Classification

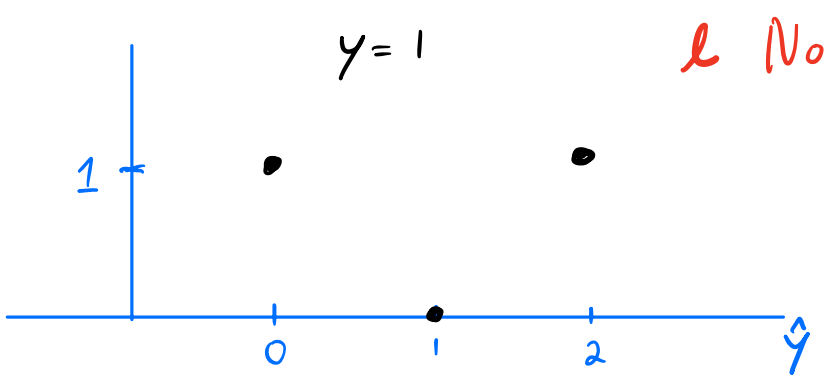
Please do SPOT Survey

Labels are unordered (and usually finite)

Ex: Types of wine $\mathcal{Y} = \{\text{Barolo}, \text{Grignolino}, \text{Barbera}\}$
$= \{0, 1, 2\}$

Loss ℓ is 0-1 loss

$$\ell(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{otherwise} \end{cases}$$



ℓ Not continuous

$$\hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(\vec{x}_i), y_i)$$ ⇒ Not continuous

$$f_{Bayes} = \underset{f \in \{f \mid f: \mathcal{X} \to \mathcal{Y}\}}{\operatorname{argmin}} \; L(f)$$

$$L(f) = \mathbb{E}[\ell(f(\vec{X}), Y)]$$

$$f_{Bayes}(\vec{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; p(y \mid \vec{x})$$

Binary Classification $y = \{0, 1\}$

MLE to estimate pmf $p(y|\vec{x})$

$D = ((\vec{X}_1, Y_1), \ldots, (\vec{X}_n, Y_n))$

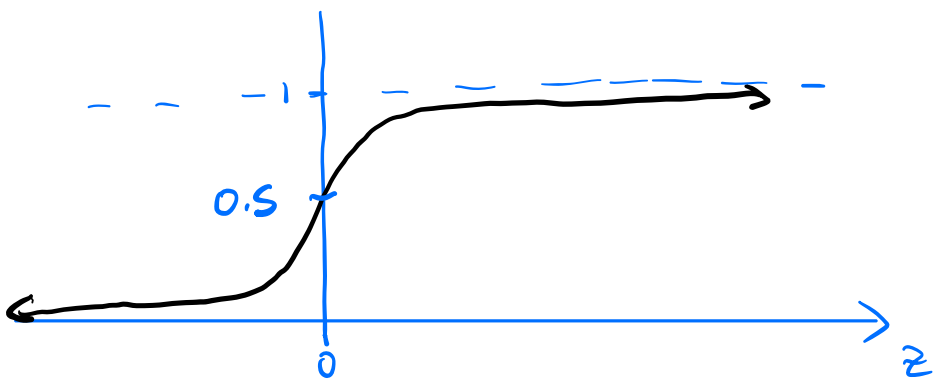
$(\vec{X}_i, Y_i)$ are i.i.d. with $\mathbb{P}_{\vec{X},Y}$, $p_{\vec{X},Y}$

Assume $Y_i | \vec{X}_i = \vec{x}_i \sim \text{Bernoulli}(\alpha^*(\vec{x}_i))$

$= \text{Bernoulli}(\underbrace{\sigma(\vec{x}_i^T \vec{w}^*)}_{\in [0,1]})$

$\sigma(z) = \frac{1}{1+e^{-z}}$ $\quad z \in \mathbb{R}$

"sigmoid" or "logistic"



$$p(y|\vec{x}, \vec{w}) = \sigma(\vec{x}^T\vec{w})^y \cdot (1 - \sigma(\vec{x}^T\vec{w}))^{(1-y)}$$

