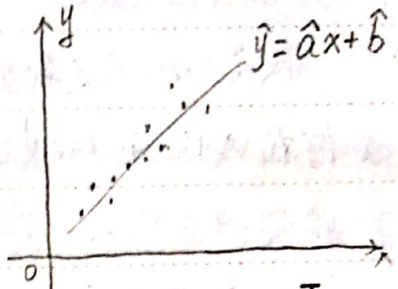
第一章 回归模型

§.1.1. 线性回归——最基本.



1. 模型背景.

如果说给定样本 $T=\{(\vec{x}_1,y_1),(\vec{x}_2,y_2),\cdots,(\vec{x}_n,y_n)\}$, 其中 $\vec{x}_i=(x_1,x_2,x_3,\cdots,x_n)^T$ 是一个属性向量, 维数为 n.

若 y 与 $\vec{x}$ 之间有线性关系, 即 $f(\vec{x})=\omega^T x+b$. $\vec{\omega}=(\omega_1,\omega_2,\omega_3,\cdots,\omega_n)^T$.

那么我们就可以使用线性回归模型去对它进行拟合.

属性向量的分量 x_i 可能是连续的取值。如果是离散的, 如高矮, 则人为数值化.

2. 评价值指标: 均方误差 $(\omega^*,b^*)=\underset{(\omega,b)}{\arg\min}\frac{1}{2}\sum_{i=1}^{m}(f(x_i)-y_i)^2$.

3. 模型求解:

(1). 一对一的特殊情况: $\vec{x}$ 只有1个分量, 目标函数 $E=\frac{1}{2}\sum_{i=1}^{m}(\omega x_i+b-y_i)^2$.

对自变量 ω、b 求偏导, 解 $\begin{cases}\frac{\partial E}{\partial \omega}=\sum_{i=1}^{m}x_i(\omega x_i+b-y_i)=0.\\ \frac{\partial E}{\partial b}=\sum_{i=1}^{m}(\omega x_i+b-y_i)=0\end{cases}$ 得极值点 $\begin{cases}\omega=\frac{cov(x,y)}{Var(x)}\\ b=\bar{y}-\bar{x}\cdot\omega.\end{cases}$

(2). 一对多的一般情况: $\vec{x}$ 有 n 个分量, 目标函数 $E=\frac{1}{2}\sum_{i=1}^{m}(\omega^T x_i+b-y_i)^2$.

这是多元线性回归问题。我们可以考虑用 $\hat{\omega}=(\omega^T;b)^T$ 代替

而 $X=\begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1n} & 1\\ x_{21} & x_{22} & \cdots & x_{2n} & 1\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ x_{m1} & x_{m2} & \cdots & x_{mn} & 1\end{pmatrix}$ m 个数据点, 一个数据 n 个属性.

那么目标函数 $E(\hat{\omega}^*)=(y-X\cdot\hat{\omega})^T(y-X\hat{\omega})$. 求解方法有两种, 这里只介绍一下

矩阵求导:

$\frac{\partial E}{\partial \hat{\omega}}=2X^T(X\hat{\omega}-y)=0$. 那么解矩阵方程: $(X^TX)\hat{\omega}=X^Ty$.

讨论: ① 若 X^TX 可逆: $\hat{\omega}^*=(X^TX)^{-1}X^Ty$. 没有争议.

② 若 X^TX 不可逆: 方程难解, 此时有多个目标解向量 $\hat{\omega}^T$

选择哪一个最好, 还是要引入正则化项 (regularization).

4. 模型扩展:

如果发现 y 与 x 有指数关系, 那么 $\ln y=\omega^T x+b$, $y=e^{\omega^T x+b}$ (对数线性回归)

……

扫描全能王 创建

8.1.2. 岭回归——改进

1. 对线性回归的改进.

假设：对于给定数据集，有一小部分数据有明显偏差，这显然会干扰 prediction. 如何在线性回归的基础上让机器自己剔除异常，就是 Ridge Regression 的优势.

2. 岭回归与Lasso回归的改进

岭回归 Ridge Regression 以及 Lasso Regression 都在线性回归的基础上加了罚项进行正则化.

Ridge：L_2 正则化 $J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(f_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{n}\theta_j^2$　θ为系数向量

Lasso：L_1 正则化 $J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(f_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{n}|\theta_j|$

Lasso 回归使损失函数中许多θ都可以为0. 计算量相对较小.

这里，λ叫正则化系数. 过大易欠，过小则过.

3. 岭回归求解：

$$\frac{\partial J(\theta)}{\partial \theta}=0 \Rightarrow 0-X^Ty-X^Ty+2X^TX\theta+2\lambda\theta=0.$$

$$\Rightarrow \theta=(X^TX+\lambda I)^{-1}X^Ty.$$

Lasso回归求解：比较复杂

$$\theta=\begin{cases}(m_j-\frac{\lambda}{2})/n_j, & m_j>\frac{\lambda}{2}\\ 0 & \text{else}\\ (m_j+\frac{\lambda}{2})/n_j, & m_j<\frac{\lambda}{2}\end{cases}\qquad \begin{cases}m_j=\sum_{i=1}^{n}x_{ij}(y_i-\sum_{k\neq j}\theta_k x_{ik})\\ n_j=\sum_{i=1}^{n}x_{ij}^2\end{cases}$$

4. λ值选取

(1). 岭迹法. 由 $\theta=(X^TX+\lambda I)^{-1}X^Ty$，θ是λ的函数

θ趋于稳定的λ就是待求

(2) 交叉验证（建议）

扫描全能王 创建