2/28/18 Lecture 9

$\mathcal{H} = \{w_0 + w_1 \tilde{x} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\} = \{w_0 + w_1 \mathbb{1}_{x=green} : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\}$

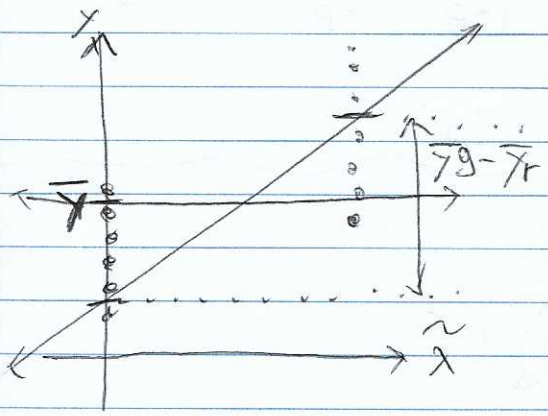
$\Rightarrow g(x) = b_0 + b_1 \mathbb{1}_{x=green}$

$g_{null} = \bar{y}$

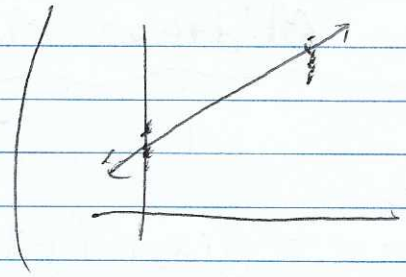
$b_0 = \bar{y}_r$

$b_1 = \bar{y}_g - \bar{y}_r$

$$g(x) = \begin{cases} \bar{y}_r \text{ if } x = \text{Red} \\ \bar{y}_g \text{ if } x = \text{Green} \end{cases} = \bar{y}_r + (\bar{y}_g - \bar{y}_r)\mathbb{1}_{x=g}$$



Here the value of R^2 at $\bar{y}$ is 0



Better R^2 if points close to $\bar{y}_g$ and $\bar{y}_r$

$b_0 = \bar{y} - b_1 \bar{x}$ Lets assume $b_1 = \bar{y}_g - \bar{y}_r$ and compare to b_0.

$b_1 = r\frac{s_y}{s_x} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$

$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{y_{g_1} + \cdots y_{g n_g} + y_{r_1} + \cdots + y_{r n_r}}{n}$ let $p = \frac{n_g}{n}$

$= \frac{\sum y_{g_i}}{n}\frac{n_g}{n_g} + \frac{\sum y_{r_i}}{n}\cdot\frac{n_r}{n_r} = \bar{y}_g \frac{n_g}{n} + \bar{y}_r \frac{n_r}{n} = p\bar{y}_g + (1-p)\bar{y}_r$

$b_0 = (p\bar{y}_g + (1-p)\bar{y}_r) - (\bar{y}_g - \bar{y}_r) = p\bar{y}_g + (1-p)\bar{y}_r - p\bar{y}_g + p\bar{y}_r$

$= \bar{y}_r((1-p) + p) = \bar{y}_r$ ✓

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{x_{g_1} + \dots + x_{gn_g} + x_{r_1} + \dots + x_{rn_r}}{n} = \frac{n_g}{n} = p$$

$$\sum x_i y_i = \sum y_{g_i} = n_g \bar{y}_g \quad n\bar{x}\bar{y} = np\bar{y} \quad \sum x_i^2 = n_g \quad n\bar{x}^2 = np^2$$

$$b_1 = r\frac{S_y}{S_x} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{n_g \bar{y}_g - np\bar{y}}{n_g - np^2} \quad \frac{\frac{1}{n}}{\frac{1}{n}} = \frac{p\bar{y}_g - p\bar{y}}{p - p^2} = \frac{\bar{y}_g - \bar{y}}{1-p}$$

$$= \frac{\bar{y}_g - (p\bar{y}_g + (1-p)(\bar{y}_r))}{1-p} = \frac{\bar{y}_g}{1-p} - \frac{p\bar{y}_g}{1-p} - \bar{y}_r = \bar{y}_g - \bar{y}_r \checkmark$$

Midterm Material End

$\hat{y} = g(x)$

$y = g(x) + e = g(x) + \underbrace{(h^*(x) - g(x))}_{\text{estimation error}} + \varepsilon$

$= g(\vec{x}) + \underbrace{(h^*(\vec{x}) - g(\vec{x}))}_{\text{misspecification error}} + (f(\vec{z}) - f(\vec{x}))$

$h^* \in \mathcal{H}$ where h^* is "closest" element $\in \mathcal{H}$ to f

$g \in \mathcal{H}$

$h^* = \beta_0 + \beta_1 x \quad g = b_0 + b_1 x$

To minimize estimation, we need a larger n (law of large numbers)
(also better $\mathcal{H}$)

To minimize misspecification, better $\mathcal{H}$ (then also $\mathcal{H}$)

To minimize ignorance, Choose better $\vec{x}$s that reflect true causal inputs

SVM model (support vector machine)

If linearly separable...

$$\min \|\vec{w}\| \text{ subj to } \forall i \ \left(y_i - \tfrac{1}{2}\right)\left(\vec{w} \cdot \vec{x}_i + b\right) \geq \tfrac{1}{2}$$

If $\mathbb{D}$ is not linearly separable...

$$\text{minimize: } \frac{1}{n}\sum_{i=1}^{n} \max\left\{0, \underbrace{-\tfrac{1}{2} - \left(y_i - \tfrac{1}{2}\right)\left(\vec{w} \cdot \vec{x}_i + b\right)}_{H_i}\right\} + \lambda \|\vec{w}\|^2$$

(λ: hyperparameter)

d_i: distance away from line on the wrong side of the line

↳ Parking Tickets

worse mistakes are more costly