$$p(y|\vec{x}) = p(y|\vec{x}, \vec{w}^*)$$

$$\vec{w}_{MLE} = \arg\max_{\vec{w} \in \mathbb{R}^{d+1}} p(D \mid \vec{w})$$

$$= \arg\max_{\vec{w} \in \mathbb{R}^{d+1}} \prod_{i=1}^{n} p(\vec{x}_i, y_i \mid \vec{w})$$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} - \sum_{i=1}^{n} \log\left(p(\vec{x}_i, y_i \mid \vec{w})\right)$$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} - \sum_{i=1}^{n} \left[\log\left(p(y_i \mid \vec{x}_i, \vec{w})\right) + \log\left(p(\vec{x}_i \mid \vec{w})\right)\right]$$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} - \sum_{i=1}^{n} \log\left(p(y_i \mid \vec{x}_i, \vec{w})\right)$$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} - \sum_{i=1}^{n} \log\left(\sigma(\vec{x}_i^T \vec{w})^{y_i} \cdot (1 - \sigma(\vec{x}_i^T \vec{w}))^{(1-y_i)}\right)$$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} - \sum_{i=1}^{n} \left[\log\left(\sigma(\vec{x}_i^T \vec{w})^{y_i}\right) + \log\left((1 - \sigma(\vec{x}_i^T \vec{w}))^{(1-y_i)}\right)\right]$$

$\log(a^b) = b \log(a)$

$$= \arg\min_{\vec{w} \in \mathbb{R}^{d+1}} - \sum_{i=1}^{n} \left[\underbrace{y_i \log\left(\sigma(\vec{x}_i^T \vec{w})\right)}_{= h_i = y_i \log(v_i)} + \underbrace{(1-y_i) \log\left((1 - \sigma(\vec{x}_i^T \vec{w}))\right)}_{= r_i = (1-y_i)\log(1-v_i)}\right]$$

$= g(\vec{w})$ convex

$u_i = \vec{x}_i^T \vec{w}$, $v_i = \sigma(u_i)$

$$\frac{\partial g}{\partial w_j} = -\sum_{i=1}^{n}\left[\frac{dh_i}{dv_i}\frac{dv_i}{du_i}\frac{\partial u_i}{\partial w_j} + \frac{dr_i}{dv_i}\frac{dv_i}{du_i}\frac{\partial u_i}{\partial w_j}\right]$$

$$\frac{dh_i}{dv_i} = y_i\frac{1}{v_i} = \frac{y_i}{\sigma(u_i)} \qquad \frac{dr_i}{dv_i} = -(1-y_i)\frac{1}{1-v_i} = -\frac{(1-y_i)}{1-\sigma(u_i)}$$

$$\frac{dv_i}{du_i} = \sigma(u_i)(1-\sigma(u_i)), \quad \frac{\partial u_i}{\partial w_j} = x_{ij}$$

$$\frac{\partial g}{\partial w_j} = -\sum_{i=1}^{n}\left[\frac{y_i}{\sigma(u_i)}\sigma(u_i)(1-\sigma(u_i))x_{ij} - \frac{(1-y_i)}{1-\sigma(u_i)}\sigma(u_i)(1-\sigma(u_i))x_{ij}\right]$$

$$= -\sum_{i=1}^{n}\left[y_i(1-\sigma(u_i))x_{ij} - (1-y_i)\sigma(u_i)x_{ij}\right]$$

$$= -\sum_{i=1}^{n}\left[y_i x_{ij} - y_i\sigma(u_i)x_{ij} - \sigma(u_i)x_{ij} + y_i\sigma(u_i)x_{ij}\right]$$

$$= -\sum_{i=1}^{n}\left[y_i x_{ij} - \sigma(u_i)x_{ij}\right]$$

$$\frac{\partial g}{\partial w_j} = \sum_{i=1}^{n}\left[(\sigma(\vec{x}_i^{\,T}\vec{w}) - y_i)x_{ij}\right] \stackrel{?}{=} 0$$

No Closed form Soln

$j \in \{0, 1, \ldots, d\}$

Gradient Descent:

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} - \eta^{(t)}\nabla g(\vec{w}^{(t)})$$

$$\nabla g(\vec{w}^{(t)}) = \sum_{i=1}^{n} (\sigma(\vec{x}_i^T \vec{w}) - y_i)\vec{x}_i \in \mathbb{R}^{d+1}$$

$$\vec{w}^{(T)} \approx \vec{w}_{MLE}$$

$$\sigma^*(\vec{x}) = \sigma(\vec{x}^T \vec{w}^*) \simeq \sigma(\vec{x}^T \vec{w}_{MLE}) = f_{MLE}(\vec{x})$$

$$\mathcal{A}(D) = \hat{f}_{Bin}$$

$$f_{Bayes}(\vec{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(y|\vec{x})$$

