Oct. 24, 2019 Pairs samples (x_i, y_i) e.g $y_i = [0\,0\,1\,0]$

k classes e.g $k=4$

Reminder: $\sigma(x^T w) = \dfrac{\exp(x^T w)}{1+\exp(x^T w)}$

Contrast (one-vs-rest): Separate log. regression predictor

$$\text{probs} = \begin{bmatrix} \sigma(x^T w_1) \\ \sigma(x^T w_2) \\ \vdots \\ \sigma(x^T w_k) \end{bmatrix} = \begin{bmatrix} \exp(x^T w_1) / (1+\exp(x^T w_1)) \\ \vdots \\ \exp(x^T w_k) / (1+\exp(x^T w_k)) \end{bmatrix}$$

probs $\in \mathbb{R}^k$

$x \in \mathbb{R}^d$

$w_1 \in \mathbb{R}^d, \ldots, w_k \in \mathbb{R}^d$ $\quad W = [w_1\ w_2\ \ldots\ w_k]$ $d \times k$

Softmax for multinomial logistic regression

$$\text{probs} = \text{softmax}(x^T W) = \begin{bmatrix} \exp(x^T w_1) / \sum_{m=1}^{k} \exp(x^T w_m) \\ \vdots \\ \exp(x^T w_k) / \sum_{m=1}^{k} \exp(x^T w_m) \end{bmatrix}$$

<u>Difference 1:</u> probs for softmax sum to 1 but not for seperate binary classifiers

e.g. $\exp(x^T w_1)$ is really big (1000) (for class 1)
what does this mean for prob of class 2

<u>Modelling assumpt.</u> $p(y|x)$ as a joint distribution

e.g $p(y = [0\ 1\ 0\ 0] \mid x)$

Separate binary case: we use $\rightarrow [p(y_1 = 1 \mid x), \ldots, p(y_k = 1 \mid x)]$

Update: for each $m = 1, \ldots, k$. $(\text{softmax}(x^T W)_m - y_m)\, x$

generic for GLMs: $(\underbrace{f(x^T W)}_{\hat{y}} - y)\, x$

Objective: $p(y \mid x) = p(y_1 = 1 \mid x)^{y_1}\, p(y_2 = 1 \mid x)^{y_2} \cdots p(y_k = 1 \mid x)^{y_k}$

minimize - log-likelihood

$$-\ln p(y \mid x) = -\sum_{m=1}^{k} \ln \left[p(y_m = 1 \mid x)^{y_m} \right]$$

s.t. only one $y_m = 1$

$$= -\sum_{m=1}^{k} y_m \ln p(y_m = 1 \mid x)$$

$$\ln p(y_m = 1 | x) = \ln \frac{\exp(x^T w_m)}{\sum_{r=1}^{k} \exp(x^T w_r)}$$

$$\Rightarrow -\ln p(y|x) = -\sum_{m=1}^{k} y_m \left[\ln \exp(x^T w_m) - \ln \sum_{r=1}^{k} \exp(x^T w_r) \right]$$

$$= -\sum_{m=1}^{k} y_m \left[x^T w_m - \ln \sum_{r=1}^{k} \exp(x^T w_r) \right]$$

<u>Exercise</u>: Show $\frac{\partial}{\partial \vec{w}_m} [-\ln p(y|x)] = \left[\frac{\exp(x^T w_m)}{\sum_{r=1}^{k} \exp(x^T w_r)} - y_m \right] x$

$$\nabla - \sum_{i=1}^{n} \ln p(y_i | x_i) \stackrel{?}{=} 0$$

Is there a closed form for W ?