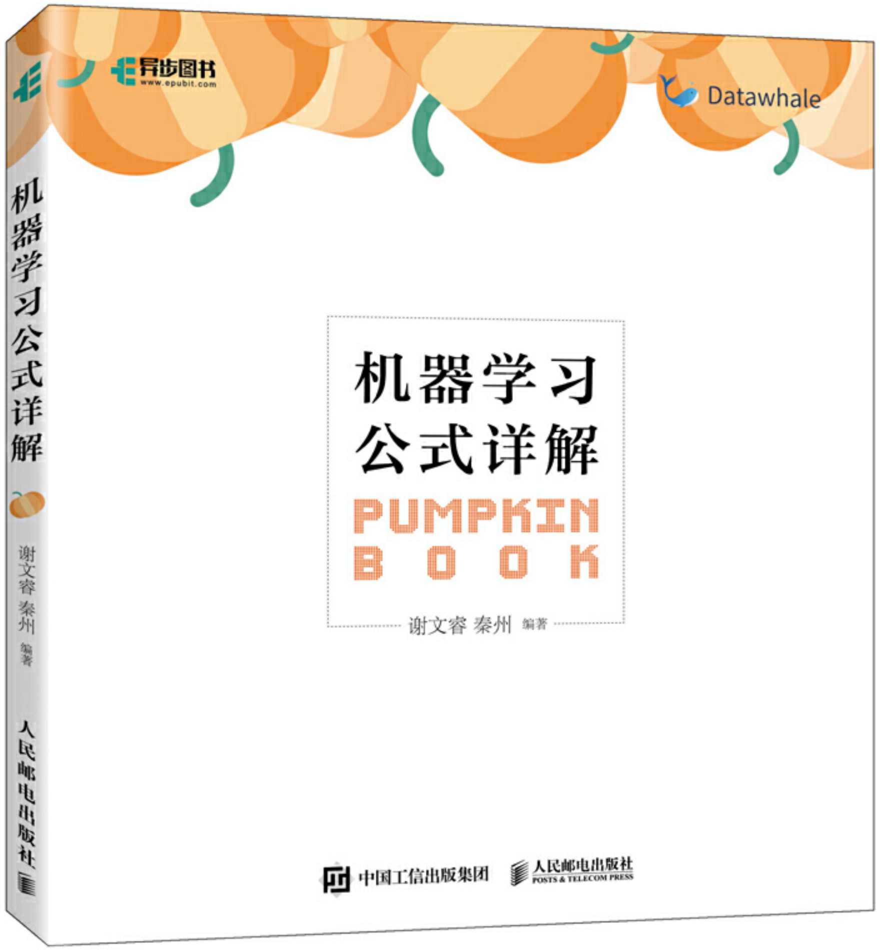

《机器学习公式详解》（南瓜书）

第3章 对数几率回归

本节主讲：谢文睿

本节大纲

Datawhale 异步社区

西瓜书对应章节：3.3

1. 算法原理
2. 损失函数的极大似然估计推导
3. 损失函数的信息论推导

算法原理

在线性模型的基础上套一个映射函数来实现分类功能

拓展阅读：https://sm1les.com/2019/01/17/logistic-regression-and-maximum-entropy/

极大似然估计

第一步：确定概率质量函数（概率密度函数）

已知离散型随机变量$y \in \{0,1\}$取值为1和0的概率分别建模为

$$p(y=1|\boldsymbol{x})=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}=\frac{e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}$$

$$p(y=0|\boldsymbol{x})=1-p(y=1|\boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}$$

为了便于讨论，令$\boldsymbol{\beta}=(\boldsymbol{w};b),\hat{\boldsymbol{x}}=(\boldsymbol{x};1)$，则上式可简写为

$$p(y=1|\hat{\boldsymbol{x}};\boldsymbol{\beta})=\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}=p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})$$

$$p(y=0|\hat{\boldsymbol{x}};\boldsymbol{\beta})=\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}=p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta})$$

极大似然估计

由以上概率取值可推得随机变量$y \in \{0,1\}$的概率质量函数为

$$p(y|\hat{\boldsymbol{x}};\boldsymbol{\beta}) = y \cdot p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta}) + (1-y) \cdot p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta}) \quad \text{此即为公式3.26}$$

或者为

$$p(y|\hat{\boldsymbol{x}};\boldsymbol{\beta}) = [p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})]^y [p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta})]^{1-y}$$

接下来的讲解采用第一种形式（西瓜书中所采用的），采用第二种形式来进行接下来的推导其实更简单，具体参见南瓜书式（3.27）的解析~

极大似然估计

第二步：写出似然函数

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} p(y_i|\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})$$

对数似然函数为

$$\ell(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{m} \ln p(y_i|\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \ln\left(y_i p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) + (1-y_i)p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)$$

极大似然估计

Datawhale 异步社区

将$p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}, p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})=\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}$代入上式可得

$$
\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{m}\ln\left(\frac{y_i e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}+\frac{1-y_i}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right) \\
&= \sum_{i=1}^{m}\ln\left(\frac{y_i e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}+1-y_i}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right) \\
&= \sum_{i=1}^{m}\left(\ln(y_i e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}+1-y_i)-\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})\right)
\end{aligned}
$$

