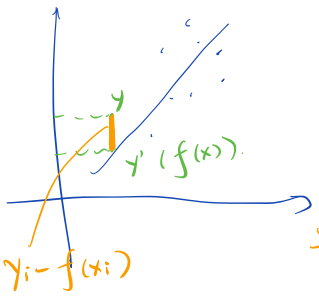
第3章 线性模型.

一. 线性回归.

- 连续值 $f(x) = wx+b$
- 二值离散值 (A: 1, B: 0).

  $f(x) = w_1x_1 + w_2x_2 + b$
- 多值
  - 有序 (A: 1, 2, 3). $f(x) = w_1x_1 + w_2x_2 + w_3x_3 + b$
  - 无序 (1,0,0) (0,1,0) ... $f(x) = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6 + b.$

1. 最小二乘估计: 基于均方误差最小化 模型求解.



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

$$= \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

$\arg\min_{(w,b)}$ → 求使 $E_{(w,b)}$ 最小的 w, b (x, y 已知)

2. 极大似然估计.

估计概率分布的参数值. 先确定分布类型 再求参数 → 极大似然.

离散(连续型)随机变量 X. 概率质量函数 $P(x;\theta)$ (概率密度函数 $p(x;\theta)$). (连续型)

θ — 待估计参数(可有多个).

$x_1, x_2, \cdots x_n$ 来自 X 的 n 个独立同分布样本.

联合概率 $L(\theta)=\prod_{i=1}^{n} P(x_i;\theta)$ —— 样本的似然函数

已知量　未知量

使 $L(\theta)$ 最大的 θ^* 即 θ 的估计值

连乘项 用对数函数化简 → 连加项.

$\ln L(u,\sigma^2)$ 和 $L(u,\sigma^2)$ 最值点相同

$\ln \prod \Rightarrow \sum \ln$

应用:

线性回归: $y=wx+b+\epsilon$ —— 不受控制的随机误差

通常假设 $\epsilon \sim N(0,\sigma^2)$　中心极限定理解释.

$$\Rightarrow p(\epsilon)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

代入 $\epsilon=y-wx-b$

$$\Rightarrow p(y)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-(wx+b))^2}{2\sigma^2}\right)$$

利用极大似然估计 $L(w,b)=\prod_{i=1}^{m} p(y_i)$

$$\ln L(w,b)=\sum_{i=1}^{m}\ln\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-(wx+b))^2}{2\sigma^2}\right)\right]$$

$$=\sum_{i=1}^{m}\ln\frac{1}{\sqrt{2\pi}\sigma}+\sum_{i=1}^{m}\left(-\frac{(y-(wx+b))^2}{2\sigma^2}\right)$$

$$=m\ln\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y-(wx+b))^2$$

m.σ为常数 故 $(w^*, b^*) = \underset{(w,b)}{\operatorname{argmax}} \ln L(w,b)$

$= \underset{(w,b)}{\operatorname{argmin}} \sum_{i=1}^{m} (y_i - wx_i - b)^2$ 公式 3.4.

求对 w.b. ⇒ 多元函数 求最值点.. $E = \sum_{i=1}^{m} (y_i - wx_i - b)^2$

↓

凸函数

凸集：集合 $D \subset \mathbf{R}^n$ $\forall x, y \in D$. 与 $\forall \alpha \in [0,1]$ 有

$\alpha x + (1-\alpha) y \in D$ ⇒ D为凸集.

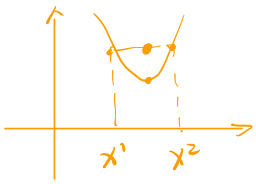
几何意义：集合内任两点连线上任意点均 $\in D$

e.g. $\emptyset$. n维欧式空间 $\mathbf{R}^n$.

凸函数：设D非空凸集. f定义在D上的函数. $\forall x^1, x^2 \in D$. $\alpha \in (0,1)$.

均有 $f(\alpha x^1 + (1-\alpha) x^2) \leq \alpha f(x^1) + (1-\alpha) f(x^2)$

则 f为D上的凸函数.



梯度（多元函数的一阶导数） （列向量）

n元函数 $f(x)$ 对自变量 $x = (x_1, \cdots, x_n)^T$ 的各分量 x_i 的偏导数都存在.

则 $f(x)$ 在 x 处一阶可导

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \Rightarrow f(x) \text{ 在 } x \text{ 处的梯度}$$

Hessian（海塞）矩阵 （多元…二阶导数）.

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & & & \vdots \\ & & & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

定理：设 $D \subset \mathbb{R}^n$ 非空开凸集. $f: D \subset \mathbb{R}^n \to \mathbb{R}$. 且 $f(x)$ 在 D 上二阶…

若 $f(x)$ 的 Hessian 矩阵在 D 上半正定，则 $f(x)$ 是 D 上的凸函数.

