3.1 基本形式

3.1.1 定义

给定由 d 个属性描述的示例 $\boldsymbol{x}=(x_1;x_2;...;x_d)$, 其中 x_i 是 $\boldsymbol{x}$ 在第 i 个属性上的取值, 线性模型(linear model)试图学得一个通过属性的线性组合来进行预测的函数,即

$$f(\boldsymbol{x})=w_1x_1+w_2x_2+...+w_dx_d+b \tag{3.1}$$

向量形式:

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \tag{3.2}$$

其中 $\boldsymbol{w}=(w_1;w_2;...;w_d)$. $\boldsymbol{w}$ 和 b 学得之后, 模型就能确定下来.

3.1.2 线性模型的意义

线性模型是基础, .许多功能更为强大的非线性模型 (nonlinear model)可在线性模型的基础上 通过引入**层级结构**或**高维映射**而得.

3.2 线性回归

给定数据集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),...,(\boldsymbol{x}_m,y_m)\}$， 其中 $\boldsymbol{x}_i=(x_{i1};x_{i2};...;x_{id}),y_i\in\mathbb{R}$. "线性回归" (linear regression)试图学得一个线性模型以尽可能准确地预测实值输出标记.

3.2.1 简单形式的线性回归

最简单的情形:输入属性的数目只有一个.此时忽略关于属性的下标 ,即 $D=\{(x_i,y_i)\}_{i=1}^m$,其中 $x_i\in\mathbb{R}$.

线性回归试图学得

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

1 属性的转化

对离散属性，若属性值间**存在"序" (order)关系**，可通过连续化将其转化为**连续值**."身高"的取值"高" "矮"可转化为 $\{1.0,0.0\}$."高" "中" "低"可转化为 $\{1.0,0.5,0.0\}$; 若属性值间**不存在序关系**，假定有 k 个属性值，则通常转化为 **k 维向量**，如属性"瓜类"的取值"西瓜" "南瓜" "黄瓜"可转化为 $(0,0,1),(0,1,0),(1,0,0)$

2 均方误差最小化

试图让均方误差最小化，即

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

对3.4分别对w 和 b 求导,得到

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

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

最后求得最优的闭式(closed-form)解

$$
w=\frac{\sum\limits_{i=1}^{m} y_{i}\left(x_{i}-\bar{x}\right)}{\sum\limits_{i=1}^{m} x_{i}^{2}-\frac{1}{m}\left(\sum\limits_{i=1}^{m} x_{i}\right)^{2}} \tag{3.7}
$$

$$
b=\frac{1}{m} \sum_{i=1}^{m}\left(y_{i}-w x_{i}\right) \tag{3.8}
$$

其中 $\bar{x}=\frac{1}{m} \sum_{i=1}^{m} x_{i}$ 为 x 的均值.

> **推导1:**
>
> $$
> w \sum_{i=1}^{m} x_{i}^{2}-\sum_{i=1}^{m}\left(y_{i}-b\right) x_{i}=0 \tag{1}
> $$
>
> $$
> m b=\sum_{i=1}^{m}\left(y_{i}-w x_{i}\right) \tag{2}
> $$
>
> 由2可以直接求得 b , 再将2带入到1中,有:
>
> $$
> w \sum_{i=1}^{m} x_{i}^{2}-\sum_{i=1}^{m}\left(y_{i}-\frac{1}{m} \sum_{i=1}^{m}\left(y_{i}-w x_{i}\right)\right) x_{i}=0
> $$
>
> $$
> w \sum_{i=1}^{m} x_{i}^{2}-\sum_{i=1}^{m} x_{i} y_{i}+\frac{1}{m} \sum_{i=1}^{m} \sum_{i=1}^{m} x_{i}\left(y_{i}-w x_{i}\right)=0
> $$
>
> $$
> w \sum_{i=1}^{m} x_{i}^{2}-\frac{1}{m} \sum_{i=1}^{m} x_{i} \cdot \sum_{i=1}^{m} x_{i} \cdot w+\frac{1}{m} \sum_{i=1}^{m} x_{i} \cdot \sum_{i=1}^{m} y_{i}-\sum_{i=1}^{m} x_{i} y_{i}
> $$

