（Unigram,不依赖前面的词）

$$P(S) = P(w_1)P(w_2)P(w_3)\cdots P(w_k)$$

- 2-gram（Bigram,依赖前面一个词）

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|w_2)\cdots P(w_k|w_{k-1})$$

- 3-gram（Trigram,依赖前面两个词）

$$P(S) = P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)\cdots P(w_k|w_{k-1}, w_{k-2})$$

文本情感分类（基于贝叶斯公式和语言模型）

- 该问题可建模为二分类问题

  1：正向情感、0：负向情感

$$Sentiment = \frac{P(Y=1|S)}{P(Y=0|S)} = \frac{P(S|Y=1)P(Y=1)}{P(S|Y=0)P(Y=0)}$$

- 使用语言模型N-gram来估计 $P(Y=1|S)$

5. 参数估计：Frequentist v.s. Bayesian

- 频率主义学派(Frequentist)
  - 认为参数未知，但客观存在一个定值
  - 通过极大似然估计（Maximum Likelihood Estimation）确定概率模型参数

- 贝叶斯学派(Bayesian)
  - 参数是未观测到的随机变量，本身可能也有分布。
  - 可假设参数服从一个分布，然后使用数据估计参数的后验分布。

两派区别：参看教材的参考文献[Efron, 2005]和[Samaniego, 2010]

5. 极大似然估计

- 概率分布的参数估计(parameter estimation)
  - 先假定某随机事件的概率符合某种确定的分布形式，可基于抽样数据（训练数据）对概率分布的参数进行估计。
  - 从机器学习的角度可以将参数估计看成模型训练
- 对于NB所需的 $P(\boldsymbol{x}|c)$，可假设其形式，且被参数 $\boldsymbol{\theta}_c$ 唯一确定，那么使用训练集训练概率模型的任务，其实就是数理统计中的估计参数 $\boldsymbol{\theta}_c$

极大似然估计

- 以估计类条件概率分布为例

令 D_c 表示训练集 D 中第 c 类样本组成的集合，假设这些样本是独立同分布的，则参数 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是

$$P(D_c \mid \boldsymbol{\theta}_c) = \prod_{\boldsymbol{x} \in D_c} P(\boldsymbol{x} \mid \boldsymbol{\theta}_c) \ . \tag{7.9}$$

对 $\boldsymbol{\theta}_c$ 进行极大似然估计，就是去寻找能最大化似然 $P(D_c \mid \boldsymbol{\theta}_c)$ 的参数值 $\hat{\boldsymbol{\theta}}_c$. 直观上看，极大似然估计是试图在 $\boldsymbol{\theta}_c$ 所有可能的取值中，找到一个能使数据出现的"可能性"最大的值.

Log-likelihood: 对似然进行对数化

式(7.9)中的连乘操作易造成下溢, 通常使用对数似然(log-likelihood)

$$\begin{aligned} LL(\boldsymbol{\theta}_c) &= \log\ P(D_c \mid \boldsymbol{\theta}_c) \\ &= \sum_{\boldsymbol{x} \in D_c} \log P(\boldsymbol{x} \mid \boldsymbol{\theta}_c)\ , \end{aligned} \tag{7.10}$$

此时参数 $\boldsymbol{\theta}_c$ 的极大似然估计 $\hat{\boldsymbol{\theta}}_c$ 为

$$\hat{\boldsymbol{\theta}}_c = \underset{\boldsymbol{\theta}_c}{\arg\max}\ LL(\boldsymbol{\theta}_c)\ . \tag{7.11}$$

