3.1 基本形式

$x=(x_1;x_2;\cdots x_d)$.

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

$f(x)=w^Tx+b$. $\quad w=(w_1;w_2;\cdots;w_d)$.

3.2 线性回归.

$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$, 其中 $x_i=(x_{i1};x_{i2};\cdots;x_{id})$, $y_i\in R$.

输入属性为1个: $D=\{(x_i,y_i)\}_{i=1}^m$. $x_i\in R$. 由 $f(x_i)=wx_i+b$. 使 $f(x_i)\simeq y_i$. ← 回归目标.

均方误差: $(w^*,b^*)=\underset{(w,b)}{\arg\min}\sum_{i=1}^m(f(x_i)-y_i)^2$ ← 欧式距离. (使均方误差最小化的w,b的值.)

基于均方误差最小化为最小二乘法.

已知 $E(w,b)=\sum_{i=1}^m(y_i-wx_i-b)^2$ 对 w,b 求导, 得 $\begin{cases}\frac{\partial E(w,b)}{\partial w}=2\left(w\sum_{i=1}^m x_i^2-\sum_{i=1}^m(y_i-b)x_i\right)\\ \frac{\partial E(w,b)}{\partial b}=2\left(mb-\sum_{i=1}^m(y_i-wx_i)\right)\end{cases}$.

令导数为0, 得解.

多元线性回归: $f(x_i)=w^Tx_i+b$. 使 $f(x_i)\simeq y_i$. $\hat{w}=(w;b)$. D变为 $m\times(d+1)$ 的矩阵 X

$$X=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1d}&1\\x_{21}&x_{22}&\cdots&x_{2d}&1\\\vdots&\vdots&\ddots&\vdots&\vdots\\x_{m1}&x_{m2}&\cdots&x_{md}&1\end{pmatrix}=\begin{pmatrix}x_1^T&1\\x_2^T&1\\\vdots&\vdots\\x_m^T&1\end{pmatrix}$$

$y=(y_1;y_2;\cdots;y_m)$.

$\hat{w}^*=\underset{\hat{w}}{\arg\min}(y-X\hat{w})^T(y-X\hat{w})$.

令 $E_{\hat{w}}=(y-X\hat{w})^T(y-X\hat{w})$ 对 $\hat{w}$ 求导, 得 $\frac{\partial E_{\hat{w}}}{\partial\hat{w}}=2X^T(X\hat{w}-y)$. 矩阵的微分公式: $\frac{\partial a^Tx}{\partial x}=\frac{\partial x^Ta}{\partial x}=a$, $\frac{\partial x^TAx}{\partial x}=(A+A^T)x$

当 X^TX 为满秩矩阵. $\hat{w}^*=(X^TX)^{-1}X^Ty$ 令 $\hat{x}_i=(x_i,1)$. 则 $f(\hat{x}_i)=\hat{x}_i^T(X^TX)^{-1}X^Ty$

当 X^TX 为非满秩矩阵. (列 >> 行). 正则化

广义线性回归: $y=g^{-1}(w^Tx+b)$. $g(\cdot)$ 联系函数. $g(\cdot)=\ln(\cdot)$ $\ln y=w^Tx+b$. 对数

3.3. 对数几率回归. (其实是分类).

二分类. $y\in\{0,1\}$. $z=w^Tx+b$.

单位阶跃函数. $y=\begin{cases}0&z<0\\0.5&z=0\\1&z>0\end{cases}$ → 变 $\frac{1}{1+e^{-z}}$ Sigmoid

$$\begin{cases} 0 & z<0 \\ 0.5 & z=0 \\ 1 & z>0 \end{cases}$$

$y=\frac{1}{1+e^{-(w^Tx+b)}}$ $\quad \ln\frac{y}{1-y}=w^Tx+b$. $\quad \ln\frac{y}{1-y}$ 对数几率

正例 $\frac{y}{1-y}$ 反例

↓ y视为后验概率估计.

$$\ln\frac{p(y=1|x)}{p(y=0|x)}=w^Tx+b.$$

$$p(y=1|x)=\frac{e^{w^Tx+b}}{1+e^{w^Tx+b}}$$

$$\therefore p(y=0|x)=\frac{1}{1+e^{w^Tx+b}}$$

用极大似然法EM估计 w,b.

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

每个样本属于其真实标记概率越大越好.

令 $\beta=(w,b)$. $\hat{x}=(x;1)$. 则 $w^Tx+b=\beta^T\hat{x}$.

