

判别模型与生成模型

Discriminative Model and Generative Model

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

机器学习简介

- ▶ **机器学习** (Machine Learning) 是一门研究如何通过计算的手段, 利用经验来改善系统自身性能的学科¹
- ▶ 机器学习任务通常可以分为监督学习 (Supervised learning)、无监督学习 (Unsupervised learning)、强化学习 (Reinforcement learning)

¹本课程讲授机器学习的一些高级主题, 而关于机器学习的基础知识, 请感兴趣的同学选修我校“大数据管理与应用”专业本科生的“智能决策与机器学习”课程, 或者在线学习B站上吴恩达老师在斯坦福大学开设的课程CS229 [“Machine Learning”](#)

机器学习简介

- ▶ **监督学习** (Supervised learning) 是指从标注数据中学习预测模型的机器学习问题，常见的任务包括分类 (Classification)、回归 (Regression)
 - ▶ $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$, 其中 $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$ 表示样本的多维特征 (属性), y_i 是标签 (y_i 取离散值时为分类问题, y_i 取连续值时为回归问题)
 - ▶ 实例: 信用违约预测、销售量预测

机器学习简介

- ▶ **无监督学习** (Unsupervised learning) 是指从无标注数据中学习预测模型的机器学习问题，又被称为知识发现 (Knowledge discovery)，常见的任务包括聚类 (Clustering)、降维 (Dimensionality reduction)
 - ▶ $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ，其中 $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$ 表示样本的多维特征 (属性)
 - ▶ 实例：精准营销中的用户市场细分
 - ▶ 注：降维通常作为机器学习算法的前驱过程

机器学习简介

- ▶ **强化学习** (Reinforcement learning) 是指智能系统在环境的连续互动中学习最优行为策略的机器学习问题
 - ▶ (1) 无人驾驶、AlphaGo、波士顿动力机器人
 - ▶ (2) 收益管理
 - ▶ 赌博机模型/多臂老虎机 (Multi-armed bandit model)
 - ▶ Bernstein, F., Modaresi, S., & Sauré, D. (2019). A dynamic clustering approach to data-driven assortment personalization. *Management Science*, 65(5), 2095-2115.
 - ▶ Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4), 500-522.
 - ▶ Russo, D., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2017). A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*.
 - ▶ (3) 组合优化
 - ▶ Bengio, Y., Lodi, A., & Prouvost, A. (2021). Machine learning for combinatorial optimization: a methodological tour d' horizon. *European Journal of Operational Research*, 290(2), 405-421.

分类问题探析

- ▶ **分类问题：**已获得一批已标注的训练数据 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ ，其中 $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)})$ 表示样本的多维特征， $y_i \in \mathcal{Y} = \{c_1, \dots, c_K\}$ 是标签， $\mathcal{Y} = \{c_1, \dots, c_K\}$ 是类集合， K 是类别数
- ▶ **目标：**通过训练数据学习得到一个分类器 (classifier) $f(\cdot)$ ，从而对新的样本 \mathbf{x} 预测其类别 $y = f(\mathbf{x}) \in \mathcal{Y}$
- ▶ **问题：**如何评价分类器 $f(\cdot)$ 的性能呢？
- ▶ 假设所有样本都来自于一个总体，其分布为 $p(X, Y)$
- ▶ 引入以下0-1损失函数：假设分类损失相同(Homogeneous classes)

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

- ▶ 期望损失函数，又称风险函数²

$$\begin{aligned} R_{\text{exp}}(f) &= E_{X, Y} [L(Y, f(X))] \\ &= \int \int p(X, Y) \cdot L(Y, f(X)) dX dY \end{aligned}$$

使得上述期望损失最小化的分类器 f^* 被称为贝叶斯最优分类器 (Bayesian optimal classifier)，与之对应的风险 $R_{\text{exp}}(f^*)$ 被称为贝叶斯风险 (Bayesian risk)，而 $1 - R_{\text{exp}}(f^*)$ 反映了分类器所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限

²决策论中将“期望损失”称为“风险” (risk)

分类问题探析

- ▶ 在基于概率的分类学习算法中, 分类器 $f(\cdot)$ 通常被实现为预测样本 \mathbf{x} 分到各个类 c_k 的概率 $p(c_k | \mathbf{x})$, $k = 1, \dots, K$
 - ▶ 该概率又被称为后验概率, 即观测到样本 \mathbf{x} 后判定其属于某个类别的概率
- ▶ 问题: (1) 为什么要根据后验概率 $p(c_k | \mathbf{x})$ 判定分类结果呢? (2) 如果获得了样本 \mathbf{x} 分到各个类 c_k 的概率 $p(c_k | \mathbf{x})$, 应当如何确定其分类结果呢?

