第四章　概率方法

张皓

https://haomood.github.io/homepage/

zhangh0214@gmail.com

摘要

本文介绍机器学习算法中的概率方法. 概率方法会对数据的分布进行假设, 对概率密度函数进行估计, 并使用这个概率密度函数进行决策. 本文介绍四种最常用的概率方法: 线性回归 (用于回归任务)、对数几率回归 (用于二分类任务)、*Softmax* 回归 (用于多分类任务) 和朴素贝叶斯分类器 (用于多分类任务).[*] 前三种方法属于判别式模型, 而朴素贝叶斯分类器属于生成式模型.

本系列文章有以下特点: *(a).* 为了减轻读者的负担并能使尽可能多的读者从中收益, 本文试图尽可能少地使用数学知识, 只要求读者有基本的微积分、线性代数和概率论基础, 并在第一节对关键的数学知识进行回顾和介绍. *(b).* 本文不省略任何推导步骤, 适时补充背景知识, 力图使本节内容是自足的, 使机器学习的初学者也能理解本文内容. *(c).* 机器学习近年来发展极其迅速, 已成为一个非常广袤的领域. 本文无法涵盖机器学习领域的方方面面, 仅就一些关键的机器学习流派的方法进行介绍. *(d).* 为了帮助读者巩固本文内容, 或引导读者扩展相关知识, 文中穿插了许多问题, 并在最后一节进行问题的 "快问快答".

[*]严格来说, 前三者兼有多种解释, 既可以看做是概率方法, 又可以看做是非概率方法.

1　准备知识

本节给出概率方法的基本流程, 后续要介绍的不同的概率方法都遵循这一基本流程.

1.1　概率方法的建模流程

(1). 对 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ 进行概率假设. 我们假定 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ 具有某种确定的概率分布形式, 其形式被参数向量 $\boldsymbol{\theta}$ 唯一地确定.

(2). 对参数 $\boldsymbol{\theta}$ 进行最大后验估计. 基于训练样例对概率分布的参数 $\boldsymbol{\theta}$ 进行最大后验估计 (maximum a posteriori, MAP), 得到需要优化的损失函数.

最大后验估计是指

$$\boldsymbol{\theta}^{\star} := \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{\theta} \mid D) = \arg\max_{\boldsymbol{\theta}} \; p(D \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \quad (1)$$

其在最大化时考虑如下两项:

- 参数的先验分布 $p(\boldsymbol{\theta})$. 最大后验估计认为参数 $\boldsymbol{\theta}$ 未知并且是一个随机变量, 其本身服从一个先验分布 $p(\boldsymbol{\theta})$. 这个先验分布蕴含了我们关于参数的领域知识.
- 基于观测数据得到的似然 (likelihood) $p(D \mid \boldsymbol{\theta})$. 最大化似然是在 $\boldsymbol{\theta}$ 的所有可能的取值中, 找到一个能使样本属于其真实标记的概率最大的值.

最大后验估计是在考虑先验分布 $p(\boldsymbol{\theta})$ 时最大化基于观测数据得到的似然 (likelihood) $p(D \mid \boldsymbol{\theta})$.

参数估计的两个不同学派的基本观点是什么? 这实际上是参数估计 (parameter estimation) 过程, 统计学中的频率主义学派 (frequentist) 和贝叶斯学派 (Bayesian) 提供了不同的解决方案 [3, 9] . 频率主义学派认为参数虽然未知, 但确实客观存在的固定值, 因此通常使用极大似然估计来确定参数值. 贝叶斯学派则

认为参数是未观察到的随机变量, 其本身也可有分布, 因此, 可假定参数服从一个先验分布, 然后基于观察到的数据来计算参数的后验分布.

定理 1. 最大后验估计的结果是优化如下形式的损失函数

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{m}\sum_{i=1}^{m}\ell(\boldsymbol{\theta}) + \lambda\Omega(\boldsymbol{\theta})\,. \tag{2}$$

Proof. 利用样例的独立同分布假设,

$$\begin{aligned}
\boldsymbol{\theta}^\star &:= \arg\max_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}\mid D)\\
&= \arg\max_{\boldsymbol{\theta}}\ p(D\mid\boldsymbol{\theta})\,p(\boldsymbol{\theta})\\
&= \arg\max_{\boldsymbol{\theta}}\ \log p(D\mid\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\\
&= \arg\max_{\boldsymbol{\theta}}\ \log\prod_{i=1}^{m} p(\boldsymbol{x}_i, y_i\mid\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\\
&= \arg\min_{\boldsymbol{\theta}}\ -\sum_{i=1}^{m}\log p(y_i\mid\boldsymbol{x}_i,\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})\,. \quad (3)
\end{aligned}$$

之后, 从 $-\sum_{i=1}^{m}\log p(y_i\mid\boldsymbol{x}_i,\boldsymbol{\theta})$ 得到 $\frac{1}{m}\sum_{i=1}^{m}\ell(\boldsymbol{\theta})$, 从 $-\log p(\boldsymbol{\theta})$ 得到 $\lambda\Omega(\boldsymbol{\theta})$. □

经验风险和结构风险的含义? $\mathcal{L}(\boldsymbol{\theta})$ 的第一项称为经验风险 (empirical risk), 用于描述模型与训练数据的契合程度. 第二项称为结构风险 (structural risk) 或正则化项 (regularization term), 源于模型的先验概率, 表述了我们希望获得何种性质的模型 (例如希望获得复杂度较小的模型). λ 称为正则化常数, 对两者进行折中.

结构风险的作用? (1). 为引入领域知识和用户意图提供了途径. (2). 有助于削减假设空间, 从而降低了最小化训练误差的过拟合风险. 这也可理解为一种 "罚函数法", 即对不希望得到的结果施以惩罚, 从而使得优化过程趋向于希望目标. ℓ_p 范数是常用的正则化项.

定理 2. 若参数 $\boldsymbol{\theta}$ 的先验分布服从高斯分布 $\boldsymbol{\theta}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$, 那么对应的结构风险是 $\Omega(\boldsymbol{\theta}) := \frac{1}{2}\|\boldsymbol{\theta}\|^2$.

Proof. 由于从 $-\log p(\boldsymbol{\theta})$ 得到 $\lambda\Omega(\boldsymbol{\theta})$,

$$\begin{aligned}
\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}}\ -\log p(\boldsymbol{\theta})\\
&= \arg\min_{\boldsymbol{\theta}}\ -\log\frac{1}{\sqrt{(2\pi)^{d+1}\det\sigma^2\boldsymbol{I}}}\exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top\sigma^2\boldsymbol{I}\boldsymbol{\theta}\right)\\
&= \arg\min_{\boldsymbol{\theta}}\ \frac{1}{2}\boldsymbol{\theta}^\top\sigma^2\boldsymbol{I}\boldsymbol{\theta}\\
&= \arg\min_{\boldsymbol{\theta}}\ \frac{1}{2}\|\boldsymbol{\theta}\|^2\,, \qquad (4)
\end{aligned}$$

其中先验分布 $\boldsymbol{\theta}\sim\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$ 的参数 σ^2 转化为正则化常数 λ. □

为什么最常假设参数的先验分布是高斯分布 (或最常使用 ℓ_2 正则化)? 这是因为高斯分布 $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 是所有均值和熵存在且协方差矩阵是 $\boldsymbol{\Sigma}$ 的分布中熵最大的分布. 最大熵分布是在特定约束下具有最大不确定性的分布. 在没有更多信息的情况下, 那些不确定的部分都是 "等可能的". 在设计先验分布 $p(\boldsymbol{\theta})$ 时, 除了我们对参数的认知 (例如均值和值域) 外, 我们不想引入任何其余的偏见 (bias). 因此最大熵先验 (对应 ℓ_2 正则化) 常被使用. 除高斯先验外, 还可以使用不提供信息的先验 (uninformative prior), 其在一定范围内均匀分布, 对应的损失函数中没有结构风险这一项.

(3). 对损失函数 $\mathcal{L}(\boldsymbol{\theta})$ 进行梯度下降优化.

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\,. \tag{5}$$

梯度下降的细节留在下一节介绍.

概率方法的优缺点各是什么? 优点: 这种参数化的概率方法使参数估计变得相对简单. 缺点: 参数估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布. 在现实应用中, 欲做出能较好地接近潜在真实分布的假设, 往往需在一定程度利用关于应用任务本身的经验知识, 否则仅凭 "猜测" 来假设概率分布形式, 很可能产生误导性的结果. 我们不一定非要概率式地解释这个世界, 在不考虑概率的情况下, 直接找到分类边界, 也被称为判别函数 (discriminant function) 有时甚至能比判别式模型产生更好的结果.

