Multiclass Classification $y = \{0, \ldots, K-1\}$

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 $Y_i | \vec{X}_i = \vec{x}_i \sim \text{Categorical}(\underbrace{\alpha^*_0(\vec{x}_i), \ldots, \alpha^*_{K-1}(\vec{x}_i)}_{\vec{\alpha}^*(\vec{x}_i)})$

$$p(y=0|\vec{x}_i) = \alpha^*_0(\vec{x}_i) \qquad \sum_{y=0}^{K-1} \alpha^*_y(\vec{x}_i) = 1$$

$$\vdots$$

$$p(y=K-1|\vec{x}_i) = \alpha^*_{K-1}(\vec{x}_i)$$

$\sigma(\vec{z}) = (\sigma_0(\vec{z}), \ldots, \sigma_{K-1}(\vec{z})) \in [0,1]^K$ "softmax"

$$\sigma_y(z_0, \ldots, z_{K-1}) = \frac{\exp(z_y)}{\sum_{q=0}^{K-1} \exp(z_q)}$$

$$p(y|\vec{x}) = \sigma_y(\vec{x}^T\vec{w}^*_0, \ldots, \vec{x}^T\vec{w}^*_{K-1})$$

$$\vec{\alpha}^*(\vec{x}_i) = \sigma(\vec{x}_i^T\vec{w}^*_0, \ldots, \vec{x}_i^T\vec{w}^*_{K-1})$$

$\vec{w}_{MLE,0}, \ldots \vec{w}_{MLE,K-1}$

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

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

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

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

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

$= g(\vec{w}_0,\dots,\vec{w}_{K-1})$ (Convex)

$\log\{\cdots\} = h_i(r_i)$

$u_{iq} = \vec{x}_i^T\vec{w}_q$, $v_{iq} = \exp(u_{iq})$, $r_i = \sum_{q=0}^{K-1}\exp(u_{iq})$

$\vec{w}_y = (w_{y0}, w_{y1}, \dots, w_{yd})^T \in \mathbb{R}^{d+1}$ for all $y \in \mathcal{Y}$

$$\frac{\partial g}{\partial w_{yj}}(\vec{w}_0,\dots,\vec{w}_{K-1}) = -\sum_{i=1}^{n}\left[\frac{\partial u_{iy_i}}{\partial w_{yj}} - \frac{dh_i}{dr_i}\sum_{q=0}^{K-1}\frac{dv_{iq}}{du_{iq}}\frac{\partial u_{iq}}{\partial w_{yj}}\right]$$

$$\frac{dh_i}{dr_i} = \frac{1}{r_i}, \quad \frac{dv_{iq}}{du_{iq}} = \exp(u_{iq}), \quad \frac{\partial u_{iq}}{\partial w_{yj}} = \begin{cases} x_{ij} & \text{if } y = q \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial u_{iy_i}}{\partial w_{yj}} = \mathbb{I}_{\{y_i\}}(y)\, x_{ij}$$

$$= \mathbb{I}_{\{q\}}(y)\, x_{ij}$$

Indicator function: $\mathbb{I}_{\mathcal{Z}}$

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

$$= -\sum_{i=1}^{n}\left[\mathbb{I}_{\{y_i\}}(y)\,x_{ij} - \frac{1}{\sum_{q=0}^{K-1}\exp(\vec{x}_i^{\,T}\vec{w}_q)}\sum_{q=0}^{K-1}\left(\exp(\vec{x}_i^{\,T}\vec{w}_q)\,\mathbb{I}_{\{q\}}(y)\,x_{ij}\right)\right]$$

$$= -\sum_{i=1}^{n}\left[\mathbb{I}_{\{y_i\}}(y)\,x_{ij} - x_{ij}\,\frac{\exp(\vec{x}_i^{\,T}\vec{w}_y)}{\sum_{q=0}^{K-1}\exp(\vec{x}_i^{\,T}\vec{w}_q)}\right]$$

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

$$\frac{\partial g}{\partial w_{yj}}(\vec{w}_0,\ldots,\vec{w}_{K-1}) = \sum_{i=1}^{n}\left(\sigma_y(\vec{x}_i^{\,T}\vec{w}_0,\ldots,\vec{x}_i^{\,T}\vec{w}_{K-1}) - \mathbb{I}_{\{y_i\}}(y)\right)x_{ij}$$

for all $y \in \mathcal{Y}$, $j \in \{0,\ldots,d\}$

No closed form solution

Use gradient descent

$$\nabla_{\vec{w}_y} g(\vec{w}_0,\ldots,\vec{w}_{K-1}) = \left(\frac{\partial g}{\partial \vec{w}_{y0}},\ldots,\frac{\partial g}{\partial \vec{w}_{yd}}\right)^T \in \mathbb{R}^{d+1} \quad \text{for } y \in \mathcal{Y}$$

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

Multiclass Classification Learner:

