线性模型

[[TOC]]

0.参考资料：

-
- ROC可以更聚焦于模型本身，降低测试集带来的干扰

00 补充数学知识：

1. 什么是方差的无偏估计

参考：

- 为什么样本方差（sample variance）的分母是 n-1
-

随机变量期望已知时，计算方差

TBD 需要补充中心极限定理

> 首先，我们假定随机变量 X 的数学期望 μ 是已知的，然而方差 σ^2 未知。在这个条件下，根据方差的定义我们有
>
> $$\mathbb{E}\Big[(X_i-\mu)^2\Big]=\sigma^2,\quad \forall i=1,\ldots,n,$$
>
> 由此可得
>
> $$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(X_i-\mu\Big)^2\Big]=\sigma^2\,.$$
>
> **因此 $\frac{1}{n}\sum_{i=1}^{n}\Big(X_i-\mu\Big)^2$ 是方差 σ^2 的一个无偏估计，注意式中的分母不偏不倚正好是 n！**
>
> 这个结果符合直觉，并且在数学上也是显而易见的。

随机变量期望未知时，方差的有偏估计

现在，我们考虑随机变量 X 的数学期望 μ 是未知的情形。这时，我们会倾向于无脑直接用样本均值 $\bar{X}$ 替换掉上面式子中的 μ。这样做有什么后果呢？后果就是，

如果直接使用 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 作为估计，那么你会倾向于低估方差！

这是因为：

eqnarray%7D&width=40" />

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\mu)+(\mu-\bar{X})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-\bar{X})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\bar{X}-\mu)(\mu-\bar{X})+(\mu-\bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\bar{X})^2
\end{aligned}$$

换言之，除非正好 $\bar{X}=\mu$，否则我们一定有

$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 < \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$，

而不等式右边的那位才是的对方差的“正确”估计！

这个不等式说明了，为什么直接使用 $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ 会导致对方差的低估。

方差无偏估计

参考为什么样本方差（sample variance）的分母是 n-1？ - 马同学的回答 - 知乎

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\cdot n\cdot(\overline{X}-\mu)+(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-2(\overline{X}-\mu)^2+(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right]-\mathrm{E}\left[(\overline{X}-\mu)^2\right] \\
&= \sigma^2-\mathrm{E}\left[(\overline{X}-\mu)^2\right]
\end{aligned}$$

其中（证明见Prove that $E (\overline{X} - \mu)^2 = \frac{1}{n}\sigma^2$）：

$$\mathrm{E}\left[(\overline{X}-\mu)^2\right]=\frac{1}{n}\sigma^2.$$

所以:

$$E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2]=\sigma^2-\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2$$

也就是说，低估了 $\frac{1}{n}\sigma^2$ ，进行一下调整:

$$\frac{n}{n-1}E[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2]=E[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2]=\sigma^2$$

因此使用下面这个式子进行估计，得到的就是无偏估计:

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

最新文章请查看（可能会有后继更新）：为什么样本方差的分母是n-1?

2. 先验后验

$$p(Y|X)=\frac{\overset{p(X,Y)}{\boxed{p(X|Y)p(Y)}}}{p(X)}$$

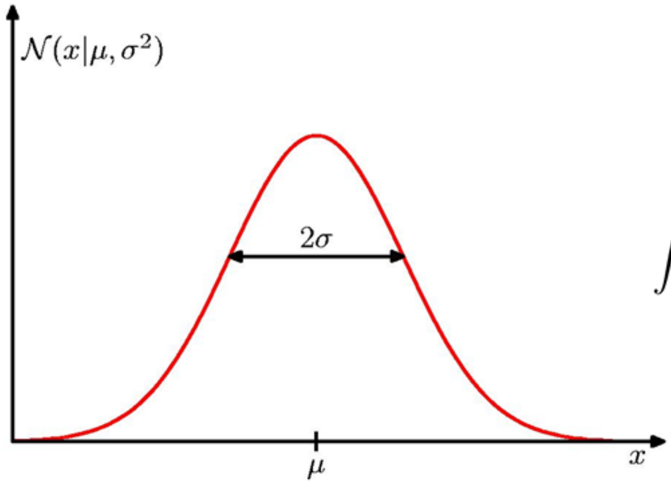
$$p(X)=\sum_{Y}p(X|Y)p(Y)$$

后验		似然		先验
posterior	$\propto$	**likelihood**	$\times$	**prior**
$p(Y\|X)$		$p(X\|Y)$		$p(Y)$

3. 概率分布

连续变量一种最重要的概率分布

$$\mathcal{N}\left(x|\mu,\sigma^2\right)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(x|\mu,\sigma^2)>0$$

$$\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x=1$$

1. 线性回归

1.1 视角1：直接选择均方误差为损失函数，最小化损失(最小二乘法)

线性回归试图学得

$$f(x_i)=wx_i+b,\ 使得\ f(x_i)\simeq y_i\ . \tag{3.3}$$

均方误差亦称平方损失 (square loss).

