Section 3 Linear Regression

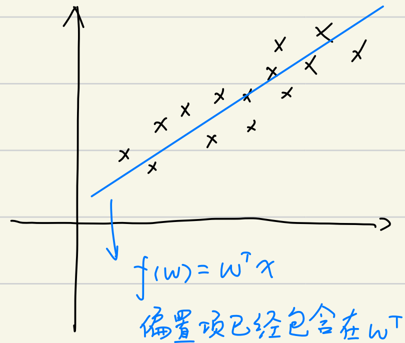
最小二乘法 (Least Squre Method)

数据集：$D=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$

其中 $x_i\in\mathbb{R}^p$, $y_i\in\mathbb{R}$, $i=1,2,\cdots,N$.

令 $X=(x_1\ x_2\cdots x_N)^T$, $X\in\mathbb{R}^{N\times p}$

$Y=(y_1\ y_2\cdots y_N)^T$, $Y\in\mathbb{R}^{N\times 1}$



$f(w)=w^Tx$

偏置项已经包含在w^T中

最小二乘估计：$L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2$

$\boxed{\text{矩阵表达}}$：

$$L(w)=\sum_{i=1}^{N}(w^Tx_i-y_i)^2$$

$$=\underbrace{(w^Tx_1-y_1\quad w^Tx_2-y_2\quad\cdots\quad w^Tx_N-y_N)}_{w^T(x_1\ x_2\cdots x_N)-(y_1\ y_2\cdots y_N)\ \Rightarrow\ w^TX^T-Y^T}\underbrace{\begin{pmatrix} w^Tx_1-y_1\\ w^Tx_2-y_2\\ \vdots\\ w^Tx_N-y_N\end{pmatrix}}_{\Rightarrow XW-Y}$$

$$=(W^TX^T-Y^T)(XW-Y)$$

$$=W^TX^TXW-\underbrace{W^TX^TY}-\underbrace{Y^TXW}+Y^TY \quad \text{每一项都为实数}$$

$$(W^TX^TY)^T=Y^TXW$$

$$=W^TX^TXW-2W^TX^TY+Y^TY$$

则 $w = \underset{w}{\arg\min}\ w^TX^TXw - 2w^TX^TY + Y^TY$

$$\frac{\partial L(w)}{\partial w} = \underbrace{2X^TXw}_{\downarrow\ \text{证明.}} - 2X^TY = 0 \Rightarrow \boxed{w = \underbrace{(X^TX)^{-1}X^T}_{X^{\dagger}}Y}$$

$\to X^{\dagger}$，称为X之伪逆

简化问题，令 $A = X^TX$，求 $\frac{\partial}{\partial w}(w^TAw)$，其中 $w \in \mathbb{R}^{p\times 1}$

假设 $w = (w_1, w_2, \cdots w_p)^T$.

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix}$$

$$\text{则 } w^TAw = \left(\sum_{i=1}^{p} w_i a_{i1} \quad \sum_{i=1}^{p} w_i a_{i2} \cdots \sum_{i=1}^{p} w_i a_{ip}\right)(w_1 \quad w_2 \cdots w_p)^T$$

$$= \sum_{j=1}^{p} w_j \sum_{i=1}^{p} w_i a_{ij}$$

$$\frac{\partial(w^TAw)}{\partial w_k} = w_1 \cdot a_{k1} + w_2 a_{k2} + \cdots + \left(\sum_{i=1}^{p} w_i a_{ik} + w_k a_{kk}\right) + \cdots$$

$$= \sum_{i=1}^{p} w_i a_{ki} + \sum_{i=1}^{p} w_i a_{ik}$$

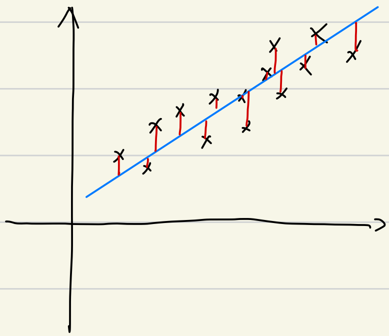
$$\text{则 } \frac{\partial(w^TAw)}{\partial w} = \begin{pmatrix} \sum_{i=1}^{p} w_i a_{1i} \\ \sum_{i=1}^{p} w_i a_{2i} \\ \vdots \\ \sum_{i=1}^{p} w_i a_{pi} \end{pmatrix} + \begin{pmatrix} \sum_{i=1}^{p} w_i a_{i1} \\ \sum_{i=1}^{p} w_i a_{i2} \\ \vdots \\ \sum_{i=1}^{p} w_i a_{ip} \end{pmatrix} = Aw + A^Tw$$

由于 $A = X^TX$ 是对称矩阵，即 $A^T = A$.