$$= \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(y|\vec{x}, \vec{w}^*)$$

$$\approx \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(y|\vec{x}, \vec{w}_{MLE})$$

$$= \hat{f}_{Bin}(\vec{x})$$

$$p(y=1|\vec{x}, \vec{w}_{MLE}) = \sigma(\vec{x}^T \vec{w}_{MLE}) = f_{MLE}(\vec{x})$$

$$p(y=0|\vec{x}, \vec{w}_{MLE}) = 1 - \sigma(\vec{x}^T \vec{w}_{MLE}) = 1 - f_{MLE}(\vec{x})$$

$$= \begin{cases} 1 & \text{if } f_{MLE}(\vec{x}) \geq 0.5 \\ 0 & \text{if } f_{MLE}(\vec{x}) < 0.5 \end{cases}$$

$$= \begin{cases} 1 & \text{if } \vec{x}^T \vec{w} \geq 0 \\ 0 & \text{if } \vec{x}^T \vec{w} < 0 \end{cases}$$

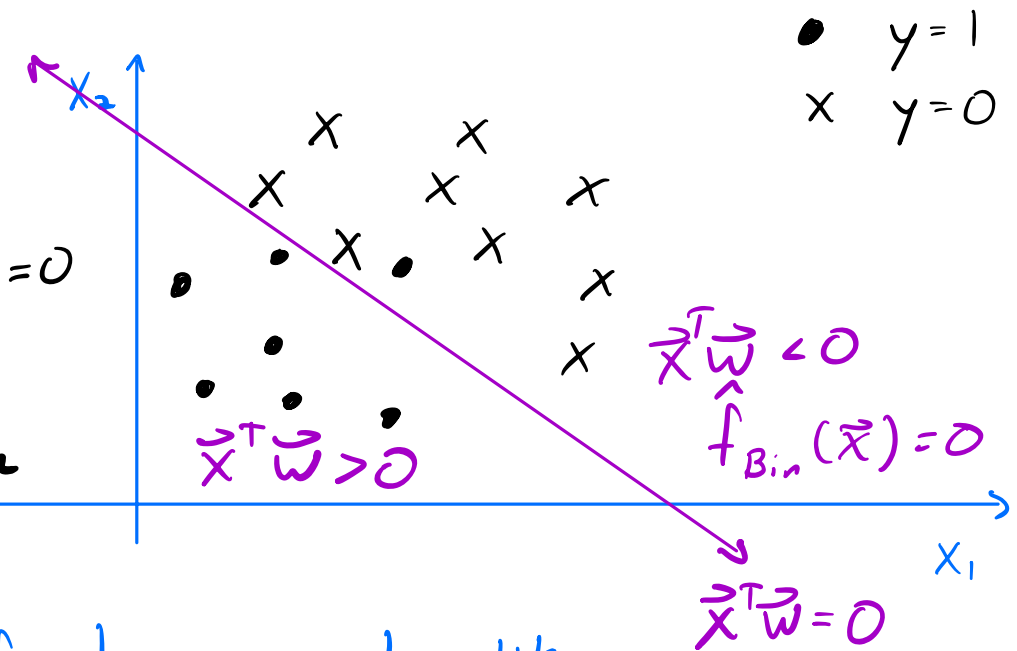
$\vec{x}^T \vec{w} = 0$ is d-1 dimensional hyperplane

Ex: d = 2

$\vec{x}^T \vec{w} = 0$

$w_0 + x_1 w_1 + x_2 w_2 = 0$

$x_2 = -x_1 \frac{w_1}{w_2} - \frac{w_0}{w_2}$



Polynomial features apply like before

MAP gives Regularization like before

Logistic Regression

$\mathcal{Y}_{Log} = [0,1]$ representing $p(y=1|\vec{x}) = \alpha^*(x)$

$$\ell(f(\vec{x}), y) = -\left(y \log(f(\vec{x})) + (1-y) \log((1 - f(\vec{x})))\right)$$

Binary Cross-Entropy

$$\mathcal{F} = \{f : \mathcal{X} \to [0,1] : f(\vec{x}) = \sigma(\vec{x}^T \vec{w}), \ \vec{w} \in \mathbb{R}^{d+1}\}$$

Learner: $A(D) = \hat{f}$

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{L}(f)$$

$$= \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} g(\vec{w})$$

$$= f_{MLE}$$

Why not just learn $p(y=1|\vec{x})$ using linear regression?

Ex:

$d=1$