$$\begin{aligned} R_{\text{exp}}(f) &= E_{X,Y}[L(Y, f(X))] \\ &= \int \int p(X, Y) L(Y, f(X)) dX dY \\ &= \int \int p(X) p(Y | X) L(Y, f(X)) dX dY \\ &= \int p(X) \left[\int p(Y | X) L(Y, f(X)) dY \right] dX \end{aligned}$$

为了使得期望损失最小, 只需对每个样本 $X = \mathbf{x}$ 逐个最小化, 由此得到

$$f^*(\mathbf{x}) = \arg \min_{f(\mathbf{x}) \in \mathcal{Y}} \int p(y | \mathbf{x}) L(y, f(\mathbf{x})) dy$$

分类问题探析

- 在分类问题中, 标签 y 是离散变量, $y \in \mathcal{Y} = \{c_1, \dots, c_K\}$

$$\begin{aligned} f^*(x) &= \arg \min_{f(x) \in \mathcal{Y}} \sum_{k=1}^K p(c_k | x) L(c_k, f(x)) \\ &= \arg \min_{f(x) \in \mathcal{Y}} \sum_{k: f(x)=c_k} p(c_k | x) L(c_k, f(x)) + \sum_{k: f(x) \neq c_k} p(c_k | x) L(c_k, f(x)) \\ &= \arg \min_{f(x) \in \mathcal{Y}} \sum_{k: f(x)=c_k} p(c_k | x) \cdot 0 + \sum_{k: f(x) \neq c_k} p(c_k | x) \cdot 1 \\ &= \arg \min_{f(x) \in \mathcal{Y}} \sum_{k: f(x) \neq c_k} p(c_k | x) \\ &= \arg \min_{f(x)=c_k \in \mathcal{Y}} 1 - p(c_k | x) \quad \text{仅有一个 } c_k \in \mathcal{Y} \text{ 与 } f(x) \text{ 相等} \\ &= \arg \max_{f(x)=c_k \in \mathcal{Y}} p(c_k | x) \end{aligned}$$

上式说明, 在分类损失相同的前提下(Homogeneous classes), 贝叶斯最优分类器对应着最大化样本的后验概率, 这就是基于概率的分类学习算法之所以通过后验概率最大化准则判定分类结果的原因

判别模型与生成模型

- ▶ 经过前面的分析，我们知道应该根据后验概率 $p(c_k | x)$ 判定样本 x 的分类结果，因此基于概率的分类学习算法的核心在于如何获得后验概率 $p(c_k | x)$
- ▶ 通常有两种方式，分别是**判别式方法**（discriminative approach）和**生成式方法**（generative approach），所学到的模型分别被称为**判别模型**（discriminative model）和**生成模型**（generative model）
 - ▶ 这是机器学习方法的另一种分类方式
 - ▶ 注：不仅仅适用于分类任务，所有的机器学习方法都适用

判别模型与生成模型

- ▶ **判别式方法**: 直接对后验概率 $p(c_k | \mathbf{x})$ 建模
 - ▶ 代表: 逻辑斯谛回归 (Logistic regression)
- ▶ **生成式方法**: 应用贝叶斯公式, 由数据学习联合概率分布 $p(\mathbf{x}, c_k)$, 然后求出后验概率 $p(c_k | \mathbf{x})$

$$p(c_k | \mathbf{x}) = \frac{p(\mathbf{x}, c_k)}{p(\mathbf{x})}$$

- ▶ 代表: 朴素贝叶斯分类器 (Naive Bayes classifier)

逻辑斯谛回归 (Logistic regression)

- ▶ 考虑二分类任务，其输出标签记为 $y \in \{0, 1\}$
- ▶ 通过Sigmoid函数将线性模型 $\mathbf{w}^\top \mathbf{x} + b$ 与类别 y 联系起来，得到后验概率

$$p(y = 1 \mid \mathbf{x}) = \frac{\exp(\mathbf{w}^\top \mathbf{x} + b)}{1 + \exp(\mathbf{w}^\top \mathbf{x} + b)}$$

$$p(y = 0 \mid \mathbf{x}) = \frac{1}{1 + \exp(\mathbf{w}^\top \mathbf{x} + b)}$$