> $$w\left(\sum_{i=1}^{m} x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2\right)=\sum_{i=1}^{m} x_i y_i-\bar{x} \cdot \sum_{i=1}^{m} y_i$$
>
> 则就可以求得:
>
> $$w=\frac{\sum_{i=1}^{m} y_i\left(x_i-\bar{x}\right)}{\sum_{i=1}^{m} x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}$$

3.2.2 一般形式

更一般的情形是如本节开头的数据集 D， 样本由 d 个属性描述.此时我们试图学得

$$f\left(\boldsymbol{x}_i\right)=\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i+b, \text{使得 } f\left(\boldsymbol{x}_i\right) \simeq y_i,$$

> **注1**: 这里的 $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i$ 和 $\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{w}$ 是**等价**的.

这也就是"多元线性回归"(multivariate linear regression)

1 形式变换

为了简便起见, 把 $\boldsymbol{w}$ 和 b 吸收乳向量形式 $\hat{\boldsymbol{w}}=(\boldsymbol{w}; b)$, 相应的, 把数据集 D 表示为一个 $m \times(d+1)$ 大小的矩阵 $\mathbf{X}$, 其中每行对应于一个示例，该行前 d 个元素对应于示例的 d 个属性值，最后一个元素恒置为 1 ，即

$$\mathbf{X}=\left(\begin{array}{ccccc}x_{11} & x_{12} & \ldots & x_{1 d} & 1 \\ x_{21} & x_{22} & \ldots & x_{2 d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m 1} & x_{m 2} & \ldots & x_{m d} & 1\end{array}\right)=\left(\begin{array}{cc}\boldsymbol{x}_1^{\mathrm{T}} & 1 \\ \boldsymbol{x}_2^{\mathrm{T}} & 1 \\ \vdots & \vdots \\ \boldsymbol{x}_m^{\mathrm{T}} & 1\end{array}\right)$$

再把标记也写成向量形式 $\boldsymbol{y}=\left(y_1 ; y_2 ; \ldots ; y_m\right)$, 则类似于式(3.4), 有

$$\hat{\boldsymbol{w}}^*=\underset{\hat{\boldsymbol{w}}}{\arg \min }(\boldsymbol{y}-\mathbf{X} \hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\mathbf{X} \hat{\boldsymbol{w}}) \tag{3.9}$$

> **注2**: 3.9就是向量的平方形式. 同时类似于注1, $\mathbf{X} \hat{w}$ 展开其中一项, 其实就是 $\boldsymbol{x}_1^{\mathrm{T}} \boldsymbol{w}+b$

2 求解过程

令 $E_{\hat{\boldsymbol{w}}}=(\boldsymbol{y}-\mathbf{X} \hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\mathbf{X} \hat{\boldsymbol{w}})$, 对 $\hat{\boldsymbol{w}}$ 求导得到

$$\frac{\partial E_{\hat{\mathbf{w}}}}{\partial \hat{\boldsymbol{w}}}=2 \mathbf{X}^{\mathrm{T}}(\mathbf{X} \hat{\boldsymbol{w}}-\boldsymbol{y}) \tag{3.10}$$

> **推导2**: 将 $E_{\hat{\boldsymbol{w}}}=(\boldsymbol{y}-\mathbf{X} \hat{\boldsymbol{w}})^{\mathrm{T}}(\boldsymbol{y}-\mathbf{X} \hat{\boldsymbol{w}})$ 展开可以得到:
>
> $E_{\hat{\boldsymbol{w}}}=\boldsymbol{y}^T \boldsymbol{y}-\boldsymbol{y}^T \mathbf{X} \hat{\boldsymbol{w}}-\hat{\boldsymbol{w}}^T \mathbf{X}^T \boldsymbol{y}+\hat{\boldsymbol{w}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{w}} \hat{\boldsymbol{w}}$,再对$\hat{\boldsymbol{w}}$ 进行求导:

> $$\frac{\partial E_{\hat{\boldsymbol{w}}}}{\partial \hat{\boldsymbol{w}}}=\frac{\partial \boldsymbol{y}^{T} \boldsymbol{y}}{\partial \hat{\boldsymbol{w}}}-\frac{\partial \boldsymbol{y}^{T} \mathbf{X} \hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}}-\frac{\partial \hat{\boldsymbol{w}}^{T} \mathbf{X}^{T} \boldsymbol{y}}{\partial \hat{\boldsymbol{w}}}+\frac{\partial \hat{\boldsymbol{w}}^{T} \mathbf{X}^{T} \mathbf{X} \hat{\boldsymbol{w}}}{\partial \hat{\boldsymbol{w}}}$$
>
> 根据向量求导公式可得:
>
> $\frac{\partial E_{\hat{w}}}{\partial \hat{w}}=0-\mathbf{X}^{T} \boldsymbol{y}-\mathbf{X}^{T} \boldsymbol{y}+\left(\mathbf{X}^{T} \mathbf{X}+\mathbf{X}^{T} \mathbf{X}\right) \hat{\boldsymbol{w}}$
>
> 进一步得到:
>
> $\frac{\partial E_{\hat{w}}}{\partial \hat{w}}=2 \mathbf{X}^{T}(\mathbf{X} \hat{\boldsymbol{w}}-\boldsymbol{y})$

> **注3**: 矩阵求导相关参考资料
>
> 矩阵求导-维基百科

进一步, 当 $\mathbf{X}^{\mathrm{T}} \mathbf{X}$ 为满秩矩阵或正定矩阵时, 令(3.10) 为零可以得到:

$$\hat{\boldsymbol{w}}^{*}=\left(\mathbf{X}^{\mathrm{T}} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y} \tag{3.11}$$

令 $\hat{\boldsymbol{x}}_{i}=\left(\boldsymbol{x}_{i}, 1\right)$,则最终的回归模型为:

$$f\left(\hat{\boldsymbol{x}}_{i}\right)=\hat{\boldsymbol{x}}_{i}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}} \mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}} \boldsymbol{y} \tag{3.12}$$

而现实任务中 $\mathbf{X}^{\mathrm{T}} \mathbf{X}$ 常常不是满秩矩阵, 常见的作法是引入正则化(regularization)项.

3 广义线性模型和联系函数

一般的, 考虑单调可微函数 $g(\cdot)$, 令

$$y=g^{-1}\left(w^{T} x+b\right) \tag{3.15}$$

这样的模型称为**"广义线性模型"**(generalized linear model), 其中函数 $g(\cdot)$ 称为**"联系函数"**(link function). 易见, 对数线性回归是广义线性模型在 $g(\cdot)=\ln (\cdot)$ 时的特例.

3.3 对数几率回归

3.3.1 单位阶跃函数

考虑二分类任务， 其输出标记 $y \in\{0,1\}$，而线性回归模型产生的预测值 $z=\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b$是实值，于是, 我们需将实值 z 转换为 0/ 1 值. 最理想的是"单位阶跃函数" (unit-step function) .

$$y=\left\{\begin{array}{cc}0, & z<0 \\ 0.5, & z=0 \\ 1, & z>0\end{array}\right. \tag{3.16}$$

即即若预测值 z 大于零就判为正例, 小于零则判为反例, 预测值为临界值零则可任意判别.

3.3.2 对数几率函数

$$y=\frac{1}{1+e^{-z}} \tag{3.17}$$

将对数几率函数作为 $g^{-}(\cdot)$ 带入到 (3.15), 就可以得到:

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

进行变换可得:

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

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

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

成为"几率"(odds), 反映了 $\boldsymbol{x}$ 作为正例的相对可能性. 对几率取对数则可以得到"对数几率"(log odds,也称logit)

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

3.3.3 对数几率回归中参数的求解

若将式(3.18)中的 y 视为类后验概率估计 $p(y=1|\boldsymbol{x})$, 则(3.19) 可以重写为:

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

由于 $p(y=1|\boldsymbol{x})+p(y=0|\boldsymbol{x})=1$, 则可以求得:

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

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

于是, 我们可通过"极大似然法" (maximum likelihood method)来估计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) .