如何确定 w 和 b 呢? 显然, 关键在于如何衡量 $f(x)$ 与 y 之间的差别. 2.3 节介绍过, 均方误差 (2.2)是回归任务中最常用的性能度量, 因此我们可试图让均方误差最小化, 即

w^*, b^* 表示 w 和 b 的解.

$$\begin{aligned}(w^*,b^*)&=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}\left(f\left(x_i\right)-y_i\right)^2\\&=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}(y_i-wx_i-b)^2\ .\end{aligned} \tag{3.4}$$

均方误差有非常好的几何意义，它对应了常用的欧几里得距离或简称“欧氏距离”(Euclidean distance). 基于均方误差最小化来进行模型求解的方法称为“最小二乘法”(least square method). 在线性回归中, 最小二乘法就是试图找到一条直线, 使所有样本到直线上的欧氏距离之和最小.

最小二乘法用途很广, 不仅限于线性回归.

求解 w 和 b 使 $E_{(w,b)}=\sum_{i=1}^{m}(y_i-wx_i-b)^2$ 最小化的过程, 称为线性回归模型的最小二乘“参数估计”(parameter estimation). 我们可将 $E_{(w,b)}$ 分别对 w 和 b 求导, 得到

这里 $E_{(w,b)}$ 是关于 w 和 b 的凸函数, 当它关于 w 和 b 的导数均为零时, 得到 w 和 b 的最优解.

对区间 $[a,b]$ 上定义的函数 f, 若它对区间中任意两点 x_1,x_2 均有 $f(\frac{x_1+x_2}{2})\leqslant\frac{f(x_1)+f(x_2)}{2}$, 则称 f 为区间 $[a,b]$ 上的凸函数.

U 形曲线的函数如 $f(x)=x^2$, 通常是凸函数.

对实数集上的函数, 可通过求二阶导数来判别: 若二阶导数在区间上非负, 则称为凸函数; 若二阶导数在区间上恒大于 0, 则称为严格凸函数.

$$\frac{\partial E_{(w,b)}}{\partial w}=2\left(w\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}(y_i-b)\,x_i\right)\ , \tag{3.5}$$

$$\frac{\partial E_{(w,b)}}{\partial b}=2\left(mb-\sum_{i=1}^{m}(y_i-wx_i)\right)\ , \tag{3.6}$$

然后令式(3.5)和(3.6)为零可得到 w 和 b 最优解的闭式(closed-form)解

$$w=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}\ , \tag{3.7}$$

1.2 视角2：假定模型输出含有高斯白噪声，高斯函数分布为似然函数，极大似然估计

重新考查曲线拟合问题

给定x的条件下t的高斯条件概率分布

$$p(t|x,\mathbf{w},\beta)=\mathcal{N}\left(t|y(x,\mathbf{w}),\beta^{-1}\right)$$

噪声

$$t=y(\mathbf{x},\mathbf{w})+\epsilon$$

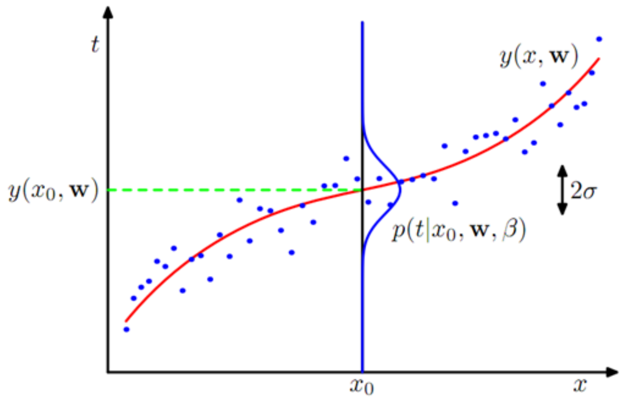


图 1.16: 用图形说明了公式（1.60）给出的给定x的条件下t的高斯条件概率分布，其中均值为多项式函数$y(x,\boldsymbol{w})$，精度由参数β给出，它与方差的关系为$\beta^{-1}=\sigma^2$。

分布的均值为**y(x, w)**

$$y(x,\boldsymbol{w})=w_0+w_1x+w_2x^2+\ldots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

最大似然 估计参数 $\mathbf{w}$ 和 β

数据：$\mathbf{x}=(x_1,\ldots,x_N)^T \quad \mathbf{t}=(t_1,\ldots,t_N)^T$

似然函数：$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}\left(t_n|y(x_n,\mathbf{w}),\beta^{-1}\right)$

取对数：$\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=\underbrace{-\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2}_{\beta E(\mathbf{w})}+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)$

Determine $\mathbf{w}_{\mathrm{ML}}$ by minimizing sum-of-squares error, $E(\mathbf{w})$.