1.2 梯度下降

我们的目标是求解下列无约束的优化问题

$$\boldsymbol{\theta}^\star := \arg\min_{\boldsymbol{\theta}}\ \mathcal{L}(\boldsymbol{\theta})\,, \tag{6}$$

其中 $\mathcal{L}(\boldsymbol{\theta})$ 是连续可微函数. 梯度下降是一种一阶 (first-order) 优化方法, 是求解无约束优化问题最简单、最经典的求解方法之一.

梯度下降的基本思路? 梯度下降贪心地迭代式地最小化 $\mathcal{L}(\boldsymbol{\theta})$. 梯度下降希望找到一个方向 (单位向量) $\boldsymbol{v}$ 使得 $\mathcal{L}$ 在这个方向下降最快, 并在这个方向前进 α 的距离

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\boldsymbol{v}\,. \tag{7}$$

定理 3. 梯度下降的更新规则是公式 5 . 重复这个过程, 可收敛到局部极小点.

Proof. 我们需要找到下降最快的方向 $\boldsymbol{v}$ 和前进的距离 α.

(1). 下降最快的方向 $\boldsymbol{v}$. 利用泰勒展开

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0)+(\boldsymbol{x}-\boldsymbol{x}_0)^\top \frac{\partial f}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}_0}+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^\top \frac{\partial^2 f}{\partial x^2}\Big|_{\boldsymbol{x}_0}(\boldsymbol{x}-\boldsymbol{x}_0) \tag{8}$$

的一阶近似,

$$\begin{aligned}
\boldsymbol{v}^\star &:= \mathop{\arg\min}_{\|\boldsymbol{v}\|=1} \ \mathcal{L}(\boldsymbol{\theta}+\alpha\boldsymbol{v}) \\
&\approx \mathop{\arg\min}_{\|\boldsymbol{v}\|=1} \ \mathcal{L}(\boldsymbol{\theta})+\alpha\boldsymbol{v}^\top\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} \\
&= \mathop{\arg\min}_{\|\boldsymbol{v}\|=1} \ \alpha\|\boldsymbol{v}\|\left\|\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right\|\cos\left\langle \boldsymbol{v}, \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right\rangle \\
&= \mathop{\arg\min}_{\|\boldsymbol{v}\|=1} \ \cos\left\langle \boldsymbol{v}, \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right\rangle \\
&= -\frac{\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}}{\left\|\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right\|},
\end{aligned} \tag{9}$$

即下降最快的方向是损失函数的负梯度方向.

(2). 前进的距离 α. 我们希望在开始的时候前进距离大一些以使得收敛比较快, 而在接近最小值时前进距离小一些以不错过最小值点. 因此, 我们设前进距离为损失函数梯度的一个倍数

$$\alpha^\star := \eta\left\|\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right\|, \tag{10}$$

其中 η 被称为学习率 (learning rate).

向公式 7 代入最优的 $\alpha^\star$ 和 $\boldsymbol{v}^\star$ 后即得. □

引理 4. 若 $\mathcal{L}(\boldsymbol{\theta})$ 满足 *L-Lipschitz* 条件, 则设置步长 $\eta := \frac{1}{2L}$ 即可确保收敛到局部极小点.

定义 1 (凸函数). 对区间 $[a,b]$ 上定义的函数 $f:\mathbb{R}^n\to\mathbb{R}$, 若它对区间中任意两点 $\boldsymbol{x}_1,\boldsymbol{x}_2$ 均有

$$f\left(\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2}\right)\le\frac{f(\boldsymbol{x}_1)+f(\boldsymbol{x}_2)}{2}, \tag{11}$$

则称 f 为区间 $[a,b]$ 上的凸函数 (convex function). 当 $<$ 成立时, 称为严格凸函数 (strict convex function). U 形曲线的函数如 $f(x):=x^2$ 通常是凸函数.

引理 5 (凸函数判定准则). 对 $f:\mathbb{R}\to\mathbb{R}$, 若 $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\ge 0 (>0)$, 则 f 是凸函数 *(严格凸函数)*. 对 $f:\mathbb{R}^d\to\mathbb{R}$, 若 $\frac{\partial^2 f}{\partial \boldsymbol{x}^2}\succeq 0 (\succ 0)$, 即 $\frac{\partial^2 f}{\partial \boldsymbol{x}^2}$ 是一个半正定 *(正定)* 矩阵, 则 f 是凸函数 *(严格凸函数)*.

引理 6. 对凸函数 $f:\mathbb{R}^d\to\mathbb{R}$, 局部极小点对应着函数的全局最小点, 此时梯度下降可确保收敛到全局最优解. 全局最小点满足 $\frac{\partial f}{\partial \boldsymbol{x}}=\boldsymbol{0}$ 的 $\boldsymbol{x}$.

2 线性回归

2.1 建模流程

线性回归 (linear regression) 回归问题 $y\in\mathbb{R}$. 其建模方法包括如下三步 (参见第 1.1 节).

(1). 对 $p(y\mid\boldsymbol{x},\boldsymbol{\theta})$ 进行概率假设.

我们假设

$$y := \boldsymbol{w}^\top\boldsymbol{x}+b+\varepsilon,\ \varepsilon\sim\mathcal{N}(0,\sigma^2). \tag{12}$$

ε 被称为误差项, 捕获了 (a). 特征向量 $\boldsymbol{x}$ 中没有包含的因素. (b). 随机噪声. 对不同的样本 $\boldsymbol{x}_i$, ε_i 是独立同分布地从 $\mathcal{N}(0,\sigma^2)$ 中进行采样得到的.

线性回归的假设函数是

$$h(\boldsymbol{x}) := \boldsymbol{w}^\top\boldsymbol{x}+b. \tag{13}$$

为了书写方便, 我们记

$$\boldsymbol{\theta} := \begin{bmatrix}\boldsymbol{w}\\ b\end{bmatrix},\ \tilde{\boldsymbol{x}} := \begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix}. \tag{14}$$

那么公式 12 等价于

$$y=\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}+\varepsilon,\ \varepsilon\sim\mathcal{N}(0,\sigma^2). \tag{15}$$

在本文其余部分我们将沿用这一简化记号. 因此,

$$p(y\mid\boldsymbol{x},\boldsymbol{\theta})=\mathcal{N}(\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}},\sigma^2). \tag{16}$$

(2). 对参数 $\boldsymbol{\theta}$ 进行最大后验估计.

定理 7. 假设参数 $\boldsymbol{\theta}$ 服从高斯先验, 对参数 $\boldsymbol{\theta}$ 进行最大后验估计等价于优化如下损失函数

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{2m}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}_i-y_i)^2+\lambda\|\boldsymbol{\theta}\|^2. \tag{17}$$

其中

$$\ell(\boldsymbol{\theta}) := \frac{1}{2}(\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}-y)^2 \tag{18}$$

被称为平方损失 *(square loss)*. 在线性回归中, 平方损失就是试图找到一个超平面 $\boldsymbol{w}^\top\boldsymbol{x}+b=0$, 使所有样本到该超平面的欧式距离 *(Euclidean distance)* 之和最小.

Table 1: 常用线性回归模型对比.

学习算法	噪声分布	参数先验	损失函数	特点
简单线性回归	高斯	无信息先验	$\frac{1}{2m}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2$	当属性数 d 大, 样本数 m 小时, 容易过拟合
岭回归 [12]	高斯	高斯	$\frac{1}{2m}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2$	相比简单线性回归, 可以显著降低过拟合风险
LASSO [10]	高斯	拉普拉斯	$\frac{1}{2m}\sum_{i=1}^{m}(\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i)^2 + \lambda\|\boldsymbol{\theta}\|_1$	降低过拟合风险, 并更容易得到稀疏解
稳健线性回归	拉普拉斯	高斯	$\frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{\theta}^\top \boldsymbol{x}_i - y_i\| + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2$	适用于数据存在异常点

Proof.

$$
\begin{aligned}
\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{m} \log p(y_i \mid \boldsymbol{x}_i, \theta) + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)^2}{2\sigma^2}\right) + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; \sum_{i=1}^{m} (y_i - \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)^2 + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; \frac{1}{2m}\sum_{i=1}^{m} (\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i)^2 + \lambda\|\boldsymbol{\theta}\|^2 .
\end{aligned}
$$

其中, 最后一行只是为了数学计算上方便, 下文推导对数几率回归和 Softmax 回归时的最后一步亦然. □

(3). 对损失函数 $\mathcal{L}(\boldsymbol{\theta})$ 进行梯度下降优化.

