Muticlass Classification $\mathcal{Y} = \{0, \dots, K-1\}$

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

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

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

$$D = \left((\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n) \right)$$

$$(\vec{X}_i, Y_i) \overset{i.i.d.}{\sim} P_{\vec{X}, Y} \qquad p(y \mid \vec{x})$$

$$Y_i \mid \vec{X}_i = \vec{x}_i \sim \text{Categorical}\Big(\overbrace{\alpha_0^*(\vec{x}_i), \dots, \alpha_{K-1}^*(\vec{x}_i)}^{\alpha^*(\vec{x}_i)}\Big)$$

$$p(y=0 \mid \vec{x}_i) = \alpha_0^*(\vec{x}_i)$$

$$\vdots$$

$$p(y=K-1 \mid \vec{x}_i) = \alpha_{K-1}^*(\vec{x}_i)$$

$$\sum_{y=0}^{K-1} \alpha_y^*(\vec{x}_i) = 1$$

$$\alpha_y^*(\vec{x}_i) = \nabla_y\Big(\vec{x}_i^T \overset{\in \mathbb{R}^{d+1}}{\vec{w}_0^*}, \dots, \vec{x}_i^T \vec{w}_{K-1}^*\Big)$$

$$= \frac{\exp(\vec{x}_i^T \vec{w}_y^*)}{\sum_{q=0}^{K-1} \exp(\vec{x}_i^T \vec{w}_q^*)} \in [0, 1] \quad \text{softmax}$$

$$\sum_{y=0}^{K-1} \alpha_y^*(\vec{x}_i) = \sum_{y=0}^{K-1} \frac{\exp(\vec{x}_i^T \vec{w}_y^*)}{\sum_{q=0}^{K-1} \exp(\vec{x}_i^T \vec{w}_q^*)}$$

$$= \frac{\sum_{y=0}^{K-1} \exp(\vec{x}_i^T \vec{w}_y^*)}{\sum_{q=0}^{K-1} \exp(\vec{x}_i^T \vec{w}_q^*)}$$

$$= 1$$

$$\vec{w}_{MLE,0}, \vec{w}_{MLE,1}, \dots, \vec{w}_{MLE,k-1}$$

$$= \underset{\vec{w}_0,\dots\vec{w}_{k-1}\in\mathbb{R}^{d+1}}{\operatorname{argmin}} -\sum_{i=1}^{n} \log\left(p(y_i \mid \vec{x}_i, \vec{w}_0, \dots, \vec{w}_{k-1})\right)$$

$$= \underset{\vec{w}_0,\dots\vec{w}_{k-1}\in\mathbb{R}^{d+1}}{\operatorname{argmin}} -\sum_{i=1}^{n} \log\left(\nabla_{y_i}(\vec{x}_i^T\vec{w}_0, \dots, \vec{x}_i^T\vec{w}_{k-1})\right)$$

$$= \underset{\vec{w}_0,\dots\vec{w}_{k-1}\in\mathbb{R}^{d+1}}{\operatorname{argmin}} -\sum_{i=1}^{n} \log\left(\frac{\exp(\vec{x}_i^T\vec{w}_{y_i})}{\sum_{q=0}^{k-1}\exp(\vec{x}_i^T\vec{w}_q)}\right)$$

$\log(a/b) = \log(a) - \log(b)$

$$= \underset{\vec{w}_0,\dots\vec{w}_{k-1}\in\mathbb{R}^{d+1}}{\operatorname{argmin}} -\sum_{i=1}^{n} \left[\log\left(\exp(\vec{x}_i^T\vec{w}_{y_i})\right) - \log\left(\sum_{q=0}^{k-1}\exp(\vec{x}_i^T\vec{w}_q)\right)\right]$$

$$= \underset{\vec{w}_0,\dots\vec{w}_{k-1}\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \underbrace{-\sum_{i=1}^{n} \left[\vec{x}_i^T\vec{w}_{y_i} - \log\left(\sum_{q=0}^{k-1}\exp(\vec{x}_i^T\vec{w}_q)\right)\right]}_{=g(\vec{w}_0,\dots,\vec{w}_{k-1}) \;=\; g(\vec{w})}$$