所以 $\frac{\partial(w^TAw)}{\partial w} = 2Aw \Rightarrow \frac{\partial(w^TX^TXw)}{\partial w} = 2X^TXw.$

几何解释

①



红色的线段就是目标函数需要最小化的对象，即将误差均摊到每一个样本点。

② 我们注意到 $X=(x_1,x_2,\cdots,x_N)^T\in\mathbb{R}^{N\times p}$，其中 $N>p$.

将 X 的列向量看作是一组基底，那么这组基底就在 N 维空间中构成了 p 维子空间，可能小于 p，则 XW 就是这 p 个列向量的线性组合，仍属于这个 p 维子空间。

线性回归问题就可以看作，寻找到 p 个 N 维列向量的一个线性组合，使其与 Y 的距离最小，即该线性组合是 Y 在 p 个向量构成空间上的投影。

这个距离其实就是误差

由此可得：线性组合为 XW，

由于线性组合是 Y 在子空间上的投影，

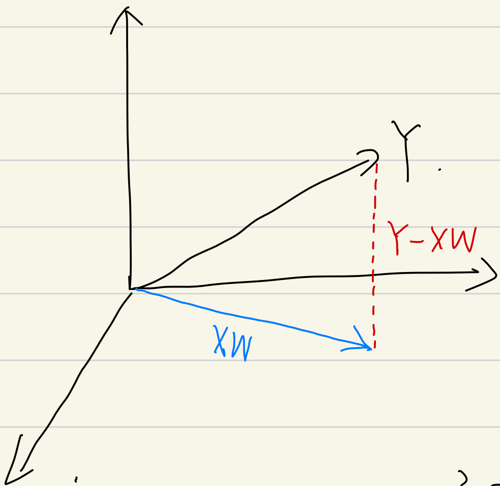
那么 $Y-XW$ 为子空间的一个法向量

则 $X^T(Y-XW)=0$　法向量与基底内积为0.

$$\boxed{\Rightarrow\ W=(X^TX)^{-1}X^TY}$$

以 $N=3$, $p=2$ 为例. 假设基底为 $(1,0,0)^T$, $(0,1,0)^T$

构成的子空间为 xoy 平面.



$$X=\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}$$

$$Y=\begin{bmatrix}3\\2\\2\end{bmatrix}$$

$$W=\begin{bmatrix}w_1\\w_2\end{bmatrix}$$

$$W=(X^TX)^{-1}X^TY=\begin{bmatrix}3\\2\end{bmatrix}$$

综上, ① 几何角度是将误差均摊到每个样本上

② 几何角度是将误差在向量空间中表示出来.

概率视角的最小二乘估计

已知：$X=(x_1,x_2,\cdots,x_N)^T \in \mathbb{R}^{N\times p}$，$Y=(y_1,y_2,\cdots,y_N)\in\mathbb{R}^{N\times 1}$

$W=(w_1,w_2,\cdots,w_p)^T\in\mathbb{R}^{p\times 1}$，$\varepsilon\sim N(0,\sigma^2)$

如何设 $y_i=f(w)+\varepsilon$，其中 $f(w)=w^Tx_i$.

则 $y_i=w^Tx_i+\varepsilon$.

即 $y_i|x_i;w\sim N(w^Tx_i,\sigma^2)$.

定义 $$\mathcal{L}(w)=\log P(Y|X;W)=\sum_{i=1}^{N}\log P(y_i|x_i;w)$$

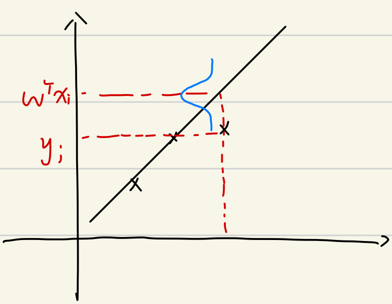
$$\hat{W}_{MLE}=\arg\max_{W}\sum_{i=1}^{N}\log p(y_i|x_i;w)$$

$$=\arg\max_{w}\sum_{i=1}^{N}\left(\frac{1}{2}\log\frac{1}{2\pi\sigma^2}-\frac{(y_i-w^Tx_i)^2}{2\sigma^2}\right)$$

$$\boxed{=\arg\min_{w}\sum_{i=1}^{N}(y_i-w^Tx_i)^2.}$$

和最小二乘法的目标函数一致.