可以容易地得到损失函数对参数的偏导数

$$
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m} (\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i)\tilde{\boldsymbol{x}}_i + \lambda\boldsymbol{\theta} . \tag{20}
$$

2.2 线性回归的闭式解

线性回归对应的平方损失的函数形式比较简单, 可以通过求 $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \mathbf{0}$ 直接得到最优解.

定理 8. *线性回归的闭式解为*

$$
\boldsymbol{\theta}^\star = (\boldsymbol{X}^\top \boldsymbol{X} + m\lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y} . \tag{21}
$$

Proof. $\mathcal{L}(\boldsymbol{\theta})$ 可等价地写作

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2m}\sum_{i=1}^{m} (\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\
&= \frac{1}{2m}\sum_{i=1}^{m} (\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - y_i)^2 + \frac{\lambda}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} \\
&= \frac{1}{2m}\left\| \begin{bmatrix} \tilde{\boldsymbol{x}}_1^\top \boldsymbol{\theta} - y_1 \\ \tilde{\boldsymbol{x}}_2^\top \boldsymbol{\theta} - y_2 \\ \vdots \\ \tilde{\boldsymbol{x}}_m^\top \boldsymbol{\theta} - y_m \end{bmatrix} \right\|^2 + \frac{\lambda}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} .
\end{aligned} \tag{22}
$$

令

$$
\boldsymbol{X} := \begin{bmatrix} \tilde{\boldsymbol{x}}_1^\top \\ \tilde{\boldsymbol{x}}_2^\top \\ \vdots \\ \tilde{\boldsymbol{x}}_m^\top \end{bmatrix} \in \mathbb{R}^{m\times(d+1)}, \; \boldsymbol{y} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^{m} , \tag{23}
$$

那么

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}) \\
&= \frac{1}{2m}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2 + \frac{\lambda}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta} \\
&= \frac{1}{2m}(\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{y}) + \frac{\lambda}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}
\end{aligned} \tag{24}
$$

求解

$$
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{1}{m}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} - \frac{1}{m}\boldsymbol{X}^\top \boldsymbol{y} + \lambda\boldsymbol{\theta} = \mathbf{0} \tag{25}
$$

即得. □

$\boldsymbol{X}^\top \boldsymbol{X}$ 不可逆的情况及解决方案? (1). 属性数 $d+1$ 多于样例数 m. (2). 属性之间线性相关. 通过正则化项 $m\lambda\boldsymbol{I}$, 即使 $\boldsymbol{X}^\top \boldsymbol{X}$ 不可逆, $\boldsymbol{X}^\top \boldsymbol{X} + m\lambda\boldsymbol{I}$ 仍是可逆的.

2.3 其他正则化回归模型

事实上, 上文介绍的线性回归模型是岭回归 (ridge regression). 根据正则化项的不同, 有三种常用的线性回归模型, 见表 1 .

基于 ℓ_0、ℓ_1 和 ℓ_2 范数正则化的效果? ℓ_2 范数倾向于 $\boldsymbol{w}$ 的分量取值尽量均衡, 即非零分量个数尽量稠密. 而 ℓ_0"范数" 和 ℓ_1 范数则倾向于 $\boldsymbol{w}$ 的分量尽量稀疏, 即非零分量个数尽量少, 优化结果得到了仅采用一部分属性的模型. 也就是说, 基于 ℓ_0"范数" 和 ℓ_1 范数正则化的学习方法是一种嵌入式 (embedding) 特征选择方法, 其特征选择过程和学习器训练过程融为一体, 两者在同一个优化过程中完成. 事实上, 对 $\boldsymbol{w}$ 施加稀疏约束最自

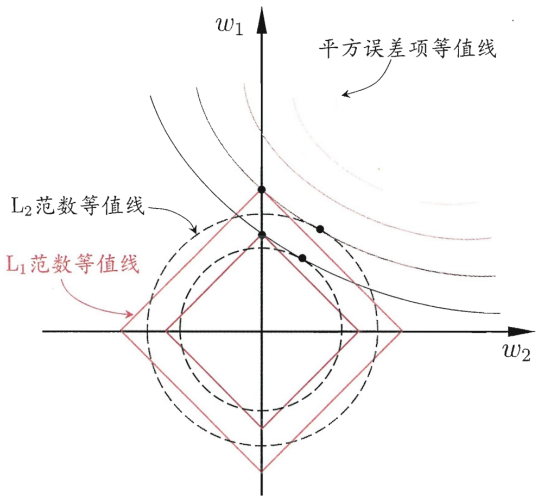

Figure 1: ℓ_1 正则化 (红色) 比 ℓ_2 正则化 (黑色) 更易于获得稀疏解. 本图源于 [17] .

然的是使用 ℓ_0“范数”. 但 ℓ_0“范数” 不连续, 难以优化求解. 因此常采用 ℓ_1 范数来近似.

为什么 ℓ_1 正则化比 ℓ_2 正则化更易于获得稀疏解? 假设 $\boldsymbol{x} \in \mathbb{R}^2$, 则 $\boldsymbol{w} \in \mathbb{R}^2$. 我们绘制出平方损失项、$\ell_1$ 范数和 ℓ_2 范数的等值线 (取值相同的点的连线), 如图 1 所示. LASSO 的解要在平方损失项和正则化项之间折中, 即出现在图中平方误差项等值线和正则化项等值线的相交处. 从图中可以看出, 采用 ℓ_1 正则化时交点常出现在坐标轴上 ($w_2 = 0$), 而采用 ℓ_2 正则化时交点常出现在某个象限中 (w_1, w_2 均不为 0).

考虑一般的带有 ℓ_1 正则化的优化目标

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1 \,, \tag{26}$$

若 $\ell(\boldsymbol{\theta})$ 满足 L-Lipschitz 条件, 即

$$\exists L > 0, \forall \boldsymbol{x}_1, \boldsymbol{x}_2. \ \left\| \frac{\partial \ell}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}_1} - \frac{\partial \ell}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}_2} \right\|^2 \leq L \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \,. \tag{27}$$

优化通常使用近端梯度下降 (proximal gradient descent, PGD) [1] . PGD 也是一种贪心地迭代式地最小化策略, 能快速地求解基于 ℓ_1 范数最小化的方法.

定理 9. *假设当前参数是 θ_t, PGD 的更新准则是*

$$\theta_{t+1,j} := \begin{cases} z_j + \dfrac{\lambda}{L} & \text{若 } z_j < -\dfrac{\lambda}{L} ; \\ 0 & \text{若 } -\dfrac{\lambda}{L} \leq z_j \leq \dfrac{\lambda}{L} ; \\ z_j - \dfrac{\lambda}{L} & \text{若 } z_j > \dfrac{\lambda}{L} , \end{cases} \tag{28}$$

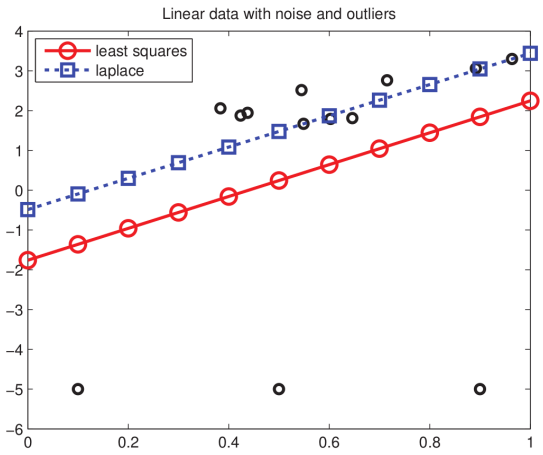

Figure 2: 存在异常点 (图下方的三个点) 时普通线性回归 (红色) 和稳健线性回归 (蓝色). 本图源于 [7] .

其中

$$\boldsymbol{z} := \boldsymbol{\theta}_t - \frac{1}{L} \frac{\partial \ell}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_t} . \tag{29}$$

Proof. 在 θ_t 附近将 $\ell(\theta)$ 进行二阶泰勒展开近似

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \arg\min_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}_t) + (\boldsymbol{\theta} - \boldsymbol{\theta}_t)^\top \frac{\partial \ell}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_t} + \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_t\|^2 \\ &\qquad + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \ \frac{L}{2} \left\| \boldsymbol{\theta} - \left(\boldsymbol{\theta}_t - \frac{1}{L} \frac{\partial \ell}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_t} \right) \right\|^2 + \lambda \|\boldsymbol{\theta}\|_1 \\ &= \arg\min_{\boldsymbol{\theta}} \ \sum_{j=1}^{d} \left(\frac{1}{2} (\theta_d - z_d)^2 + \frac{\lambda}{L} |\theta_d| \right) . \end{aligned} \tag{30}$$

由于 $\boldsymbol{\theta}$ 各维互不影响 (不存在交叉项), 因此可以独立求解各维. □

在 LASSO 的基础上进一步发展出考虑特征分组结构的 Group LASSO [14] 、考虑特征序结构的 Fused LASSO [11] 等变体. 由于凸性不严格, LASSO 类方法可能产生多个解, 该问题通过弹性网 (elastic net) 得以解决 [16] .