（类似一元：判断凹凸性）.

证明 $E(w,b)$ 是关于 w 和 b 的凸函数 $\Rightarrow \nabla^2 E_{(w,b)}$ 正定.

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

$$= 2 \sum_{i=1}^{m} (-x_i) \cdot (y_i - wx_i - b)$$

$$= 2 \left[w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b) x_i \right] \quad 2.5$$

$$\frac{\partial^2 E(w,b)}{\partial w^2} = 2 \frac{\partial}{\partial w} \left[w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b) x_i \right] \text{ 常数项.}$$

$$= 2 \sum_{i=1}^{m} x_i^2$$

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

$$= 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)$$

$$\frac{\partial^2 E(w,b)}{\partial b^2} = 2m$$

$$\frac{\partial^2 E(w,b)}{\partial w \partial b} = \frac{\partial}{\partial b}\left(\frac{\partial E(w,b)}{\partial w}\right)$$

$$= 2\sum_{i=1}^{m} x_i$$

$$\frac{\partial^2 E(w,b)}{\partial b \partial w} = 2\sum_{i=1}^{m} x_i$$

$$\therefore \nabla^2 E(w,b) = \begin{bmatrix} 2\sum_{i=1}^{m} x_i^2 & 2\sum_{i=1}^{m} x_i \\ 2\sum_{i=1}^{m} x_i & 2m \end{bmatrix}$$

半正定矩阵：实对称矩阵所有顺序主子式均为非负.

$$\left|2\sum_{i=1}^{m} x_i^2\right| > 0$$

$$\begin{vmatrix} 2\sum_{i=1}^{m} x_i^2 & 2\sum_{i=1}^{m} x_i \\ 2\sum_{i=1}^{m} x_i & 2m \end{vmatrix} = 4m\sum_{i=1}^{m} x_i^2 - 4\left(\sum_{i=1}^{m} x_i\right)^2$$

$$= 4m\sum_{i=1}^{m} x_i^2 - 4 \cdot m \cdot \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2$$

$$= 4m\left(\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i \bar{x}\right)$$

$$= 4m\sum_{i=1}^{m}(x_i^2 - 2x_i\bar{x} + x_i\bar{x})$$

$$\left(\sum_{i=1}^{m} x_i\bar{x} = \bar{x} \cdot m \cdot \frac{1}{m}\sum_{i=1}^{m} x_i = m \cdot \bar{x}^2 = \sum_{i=1}^{m} \bar{x}^2\right)$$

$$= 4m\sum_{i=1}^{m}(x_i - \bar{x})^2 \geq 0.$$

$\therefore \nabla^2 E(w,b)$ 是半正定矩阵 $\Rightarrow E(w,b)$ 凸函数.

凸充分性定理（凸函数极值）：$f: \mathbb{R}^n \to \mathbb{R}$ 凸函数.

x^* 全局对 $\Leftrightarrow \nabla f(x^*) = 0$

$$\nabla E_{(w,b)} = \begin{bmatrix} \frac{\partial E_{(w,b)}}{\partial w} \\ \frac{\partial E_{(w,b)}}{\partial b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{cases} 2\left[w \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} (y_i - b) x_i \right] = 0 \\ 2\left(mb - \sum_{i=1}^{m} (y_i - w x_i) \right) = 0 \end{cases}$$

$$\hookrightarrow \quad b = \frac{1}{m} \sum_{i=1}^{m} (y_i - w x_i) \qquad 3.8$$

$$\Rightarrow \quad b = \frac{1}{m} \sum_{i=1}^{m} y_i - w\bar{x} = \bar{y} - w\bar{x}$$

$$w \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} (y_i - b) x_i$$

$$w \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} y_i x_i - b \sum_{i=1}^{m} x_i$$

$$w \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} [x_i y_i - (\bar{y} - w\bar{x}) x_i]$$

$$w \sum_{i=1}^{m} x_i^2 - w\bar{x} \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i x_i - \bar{y} \sum_{i=1}^{m} x_i$$

$$\Rightarrow \quad w = \frac{\sum_{i=1}^{m} y_i x_i - \bar{y} \sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i^2 - \bar{x} \sum_{i=1}^{m} x_i}$$

其中. $\bar{y} \sum_{i=1}^{m} x_i = \frac{1}{m} \sum y_i \sum x_i = \bar{x} \sum y_i$

$\bar{x} \sum x_i = \frac{1}{m} (\sum x_i)^2$

$$\therefore \quad w = \frac{\sum_{i=1}^{m} y_i (x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m} (\sum_{i=1}^{m} x_i)^2} \qquad 3.7$$

↓ 三项转化矩阵 numpy.

w 向量化.