极大似然估计的特点

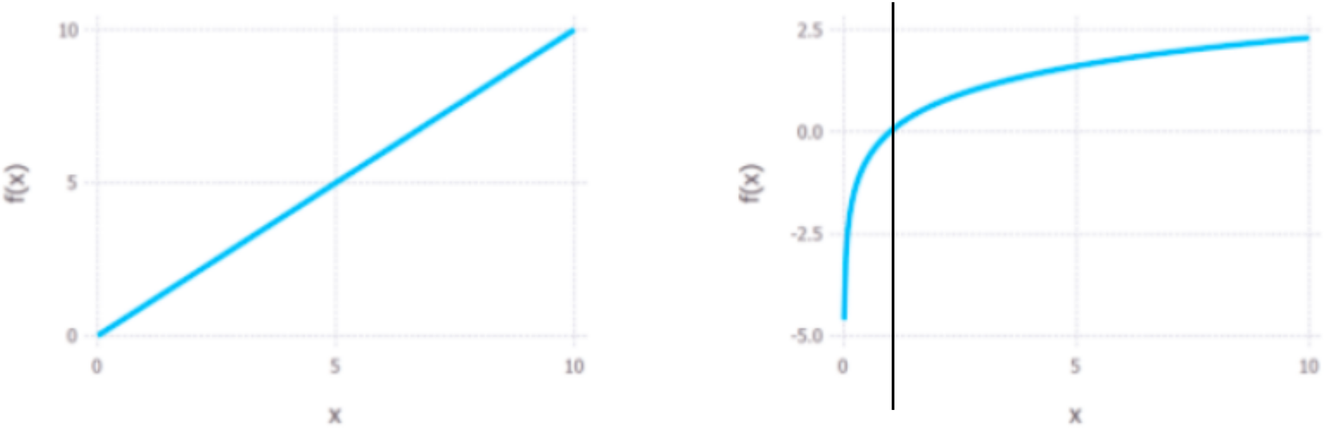
- 每一个样本均是由同一个分布相互独立生成。
  - 生成整个数据集的联合概率是生成每个样本概率的乘积
- 使用对数似然可以将连乘化为连加

$$p(X \mid \Theta) = \prod_{i=1}^{N} p(x_i \mid \Theta) \Rightarrow \ln p(X \mid \Theta) = \sum_{i=1}^{N} \ln p(x_i \mid \Theta)$$

- 遏制概率连乘导致的浮点数下溢 $x < e^{-1}$ 时 $|\ln x| > 1$
- 概率密度中如有指数项，使用对数似然有利计算

对数似然不影响单调，最大似然处参数相同

$$p(x \mid \Theta_1) > p(x \mid \Theta_2) \Leftrightarrow \ln p(x \mid \Theta_1) > \ln p(x \mid \Theta_2)$$



(a) $f(x) = x$

(b) $f(x) = \ln(x)$

最大似然估计的其他应用

- 估计正态分布的参数：$\mu \quad \sigma$
- 教材“3.3 Logistic回归”中，使用极大似然法估计代价函数的参数
  - P59
  - 对比本课1.2节“基于交叉熵的代价函数”。

6. EM算法

- 最初解决数据缺失条件下的参数估计问题
- 思路
  - 用现有数据估计分布（模型参数）
  - 用估计的分布（模型参数）估计缺失数据
  - 整合原有数据和补齐数据重新估计分布
  - 反复迭代直至收敛
- 目前：认为分布未被现有观测数据完全体现，存在隐含变量，因此，在分布参数估计时要考虑隐含变量的存在。

Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the royal statistical society. Series B (methodological), 1977: 1-38.

6. EM算法

- 分布生成的未被观测的随机事件用Z表示

$$P(X|\theta) = P(X, Z|\theta) = \sum_{Z} P(X|Z, \theta)P(Z|\theta)$$

- 使用极大似然法估计模型参数

$$\hat{\theta} = \arg\max_{\theta} \ln P(X, Z|\theta)$$

6. EM算法

交替执行E步和M步，直至收敛到局部最优解

- E步(Expectation)：
  - 基于Θ^t推断因变量Z的期望，记为Z^t

    $$Q(\Theta|\Theta^t) = \mathbb{E}_{Z|X,\Theta^t} LL(\Theta|X, Z)$$

- M步(Maximization):
  - 寻找参数最大化期望似然

    $$\Theta^{t+1} = \arg\max_{\Theta} Q(\Theta|\Theta^t)$$

EM算法用途

- EM算法多用于高斯混合模型的参数估计
  – 隐变量是样本属于混合模型中的拿个高斯分布
- 隐变量估计
  – 在极大似然估计因隐变量存在失效时，可以考虑用EM算法

总结

- 贝叶斯理论
- 朴素贝叶斯分类器 (Naïve Bayes Classifier)
- 极大似然估计(Maximum Likelihood Estimation, MLE)
- 贝叶斯信念网 (Bayes Belief Network)
- 贝叶斯与信息检索语言模型(Language Model)
- EM（期望最大化）算法