2.4 存在异常点数据的线性回归

一旦数据中存在异常点 (outlier), 由于平方损失计算的是样本点到超平面距离的平方, 远离超平面的点会对回归结果产生更大的影响, 如图 2 所示. 平方损失对应于假设噪声服从高斯分布 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, 一种应对

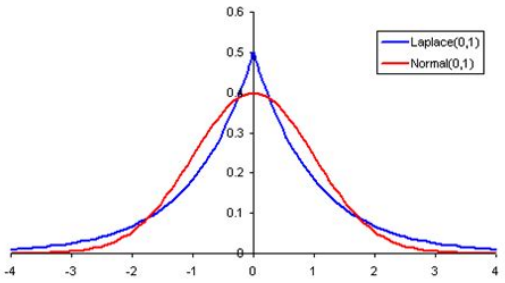

Figure 3: 高斯分布 $\mathcal{N}(0,1)$ (红色) 和拉普拉斯分布 $\mathrm{Lap}(0,1)$ (蓝色). 本图源于https://www.epixanalytics.com/modelassist/AtRisk/images/15/image632.gif.

异常点的方法是取代高斯分布为其他更加重尾 (heavy tail) 的分布, 使其对异常点的容忍能力更强, 例如使用拉普拉斯分布 $\varepsilon \sim \mathrm{Lap}(0,b)$, 如图 3 所示.

定义 2 (拉普拉斯分布 (Laplace distribution) $\mathrm{Lap}(\mu,b)$). 又称为双边指数分布 (double sided exponential distribution), 具有如下的概率密度函数

$$p(x) := \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right). \tag{31}$$

该分布均值为 μ, 方差为 $2b^2$.

定理 10. 假设参数服从高斯先验,

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Lap}(\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}, b)\,, \tag{32}$$

对参数 $\boldsymbol{\theta}$ 进行最大后验估计等价于优化如下损失函数

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{m}\sum_{i=1}^{m} |\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i| + \lambda\|\boldsymbol{\theta}\|^2\,. \tag{33}$$

Proof.

$$\begin{aligned}
\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{m} \log p(y_i \mid \boldsymbol{x}_i, \theta) + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^{m} \log \frac{1}{2b}\exp\left(-\frac{|y_i - \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i|}{b}\right) + \lambda\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; \frac{1}{m}\sum_{i=1}^{m} |\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i| + \lambda\|\boldsymbol{\theta}\|^2\,.
\end{aligned} \tag{34}$$

□

由于绝对值函数不光滑, 不便基于梯度下降对公式 33 进行优化. 通过分离变量技巧, 可将其转化为二次规划 (quadratic programming) 问题, 随后调用现有的软件包进行求解. 我们在下一章形式化 SVR 时还会再使用这个技巧.

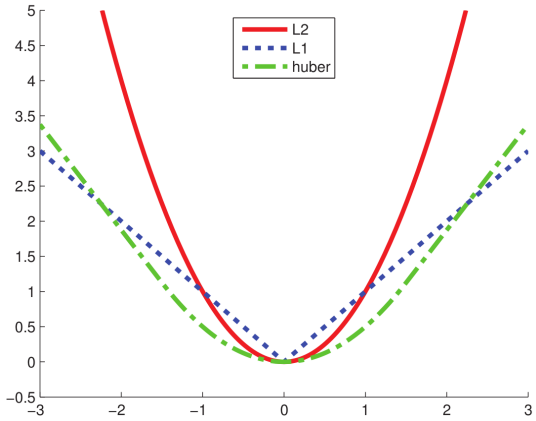

Figure 4: ℓ_2 损失 (红色)、ℓ_1 损失 (蓝色) 和 Huber 损失 (绿色). 本图源于 [7] .

定理 11. 最小化公式 33 等价于如下二次规划问题, 其包含 $d+1+2m$ 个变量, $3m$ 个约束:

$$\begin{aligned}
\arg\min_{\boldsymbol{\theta},\boldsymbol{\varepsilon}^+,\boldsymbol{\varepsilon}^-} \quad & \frac{1}{m}\sum_{i=1}^{m}(\varepsilon_i^+ - \varepsilon_i^-) + \lambda\|\boldsymbol{\theta}\|^2 \\
\text{s.t.} \quad & \varepsilon_i^+ \geq 0, && i = 1,2,\ldots,m\,, \\
& \varepsilon_i^- \geq 0, && i = 1,2,\ldots,m\,, \\
& \boldsymbol{\varepsilon}_i^+ - \boldsymbol{\varepsilon}_i^- + y_i - \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i = 0 && i = 1,2,\ldots,m\,.
\end{aligned} \tag{35}$$

Proof. 令 $\varepsilon_i^+ - \varepsilon_i^- := \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i$, 并约束 $\varepsilon_i^+ \geq 0, \varepsilon_i^- \geq 0$. □

此外, 为了结合高斯分布 (对应平凡损失) 容易优化和拉普拉斯分布 (对应 ℓ_1 损失) 可以应对异常值的优点, Huber 损失 [5]在误差接近 0 时为平方损失, 在误差比较大时接近 ℓ_1 损失, 如图 4 所示.

$$\ell(\boldsymbol{\theta}) := \begin{cases} \dfrac{1}{2}(\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} - y)^2 & \text{若 } |\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} - y| \leq \tau\,; \\ \tau|\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} - y| - \dfrac{\tau^2}{2} & \text{若 } |\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} - y| > \tau\,. \end{cases} \tag{36}$$

Huber 损失处处可微, 使用基于梯度的方法对 Huber 损失进行优化会比使用拉普拉斯分布更快.

2.5 广义线性模型

线性回归利用属性的线性组合 $\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}$ 进行预测. 除了直接利用 $\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}$ 逼近 y 外, 还可以使模型的预测值逼

Table 2: 线性回归、对数几率回归和 Softmax 回归对比.

学习算法	标记空间 $\mathcal{Y}$	假设函数 $h(\boldsymbol{x})$	概率假设 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$	经验损失	损失形式 $\ell(\boldsymbol{\theta})$
线性回归	$\mathbb{R}$	$\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}$	$\mathcal{N}(h(\boldsymbol{x}), \sigma^2)$	平方损失	$\frac{1}{2}(\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} - y)^2$
对数几率回归	$\{0, 1\}$	$\frac{1}{1+\exp(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})}$	$h(\boldsymbol{x})^y (1 - h(\boldsymbol{x}))^{1-y}$	对数几率损失	$-y\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} + \log(1 + \exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})$
对数几率回归	$\{-1, 1\}$	$\frac{1}{1+\exp(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})}$	$h(y\boldsymbol{x})$	对数几率损失	$\log(1 + \exp(-y\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}))$
Softmax 回归	$\{1, 2, \ldots, C\}$	$\frac{\exp \boldsymbol{\theta}_k^\top \tilde{\boldsymbol{x}}}{\sum_{c=1}^C \exp \boldsymbol{\theta}_c^\top \tilde{\boldsymbol{x}}}$	$h(\boldsymbol{x})_y$	交叉熵损失	$-\boldsymbol{\theta}_y^\top \tilde{\boldsymbol{x}} + \log \sum_{c=1}^C \exp \boldsymbol{\theta}_c^\top \tilde{\boldsymbol{x}}$

近 y 的衍生物. 考虑单调可微函数 g, 令

$$h(\boldsymbol{x}) := g^{-1}(\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}), \tag{37}$$

这样得到的模型称为广义线性模型 (generalized linear model), 其中函数 g 被称为联系函数 (link function). 本文介绍的线性回归、对数几率回归和 Softmax 回归都属于广义线性模型, 如表 2 所示.

广义线性模型的优点? (1). 形式简单、易于建模. (2). 很好的可解释性. $|w_j|$ 直观表达了各属性在预测中的重要性.

