Response Space ($\mathcal{Y}$)	Class of Model
$\subseteq \mathbb{R}$	Regression, Survival
$\{C_1, C_2, \dots, C_k\}$	Classification, probability estimation
$K=2$	Binary Classification, prob. est.

$\{C_1, C_2\}$
$\quad = 0 \quad = 1$

No | Count Model

If $\mathcal{Y} = \{0, 1\}$ $\Leftrightarrow$ $Y \sim \text{Bern}(t(\vec{z}))$ degenerate

$$y = t(\vec{z})$$
$$= f(\vec{x}) + \delta, \quad \delta \in \{0, 1, -1\}$$
$$= h^*(\vec{x}) + \mathcal{E}, \quad \mathcal{E} \in \{0, 1, -1\}$$
$$= g(\vec{x}) + e, \quad e \in \{0, 1, -1\}$$

(I) Assume a function

$$f_{pr}(\vec{x}) : \mathbb{R}^{p+1} \to (0,1)$$

$=$ best guess with $\vec{x}$ of $P(Y=1 \mid \vec{x})$

(II) Assume $Y_i \mid \vec{x}_i$ independent $\forall i$.

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

$$\Rightarrow P(\mathbb{D}) = P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \mid \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)$$
$$\overset{(II)}{=} \prod_{i=1}^{n} f_{pr}(\vec{x}_i)^{y_i} (1 - f_{pr}(\vec{x}_i))^{1-y_i} \quad \text{likelihood}$$

Is it possible to $\underset{f \in \mathcal{F}}{\text{argmax}} \left\{ \prod_{i=1}^{n} f_{pr}(\vec{x}_i)^{y_i} (1 - f_{pr}(\vec{x}_i))^{1-y_i} \right\}$? No.... because it is impossible $\mathcal{F}$ = all functions

Instead assume $\mathcal{H}_{pr} = \{ \vec{w} \cdot \vec{x} : \vec{w} \in \mathbb{R}^{p+1} \}$. 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).

$$\Rightarrow \mathcal{H}_{pr} = \{ \phi(\vec{w} \cdot \vec{x}) : \vec{w} \in \mathbb{R}^{p+1} \} \quad \phi : \mathbb{R} \to (0,1)$$

and strictly increasing

called a "link function".

3 common choices in order of popularity:

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

(2) Probit $\phi(u) = F_Z^{-1}(u)$ ← 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: $\mathcal{H}_{pr} = \left\{ \dfrac{1}{1+e^{-\vec{w} \cdot \vec{x}}} : \vec{w} \in \mathbb{R}^{p+1} \right\}$

Maximum Likelihood

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

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

↓

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

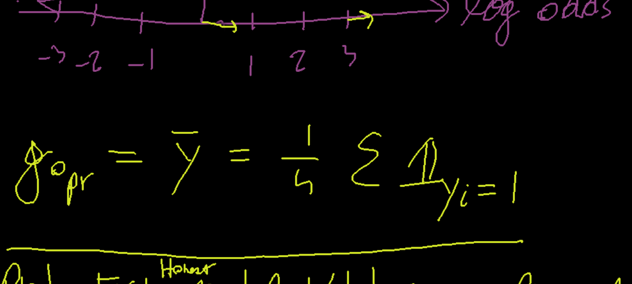
$$g_{pr}(\vec{x}_*) = \frac{1}{1+e^{-\vec{b} \cdot \vec{x}}} = \hat{p}(Y_* = 1 \mid \vec{x}_*) = \hat{p}_*$$
$$\in \mathcal{H}_{pr} \qquad \approx P(Y_* = 1 \mid \vec{x}_*) = \underline{\underline{f_{pr}(\vec{x}_*)}}$$

What is the interpretation of b_j's. OddsAgainst$(Y=1 \mid \vec{x})$

$$\hat{p} = \frac{1}{1+e^{-\vec{b} \cdot \vec{x}}} \Rightarrow \frac{1}{\hat{p}} = 1 + e^{-\vec{b} \cdot \vec{x}} \Rightarrow \frac{1-\hat{p}}{\hat{p}} = \frac{1}{\hat{p}} - 1 = e^{-\vec{b} \cdot \vec{x}}$$

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

$b_3 = 0.7$, x_3 increases by 1 $\Rightarrow$ LogOdds$(Y=1 \mid \vec{x})$ increases by 0.7.



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

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

Prob. Est. Model Validation (Honest). Previously.... $SSE = \sum (y_i - \hat{y}_i)^2$

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...

"Scoring Rule" $S(y, \hat{p})$ ↗ higher the score, the better the model $\approx R^2$

A "proper scoring" rule:

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

Two popular proper scoring rules:

(1) Log Scoring Rule: $s_i = y_i \ln(\hat{p}_i) + (1 - y_i) \ln(1 - \hat{p}_i)$

(2) Brier Score: $s_i = -(y_i - \hat{p}_i)^2$

overall score is average.

$$\Rightarrow S = \frac{1}{n} \sum s_i$$