For w:

min: $\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2$

For β: $\frac{1}{\beta_{\mathrm{ML}}}=\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}_{\mathrm{ML}})-t_n\}^2$

1.3 正则化其实是对模型参数的最大后验

最大后验MAP 向贝叶斯迈进

引入w上的先验分布: α是先验分布的精度

$$p(\mathbf{w}|\alpha)=\mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I})=\left(\frac{\alpha}{2\pi}\right)^{(M+1)/2}\exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

max: $p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta)\propto p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$ 最大后验

min: $\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w})-t_n\}^2+\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$

最大化后验概率 等价于 最小化 正则化的平方和误差函数

正则化参数为$\lambda=\frac{\alpha}{\beta}$

2. 逻辑回归--对数几率函数为输出，极大似然法

核心在于，如何从对数几率函数推导出似然函数，再推导出损失函数与导数

- 逻辑回归其实是想用线性模型去完成一个分类任务
- 因此，选择对数几率函数（sigmoid是形似S的函数，对数几率函数是代表）作为模型的输出，就可以起到二分类的作用
- 详细的推导见《机器学习》--周志华

2.1 模型的输出

Sigmoid 函数即形似 S 的函数. 对率函数是 Sigmoid 函数最重要的代表, 在第 5 章将看到它在神经网络中的重要作用.

从图 3.2 可看出, 对数几率函数是一种“Sigmoid 函数”, 它将 z 值转化为一个接近 0 或 1 的 y 值, 并且其输出值在 $z=0$ 附近变化很陡. 将对数几率函数作为 $g^{-}(\cdot)$ 代入式(3.15), 得到

$$y=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}} . \tag{3.18}$$

类似于式(3.14), 式(3.18)可变化为

$$\ln\frac{y}{1-y}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b . \tag{3.19}$$

若将 y 视为样本 $\boldsymbol{x}$ 作为正例的可能性, 则 $1-y$ 是其反例可能性, 两者的比值

$$\frac{y}{1-y} \tag{3.20}$$

称为“几率”(odds), 反映了 $\boldsymbol{x}$ 作为正例的相对可能性. 对几率取对数则得到“对数几率”(log odds, 亦称 logit)

$$\ln\frac{y}{1-y} . \tag{3.21}$$

把式子中的y视为类别概率的话，我们就可以得出模型的似然函数

下面我们来看看如何确定式(3.18)中的 $\boldsymbol{w}$ 和 b. 若将式(3.18)中的 y 视为类后验概率估计 $p(y=1\mid\boldsymbol{x})$, 则式(3.19)可重写为

$$\ln\frac{p(y=1\mid\boldsymbol{x})}{p(y=0\mid\boldsymbol{x})}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b . \tag{3.22}$$

显然有

$$p(y=1\mid\boldsymbol{x})=\frac{e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}} , \tag{3.23}$$

$$p(y=0\mid\boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}} . \tag{3.24}$$

2.2 模型的损失函数及梯度

使用极大似然法，最大化模型的对数似然

于是，我们可通过"极大似然法"(maximum likelihood method)来估计 $\boldsymbol{w}$ 和 b. 给定数据集 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$，对率回归模型最大化"对数似然"(log-likelihood)

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) \ , \tag{3.25}$$

即令每个样本属于其真实标记的概率越大越好. 为便于讨论，令 $\boldsymbol{\beta} = (\boldsymbol{w}; b)$, $\hat{\boldsymbol{x}} = (\boldsymbol{x}; 1)$，则 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$ 可简写为 $\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$. 再令 $p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) = p(y = 1 \mid \hat{\boldsymbol{x}}; \boldsymbol{\beta})$, $p_0(\hat{\boldsymbol{x}}; \boldsymbol{\beta}) = p(y = 0 \mid \hat{\boldsymbol{x}}; \boldsymbol{\beta}) = 1 - p_1(\hat{\boldsymbol{x}}; \boldsymbol{\beta})$，则式(3.25) 中的似然项可重写为

$$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) = y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) + (1 - y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) \ . \tag{3.26}$$

将式(3.26)代入(3.25)，并根据式(3.23)和(3.24)可知，最大化式(3.25)等价于最小化

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i + \ln \left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_i} \right) \right) \ . \tag{3.27}$$

而后求出导数就行了

其中关于 $\boldsymbol{\beta}$ 的一阶、二阶导数分别为

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i (y_i - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})) \ ,$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\mathrm{T}} p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})(1 - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})) \ .$$