如何利用广义线性模型解决非线性问题? (1). 引入层级结构. 例如深度学习是对样本 $\boldsymbol{x}$ 进行逐层加工, 将初始的低层表示转化为高层特征表示后使用线性分类器. (2). 高维映射. 例如核方法将 $\boldsymbol{x}$ 映射到一个高维空间 $\phi(\boldsymbol{x})$ 后使用线性分类器.

3 对数几率回归

3.1 建模流程

对数几率回归 (logistic regression) 应对二分类问题. 其建模方法包括如下三步 (参见第 1.1 节).

(1). 对 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ 进行概率假设.

对二分类任务, 标记 $y \in \{0, 1\}$, 而 $\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}$ 产生的是实数值, 于是, 我们需要找到一个单调可微函数 g 将 $\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}$ 转化为 $\{0, 1\}$. 最理想的是用单位阶跃函数

$$h(\boldsymbol{x}) := \frac{1}{2} \operatorname{sign} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} + \frac{1}{2}. \tag{38}$$

当 $\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}$ 大于 0 时输出 1, 小于 0 时输出 0. 但是, 单位阶跃函数不连续不可微, 无法利用梯度下降方法进行优化. 因此, 我们希望找到一个能在一定程度上近似单位阶跃函数并单调可微的替代函数 (surrogate function).

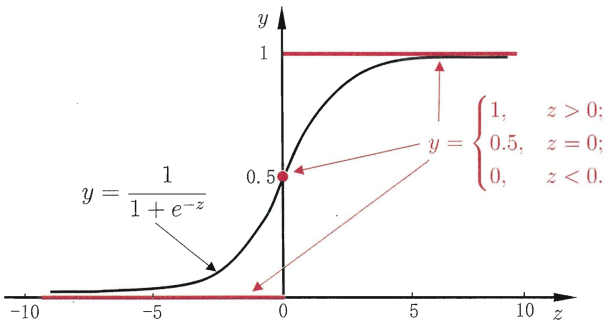

Figure 5: 单位阶跃函数 (红色) 与对数几率函数 (黑色). 本图源于 [17] .

如图 5 所示, 对数几率函数 (sigmoid function) 正是这样一个常用的替代函数

$$h(\boldsymbol{x}) := \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})}. \tag{39}$$

我们将其视为后验概率估计, 即

$$\Pr(y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) := \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}, \tag{40}$$

那么

$$\Pr(y = 0 \mid \boldsymbol{x}, \boldsymbol{\theta}) = 1 - p(y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = 1 - \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}. \tag{41}$$

两者可以合并写作

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = (\operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})^y (1 - \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})^{1-y}. \tag{42}$$

也就是说, $y \mid \boldsymbol{x}, \boldsymbol{\theta}$ 服从伯努利分布 $\operatorname{Ber}(\operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})$.

(2). 对参数 $\boldsymbol{\theta}$ 进行最大后验估计.

定理 12. 假设参数 $\boldsymbol{\theta}$ 服从高斯先验, 对参数 $\boldsymbol{\theta}$ 进行最大后验估计等价于优化如下损失函数

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{m} \sum_{i=1}^m \Big(-y_i \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i + \log(1 + \exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i) \Big) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2. \tag{43}$$

其中

$$\ell(\boldsymbol{\theta}) := -y\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} + \log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}) \tag{44}$$

称为对数几率损失 *(logistic loss)*.

Proof.

$$\begin{aligned}\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^m \log p(y_i \mid \boldsymbol{x}_i, \theta) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^m \Big(y_i \log \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i \\ &\qquad\qquad + (1-y_i)\log(1-\operatorname{sigm}\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)\Big) \\ &\qquad + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 . \end{aligned} \tag{45}$$

注意到

$$\begin{aligned}\log \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i &= \log \frac{1}{1+\exp(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)} \\ &= \log \frac{\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i}{1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i} \\ &= \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - \log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i) , \end{aligned} \tag{46}$$

$$\begin{aligned}\log(1-\operatorname{sigm}\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i) &= \log\Big(1 - \frac{1}{1+\exp(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)}\Big) \\ &= \log \frac{1}{1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i} \\ &= -\log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i) . \end{aligned} \tag{47}$$

因此

$$\begin{aligned}\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \; -\sum_{i=1}^m \Big(y_i(\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - \log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)) \\ &\qquad\qquad -(1-y_i)\log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)\Big) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \; \sum_{i=1}^m \Big(-y_i \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i + \log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)\Big) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \; \frac{1}{m}\sum_{i=1}^m \Big(-y_i \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i + \log(1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i)\Big) \\ &\qquad + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 . \end{aligned} \tag{48}$$

□

(3). 对损失函数 $\mathcal{L}(\theta)$ 进行梯度下降优化.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} &= \frac{1}{m}\sum_{i=1}^m \left(-y_i \boldsymbol{x}_i + \frac{\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i}{1+\exp \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i} \boldsymbol{x}_i \right) + \lambda \boldsymbol{\theta} \\ &= \frac{1}{m}\sum_{i=1}^m \Big(\operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i - y_i \Big) \boldsymbol{x}_i + \lambda \boldsymbol{\theta} . \end{aligned} \tag{49}$$

3.2 与广义线性模型的关系

对数几率回归的假设函数 $h(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}})}$ 等价于

$$\log \frac{h(\boldsymbol{x})}{1-h(\boldsymbol{x})} = \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} , \tag{50}$$

其中

$$\frac{h(\boldsymbol{x})}{1-h(\boldsymbol{x})} = \frac{p(y=1 \mid \boldsymbol{x}, \boldsymbol{\theta})}{p(y=0 \mid \boldsymbol{x}, \boldsymbol{\theta})} \tag{51}$$

被称为几率 (odds), 反映 $\boldsymbol{x}$ 作为正例的相对可能性. $\log \frac{h(\boldsymbol{x})}{1-h(\boldsymbol{x})}$ 被称为对数几率 (log odds, logit), 公式 50 实际上在用线性回归模型的预测结果逼近真实标记的对数几率, 这是对数几率回归名称的由来.

对数几率回归的优点? (1). 直接对分类的可能性进行建模 (假设 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ 服从伯努利分布), 无需事先假设样本 $\boldsymbol{x}$ 的分布, 这样避免了假设分布不准确所带来的问题. (2). 不仅能预测出类别, 还可以得到近似概率预测, 对许多需要概率辅助决策的任务很有用. (3). 对数几率的目标函数是凸函数, 有很好的数学性质.

引理 13. 对数几率损失函数是凸函数.

Proof. 在 $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$ 的基础上, 进一步可求得

$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}^2} = \frac{1}{m}\sum_{i=1}^m \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i (1-\operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}_i) \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top + \lambda \tag{52}$$

是一个半正定矩阵. □

3.3 $y \in \{-1, 1\}$ 的对数几率回归

为了概率假设方便, 我们令二分类问题的标记 $y \in \{0, 1\}$. 有时, 我们需要处理 $y \in \{-1, 1\}$ 形式的分类问题. 对数几率损失函数需要进行相应的改动.

(1). 对 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ 进行概率假设.

我们假设

$$\Pr(y=1 \mid \boldsymbol{x}, \boldsymbol{\theta}) := \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} , \tag{53}$$

那么

$$\Pr(y=0 \mid \boldsymbol{x}, \boldsymbol{\theta}) = 1 - \operatorname{sigm} \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} = \operatorname{sigm}(-\boldsymbol{\theta}^\top \tilde{\boldsymbol{x}}) . \tag{54}$$

两者可以合并写作

$$p(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \operatorname{sigm} y_i \boldsymbol{\theta}^\top \tilde{\boldsymbol{x}} \tag{55}$$

(2). 对参数 θ 进行最大后验估计.

定理 14. 假设参数 $\boldsymbol{\theta}$ 服从高斯先验, 对参数 $\boldsymbol{\theta}$ 进行最大后验估计等价于优化如下损失函数

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}_i)) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 . \quad (56)$$

其中

$$\ell(\boldsymbol{\theta}) := \log(1+\exp(-y\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}})) \quad (57)$$

称为对数几率损失 *(logistic loss)*.

Proof.

$$\begin{aligned}\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \ -\sum_{i=1}^{m}\log p(y_i \mid \boldsymbol{x}_i,\theta) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \ -\sum_{i=1}^{m}\log \operatorname{sigm} y_i\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}} + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \ \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}_i)) \\ &\qquad + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 . \end{aligned} \quad (58)$$

□

(3). 对损失函数 $\mathcal{L}(\boldsymbol{\theta})$ 进行梯度下降优化.