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

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

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

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

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

Softmax Regression

$y_{soft} = [0,1]^K$ representing values of

$$\left(p(y=0|\vec{x}), \ldots, p(y=K-1|\vec{x})\right)$$

Labels: $K=3$, $\vec{y}_i = (1,0,0) \Leftrightarrow$ class 0 from before

$= (0,1,0) \Leftrightarrow$ Class 1

$= (0,0,1) \Leftrightarrow$ 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"

$$\tilde{F} = \{f \mid f: \mathbb{R}^{d+1} \to [0,1]^K \text{ and } f(\vec{x}) = \sigma(\vec{x}^T\vec{w}_0, \ldots, \vec{x}^T\vec{w}_{K-1})$$

where $\vec{w}_0, \ldots, \vec{w}_{K-1} \in \mathbb{R}^{d+1}\}$

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

$$\hat{f}_{ERM} = \arg\min_{f \in \tilde{F}} \hat{L}(f)$$

$$= \arg\min_{f \in \tilde{F}} \frac{1}{n}\sum_{i=1}^{n} -\sum_{q=0}^{K-1}\left[y_{iq}\log(f_q(\vec{x}_i))\right]$$

$$= \arg\min_{f \in \tilde{F}} -\sum_{i=1}^{n}\sum_{q=0}^{K-1}\left[y_{iq}\log(f_q(\vec{x}_i))\right]$$

$$= f_{MLE}$$

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

Comparison to Logistic Regression

If $K = 2$ and $\mathcal{Y} = \{0, 1\}$

Assume $Y_i | \vec{X}_i = \vec{x}_i \sim \text{Categorical}\left(\alpha_0^*(\vec{x}_i), \alpha_1^*(\vec{x}_i)\right)$

$= \text{Bernoulli}\left(\alpha^*(\vec{x}_i) = \alpha_1^*(\vec{x}_i)\right)$

$$p(y=0|\vec{x}_i) = \alpha_0^*(\vec{x}_i), \quad p(y=1|\vec{x}_i) = \alpha_1^*(\vec{x}_i)$$

$$\alpha_0^*(\vec{x}_i) + \alpha_1^*(\vec{x}_i) = 1 \Rightarrow \alpha_0^*(\vec{x}_i) = 1 - \alpha_1^*(\vec{x}_i)$$

$$\alpha_1^*(\vec{x}) = \sigma_1(\vec{x}^T\vec{w}_0^*, \vec{x}^T\vec{w}_1^*), \quad \alpha^*(\vec{x}) = \sigma(\vec{x}^T\vec{w}^*)$$

Goal: Show $\sigma_1(\vec{x}^T\vec{w}_{MLE,0}, \vec{x}^T\vec{w}_{MLE,1}) = \sigma(\vec{x}^T\vec{w}_{MLE})$

$$\sigma_1(z_0, z_1) = \frac{\exp(z_1)}{\exp(z_1) + \exp(z_0)}$$

$$= \frac{\exp(z_1)}{\exp(z_0) + \exp(z_1)} \cdot \frac{\exp(-z_1)}{\exp(-z_1)}$$

$$= \frac{\exp(z_1 - z_1)}{\exp(z_0 - z_1) + \exp(z_0 - z_1)}$$

$\exp(a)\exp(b) = \exp(a+b)$

$$= \frac{1}{\exp(z_0 - z_1) + 1}$$

$$\sigma_1(\vec{x}^T\vec{w}_0, \vec{x}^T\vec{w}_1) = \frac{1}{\exp(\vec{x}^T\vec{w}_0 - \vec{x}^T\vec{w}_1) + 1}$$

$$= \frac{1}{1 + \exp(-\vec{x}^T(\underbrace{\vec{w}_1 - \vec{w}_0}_{= \vec{w}}))} = \sigma(\vec{x}^T\vec{w})$$

$$\sigma_0(z_0, z_1) = 1 - \sigma_1(z_0, z_1)$$

Multiclass

$$\vec{w}_{MLE,0}, \vec{w}_{MLE,1}$$

$$= \underset{\vec{w}_0, \vec{w}_1 \in \mathbb{R}^{d+1}}{\arg\min} - \sum_{i=1}^{n} \log\left(p(y_i | x_i, \vec{w}_0, \vec{w}_1)\right)$$

$$= \underset{\vec{w}_0, \vec{w}_1 \in \mathbb{R}^{d+1}}{\arg\min} - \sum_{i=1}^{n} \log\left(\sigma_{y_i}(\vec{x}_i^T\vec{w}_0, \vec{x}_i^T\vec{w}_1)\right)$$

$$= \underset{\vec{w}_0, \vec{w}_1 \in \mathbb{R}^{d+1}}{\arg\min} - \sum_{i=1}^{n} \left[y_i \log\left(\sigma_1(\vec{x}_i^T\vec{w}_0, \vec{x}_i^T\vec{w}_1)\right) + (1 - y_i)\log\left(\sigma_0(\vec{x}_i^T\vec{w}_0, \vec{x}_i^T\vec{w}_1)\right) \right]$$

$$= \underset{\vec{w}_0, \vec{w}_1 \in \mathbb{R}^{d+1}}{\arg\min} \underbrace{- \sum_{i=1}^{n} \left[y_i \log\left(\sigma(\vec{x}_i^T(\vec{w}_1 - \vec{w}_0))\right) + (1 - y_i)\log\left(\sigma_0(\vec{x}_i^T(\vec{w}_1 - \vec{w}_0))\right) \right]}_{= g(\vec{w}_0, \vec{w}_1) = h(\vec{w}) \text{ where } \vec{w} = \vec{w}_1 - \vec{w}_0}$$

Binary

$$\vec{w}_{MLE} = \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} - \sum_{i=1}^{n} \left[y_i \log\left(\sigma(\vec{x}_i^T \vec{w})\right) + (1-y_i) \log\left(\sigma_0(\vec{x}_i^T \vec{w})\right) \right]$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} \; h(\vec{w})$$

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

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