- ▶ 利用极大似然估计法学习模型的参数 \mathbf{w} 和 b

$$\begin{aligned} \max L &= \prod_{i=1}^N p(y = 1 \mid \mathbf{x})^{y_i} \cdot p(y = 0 \mid \mathbf{x})^{1-y_i} \\ &= \prod_{i=1}^N \left[\frac{\exp(\mathbf{w}^\top \mathbf{x} + b)}{1 + \exp(\mathbf{w}^\top \mathbf{x} + b)} \right]^{y_i} \cdot \left[\frac{1}{1 + \exp(\mathbf{w}^\top \mathbf{x} + b)} \right]^{1-y_i} \end{aligned}$$

- ▶ 这是一个凸优化问题，经典的数值优化算法如梯度下降法、牛顿法都可以求解

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 已获得一批已标注的训练数据

$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$, 其中

$\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)})$ 表示样本的多维特征,

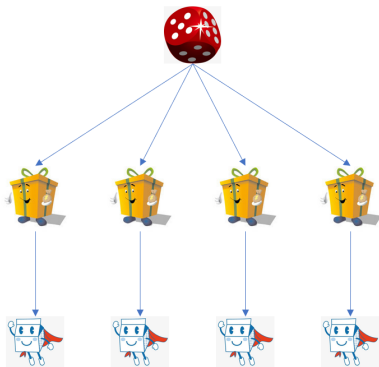
$y_i \in \mathcal{Y} = \{c_1, \dots, c_K\}$ 是标签, $\mathcal{Y} = \{c_1, \dots, c_K\}$ 是类集合,
 K 是类别数

$$\begin{aligned} p(c_k | \mathbf{x}) &= \frac{p(\mathbf{x}, c_k)}{p(\mathbf{x})} \\ &= \frac{p(c_k)p(\mathbf{x} | c_k)}{p(\mathbf{x})} \\ &= \frac{p(c_k)p(\mathbf{x} | c_k)}{\sum_{k'=1}^K p(c_{k'})p(\mathbf{x} | c_{k'})} \end{aligned}$$

- ▶ $p(c_k)$: 类 c_k 的先验概率(prior)
- ▶ $p(\mathbf{x} | c_k)$: 样本 \mathbf{x} 相对于类 c_k 的条件概率(class-conditional probability), 亦被称为似然(likelihood)
- ▶ $p(\mathbf{x})$: 用于归一化的“证据”(evidence)因子

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 如何理解朴素贝叶斯分类器是生成式方法？



- ▶ 生成式方法揭示了数据生成的过程，在朴素贝叶斯分类器中，先以概率 $p(c_k)$ 选择类 c_k ，再以类条件概率 $p(\mathbf{x} | c_k)$ 生成样本 \mathbf{x}

朴素贝叶斯分类器 (Naive Bayes classifier)

$$\begin{aligned} p(c_k | \mathbf{x}) &= \frac{p(\mathbf{x}, c_k)}{p(\mathbf{x})} \\ &= \frac{p(c_k)p(\mathbf{x} | c_k)}{p(\mathbf{x})} \\ &= \frac{p(c_k)p(\mathbf{x} | c_k)}{\sum_{k'=1}^K p(c_{k'})p(\mathbf{x} | c_{k'})} \end{aligned}$$

- ▶ 根据后验概率最大化原则，应当选择使得 $p(c_k | \mathbf{x})$ 最大的类 c_k 作为样本 \mathbf{x} 的分类结果， $k \in \{1, \dots, K\}$

$$p(c_k | \mathbf{x}) \propto p(c_k)p(\mathbf{x} | c_k)$$

符号 \propto 表示“正比于”

- ▶ 关键：如何获得 $p(c_k)$ 和 $p(\mathbf{x} | c_k)$?