$$\bar{y} \sum_{i=1}^{m} x_i = \bar{x} \sum_{i=1}^{m} y_i = m\bar{x}\bar{y} = \sum_{i=1}^{m} \bar{x}\bar{y}$$

$$\sum_{i=1}^{m} x_i \bar{x} = m\bar{x}^2 = \sum_{i=1}^{m} \bar{x}^2$$

$$\therefore \quad w = \frac{\sum_{i=1}^{m} (y_i x_i - y_i \bar{x} - x_i \bar{y} + \bar{x}\bar{y})}{\sum_{i=1}^{m} (x_i^2 - \bar{x} x_i - x_i \bar{x} + \bar{x}^2)}$$

$$= \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m} (x_i - \bar{x})^2}$$

令 $\boldsymbol{x} = (x_1, \cdots, x_m)^T$, $\boldsymbol{x}_d = (x_1 - \bar{x}, \cdots, x_m - \bar{x})^T$

$\boldsymbol{y} = (y_1, \cdots, y_m)^T$, $\boldsymbol{y}_d = (y_1 - \bar{y}, \cdots, y_m - \bar{y})^T$.

则 $w = \frac{\boldsymbol{x}_d^T \boldsymbol{y}_d}{\boldsymbol{x}_d^T \boldsymbol{x}_d}$

机器学习三要素：

- 模型：确定假设空间
- 策略：损失函数 $E(w, b)$
- 算法：求解损失函数，确定最优模型.

多元线性回归.

$w \in \mathbb{R}^d$

$$f(x_i) = w^T x_i + b$$

$$= (w_1\ w_2\ \cdots\ w_d) \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} + b$$

$$= w_1 x_{i1} + w_2 x_{i2} + \cdots + w_d x_{id} + b$$

将 w 扩充到 $\hat{w} \in \mathbb{R}^{d+1}$

b 替换为 $w_{d+1} \cdot 1$

$$\Rightarrow f(x_i) = w_1 x_{i1} + w_2 \cdot x_{i2} + \cdots + w_d x_{id} + w_{d+1} \cdot 1$$

$$f(x_i) = (w_1\ w_2\ \cdots\ w_d\ w_{d+1}) \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \\ 1 \end{pmatrix}$$

$$\Rightarrow f(\hat{x}_i) = \hat{w}^T \hat{x}_i$$

由最小二乘法, $E_{\hat{w}} = \sum_{i=1}^{m} (y_i - f(\hat{x}_i))^2$

$$= \sum_{i=1}^{m} (y_i - \hat{w}^T \hat{x}_i)^2$$

↓ 向量化

$$E_{\hat{w}} = \begin{pmatrix} y_1 - \hat{w}^T\hat{x}_1 & y_2 - \hat{w}^T\hat{x}_2 & \cdots & y_m - \hat{w}^T\hat{x}_m \end{pmatrix} \begin{pmatrix} y_1 - \hat{w}^T\hat{x}_1 \\ \vdots \\ y_m - \hat{w}^T\hat{x}_m \end{pmatrix}$$

$$\begin{pmatrix} y_1 - \hat{w}^T\hat{x}_1 \\ \vdots \\ y_m - \hat{w}^T\hat{x}_m \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} - \begin{pmatrix} \hat{x}_1^T\hat{w} \\ \hat{x}_2^T\hat{w} \\ \vdots \\ \hat{x}_m^T\hat{w} \end{pmatrix}$$

令 $y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$, $\begin{pmatrix} \hat{x}_1^T\hat{w} \\ \hat{x}_2^T\hat{w} \\ \vdots \\ \hat{x}_m^T\hat{w} \end{pmatrix} = \begin{pmatrix} \hat{x}_1^T \\ \hat{x}_2^T \\ \vdots \\ \hat{x}_m^T \end{pmatrix} \hat{w} = X \cdot \hat{w}$

$y - X\hat{w}$

$\therefore E_{\hat{w}} = (y - X\hat{w})^T (y - X\hat{w})$ 凸函数

求解 $\hat{w}^*$. → 多元函数最值.

1) 证明 $E_{\hat{w}} = \cdots$ 凸函数

↳ Hessian 矩阵

2) 凸函数求最值求解 $\hat{w}$.

$$\frac{\partial E_{\hat{w}}}{\partial \hat{w}} = \frac{\partial}{\partial \hat{w}} [(y - X\hat{w})^T (y - X\hat{w})]$$

$$= \frac{\partial}{\partial \hat{w}} [(y^T - \hat{w}^T X^T)(y - X\hat{w})]$$

$$= \frac{\partial}{\partial \hat{w}} (y^T y - y^T X\hat{w} - \hat{w}^T X^T y + \hat{w}^T X^T X\hat{w})$$

与 $\hat{w}$ 无关

$$= \frac{\partial}{\partial \hat{w}} \left(-y^T X \hat{w} - \hat{w}^T X^T y + \hat{w}^T X^T X \hat{w}\right)$$