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

和上面一样, 为便于讨论和简写, 令 $\boldsymbol{\beta}=(\boldsymbol{w};b)$, $\hat{\boldsymbol{x}}=(\boldsymbol{x};1)$, 则 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ 可简写为 $\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}$. 同时再令 $p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})=p(y=1|\hat{\boldsymbol{x}};\boldsymbol{\beta})$, $p_0(\hat{\boldsymbol{x}};\boldsymbol{\beta})=p(y=0|\hat{\boldsymbol{x}};\boldsymbol{\beta})=1-p_1(\hat{\boldsymbol{x}};\boldsymbol{\beta})$, 则式 (3.25)中的似然项可以写成:

$$p\left(y_i|\boldsymbol{x}_i;\boldsymbol{w},b\right)=y_ip_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right)+\left(1-y_i\right)p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right) \tag{3.26}$$

> **推导3**: y_i 只能取0或1, 分别带入即可:
>
> $$p\left(y_i|\boldsymbol{x}_i;\boldsymbol{w},b\right)=\begin{cases}p_1\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right) & \text{if } y_i=1\\ p_0\left(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta}\right) & \text{if } y_i=0\end{cases}$$

将式(3.26)代入到(3.25)中, 且根据(3.23)和(3.24), 则最大化式 (3.25) 等价于**最小化**:

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

> **推导4**: 将式 (3.26) 带入式 (3.25) 可以得到:
>
> $l(\beta)=\sum_{i=1}^{m}\ln(y_ip_1\left(\hat{\boldsymbol{x}}_i;\beta\right)+\left(1-y_i\right)p_0\left(\hat{\boldsymbol{x}}_i;\beta\right))$
>
> 同时, $p_1\left(\hat{\boldsymbol{x}}_i;\beta\right)=\frac{e^{\beta^T\hat{\boldsymbol{x}}_i}}{1+e^{\beta^T\hat{\boldsymbol{x}}_i}}, p_0\left(\hat{\boldsymbol{x}}_i;\beta\right)=\frac{1}{1+e^{\beta^T\hat{\boldsymbol{x}}_i}}$, 代入到上式可得:
>
> $$\begin{aligned}l(\beta)&=\sum_{i=1}^{m}\ln\left(\frac{y_ie^{\beta^T\hat{\boldsymbol{x}}_i}+1-y_i}{1+e^{\beta^T\hat{\boldsymbol{x}}_i}}\right)\\&=\sum_{i=1}^{m}\left(\ln\left(y_ie^{\beta^T\hat{\boldsymbol{x}}_i}+1-y_i\right)-\ln\left(1+e^{\beta^T\hat{\boldsymbol{x}}_i}\right)\right)\end{aligned}$$
>
> 由于 $y_i=0\text{或}1$, 则有:
>
> $$l(\beta)=\begin{cases}\sum_{i=1}^{m}\left(-\ln\left(1+e^{\beta^T\hat{\boldsymbol{x}}_i}\right)\right), & y_i=0\\ \sum_{i=1}^{m}\left(\beta^T\hat{\boldsymbol{x}}_i-\ln\left(1+e^{\beta^T\hat{\boldsymbol{x}}_i}\right)\right), & y_i=1\end{cases}$$
>
> 把两式综合可得:
>
> $l(\beta)=\sum_{i=1}^{m}\left(y_i\beta^T\hat{\boldsymbol{x}}_i-\ln\left(1+e^{\beta^T\hat{\boldsymbol{x}}_i}\right)\right)$
>
> 添加负号即是式(3.27), 也即是最小化

式 (3.27)是关于 β 的高阶可导连续凸函数 , 根据凸优化理论, 用梯度下降法,牛顿法都可以求得最优解. 则可以得到

$$\boldsymbol{\beta}^*=\underset{\boldsymbol{\beta}}{\arg\min}\,\ell(\boldsymbol{\beta}) \tag{3.28}$$

3.4 线性判别分析

暂放

3.5 多分类学习

暂放