Lecture - 20 04/21/2020

Response Space(y) | Class of model

$\subseteq \mathbb{R}$ - Regression, Survival

$\{C_1, C_2, \ldots, C_K\}$ Classification, Probability estimation

$K = 2$

$[C_1, C_2]$ - Binary Classification, prob. est.

$C_1 = 0$, $C_2 = 1$

If $y = \{0, 1\}$

$y = t(\vec{z})$ $\iff Y \sim \text{Bern}(t(\vec{z}))$ ← degenerate

$= f(\vec{x}) + \delta,\ \delta \in \{0, 1, -1\}$ ⓘ Assume a function

$= h^*(\vec{x}) + \mathcal{E},\ \mathcal{E} \in \{0, 1, -1\}$ $f_{pr}(\vec{x}): \mathbb{R}^{p+1} \longrightarrow (0,1)$

$= g(\vec{x}) + e,\ e \in \{0, 1, -1\}$ $\overset{=}{\text{best}}$ guess with $\vec{x}$ of $p(Y=1|\vec{x})$

Ⓘ Assume $Y_i|\vec{x_i}$ independent Y_i

$\Rightarrow Y_i = y_i | \vec{x_i} \overset{ind}{\sim} \text{Bern}(f_{pr}(\vec{x_i})) = f_{pr}(\vec{x_i})^{y_i}(1 - f_{pr}(\vec{x_i}))^{1-y_i}$

$\Rightarrow P(\mathbb{D}) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n | \vec{x_1}, \vec{x_2}, \ldots, \vec{x_n})$

$\overset{\text{II}}{=} \prod_{i=1}^{n} f_{pr}(\vec{x_i})^{y_i} (1 - f_{pr}(\vec{x_i}))^{1-y_i}$ likelihood.

Is it possible to $\underset{f \in F}{\text{argmax}} \left\{ \prod_{i=1}^{n} f_{pr}(\vec{x_i})^{y_i} (1 - f_{pr}(\vec{x_i}))^{1-y_i} \right\}$?

NO, b/c it is impossible. F = all functions.

Instead assume $H_{pr} = \{\vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1}\}$. Is this okay?

NO, Since $\vec{w} \cdot \vec{x} \notin (0,1)$

Let's say we want to retain the term $\vec{w} \cdot \vec{x}$ in our hypothesis set. Why? Linear model is very interpretable, monotonic, simple.

generalized linear model (glm)

$\Rightarrow H_{pr} = \{\phi(\vec{w} \cdot \vec{x}) : \vec{w} \in \mathbb{R}^{p+1}\}$ $\phi : \mathbb{R} \longrightarrow (0,1)$

called a "link function"

3 common choices in order of popularity:

① Logistic $\phi(u) = \frac{e^u}{1+e^u} = \frac{1}{1+e^{-u}}$, $1-\phi(u) = \frac{1}{1+e^u}$

② Probit $\phi(u) = F_Z^{-1}(u) \overset{\text{inverse CDF of the standard normal r.v.}}{=} \Phi^{-1}(u)$

③ Complementary Log-Log $\phi(u) = 1 - e^{-e^u}$, $1-\phi(u) = e^{-e^u}$

Logistic Regression: $\mathcal{H}_{pr} = \left\{ \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}} : \vec{w} \in \mathbb{R}^{p+1} \right\}$

$$A: \vec{b} = \underset{\vec{w} \in \mathbb{R}^{p+1}}{\text{argmax}} \underbrace{\left\{ \overset{\text{maximum likelihood}}{\prod_{i=1}^{n}} \left(\frac{1}{1+e^{-\vec{w}\cdot\vec{x}_i}}\right)^{y_i} \left(\frac{1}{1+e^{\vec{w}\cdot\vec{x}_i}}\right)^{1-y_i} \right\}}_{P(\mathbb{D})}$$

no closed form solution i.e. $\nabla P(\mathbb{D}) \overset{\text{set}}{=} \vec{0}_{p+1}$ is not soluable

↓ need to use numerical optimization to approximate $\vec{b}$. That is called "fitting a logistic regression".

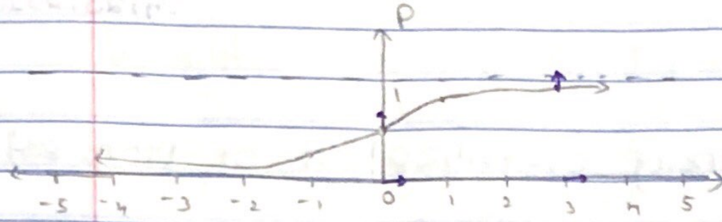
$$g_{pr}(\vec{x}_*) = \frac{1}{1+e^{-\vec{b}\cdot\vec{x}}} = \hat{p}(Y_* = 1 \mid \vec{x}_*) = \hat{p}_*$$
$\in \mathcal{H}_{pr}$
$$\approx P(Y_* = 1 \mid \vec{x}_*) = f_{pr}(\vec{x}_*)$$

What is the interpretation of b_j's.

$$\hat{p} = \frac{1}{1+e^{-\vec{b}\cdot\vec{x}}} \implies \frac{1}{\hat{p}} = 1 + e^{-\vec{b}\cdot\vec{x}} \implies \underbrace{\frac{1-\hat{p}}{\hat{p}}}_{\text{Odds Against } (Y=1|\vec{x})} = \frac{1}{\hat{p}} - 1 = e^{-\vec{b}\cdot\vec{x}}$$

$$\implies \underbrace{\frac{\hat{p}}{1-\hat{p}}}_{\text{Odds}(Y=1|\vec{x})} = e^{\vec{b}\cdot\vec{x}} \implies \vec{b}\cdot\vec{x} = \underbrace{\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)}_{\text{Log Odds}(Y=1|\vec{x}_*)} \in \mathbb{R}$$

$b_3 = 0.7$, x_3 increases by $1 \implies$ LogOdds$(Y=1|\vec{x})$ increases by 0.7



logodds	prob
0	0.5
-1	0.27
+1	0.73
-2	0.12
+2.	0.88
⋮	⋮
$-\alpha$	0
$+\alpha$	1

$g_{0_{pr}} = \bar{y} = \frac{1}{n} \sum \mathbb{1}_{y_i = 1}$

Prob. Est. Honest model validation:

previously ... $SSE = \sum (y_i - \hat{y}_i)^2$

sum of square prob

Maybe ... $SSP = \sum (p_i - \hat{p}_i)^2$ $\quad p_i = p(y_i = 1 \mid \vec{x}_i) = f_{pr}(\vec{x}_i)$

Impossible ...

higher the score, the bettor the model $\approx R^2$

"Scoring Rule" $S(y, \hat{p})$, A "proper scoring" rule:

$\forall_i \; f_{pr}(\vec{x}_i) = \underset{\hat{p} \in (0,1)}{\text{argmax}} \{ S(y_i, \hat{p}) \}$

Two populor proper scoring rule:

① log scoring rule: $S_i = y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)$

overall score is average

② Briar score: $S_i = -(y_i - \hat{p}_i)^2 \implies S = \frac{1}{n} \sum S_i$