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 类条件概率 $p(\mathbf{x} | c_k)$ 有指数级数量的参数，从数据中直接估计 $p(\mathbf{x} | c_k)$ 是不可行的
 - ▶ **维数灾难**(Curse of dimensionality): 在高维空间中，样本将变得非常稀疏，给机器学习方法带来了严重障碍
- ▶ 朴素贝叶斯分类器采用了**条件独立性**假设：对已知类别，假设所有属性相互独立
 - ▶ 换言之，假设每个属性独立地对分类结果发生影响
 - ▶ 虽然这是一个较强的假设，但在很多问题中取得了良好的分类效果

$$\begin{aligned} p(\mathbf{x} | c_k) &= p\left(\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(d)}) | c_k\right) \\ &= \prod_{j=1}^d p(x^{(j)} | c_k) \end{aligned}$$


- ▶ 因此，问题变成了如何获得 $p(c_k)$ 和 $p(x^{(j)} | c_k)$, $j = 1, \dots, d$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 如何获得 $p(c_k)$ 和 $p(x^{(j)} | c_k), j = 1, \dots, d$?
- ▶ 这是参数估计问题，可以采用极大似然估计或者贝叶斯估计

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 极大似然估计
- ▶ (1) 估计先验概率 $p(c_k)$
- ▶ 概率 $p(c_k)$ 表示从总体中随机选取一个类别，其结果是 c_k 的概率
 - ▶ 满足 $p(c_k) > 0, k = 1, \dots, K$ 且 $\sum_{k=1}^K p(c_k) = 1$
 - ▶ 这是一个多项分布(multinomial distribution)，相当于掷有一个有 K 个面的骰子，各个面向上的概率分别是 $p(c_k), k = 1, \dots, K$ ³

³多项分布的定义见李航《统计学习方法》(第2版) P385 

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ **多项分布：**若多元离散随机变量 $X = (X_1, X_2, \dots, X_k)$ 的概率质量函数为

$$\begin{aligned} P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) &= \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} \\ &= \frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i} \end{aligned}$$

其中 $p = (p_1, p_2, \dots, p_k)$, $p_i \geq 0, i = 1, 2, \dots, k$,
 $\sum_{i=1}^k p_i = 1, \sum_{i=1}^k n_i = n$, 则称随机变量 X 服从参数
为 (n, p) 的多项分布, 记作 $X \sim \text{Mult}(n, p)$

- ▶ 特别地, 当试验的次数 n 为 1 时, 多项分布变成类别分布 (categorical distribution) 类别分布表示试验可能出现的 k 种结果的概率; 当 k 为 2 时, 多项分布变为二项分布

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 从 N 个训练样本中估计先验概率 $p(c_k)$
 - ▶ $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$, 其中
$$\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)}), y_i \in \mathcal{Y} = \{c_1, \dots, c_K\}$$
- ▶ 极大似然估计

$$L = \frac{N!}{\prod_{k=1}^K N_k!} \prod_{k=1}^K [p(c_k)]^{N_k}$$

其中 $N_k = \sum_{i=1}^N \mathbb{I}(y_i = c_k)$ 表示 N 个训练样本中属于类 c_k 的样本的数目

- ▶ 取对数, 求偏导 (需要考虑约束 $\sum_{k=1}^K p(c_k) = 1$), 置零, 得到先验概率 $p(c_k)$ 的估计值

$$p(c_k)_{MLE} = \frac{N_k}{N} = \frac{\sum_{i=1}^N \mathbb{I}(y_i = c_k)}{N}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ (2) 估计类条件概率 $p(x^{(j)} | c_k)$
- ▶ 假设第 j 个属性是离散型变量, 可能的取值集合为 $\{a_{j1}, a_{j2}, \dots, a_{jS_j}\}$, 共有 S_j 个可能的取值⁴
- ▶ 估计类条件概率 $p(x^{(j)} | c_k)$ 就是估计 $p(x^{(j)} = a_{jl} | c_k)$, $l = 1, \dots, S_j$
 - ▶ 满足 $p(x^{(j)} = a_{jl} | c_k) > 0$, $l = 1, \dots, S_j$ 且
$$\sum_{l=1}^{S_j} p(x^{(j)} = a_{jl} | c_k) = 1$$
 - ▶ 这是一个多项分布(multinomial distribution), 相当于掷有一个有 S_j 个面的骰子, 各个面向上的概率分别是 $p(x^{(j)} = a_{jl} | c_k)$, $l = 1, \dots, S_j$

⁴连续型变量的处理参考周志华《机器学习》P151

朴素贝叶斯分类器 (Naive Bayes classifier)