令 $p_1(\hat{x};\beta)=p(y=1|\hat{x};\beta)$. $p_0(\hat{x};\beta)=p(y=0|\hat{x};\beta)=1-p_1(\hat{x};\beta)$.

$=[p_1(x_i;\beta)]^{y_i}[p_0(\hat{x_i};\beta)]^{1-y_i}$ 也可得解.

则 $p(y_i|x_i;w,b)=y_i p_1(\hat{x}_i;\beta)+(1-y_i)p_0(\hat{x}_i;\beta)$.

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

梯度下降. $\beta^*=\arg\min_\beta \ell(\beta)$

3.4. 线性判别分析. LDA.

给定训练集, 同类在直线的投影点尽可能接近, 异类样例在投影点尽可能远.

$D=\{(x_i,y_i)\}_{i=1}^m$, $y_i\in\{0,1\}$. X_i 集合. μ_i 均值. Σ_i 协方差矩阵.

将数据投影至直线 w. 则两类样本协方差为 $w^T\Sigma_0 w$, $w^T\Sigma_1 w$.

$w^T\Sigma_0 w+w^T\Sigma_1 w$ 小, 而 $\|w^T\mu_0-w^T\mu_1\|_2^2$ 尽可能大.

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

定义类间散度矩阵 $S_b=(\mu_0-\mu_1)(\mu_0-\mu_1)^T$

$$\therefore J=\frac{w^TS_bw}{w^TS_ww}.$$ 即 S_b, S_w 的广义瑞利商.

由于 w 在分子分母都为二次项, 令 $w^TS_ww=1$. 原式 $=\min_w -w^TS_bw$ s.t. $w^TS_ww=1$.

由拉格朗日乘法, 上式等价于 $S_bw=\lambda S_ww$. 由于 S_bw 方向为 $\mu_0-\mu_1$. $\therefore S_bw=\lambda(\mu_0-\mu_1)$

$\therefore w = S_w^{-1}(\mu_0 - \mu_1)$. 对 S_w 奇异值分解. $S_w = U\Sigma V^T$. 再由 $S_w^{-1} = U\Sigma^{-1}U^T$ 得 S_w^{-1}.

当两类数据同先验. 满足高斯分布且协方差相同. LDA分类最好.

推广: 多分类. N类. 类别数 m_i. 则全局散度矩阵 $S_t = S_b + S_w = \sum_{i=1}^{m}(x_i-\mu)(x_i-\mu)^T$. $S_w = \sum_{i=1}^{N} S_{wi}$. $S_{wi} = \sum_{x\in X_i}(x-\mu_i)(x-\mu_i)^T$

$$S_b = S_t - S_w = \sum_{i=1}^{N} m_i(\mu_i-\mu)(\mu_i-\mu)^T.$$

采用目标 $\max_W \frac{tr(W^T S_b W)}{tr(W^T S_w W)}$. $W \in R^{d\times(N-1)}$. $S_b W = \lambda S_w W$.

↓

投影矩阵. ⟶ 投影到 N-1 维空间 (<<样本属性数).

LDA为经典监督降维.

3.5 多分类学习.

N类. $C_1 \cdots C_N$. 拆解法 → 若干二分类.

- 一对一 → $\frac{N(N-1)}{2}$ 个二分类. 投票产生最终结果.
- 一对其余 → N 个二分类. 一正vs其余反. 置信度最大为分类结果.
- 多对多 → 纠错输出码 ECOC. (有一定容错能力).
  - 编码: 对N类划分 M 次. 每次二分训练. 共 M 个分类器.
  - 解码: 预测样例组成编码. 其中与各自编码比较距离最小的类别为最终结果. (编码矩阵)

$D = \{(x_1,y_1)\cdots(x_m,y_m)\}$. $y_i \in \{C_1,\cdots C_N\}$.

(→ DAG).

3.6 类别不平衡问题.

m^+, m^- 训练集中正反例子数目. $\frac{y}{1-y} > \frac{m^+}{m^-}$ 预测为正例.

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

由于无偏采样假设不成立.
- 欠采样. 去除一些反例. EasyEnsemble. 将反例划为不同集合. 减少信息丢失.
- 过采样. 增加一些正例. 不可简单重复采样. 会造成大合. better: SMOTE
- 阈值移动. 用 $\frac{y'}{1-y'}$ 替代 $\frac{y}{1-y}$. 代价敏感. 将 m^-/m^+ 换为 $cost^+/cost^-$.