极大似然估计

由于$y_i \in \{0,1\}$，则

$$\ell(\boldsymbol{\beta})=\begin{cases}\sum_{i=1}^{m}(-\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})), & y_i=0\\ \sum_{i=1}^{m}(\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})), & y_i=1\end{cases}$$

两式综合可得

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{m}\left(y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})\right)$$

由于损失函数通常是以最小化为优化目标，因此可以将最大化$\ell(\boldsymbol{\beta})$等价转化为最小化$\ell(\boldsymbol{\beta})$的相反数$-\ell(\boldsymbol{\beta})$，此即为公式(3.27)

信息论

信息论：以概率论、随机过程为基本研究工具，研究广义通信系统的整个过程。常见的应用有无损数据压缩（如ZIP文件）、有损数据压缩（如MP3和JPEG）等，本节仅引用部分精华内容。

自信息：

$$I(X) = -\log_b p(x)$$

当$b=2$时单位为bit，当$b=e$时单位为nat

信息熵（自信息的期望）：度量随机变量X的不确定性，信息熵越大越不确定

$$H(X) = E[I(X)] = -\sum_x p(x)\log_b p(x) \qquad \text{（此处以离散型为例）}$$

计算信息熵时约定：若$p(x)=0$，则$p(x)\log_b p(x) = 0$（最大值和最小值的严格数学分析留到决策树讲解）

信息论

相对熵（KL散度）：度量两个分布的差异，其典型使用场景是用来度量理想分布$p(x)$和模拟分布$q(x)$之间的差异。

$$
\begin{aligned}
D_{KL}(p\|q) &= \sum_x p(x)\log_b\left(\frac{p(x)}{q(x)}\right) \\
&= \sum_x p(x)\left(\log_b p(x) - \log_b q(x)\right) \\
&= \sum_x p(x)\log_b p(x) - \sum_x p(x)\log_b q(x)
\end{aligned}
$$

其中$-\sum_x p(x)\log_b q(x)$称为交叉熵。

信息论

从机器学习三要素中"策略"的角度来说，与理想分布最接近的模拟分布即为最优分布，因此可以通过最小化相对熵这个策略来求出最优分布。由于理想分布$p(x)$是未知但固定的分布（频率学派的角度），所以$\sum_x p(x)\log_b p(x)$为常量，那么最小化相对熵就等价于最小化交叉熵$-\sum_x p(x)\log_b q(x)$

信息论

以对数几率回归为例，对单个样本y_i来说，它的理想分布是

$$p(y_i)=\begin{cases}p(1)=1,p(0)=0, & y_i=1\\ p(1)=0,p(0)=1, & y_i=0\end{cases}$$

它现在的模拟分布是

$$q(y_i)=\begin{cases}\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}=p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta}), & y_i=1\\ \frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}=p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta}), & y_i=0\end{cases}$$

信息论

那么单个样本y_i的交叉熵为

$$-\sum_{y_i} p(y_i) \log_b q(y_i)$$

$$-p(1) \cdot \log_b p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) - p(0) \cdot \log_b p_0(\hat{\boldsymbol{x}}; \boldsymbol{\beta})$$

$$-y_i \cdot \log_b p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) - (1 - y_i) \cdot \log_b p_0(\hat{\boldsymbol{x}}; \boldsymbol{\beta})$$

令$b = e$

$$-y_i \ln p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) - (1 - y_i) \ln p_0(\hat{\boldsymbol{x}}; \boldsymbol{\beta})$$

信息论

Datawhale 异步社区

全体训练样本的交叉熵为

$$\sum_{i=1}^{m}\left[-y_i \ln p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) - (1-y_i)\ln p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right]$$

$$\sum_{i=1}^{m}\left\{-y_i\left[\ln p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}) - \ln p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right] - \ln\left(p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)\right\}$$

$$\sum_{i=1}^{m}\left[-y_i \ln\left(\frac{p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})}{p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})}\right) - \ln\left(p_0(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)\right]$$

$$\sum_{i=1}^{m}\left[-y_i \ln\left(\frac{\frac{e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}}{\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}}\right) - \ln\left(\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}}}\right)\right]$$

$$\sum_{i=1}^{m}\left[-y_i \ln\left(e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)-\ln\left(\frac{1}{1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}}\right)\right]$$

$$\sum_{i=1}^{m}\left(-y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i+\ln(1+e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i})\right)$$

此即为公式3.27

BTW

对数几率回归算法的机器学习三要素：

1. 模型：线性模型，输出值的范围为$[0,1]$，近似阶跃的单调可微函数
2. 策略：极大似然估计，信息论
3. 算法：梯度下降，牛顿法

预告

下一节：二分类线性判别分析

西瓜书对应章节：3.4

结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 关注公众号【Datawhale】，发送【南瓜书】三个字获取机器人“小豚”的微信二维码
2. 添加“小豚”为微信好友，然后对“小豚”发送【南瓜书】三个字即可自动邀请进群、