- 从 N 个训练样本中估计类条件概率 $p(x^{(j)} = a_{jl} | c_k)$
 - $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$, 其中
$$\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)})$$
, $y_i \in \mathcal{Y} = \{c_1, \dots, c_K\}$
- 极大似然估计

$$L = \frac{N!}{\prod_{k=1}^K N_k!} \prod_{k=1}^K [p(c_k)]^{N_k}$$

其中 $N_k = \sum_{i=1}^N \mathbb{I}(y_i = c_k)$ 表示 N 个训练样本中属于类 c_k 的样本的数目,

$N_{k,a_{jl}} = \sum_{i=1}^N \mathbb{I}(y_i = c_k, x_i^{(j)} = a_{jl})$ 表示类 c_k 中第 j 个属性取值为 a_{jl} 的样本的数目

- 取对数, 求偏导 (需要考虑约束 $\sum_{l=1}^{S_j} p(x^{(j)} = a_{jl} | c_k) = 1$), 置零, 得到先验概率 $p(x^{(j)} = a_{jl} | c_k)$ 的估计值

$$p(x^{(j)} = a_{jl} | c_k)_{MLE} = \frac{N_{k,a_{jl}}}{N_k} = \frac{\sum_{i=1}^N \mathbb{I}(y_i = c_k, x_i^{(j)} = a_{jl})}{\sum_{i=1}^N \mathbb{I}(y_i = c_k)}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

▶ 朴素贝叶斯算法

- ▶ **输入:** 训练数据 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 $x_i = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(d)})$, $x_i^{(j)}$ 是第 i 个样本的第 j 个特征, $x_i^{(j)}$ 可能的取值集合为 $\{a_{j1}, a_{j2}, \dots, a_{jS_j}\}$, $y_i \in \mathcal{Y} = \{c_1, \dots, c_K\}$, $i = 1, \dots, N$; 待分类样本 $\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(d)})$
- ▶ **输出:** 待分类样本 \mathbf{x} 的分类结果 $y \in \mathcal{Y} = \{c_1, \dots, c_K\}$, $i = 1, \dots, N$
- ▶ (1) 计算先验概率 $p(c_k)$, $k = 1, \dots, K$

$$p(c_k) = \frac{N_k}{N} = \frac{\sum_{i=1}^N \mathbb{I}(y_i = c_k)}{N}$$

- ▶ (2) 计算类条件概率 $p(x^{(j)} = a_{jl} | c_k)$, $j = 1, \dots, d$, $l = 1, \dots, S_j$

$$p(x^{(j)} = a_{jl} | c_k) = \frac{N_{k, a_{jl}}}{N_k} = \frac{\sum_{i=1}^N \mathbb{I}(y_i = c_k, x_i^{(j)} = a_{jl})}{\sum_{i=1}^N \mathbb{I}(y_i = c_k)}$$

- ▶ (3) 对于待分类样本 $\mathbf{x} = (x^{(1)}, x^{(2)}, \dots, x^{(d)})$, 对于 $k = 1, \dots, K$ 分别计算

$$p(c_k) \prod_{j=1}^d p(x^{(j)} = a_{jl} | c_k)$$

- ▶ (4) 确定样本 \mathbf{x} 的分类结果

$$y = \arg \max_{c_k} p(c_k) \prod_{j=1}^d p(x^{(j)} = a_{jl} | c_k)$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- 示例：试由下表所示的训练数据学习一个朴素贝叶斯分类器并确定 $\mathbf{x} = (2, S)^T$ 的类标记 y 。表中 $X^{(1)}, X^{(2)}$ 为特征，取值的集合分别为 $A_1 = \{1, 2, 3\}, A_2 = \{S, M, L\}$ ， Y 为类标记， $Y \in C = \{1, -1\}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	<i>S</i>	<i>M</i>	<i>M</i>	<i>S</i>	<i>S</i>	<i>S</i>	<i>M</i>	<i>M</i>	<i>L</i>	<i>L</i>	<i>L</i>	<i>M</i>	<i>M</i>	<i>L</i>	<i>L</i>
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

朴素贝叶斯分类器 (Naive Bayes classifier)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

