Notes on Linear Regression

- univarite linear regression: $\hat{y}_i = w_0 + w_1 x_i$ where $\hat{y}_i$ is the prediction
- "learn" parameters: w_0, w_1
- To learn, need to minimize a loss: $\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{(\overset{\text{true - predicted}}{y_i - \hat{y}_i})}_{\text{residuals}}^2$
  $= \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$

- objective

$$\min_{w_0, w_1} \mathcal{L}^{(a,b)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

$$\frac{\partial \mathcal{L}}{\partial w_0} = \frac{\partial}{\partial w_0}\frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2 = \frac{1}{n}\sum_{i=1}^{n} -(y_i - w_0 - w_1 x_i) \overset{(set)}{=} 0$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial w_0}(y_i - w_0 - w_1 x_i)^2 =$$

so: $\sum(y_i - w_0 - w_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} w_0 - \sum_{i=1}^{n} w_1 x_i = 0$

$$\Rightarrow n\bar{y} - n w_0 - w_1 n\bar{x} = 0$$

$$\Rightarrow \bar{y} - w_0 - w_1\bar{x} = 0$$

$$\Rightarrow \boxed{w_0 = \bar{y} - w_1\bar{x}} \quad (1)$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \frac{1}{n}\sum\frac{\partial}{\partial w_1}(y_i - w_0 - w_1 x_i)^2$$

$$= \frac{1}{n}\sum(-x_i(y_i - w_0 - w_1 x_i)) \overset{(set)}{=} 0$$

So: $\sum x_i y_i - w_0\sum x_i - w_1\sum x_i^2 = 0$

$$\Rightarrow n\overline{xy} - n w_0\bar{x} - n w_1\overline{x^2} = 0$$

$$\Rightarrow \boxed{w_1 = \frac{\overline{xy} - w_0\bar{x}}{\overline{x^2}}} \quad (2)$$

$$\boxed{\begin{array}{l} * \ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \text{so } \sum_{i=1}^{n} x_i = n\bar{x} \end{array}}$$

$$\boxed{w_0 = \bar{y} - w_1\bar{x} \qquad w_1 = \frac{\overline{xy} - w_0\bar{x}}{\overline{x^2}}}$$ system of "normal" equations.

plug w_0 into w_1:

$$w_1 = \frac{\overline{xy} - (\bar{y} - w_1\bar{x})\,\bar{x}}{\overline{x^2}} = \frac{\overline{xy} - \bar{x}\bar{y} + w_1\bar{x}^2}{\overline{x^2}}$$

$$\Rightarrow w_1\left(1 - \frac{\bar{x}^2}{\overline{x^2}}\right) = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2}}$$

$$\Rightarrow w_1\left(\frac{\overline{x^2} - \bar{x}^2}{\overline{x^2}}\right) = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2}}$$

$$\Rightarrow w_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

now, recall that

$$\begin{aligned} cov(x,y) &= \frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1}\sum(x_iy_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y}) \\ &= \frac{1}{n-1}(n\overline{xy} - n\bar{x}\bar{y} - \cancel{n\bar{x}\bar{y}} + \cancel{n\bar{x}\bar{y}}) \\ &= \frac{n}{n-1}(\overline{xy} - \bar{x}\bar{y}) \end{aligned}$$

and

$$\begin{aligned} Var(x) &= \frac{1}{n-1}\sum(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\sum(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n-1}(n\overline{x^2} - 2n\bar{x}^2 + n\bar{x}^2) \\ &= \frac{1}{n-1}(n\overline{x^2} - n\bar{x}^2) \\ &= \frac{n}{n-1}(\overline{x^2} - n\bar{x}^2) \end{aligned}$$

we shouldn't call the variance/covariance, instead the sample covariance/variance. cov(x,y), v(x) should be reserved for populations!

$$\frac{cov(x,y)}{v(x)} = \frac{\frac{n}{n-1}(\overline{xy} - \bar{x}\bar{y})}{\frac{n}{n-1}(\overline{x^2} - \bar{x}^2)} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = w_1.$$

What if we had a multivariate model?

$$\hat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + \dots + w_p x_{ip}$$

now we have p parameters? We could do:

$\mathcal{L} = \frac{1}{n} \sum (\hat{y}_i - y_i)^2$ and find $\frac{\partial \mathcal{L}}{\partial w_0}, \frac{\partial \mathcal{L}}{\partial w_1}, \dots, \frac{\partial \mathcal{L}}{\partial w_p}$ $\boxed{\text{Annoying!}}$

lets write (n = # data points, p = # variables)

- $w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{(p+1) \times 1}$
- $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n \times 1}$
- $x_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \in \mathbb{R}^{(p+1) \times 1}$
- $X = \begin{bmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_n^T - \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$

Now note:

$$\hat{y}_i = w_0 + w_1 x_{i1} + \dots + w_p x_{ip} = w^T x_i = x_i^T w$$

So $\mathcal{L}(w_0, w_1, \dots, w_p) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - w^T x_i)$

which can then be written as

$$\mathcal{L}(w) = (y - Xw)^T (y - Xw)$$

with objective: $\min_{w} \mathcal{L}(w)$.

now

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

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

$$= -2 X^T y + 2 X^T X w \overset{(\text{set})}{=} 0$$

$$\Rightarrow \boxed{w = (X^T X)^{-1} X^T y}$$

$$w = (X^T X)^{-1} X^T y$$

if $p = 1$ as before: $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n \times 2}$ $\quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$

$w = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \in \mathbb{R}^2$

- $$X^T X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{pmatrix}$$

$$\boxed{\begin{array}{l} (*) \\ A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \\ A^{-1} = \dfrac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \\ \text{if } ad-bc \neq 0 \end{array}}$$

- $$(X^T X)^{-1} \overset{(*)}{=} \frac{1}{n^2 \overline{x^2} - n^2 \bar{x}^2} \begin{pmatrix} n\overline{x^2} & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$$

$$= \frac{1}{n^2(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} n\overline{x^2} & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$$

- $$X^T y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ n\overline{xy} \end{pmatrix}$$

$$w = (X^T X)^{-1} X^T y = \frac{1}{n^2(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} n\overline{x^2} & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{y} \\ n\overline{xy} \end{pmatrix}$$

$$= \frac{1}{n^2(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} n^2 \overline{x^2}\,\bar{y} - n^2 \bar{x}\,\overline{xy} \\ -n^2 \bar{x}\bar{y} + n^2 \overline{xy} \end{pmatrix}$$

$w_0 = (\underline{\qquad})$

So $$w_1 = \frac{n^2(\overline{xy} - \bar{x}\bar{y})}{n^2(\overline{x^2} - \bar{x}^2)} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$$

as before! (check out numpy version).

Regularization (multivariate)

$$\mathcal{L}(w) = (y - Xw)^T (y - Xw) + \lambda w^T w$$

$$= \|y - Xw\|_2^2 + \lambda \|w\|_2^2$$

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \frac{\partial}{\partial w} \left[y^T y - 2y^T X w + w^T X^T X w + \lambda w^T w \right]$$

$$= -2X^T y + 2 X^T X w + 2\lambda w \overset{(set)}{=} 0$$

$$\Rightarrow \cancel{2} X^T X w + \cancel{2} \lambda w = \cancel{2} X^T y$$

$$\Rightarrow (X^T X + \lambda I) w = X^T y$$

$$w = (X^T X + \lambda I)^{-1} X^T y .$$