§3 线性模型

1. 基本形式

$$f(x)=w_1x_1+w_2x_2+\cdots+w_dx_d+b$$

$$f(x)=w^Tx+b$$

向量形式

2. 线性回归

均方误差/平方损失(square loss) — 性能度量　对应欧氏距离

$$(w^*,b^*)=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}(f(x_i)-y_i)^2=\underset{(w,b)}{\arg\min}\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

模型求解：最小二乘法(least square method) — 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。

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

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

$$\Rightarrow\begin{cases}w=\dfrac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}\\[2ex] b=\dfrac{1}{m}\sum_{i=1}^{m}(y_i-wx_i)\end{cases}$$

一般情形　$f(x_i)=w^Tx_i+b$

$$\hat{w}^*=\underset{\hat{w}}{\arg\min}\,(y-X\hat{w})^T(y-X\hat{w})\qquad \frac{\partial E_{\hat{w}}}{\partial \hat{w}}=2X^T(X\hat{w}-y)$$

$$\Rightarrow \hat{w}^*=(X^TX)^{-1}X^Ty$$

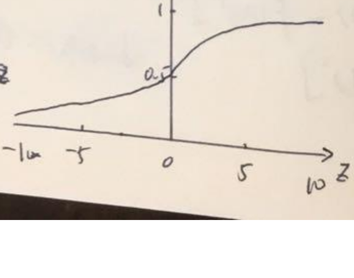
$$f(\hat{x}_i)=\hat{x}_i^T(X^TX)^{-1}X^Ty$$

解出多个$\hat{w}_i$，选择哪个作为输出，引入正则化(regularization)

对数线性回归　$\ln y=w^Tx+b$　　$y=g^{-1}(w^Tx+b)$ — 广义线性模型 GLM

3. 对数几率回归

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



$$y=\frac{1}{1+e^{-(w^Tx+b)}} \qquad \ln\frac{y}{1-y}=w^Tx+b$$

$$\ln\frac{p(y=1|x)}{p(y=0|x)}=w^Tx+b \Rightarrow \begin{cases} p(y=1|x)=\dfrac{e^{w^Tx+b}}{1+e^{w^Tx+b}} \\ p(y=0|x)=\dfrac{1}{1+e^{w^Tx+b}} \end{cases}$$

极大似然法 maximum likelihood method 估计 w. b

$$l(w,b)=\sum_{i=1}^{m}\ln p(y_i|x_i;w,b)$$

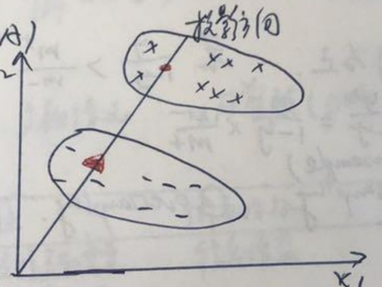
$$l(\beta)=\sum_{i=1}^{m}\left(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i})\right)$$

$$\beta^*=\arg\min_\beta l(\beta) \qquad \beta^{t+1}=\beta^t-\left(\frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta^T}\right)^{-1}\cdot\frac{\partial l(\beta)}{\partial\beta}$$

$$\frac{\partial l(\beta)}{\partial\beta}=-\sum_{i=1}^{m}\hat{x}_i\left(y_i-p_1(\hat{x}_i;\beta)\right)$$

$$\frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta^T}=\sum_{i=1}^{m}\hat{x}_i\hat{x}_i^Tp_1(\hat{x}_i;\beta)(1-p_1(\hat{x}_i;\beta))$$

4. 线性判别分析 (LDA)



• ▲ 两中心点尽可能远

x — 尽可能靠近中心点近

$$J=\frac{\|w^T\mu_0-w^T\mu_1\|_2^2}{w^T\Sigma_0w+w^T\Sigma_1w}=\frac{w^T(\mu_0-\mu_1)(\mu_0-\mu_1)^Tw}{w^T(\Sigma_0+\Sigma_1)w}$$

异类尽可能距离大
同类尽可能距离小

$$S_w=\Sigma_0+\Sigma_1=\sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T+\sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T$$ within class

$$S_b=(\mu_0-\mu_1)(\mu_0-\mu_1)^T \Rightarrow J=\frac{w^TS_bw}{w^TS_ww}$$
between class

S_b/S_w：广义瑞利商 generalized Rayleigh quotient

确定w. 令 $w^TS_ww=1$　　$\min -w^TS_bw$.

引入拉格朗日乘子法.　$S_bw=\lambda S_ww$　$\Rightarrow w=S_w^{-1}(\mu_0-\mu_1)$

$S_bw=\lambda(\mu_0-\mu_1)$

$S_b W = \lambda(\mu_0 - \mu_1)$ $W = S_W^{-1}(\mu_0 - \mu_1)$

$S_W \to$ SVD分解 $S_W = U\Sigma V^T$ $S_W^{-1} = V\Sigma^{-1}U^T$

$S_t = S_b + S_W = \sum_{i=1}^{m}(x_i-\mu)(x_i-\mu)^T$ $S_W = \sum_{i=1}^{N} S_{W_i}$

$S_{W_i} = \sum_{x \in X_i}(x-\mu_i)(x-\mu_i)^T$ $S_b = S_t - S_W = \sum_{i=1}^{N} m_i(\mu_i-\mu)(\mu_i-\mu)^T$

$\max_W \frac{tr(W^T S_b W)}{tr(W^T S_W W)}$ $S_b W = \lambda S_W W$

5. 多分类学习

OvO: one Vs. one OvR: One Vs Rest MvM: Many Vs. Many

$N(N-1)/2$ 分类器 N个分类器

MvM技术: Error Correcting Output Codes (ECOC)

6 类别不平衡问题

若 $\frac{y}{1-y} > 1$ 则预测为正. 若 $\frac{y}{1-y} > \frac{m^+}{m^-}$ 则预测为正

rescaling 再缩放 $\frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+}$

① 去除一些反例. undersampling (easy ensemple) ② oversampling. 正例进行插值 (SMOTE)

③ threshold moving

防止过拟合 ridge, Lasso 正则

ridge 是高斯先验下的最大后验估计. lasso 是拉普拉斯先验下的最大后验估计

ridge: $\min_\theta \sum_i (y_i - f_\theta(x_i))^2$ 且 $\theta^2 \le r$ lasso: $\min_\theta \sum_i (y_i - f_\theta(x_i))^2$ 且 $|\theta| \le r$