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}} &= \frac{1}{m}\sum_{i=1}^{m}\frac{\exp(-y_i\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}_i)}{1+\exp(-y_i\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}_i)}(-y_i\boldsymbol{x}_i) + \lambda\boldsymbol{\theta} \\ &= -\frac{1}{m}\sum_{i=1}^{m}\operatorname{sigm}(-y_i\boldsymbol{\theta}^\top\tilde{\boldsymbol{x}}_i)y_i\boldsymbol{x}_i + \lambda\boldsymbol{\theta} . \end{aligned} \quad (59)$$

4 Softmax 回归

4.1 建模流程

Softmax 回归应对多分类问题, 它是对数几率回归向多分类问题的推广. 其建模方法包括如下三步 (参见第 1.1 节).

(1). 对 $p(y \mid \boldsymbol{x},\boldsymbol{\theta})$ 进行概率假设.

$$h(\boldsymbol{x})_k := p(y=k \mid \boldsymbol{x},\boldsymbol{\theta}) := \frac{\exp\boldsymbol{\theta}_k^\top\tilde{\boldsymbol{x}}}{\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}} \in (0,1) . \quad (60)$$

对数几率回归假设 $p(y \mid \boldsymbol{x},\boldsymbol{\theta})$ 服从伯努利分布, Softmax 回归假设 $p(y \mid \boldsymbol{x},\boldsymbol{\theta})$ 服从如下分布

$$p(y \mid \boldsymbol{x},\boldsymbol{\theta}) = \frac{\exp\boldsymbol{\theta}_y^\top\tilde{\boldsymbol{x}}}{\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}} . \quad (61)$$

令

$$\boldsymbol{\Theta} := \begin{bmatrix}\boldsymbol{\theta}_1^\top \\ \boldsymbol{\theta}_2^\top \\ \vdots \\ \boldsymbol{\theta}_C^\top\end{bmatrix} \in \mathbb{R}^{C\times(d+1)} , \quad (62)$$

假设函数可以写成矩阵的形式

$$p(y \mid \boldsymbol{x},\boldsymbol{\Theta}) = \frac{(\exp\boldsymbol{\Theta}\tilde{\boldsymbol{x}})_y}{(\exp\boldsymbol{\Theta}\tilde{\boldsymbol{x}})^\top\mathbf{1}} . \quad (63)$$

(2). 对参数 $\boldsymbol{\theta}$ 进行最大后验估计.

定理 15. 假设参数 $\boldsymbol{\theta}$ 服从高斯先验, 对参数 $\boldsymbol{\theta}$ 进行最大后验估计等价于优化如下损失函数

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(-\boldsymbol{\theta}_{y_i}^\top\tilde{\boldsymbol{x}}_i + \log\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}_i\right) + \sum_{c=1}^{C}\frac{\lambda}{2}\|\boldsymbol{\theta}_c\|^2 . \quad (64)$$

其中

$$\ell(\boldsymbol{\theta}) := -\boldsymbol{\theta}_y^\top\tilde{\boldsymbol{x}} + \log\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}} \quad (65)$$

称为交叉熵损失 *(cross-entropy loss)*.

Proof.

$$\begin{aligned}\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \ -\sum_{i=1}^{m}\log p(y_i \mid \boldsymbol{x}_i,\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \ -\sum_{i=1}^{m}\log\frac{\exp\boldsymbol{\theta}_{y_i}^\top\tilde{\boldsymbol{x}}_i}{\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}_i} + \sum_{c=1}^{C}\frac{\lambda}{2}\|\boldsymbol{\theta}_c\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \ -\sum_{i=1}^{m}\left(\boldsymbol{\theta}_{y_i}^\top\tilde{\boldsymbol{x}}_i - \log\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}_i\right) \\ &\qquad + \sum_{c=1}^{C}\frac{\lambda}{2}\|\boldsymbol{\theta}_c\|^2 \\ &= \arg\min_{\boldsymbol{\theta}} \ \frac{1}{m}\sum_{i=1}^{m}\left(-\boldsymbol{\theta}_{y_i}^\top\tilde{\boldsymbol{x}}_i + \log\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}_i\right) \\ &\qquad + \sum_{c=1}^{C}\frac{\lambda}{2}\|\boldsymbol{\theta}_c\|^2 . \end{aligned}$$

□

(3). 对损失函数 $\mathcal{L}(\boldsymbol{\theta})$ 进行梯度下降优化.

损失函数对应于类别 k 的参数 $\boldsymbol{\theta}_k$ 的导数是

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_k} &= \frac{1}{m}\sum_{i=1}^{m}\left(-\mathbb{I}(y_i=k)\boldsymbol{x}_i + \frac{\exp\boldsymbol{\theta}_k^\top\tilde{\boldsymbol{x}}_i}{\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}_i}\boldsymbol{x}_i\right) + \lambda\boldsymbol{\theta}_k \\ &= \frac{1}{m}\sum_{i=1}^{m}\left(\frac{\exp\boldsymbol{\theta}_k^\top\tilde{\boldsymbol{x}}_i}{\sum_{c=1}^{C}\exp\boldsymbol{\theta}_c^\top\tilde{\boldsymbol{x}}_i} - \mathbb{I}(y_i=k)\right)\boldsymbol{x}_i + \lambda\boldsymbol{\theta}_k \end{aligned} \quad (66)$$

写成矩阵的形式是

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\Theta}} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{(\exp \boldsymbol{\Theta}\tilde{\boldsymbol{x}}_i)_{y_i}}{(\exp \boldsymbol{\Theta}\tilde{\boldsymbol{x}}_i)^\top \mathbf{1}} - \mathbf{e}_{y_i}\right)\boldsymbol{x}_i^\top + \lambda\boldsymbol{\Theta}\,, \quad (67)$$

其中 $\mathbf{e}_k \in \mathbb{R}^C$ 的第 k 个元素是 1, 其余元素均为 0.

对比公式 20 、 49 和 67 , 损失函数的梯度有相同的数学形式

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}(h(\boldsymbol{x}_i) - y_i)\boldsymbol{x}_i + \lambda\boldsymbol{\theta}\,. \quad (68)$$

区别在于假设函数 $h(\boldsymbol{x}_i)$ 的形式不同. 事实上, 所有的广义线性模型都有类似于公式 68 的更新准则.

4.2 交叉熵

定义由训练集观察得到的分布, 称为经验分布 (empirical distribution). 经验分布 p_i 对应于第 i 个样例, 定义 $p_{ic} := \mathbb{I}(y_i = c)$. 另一方面, $h(\boldsymbol{x}_i)$ 是由模型估计出的概率.

定理 16. *交叉熵损失旨在最小化经验分布 p_i 和学得分布 $h(\boldsymbol{x}_i)$ 之间的交叉熵. 这等价于最小化 p_i 和 $h(\boldsymbol{x}_i)$ 之间的 KL 散度, 迫使估计的分布 $h(\boldsymbol{x}_i)$ 近似目标分布 p_i.*

Proof.

$$\begin{aligned}
\boldsymbol{\theta}^\star &:= \arg\min_{\boldsymbol{\theta}} \; -\frac{1}{m}\sum_{i=1}^{m}\log h(\boldsymbol{x}_i)_y + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; -\frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\log h(\boldsymbol{x}_i)_c + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; -\frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C} p_{ic}\log h(\boldsymbol{x}_i)_c + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C}(p_{ic}\log p_{ic} - p_{ic}\log h(\boldsymbol{x}_i)_c) \\
&\qquad + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; \frac{1}{m}\sum_{i=1}^{m}\sum_{c=1}^{C} p_{ic}\log\frac{p_{ic}}{h(\boldsymbol{x}_i)_c} + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \\
&= \arg\min_{\boldsymbol{\theta}} \; \frac{1}{m}\sum_{i=1}^{m}\mathrm{KL}(p_i \,||\, h(\boldsymbol{x}_i)) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2\,. \quad (69)
\end{aligned}$$

□

5 朴素贝叶斯分类器

朴素贝叶斯分类器 (naive Bayes classifier) 也是一种概率方法, 但它是一种生成式模型. 在本节, 我们首先回顾生成式模型, 之后介绍朴素贝叶斯分类器的建模流程.