► 根据朴素贝叶斯算法, 可以获得以下概率

$$\begin{aligned} P(Y = 1) &= \frac{9}{15}, & P(Y = -1) &= \frac{6}{15} \\ P(X^{(1)} = 1 | Y = 1) &= \frac{2}{9}, & P(X^{(1)} = 2 | Y = 1) &= \frac{3}{9}, & P(X^{(1)} = 3 | Y = 1) &= \frac{4}{9} \\ P(X^{(2)} = S | Y = 1) &= \frac{1}{9}, & P(X^{(2)} = M | Y = 1) &= \frac{4}{9}, & P(X^{(2)} = L | Y = 1) &= \frac{4}{9} \\ P(X^{(1)} = 1 | Y = -1) &= \frac{3}{6}, & P(X^{(1)} = 2 | Y = -1) &= \frac{2}{6}, & P(X^{(1)} = 3 | Y = -1) &= \frac{1}{6} \\ P(X^{(2)} = S | Y = -1) &= \frac{3}{6}, & P(X^{(2)} = M | Y = -1) &= \frac{2}{6}, & P(X^{(2)} = L | Y = -1) &= \frac{1}{6} \end{aligned}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 对于待分类样本 $\mathbf{x} = (2, S)^T$ 计算得到

$$\begin{aligned} P(Y = 1)P(X^{(1)} = 2 \mid Y = 1)P(X^{(2)} = S \mid Y = 1) &= \frac{9}{15} \cdot \frac{3}{9} \cdot \frac{1}{9} = \frac{1}{45} \\ P(Y = -1)P(X^{(1)} = 2 \mid Y = -1)P(X^{(2)} = S \mid Y = -1) &= \frac{6}{15} \cdot \frac{2}{6} \cdot \frac{3}{6} = \frac{1}{15} \end{aligned}$$

因为 $P(Y = -1)P(X^{(1)} = 2 \mid Y = -1)P(X^{(2)} = S \mid Y = -1)$ 最大, 所以 $y = -1$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 贝叶斯参数估计
- ▶ 用极大似然估计可能会出现所要估计的概率值为 0 的情况
 - ▶ 某些属性可能的取值较多，而样本数相对较少，使得某些类条件概率 $p(x^{(j)} = a_{jl} | c_k)$ 的极大似然估计结果为 0
 - ▶ 解决这一问题的方法是采用贝叶斯估计

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 狄利克雷分布 (Dirichlet distribution) 是一种多元连续随机变量的概率分布, 是贝塔分布 (beta distribution) 的扩展。在贝叶斯学习中, 狄利克雷分布常作为多项分布的先验分布使用
- ▶ **定义 (狄利克雷分布):** 若多元连续随机变量 $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ 的概率密度函数为

$$p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i - 1}$$

其中 $\sum_{i=1}^k \theta_i = 1$, $\theta_i \geq 0$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$, $\alpha_i > 0, i = 1, 2, \dots, k$, 则称随机变量 θ 服从参数为 α 的狄利克雷分布, 记作 $\theta \sim \text{Dir}(\alpha)$

朴素贝叶斯分类器 (Naive Bayes classifier)

- 式中 $\Gamma(s)$ 是伽马函数, 定义为

$$\Gamma(s) = \int_0^{\infty} x^{s-1} e^{-x} dx, \quad s > 0$$

具有性质

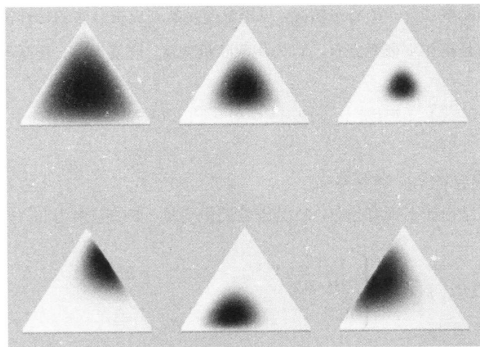
$$\Gamma(s+1) = s\Gamma(s)$$

当 s 是自然数时, 有

$$\Gamma(s+1) = s!$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 由于满足条件 $\theta_i \geq 0$, $\sum_{i=1}^k \theta_i = 1$, 所以狄利克雷分布 θ 存在于 $(k-1)$ 维单纯形上



- ▶ 上图为二维单纯型上的狄利克雷分布，参数分别为 $\alpha = (3, 3, 3)$, $\alpha = (7, 7, 7)$, $\alpha = (20, 20, 20)$, $\alpha = (2, 6, 11)$, $\alpha = (14, 9, 5)$, $\alpha = (6, 2, 6)$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 狄利克雷分布的性质