Convex

$$\vec{w} = \begin{pmatrix} | \\ \vec{w}_0 \\ | \\ | \\ \vec{w}_1 \\ | \\ \vdots \\ | \\ w_{k-1} \\ | \end{pmatrix} \in \mathbb{R}^{(d+1)k}$$

$$\frac{\partial g}{\partial \vec{w}_{y_j}} = \text{see course notes}$$

$y \in \{0, \dots, k-1\}$

$j \in \{0, \dots, d\}$

$$= -\sum_{i=1}^{n} \left(\mathbb{I}_{\{y_i\}}(y)\, x_{ij} - \nabla_y(\vec{x}_i^T \vec{w}_0, \dots, \vec{x}_i^T \vec{w}_{k-1})\, x_{ij} \right)$$

$$\mathbb{I}_{\mathcal{Z}}(z) = \begin{cases} 1 & \text{if } z \in \mathcal{Z} \\ 0 & \text{otherwise} \end{cases}$$

$$= \sum_{i=1}^{n} \left(\nabla_y(\vec{x}_i^T \vec{w}_0, \dots, \vec{x}_i^T \vec{w}_{k-1}) - \mathbb{I}_{\{y_i\}}(y) \right) x_{ij}$$

No closed form solution

Gradient Descent

$$\nabla_{\vec{w}_y} g = \left(\frac{\partial g}{\partial w_{y0}}, \dots, \frac{\partial g}{\partial w_{yd}} \right)^T \in \mathbb{R}^{d+1} \qquad y \in \mathcal{Y}$$

$$\vec{w}_y^{(t+1)} = \vec{w}_y^{(t)} - \eta^{(t)} \nabla_{\vec{w}_y} g(\vec{w}_0^{(t)}, \dots, \vec{w}_{K-1}^{(t)})$$