5.1 生成式模型

判别式模型和生成式模型各是什么? 判别式模型 (discriminant model) 直接对 $p(y \mid \boldsymbol{x})$ 进行建模, 生成式模型 (generative model) 先对联合分布 $p(\boldsymbol{x}, y) = p(\boldsymbol{x} \mid y)p(y)$ 进行建模, 然后再得到

$$p(y \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid y)p(y)}{p(\boldsymbol{x})}\,. \quad (70)$$

其中, $p(y)$ 是类先验 (prior) 概率, 表达了样本空间中各类样本所占的比例. $p(\boldsymbol{x} \mid y)$ 称为似然 (likelihood). $p(\boldsymbol{x})$ 是用于归一化的证据 (evidence). 由于其和类标记无关, 该项不影响 $p(y \mid \boldsymbol{x})$ 的估计

$$\begin{aligned}
h(\boldsymbol{x}) &:= \arg\max_c \; p(y = c \mid \boldsymbol{x}) \\
&= \arg\max_c \; p(\boldsymbol{x} \mid y = c)p(y = c) \\
&= \arg\max_c \; \log p(\boldsymbol{x} \mid y = c) + \log p(y = c) \quad (71)
\end{aligned}$$

如何对类先验概率和似然进行估计? 根据大数定律, 当训练集包含充足的独立同分布样本时, $p(y)$ 可通过各类样本出现的频率来进行估计

$$p(y = c) := \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(y_i = c)\,. \quad (72)$$

而对似然 $p(\boldsymbol{x} \mid y)$, 由于其涉及 $\boldsymbol{x}$ 所有属性的联合概率, 如果基于有限训练样本直接估计联合概率, (1). 在计算上将会遭遇组合爆炸问题. (2). 在数据上将会遭遇样本稀疏问题, 很多样本取值在训练集中根本没有出现, 而 "未被观测到"' 与 "出现概率为零" 通常是不同的. 直接按样本出现的频率来估计会有严重的困难, 属性数越多, 困难越严重.

判别式模型和生成式模型的优缺点? 优缺点对比如表 3 所示.

5.2 建模流程

(1). 对 $p(\boldsymbol{x} \mid y, \boldsymbol{\theta})$ 进行概率假设.

Table 3: 判别式模型和生成式模型的优缺点对比.

算法类型	算法性能	收敛速度	采样/生成服从潜在分布样例 $(\boldsymbol{x}, y)$	存在隐变量时进行学习
判别式模型	通常较高	通常较慢	不能	不能
生成式模型	通常较低	通常较快	能	能

生成式模型的主要困难在于, 类条件概率 $p(\boldsymbol{x} \mid y)$ 是所有属性的联合概率, 难以从有限的训练样本直接估计而得. 为避开这个障碍, 朴素贝叶斯分类器采用了*属性条件独立性假设*: 对已知类别, 假设所有属性相互独立. 也就是说, 假设每个属性独立地对分类结果发生影响

$$p(x_j, x_d \mid y) = p(x_j \mid y)\, p(x_d \mid y)\,. \tag{73}$$

此外, 对连续属性, 进一步假设

$$p(x_j \mid y = c) \sim \mathcal{N}(\mu_{cj}, \sigma_{cj}^2)\,. \tag{74}$$

因此, 朴素贝叶斯分类器的假设函数是

$$\begin{aligned} h(\boldsymbol{x}) &:= \arg\max_c \sum_{j=1}^{d} \log p(x_j \mid y = c) + \log p(y = c) \\ &= \arg\max_c \sum_{j=1}^{d} \log \frac{1}{\sqrt{2\pi\sigma_{cj}^2}} \exp\left(-\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2}\right) \\ &\quad + \log \sum_{i=1}^{m} \mathbb{I}(y_i = c) \\ &= \arg\min_c \sum_{j=1}^{d} \left(\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2} + \frac{1}{2}\log \sigma_{cj}^2\right) \\ &\quad - \log \sum_{i=1}^{m} \mathbb{I}(y_i = c)\,. \end{aligned} \tag{75, 76}$$

(2). 对参数 $\boldsymbol{\theta}$ 进行最大后验估计. 参数 $\boldsymbol{\theta}$ 包括了第 c 类样本在第 j 个属性上的高斯分布的均值 μ_{cj} 和方差 σ_{cj}^2.

定理 17. *假设参数 $\boldsymbol{\theta}$ 服从不提供信息的先验, 对参数 $\boldsymbol{\theta}$ 进行最大后验估计的结果是*

$$\mu_{cj}^{\star} := \frac{\sum_{i=1}^{m} \mathbb{I}(y_i = c)\, x_{ij}}{\sum_{i=1}^{m} \mathbb{I}(y_i = c)}\,; \tag{77}$$

$$\sigma_{cj}^{2\,\star} := \frac{\sum_{i=1}^{m} \mathbb{I}(y_i = c)\, (x_{ij} - \mu_{cj}^{\star})^2}{\sum_{i=1}^{m} \mathbb{I}(y_i = c)}\,. \tag{78}$$

Proof. 代入公式 76 ,

$$\begin{aligned} \mu_{cj}^{\star} &:= \arg\max_{\mu_{cj}} \sum_{i=1}^{m} \mathbb{I}(y_i = c)\, p(\boldsymbol{x}_i, y_i = c) \\ &= \arg\min_{\mu_{cj}} \sum_{i=1}^{m} \mathbb{I}(y_i = c)\, (x_j - \mu_{cj})^2\,. \end{aligned} \tag{79}$$

令 $\mathcal{L}(\mu_{cj}) := \sum_{i=1}^{m} \mathbb{I}(y_i = c)\, (x_j - \mu_{cj})^2$. 通过解

$$\frac{\partial \mathcal{L}}{\partial \mu_{cj}} = -\sum_{i=1}^{m} 2\, \mathbb{I}(y_i = c)\, (x_j - \mu_{cj}) = 0 \tag{80}$$

可以得到 $\mu_{cj}^{\star}$.

$$\begin{aligned} \sigma_{cj}^2 &:= \arg\max_{\mu_{cj}} \sum_{i=1}^{m} \mathbb{I}(y_i = c)\, p(\boldsymbol{x}_i, y_i = c) \\ &= \arg\min_{\mu_{cj}} \sum_{i=1}^{m} \mathbb{I}(y_i = c) \left(\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2} + \frac{1}{2}\log \sigma_{cj}^2\right). \end{aligned}$$

令 $\mathcal{L}(\sigma_{cj}^2) := \sum_{i=1}^{m} \mathbb{I}(y_i = c)\left(\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2} + \frac{1}{2}\log \sigma_{cj}^2\right)$. 通过解

$$\frac{\partial \mathcal{L}}{\partial \sigma_{cj}^2} = \sum_{i=1}^{m} \mathbb{I}(y_i = c) \left(-\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^4} + \frac{1}{2}\frac{1}{\sigma_{cj}^2}\right) = 0 \tag{81}$$

可以得到 σ_{cj}^2. □

5.3 离散属性的参数估计

朴素贝叶斯分类器可以很容易地处理离散属性. $p(x_j \mid y)$ 可估计为

$$p(x_j = a \mid y = c) := \frac{\sum_{i=1}^{m} \mathbb{I}(x_j = a \wedge y_i = c)}{\sum_{i=1}^{m} \mathbb{I}(y_i = c)}\,. \tag{82}$$

然而, 若某个属性值在训练集中没有与某个类同时出现过, 则根据公式 82 估计得到 0. 代入公式 75 得到 $-\infty$. 因此, 无论该样本的其他属性是什么, 分类结果都不会是 $y = c$, 这显然不太合理.

为了避免其他属性携带的信息被训练集中未出现的属性值 “抹去”, 在估计概率值时通常要进行平滑

(smoothing), 常用拉普拉斯修正 (Laplacian correction). 具体的说, 令 K 表示训练集 D 中可能的类别数, n_j 表示第 j 个属性可能的取值数, 则概率估计修正为

$$p(y=c) := \frac{\sum_{i=1}^{m} \mathbb{I}(y_i = c) + 1}{m + K}; \tag{83}$$

$$p(x_j = a \mid y = c) := \frac{\sum_{i=1}^{m} \mathbb{I}(x_j = a \wedge y_i = c) + 1}{\sum_{i=1}^{m} \mathbb{I}(y_i = c) + n_j}. \tag{84}$$

拉普拉斯修正实际上假设了属性值与类别均匀分布, 这是在朴素贝叶斯学习中额外引入的关于数据的先验. 在训练集变大时, 修正过程所引入的先验的影响也会逐渐变得可忽略, 使得估值渐趋向于实际概率值.

在现实任务中朴素贝叶斯有多种实现方式. 例如, 若任务对预测速度要求较高, 则对给定训练集, 可将朴素贝叶斯分类器涉及的所有概率估值事先计算好存储起来, 这样在进行预测时只需查表即可进行判别. 若任务数据更替频繁, 则可采用懒惰学习方式, 先不进行任何训练, 待收到预测请求时再根据当前数据集进行概率估值. 若数据不断增加, 则可在现有估值基础上, 仅对新增样本的属性值所涉及的概率估值进行计数修正即可实现增量学习.