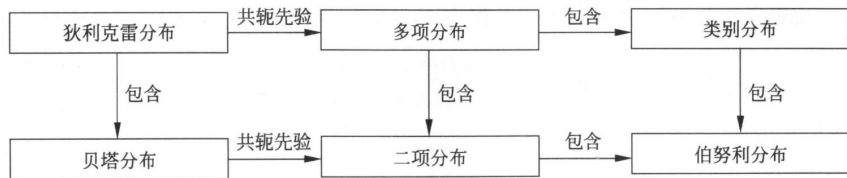
$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i-1}$$

- ▶ $E[\theta_i] = \frac{\alpha_i}{\alpha_0}$ $\text{mode}[\theta_i] = \frac{\alpha_i-1}{\alpha_0-K}$ $\text{var}[\theta_i] = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$

其中 $\alpha_0 = \sum_{i=1}^k \alpha_i$

朴素贝叶斯分类器 (Naive Bayes classifier)

► 几种概率分布之间的关系



朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 狄利克雷分布有一些重要性质:
 - ▶ (1) 狄利克雷分布属于指数分布族
 - ▶ (2) 狄利克雷分布是多项分布的共轭先验 (conjugate prior)
- ▶ **共轭先验**: 如果后验分布与先验分布属于同类, 则先验分布与后验分布称为共轭分布 (conjugate distributions), 先验分布称为似然函数的共轭先验 (conjugate prior)
- ▶ 例如: 如果多项分布的先验分布是狄利克雷分布, 则其后验分布也为狄利克雷分布, 两者构成共轭分布。作为先验分布的狄利克雷分布的参数又称为超参数
- ▶ 使用共轭分布的好处是便于从先验分布计算后验分布

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 设 $\mathcal{W} = \{w_1, w_2, \dots, w_k\}$ 是由 k 个元素组成的集合
- ▶ 随机变量 X 服从 \mathcal{W} 上的多项分布, $X \sim \text{Mult}(n, \theta)$, 其中 $n = (n_1, n_2, \dots, n_k)$ 和 $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ 是参数
- ▶ 参数 n 为从 \mathcal{W} 中重复独立抽取样本的次数, n_i 为样本中 w_i 出现的次数, $i = 1, 2, \dots, k$
- ▶ 参数 θ_i 为 w_i 出现的概率 ($i = 1, 2, \dots, k$)

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 将样本数据表示为 D , 目标是计算在样本数据 D 给定条件下参数 θ 的后验概率 $p(\theta | D)$
- ▶ 对于给定的样本数据 D , 似然函数是

$$p(D | \theta) = \frac{n!}{n_1! n_2! \cdots n_k!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k} = \frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k \theta_i^{n_i}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 假设随机变量 θ 服从狄利克雷分布 $p(\theta | \alpha)$, 其中 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ 为参数, 则 θ 的先验分布为

$$\begin{aligned} p(\theta | \alpha) &= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i-1} \\ &= \frac{1}{B(\alpha)} \prod_{i=1}^k \theta_i^{\alpha_i-1} \\ &= \text{Dir}(\theta | \alpha) \end{aligned}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- 根据贝叶斯规则, 在给定样本数据 D 和参数 α 条件下, θ 的后验概率分布是

$$\begin{aligned} p(\theta \mid D, \alpha) &= \frac{p(D \mid \theta)p(\theta \mid \alpha)}{p(D \mid \alpha)} \\ &= \frac{\frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k \theta_i^{n_i} \frac{1}{B(\alpha)} \theta_i^{\alpha_i-1}}{\int \frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k \theta_i^{n_i} \frac{1}{B(\alpha)} \theta_i^{\alpha_i-1} d\theta} \\ &= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i + n_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i + n_i)} \prod_{i=1}^k \theta_i^{\alpha_i + n_i - 1} \\ &= \text{Dir}(\theta \mid \alpha + n) \end{aligned}$$

其中 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$, $n = (n_1, n_2, \dots, n_k)$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 可以看出先验分布和后验分布都是狄利克雷分布, 两者有不同的参数, 所以狄利克雷分布是多项分布的共轭先验
- ▶ 狄利克雷后验分布的参数等于狄利克雷先验分布参数 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ 加上多项分布的观测计数 $n = (n_1, n_2, \dots, n_k)$, 好像试验之前就已经观察到计数 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$, 因此也把 α 叫做先验伪计数 (prior pseudo-counts)