for each step t you do the above calculation for all $y \in \mathcal{Y}$

Multiclass Classification Learner

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

$$p(y \mid \vec{x}, \vec{w}_0^*, \dots, \vec{w}_{K-1}^*)$$

$$//$$

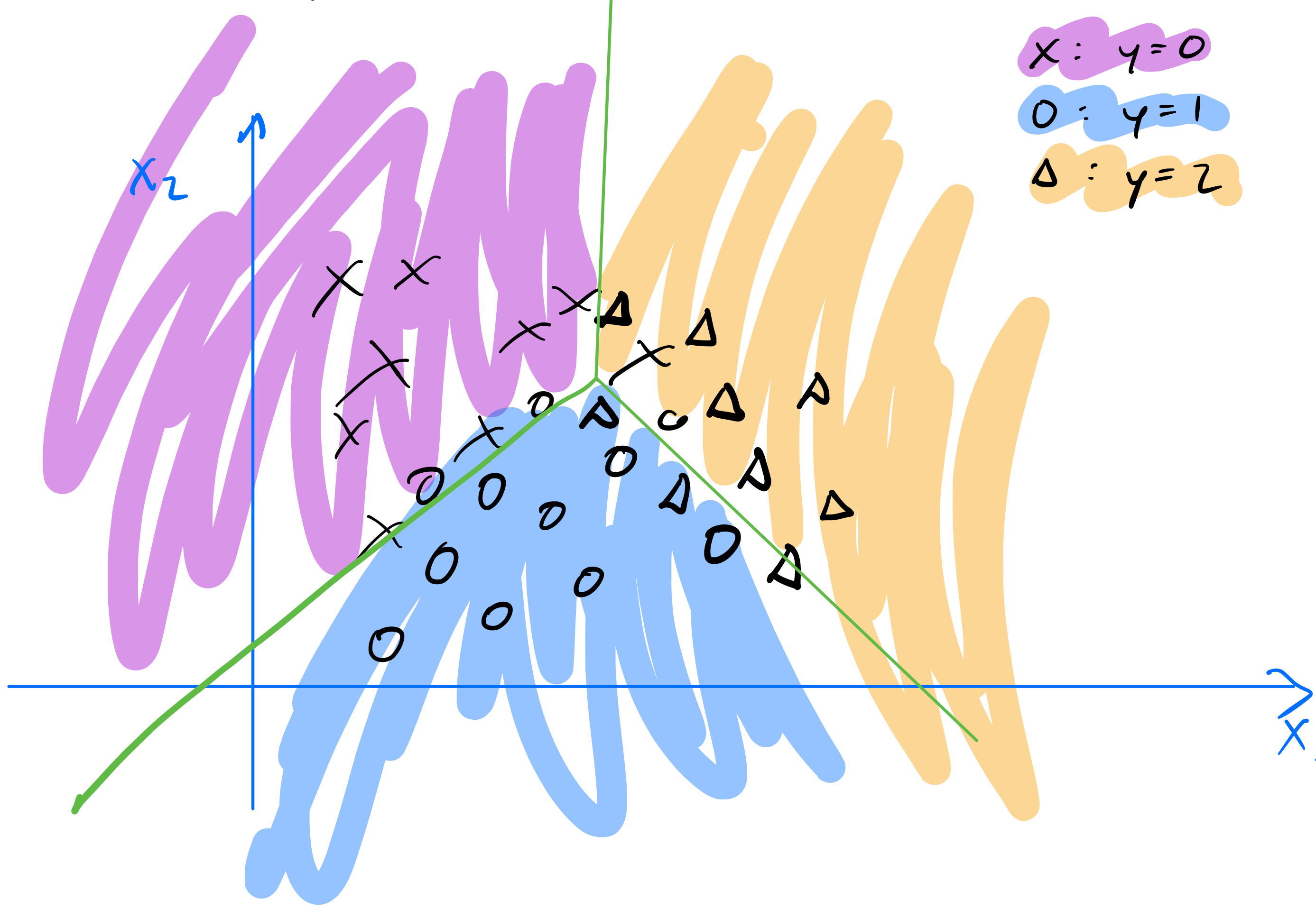
$$f_{Bayes}(\vec{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ p(y \mid \vec{x})$$

$$\approx \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ \underbrace{p(y \mid \vec{x}, \vec{w}_{MLE,0}, \dots, \vec{w}_{MLE,K-1})}_{=f_{MLE}}$$

$$\approx \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ \sigma_y(\vec{x}_i^T \vec{w}_{MLE,0}, \dots, \vec{x}_i^T \vec{w}_{MLE,K-1})$$

$$= \hat{f}_{Mul}$$

$p(y=0 \mid \vec{x}_i, \dots) = p(y=2, \vec{x}_i, \dots)$



Softmax Regression

$y_{soft} = [0,1]^K$ representing $p(y=0|\vec{x}), \dots, p(y=k-1|\vec{x})$

Multi-output Regression

labels: Ex: $k=3$ $y = (1,0,0)^T$: Class 0

$(0\ 1,0)^T$: Class 1

$(0,0,1)$: Class 2

$$\ell(\hat{\vec{y}}, \vec{y}) = -\sum_{q=0}^{k-1} \left[y_q \log(\hat{y}_q) \right]$$

Multiclass Cross-Entropy loss or Log loss

$$\mathcal{F} = \left\{ f \mid f : \mathbb{R}^{d+1} \Rightarrow [0,1]^K \text{ and } f(\vec{x}) = \sigma(\vec{x}^T \vec{w}_0, \dots, \vec{x}^T \vec{w}_{K-1}),\ \vec{w}_0, \dots, \vec{w}_{K-1} \in \mathbb{R}^{d+1} \right\}$$

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

$= f_{MLE}$

Softmax Regression $(K=2)$ = Logistic Regression

course notes for details

$$\sigma_1\left(\vec{x}^T\vec{w}_{MLE,0}, \vec{x}^T\vec{w}_{MLE,1}\right) = \sigma\left(\vec{x}^T\vec{w}_{MLE}\right)$$