标量函数

$\hat{w}$: $d+1$ 元的 向量

标量函数对向量求导

矩阵微分公式:

$x \in R^{n+1}$, $f: R^n \to R$ 关于 x 的实值标量函数

$$\frac{\partial f(x)}{\partial x} = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix} \qquad \frac{\partial f(x)}{\partial x^T} = (\quad)$$

分母布局 分子布局

$$\frac{\partial E_{\hat{w}}}{\partial \hat{w}} = -y^T X - \frac{\partial \hat{w}^T X^T y}{\partial \hat{w}} + \hat{w}^T X^T X \qquad \frac{\partial x^T A x}{\partial x} = (A + A^T) x$$

$$= -y^T X - y^T X + (X^T X + X^T X)\hat{w}$$

$$= -2X^T y + 2X^T X \hat{w}$$

$$= 2X^T (X\hat{w} - y). \qquad \text{式 10.}$$

$$\frac{\partial^2 E_{\hat{w}}}{\partial \hat{w}^2} = \frac{\partial}{\partial \hat{w}} \left(2X^T (X\hat{w} - y)\right)$$

$$= 2X^T X$$

$\Rightarrow \nabla^2 E_{\hat{w}} = 2X^T X.$

假定 $X^T X$ 为正定矩阵, 则 $E_{\hat{w}}$ 为关于 $\hat{w}$ 的凸函数.

$$\nabla E_{\hat{w}} = \frac{\partial E_{\hat{w}}}{\partial \hat{w}} = 2X^T(X\hat{w} - y) = 0$$

$$X^T X \hat{w} = X^T y$$

$$\Rightarrow \hat{w} = (X^T X)^{-1} X^T y \quad 3.11$$

二. 对数几率回归

└ 线性模型基础上套一个映射函数 进行分类.

sigmeid : $\frac{1}{1+e^{-z}}$ $\quad \mathbb{R} \to (0,1)$.

$\downarrow$

$f(x)$.

1) 确定概率质量(密度)函数.

2) 似然函数.

离散型随变 $y \in \{0,1\}$ 重建模.

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

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

令 $\beta = (w;b)$. $\hat{x} = (x;1)$.

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

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

⇒ 随变 y 概质 $p(y|\hat{x};\beta) = y\cdot p_1(\hat{x};\beta) + (1-y)\cdot p_0(\hat{x};\beta)$ 3.26

或 $[p_1(\hat{x};\beta)]^y [p_0(\hat{x};\beta)]^{1-y}$

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

对数似… $l(\beta) = \sum_{i=1}^{m} \ln p(\quad)$

$$l(\beta) = \sum \ln(y_i p_1(\quad) + (1-y_i) p_0(\quad))$$

$$= \sum \ln\left(\frac{y_i e^{\beta^T\hat{x}_i} + 1 - y_i}{1+e^{\beta^T\hat{x}_i}}\right)$$

$$= \sum \ln(y_i e^{\cdots} + 1 - y_i) - \ln(1+e^{\beta^T\hat{x}_i})$$

$\because y_i \in \{0,1\}$

$$\therefore l(\beta) = \begin{cases} \sum(-\ln(1+e^{\beta^T\hat{x}_i}) & y_i=0 \\ \sum(\beta^T\hat{x}_i - \ln(1+e^{\cdots})) & y_i=1 \end{cases}$$

综合 $l(\beta) = \sum (y_i \beta^T x_i - \ln(1+e^{\cdots}))$

损失函数 → 目标最小化.

最大化 $l(\beta)$ ⟺ 最小化 $-l(\beta)$ 3.2].

信息论:

自信息: $I(x) = -\log_b p(x)$.

$b=2$ 单位 bit $b=e$ 单 nat.

信息熵(自信息期望): 度量 x 不确定性 ↑ 不确定 ↑

$H(x) = E[I(x)] = -\sum_x p(x) \log_b p(x)$. (离散型)

约定: $p(x)=0 \Rightarrow p(x)\log_b p(x) = 0$

相对熵(KL散度): 度量分布差异. 理想分布 $p(x)$

模拟 $q(x)$.

$D_{KL}(p\|q) = \sum_x p(x) \log_b \left(\frac{p(x)}{q(x)}\right)$

$= \sum_x p(x) (\log_b p(x) - \log_b q(x))$

$= \sum_x p(x) \log_b p(x) - \sum_x p(x) \log_b q(x)$.

最小化 ⇕ 最小化交叉熵.

$\sum_x p(x) \log_b p(x)$: 理想分布信息熵

$-\sum_x p(x) \log_b q(x)$: 交叉熵.

Summary:

对数几率回归 三要素
- 模型：线性模型，输出范围 [0,1] 近似阶跃的单调可微函数
- 策略：极大似然估计 / 信息论
- 算法：梯度下降 / 牛顿法

三、二分类线性判别分析