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ **贝叶斯参数估计** 多项分布的共轭先验是Dir分布
- ▶ 在朴素贝叶斯分类器中, 对先验概率 $p(c_k)$ 对应的多项分布 $(p(c_1), p(c_2), \dots, p(c_K))$ 指定参数为 $\lambda = (\lambda, \lambda, \dots, \lambda)$ 的狄利克雷分布 $\text{Dir}(\lambda)$ 作为其先验分布
- ▶ 根据数据 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ 得到多项分布 $(p(c_1), p(c_2), \dots, p(c_K))$ 的后验分布

$$\text{Dir}(\lambda')$$

其中 $\lambda' = (\lambda'_1, \lambda'_2, \dots, \lambda'_K)$, $\lambda'_k = \lambda + \sum_{i=1}^N \mathbb{I}(y_i = c_k)$

- ▶ 取上述狄利克雷分布的均值作为先验概率 $p(c_k)$ 的贝叶斯估计的结果

$$p(c_k)_{\text{Bayes}} = \frac{\lambda + \sum_{i=1}^N \mathbb{I}(y_i = c_k)}{K\lambda + N}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 类似地，类条件概率 $p(x^{(j)} = a_{jl} | c_k)$ 的贝叶斯估计结果为

$$p(x^{(j)} = a_{jl} | c_k)_{Bayes} = \frac{\lambda + \sum_{i=1}^N \mathbb{I}(y_i = c_k, x^{(j)} = a_{jl})}{S_j \lambda + \sum_{i=1}^N \mathbb{I}(y_i = c_k)}$$

- ▶ 通常取 $\lambda = 1$ ，这时称为拉普拉斯平滑 (Laplacian smoothing)

朴素贝叶斯分类器 (Naive Bayes classifier)

- 示例：试由下表所示的训练数据学习一个朴素贝叶斯分类器并确定 $\mathbf{x} = (2, S)^T$ 的类标记 y 。表中 $X^{(1)}, X^{(2)}$ 为特征，取值的集合分别为 $A_1 = \{1, 2, 3\}, A_2 = \{S, M, L\}$ ， Y 为类标记， $Y \in C = \{1, -1\}$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	<i>S</i>	<i>M</i>	<i>M</i>	<i>S</i>	<i>S</i>	<i>S</i>	<i>M</i>	<i>M</i>	<i>L</i>	<i>L</i>	<i>L</i>	<i>M</i>	<i>M</i>	<i>L</i>	<i>L</i>
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

- 按照拉普拉斯平滑估计概率， $\lambda = 1$

朴素贝叶斯分类器 (Naive Bayes classifier)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

$$\begin{aligned} P(Y=1) &= \frac{10}{17}, & P(Y=-1) &= \frac{7}{17} \\ P(X^{(1)}=1|Y=1) &= \frac{3}{12}, & P(X^{(1)}=2|Y=1) &= \frac{4}{12}, & P(X^{(1)}=3|Y=1) &= \frac{5}{12} \\ P(X^{(2)}=S|Y=1) &= \frac{2}{12}, & P(X^{(2)}=M|Y=1) &= \frac{5}{12}, & P(X^{(2)}=L|Y=1) &= \frac{5}{12} \\ P(X^{(1)}=1|Y=-1) &= \frac{4}{9}, & P(X^{(1)}=2|Y=-1) &= \frac{3}{9}, & P(X^{(1)}=3|Y=-1) &= \frac{2}{9} \\ P(X^{(2)}=S|Y=-1) &= \frac{4}{9}, & P(X^{(2)}=M|Y=-1) &= \frac{3}{9}, & P(X^{(2)}=L|Y=-1) &= \frac{2}{9} \end{aligned}$$

朴素贝叶斯分类器 (Naive Bayes classifier)

- ▶ 对于待分类样本 $\mathbf{x} = (2, S)^T$ 计算得到

$$\begin{aligned} P(Y = 1)P(X^{(1)} = 2 | Y = 1)P(X^{(2)} = S | Y = 1) &= \frac{10}{17} \cdot \frac{4}{12} \cdot \frac{2}{12} = \frac{5}{153} = 0.0327 \\ P(Y = -1)P(X^{(1)} = 2 | Y = -1)P(X^{(2)} = S | Y = -1) &= \frac{7}{17} \cdot \frac{3}{9} \cdot \frac{4}{9} = \frac{28}{459} = 0.0610 \end{aligned}$$

因为 $P(Y = -1)P(X^{(1)} = 2 | Y = -1)P(X^{(2)} = S | Y = -1)$ 最大, 所以 $y = -1$