定义 3 (懒惰学习 (lazy learning)). 这类学习技术在训练阶段仅仅是把样本保存起来, 训练时间开销是 0, 待收到测试样本后再进行处理. 相应的, 那些在训练阶段就对样本进行学习处理的方法称为急切学习 (eager learning).

定义 4 (增量学习 (incremental learning)). 在学得模型后, 再接收到训练样例时, 仅需根据新样例对模型进行更新, 不必重新训练整个模型, 并且先前学得的有效信息不会被 "冲掉".

5.4 朴素贝叶斯分类器的推广

朴素贝叶斯分类器采用了属性条件独立性假设, 但在现实任务中这个假设往往很难成立. 于是, 人们尝试对属性条件独立性假设进行一定程度的放松, 适当考虑一部分属性间的相互依赖关系, 这样既不需要进行完全联合概率计算, 又不至于彻底忽略了比较强的属性依赖关系, 由此产生一类半朴素贝叶斯分类器 (semi-naive Bayes classifiers) 的学习方法.

独依赖估计 (one-dependent estimator, ODE) 是最常用的一种策略, 其假设每个属性在类别之外最多依赖

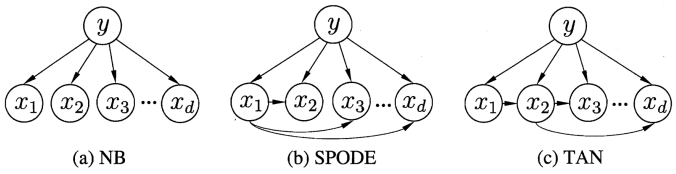

Figure 6: 朴素贝叶斯 (NB) 和两种半朴素贝叶斯分类器 (SPODE、TAN) 所考虑的属性依赖关系. 本图源于 [17].

于一个其他属性 (称为父属性). 问题的关键在于如何确定每个属性的父属性. SPODE (super-parent ODE) 假设所有属性都依赖于同一个属性, 称为超父 (super-parent). TAN (tree augmented naive Bayes) [4] 以属性节点构建完全图, 任意两结点之间边的权重设为这两个属性之间的条件互信息 $I(x_j, x_k \mid y)$. 之后构建此图的最大带权生成树, 挑选根变量, 将边置为有向, 以将属性间依赖关系约简为树形结构. 最后加入类别结点 y, 增加从 y 到每个属性的有向边. TAN 通过条件互信息刻画两属性的条件相关性, 最终保留了强相关属性之间的依赖性. AODE (averaged ODE) [13] 尝试将每个属性作为超父来构建 SPODE, 之后将那些具有足够训练数据支撑的 SPODE 集成作为最终结果. AODE 的训练过程也是 "计数", 因此具有朴素贝叶斯分类器无需模型选择、可预计算节省预测时间、也能懒惰学习、并且易于实现增量学习.

能否通过考虑属性间高阶依赖进一步提升泛化性能? 相比 ODE, kDE 考虑最多 k 个父属性. 随着依赖的属性个数 k 的增加, 准确进行概率估计所需的训练样本数量将以指数级增加. 因此, 若训练数据非常充分, 泛化性能有可能提升. 但在有限样本条件下, 则又陷入高阶联合概率的泥沼.

更进一步, 贝叶斯网 (Bayesian network), 也成为信念网 (belief network), 能表示任意属性间的依赖性. 贝叶斯网是一种概率图模型, 借助有向无环图刻画属性间的依赖关系.

事实上, 虽然朴素贝叶斯的属性条件独立假设在现实应用中往往很难成立, 但在很多情形下都能获得相当好的性能 [2, 8] . 一种解释是对分类任务来说, 只需各类别的条件概率排序正确, 无须精准概率值即可导致正确分类结果 [2] . 另一种解释是, 若属性间依赖对所有类别影响相同, 或依赖关系能相互抵消, 则属性条件独立性假设在降低计算开销的同时不会对性能产生负面

影响 [15] . 朴素贝叶斯分类器在信息检索领域尤为常用 [6] .

6 快问快答

随机梯度下降和标准梯度下降的优缺点各是什么?

- 参数更新速度. 标准梯度下降需要遍历整个训练集才能计算出梯度, 更新较慢. 随机梯度下降只需要一个训练样例即可计算出梯度, 更新较快.
- 冗余计算. 当训练集样本存在冗余时, 随机梯度下降能避免在相似样例上计算梯度的冗余.
- 梯度中的随机因素/噪声. 标准梯度下降计算得到的梯度没有随机因素, 一旦陷入局部极小将无法跳出. 随机梯度下降计算得到的梯度有随机因素, 有机会跳出局部极小继续优化.

实际应用时, 常采用随机梯度下降和标准梯度下降的折中, 即使用一部分样例进行小批量梯度下降. 此外, 相比随机梯度下降, 小批量梯度下降还可以更好利用矩阵的向量化计算的优势.

梯度下降和牛顿法的优缺点各是什么?

- 导数阶数. 梯度下降只需要计算一阶导数, 而牛顿法需要计算二阶导数. 一阶导数提供了方向信息(下降最快的方向), 二阶导数还提供了函数的形状信息.
- 计算和存储开销. 牛顿法在参数更新时需要计算 Hessian 矩阵的逆, 计算和存储开销比梯度下降更高.
- 学习率. 梯度下降对学习率很敏感, 而标准的牛顿法不需要设置学习率.
- 收敛速度. 牛顿法的收敛速度比梯度下降更快.
- 牛顿法不适合小批量或随机样本.

实际应用时, 有许多拟牛顿法旨在以较低的计算和存储开销近似 Hessian 矩阵.

线性回归的损失函数及梯度推导.

答案见上文.

为什么要使用正则化, ℓ_1 和 ℓ_2 正则化各自对应什么分布, 各有什么作用?

答案见上文.

对数几率回归的损失函数及梯度推导.

答案见上文.

对数几率回归和支持向量机的相同点和不同点?

相同点:

- 对数几率回归和支持向量机的主问题均为参数模型.
- 损失函数均为经验风险 + 结构风险 (正则化项) 的形式, 其经验风险均是 0/1 损失函数的替代损失, 结构风险都可选用 ℓ_1、ℓ_2 正则化项.
- 它们均为凸优化问题, 能有收敛到全局最优的保证. 对数几率回归和支持向量机的主问题均常采用基于梯度的方法进行优化.
- 均为二分类线性学习算法, 均可扩展到处理非线性、多分类问题.

不同点:

- 支持向量机的对偶问题是非参数模型, 需要利用训练集中的支持向量.
- 经验风险的函数形式不同 (对数几率损失和合页损失).
- 支持向量机的对偶问题采用二次规划进行优化, 或使用专门设计的 SMO 优化算法.
- 支持向量机不是概率方法, 不需要对 $p(y \mid \boldsymbol{x}, \boldsymbol{\theta})$ 进行概率假设. 对数几率回归是判别式模型, 而支持向量机直接得到判别函数.
- 对数几率回归是广义线性模型, 而支持向量机不是广义线性模型.

线性分类器如何扩展为非线性分类器?

答案见上文.

判别式模型和生成式模型各是什么, 各自优缺点是什么, 常见算法中哪些是判别式模型, 哪些是生成式模型?

答案见上文.

贝叶斯定理各项的含义?

答案见上文.

朴素贝叶斯为什么叫 "朴素" 贝叶斯?

为了避开从有限的训练样本直接估计 $p(\boldsymbol{x} \mid y)$ 的障碍, 朴素贝叶斯做出了属性条件独立假设, 该假设在现

实应用中往往很难成立.

References

[1] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005. 5

[2] P. M. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997. 12

[3] B. Efron. Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100(469):1–5, 2005. 1

[4] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997. 12

[5] P. J. Huber. Robust estimation of a location parameter. *Annals of Statistics*, 53(1):492–518, 1964. 6

[6] D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning (ECML)*, pages 4–15, 1998. 13

[7] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. 5, 6

[8] A. Y. Ng and M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 841–848, 2001. 12

[9] F. J. Samaniegos. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer Science & Business Media, 2010. 1

[10] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 4

[11] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. 5

[12] A. N. Tikhonov and V. I. Arsenin. *Solutions of Ill-posed Problems*. Winston, 1977. 4

[13] G. I. Webb, J. R. Boughton, and Z. Wang. Not so naive bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24, 2005. 12

[14] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. 5

[15] H. Zhang. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 562–567, 2004. 13

[16] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. 5

[17] 周志华. 机器学习. 清华大学出版社, 2016. 5, 7, 12