所以最小二乘估计其实等价于噪声为高斯分布的极大似然估计



如左图，表示了两者等价的图形表达.
蓝色为高斯分布 $N(0,\sigma^2)$，误差就是其造成的.
误差：$y_i-w^Tx_i$

正则化(Regularization)

岭回归(Ridge Regression)

最小二乘估计的目标函数: $\mathcal{L}(w) = \sum_{i=1}^{N} \|w^T x_i - y_i\|^2$

解析解: $\hat{w} = (X^T X)^{-1} X^T Y$.

引入正则化的原因:

这里需要注意 $(X^T X)^{-1}$.

由于 $X \in \mathbb{R}^{N \times p}$, 正常情况下 $N \gg p$.

如果 $N < p$.

由于 $r(X^T X) = r(X)$.

则 $r(X^T X)$ 必定小于 p. 但是 $X^T X \in \mathbb{R}^{p \times p}$

$\therefore$ $X^T X$ 不可逆 数学上不可逆. 线性回归角度容易造成过拟合.

$\therefore$ 解析解无法直接得到.

如何解决过拟合 $\rightarrow$
① 增加数据量
② 特征选择 / 特征提取(降维).
③ 正则化 对参数空间的约束.

正则化框架: $\arg\min\limits_{w} [\underbrace{\mathcal{L}(w)}_{\text{Loss Function}} + \lambda \underbrace{P(w)}_{\text{Penalty}}]$

正则化较为常用的为 L_1 正则化与 L_2 正则化

L_1: Lasso. $P(W)=\|W\|$, 1-范数

L_2: Ridge. 岭回归 $P(W)=\|W\|_2^2=W^TW$. 2-范数的平方.

权值衰减.

应用 L_2 后.

目标为. $\underset{W}{\operatorname{argmin}}\left[\sum_{i=1}^{N}\|W^Tx_i-y_i\|^2+\lambda W^TW\right]$

$$=\underset{W}{\operatorname{argmin}}\left[(W^TX^T-Y^T)(XW-Y)+\lambda W^TW\right]$$

$$=\underset{W}{\operatorname{argmin}}\left[W^TX^TXW-2W^TX^TY+Y^TY+\lambda W^TW\right].$$

$$=\underset{W}{\operatorname{argmin}}\underbrace{\left[W^T(X^TX+\lambda I)W-2W^TX^TY+Y^TY\right]}_{G(W)}.$$

$$\frac{\partial}{\partial W}G(W)=2(X^TX+\lambda I)W-2X^TY=0.$$

$$\boxed{\Rightarrow W=\underbrace{(X^TX+\lambda I)}_{\text{一定可逆}}{}^{-1}X^TY.}$$

一定可逆　因为 X^TX 半正定，加上 I，为正定，一定可逆.

从贝叶斯角度看待岭回归

假设 w 服从一个先验分布：$w \sim \mathcal{N}(0, \Sigma)$，且 Σ 为对角，各向同性

则由贝叶斯定理：$p(w|Y) = \dfrac{p(Y|w) \cdot p(w)}{p(y)}$.

MAP（最大后验估计）：
$$\hat{w} = \arg\max_w p(w|Y)$$
$$= \arg\max_w p(Y|w)\, p(w)$$

由之前概率角度看待最小二乘估计可得：

$$p(y|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-w^Tx)^2}{2\sigma^2}\right\}. \Rightarrow p(Y|w) = \prod_{i=1}^{N} p(y_i|w)$$

y_i 之间相互独立，联合密度是边缘密度之乘积

则 $\hat{w} = \arg\max_w P(Y|w)\, p(w)$

$$= \arg\max_w \log\left[p(Y|w)\, p(w)\right]$$

$$= \arg\max_w \sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i-w^Tx_i)^2}{2\sigma^2}\right] + \log\frac{1}{\sqrt{2\pi}|\Sigma|^{\frac{1}{2}}} - \frac{1}{2}w^T\Sigma^{-1}w$$

$$= \arg\min_w \sum_{i=1}^{N}(y_i - w^Tx_i)^2 + \sigma^2 w^T \Sigma^{-1} w.$$

由于 Σ^{-1} 为对角且各向同性，则 $\Sigma = \sigma_w^2 I$.

$$\boxed{\therefore \hat{w} = \arg\min_w \sum_{i=1}^{N}(y_i - w^Tx_i)^2 + \underbrace{\frac{\sigma^2}{\sigma_w^2}}_{\lambda}\underbrace{w^Tw}_{L_2}.}$$

和上述 L_2 正则化后的目标函数等同.

最小二乘估计

综上 LSE $\Longleftrightarrow$ MLE (noise ~ Gaussian Distribution)

Regularized LSE $\Longleftrightarrow$ MAP (noise ~ Gaussian Distribution

prior ~ Gaussian Distribution)

这个分布也是各向同性.