1 Delta Function

1.1 Stochastic gradient descent

n features – k classes – m samples

$$\theta = \begin{bmatrix} b_1 & b_2 & \cdots & b_k \\ w_{11} & w_{21} & \cdots & w_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n} & w_{21} & \cdots & w_{kn} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_k \end{bmatrix}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$z = \theta^T x = \begin{bmatrix} b_1 + w_{11}x_1 + w_{12}x_2 + \cdots + w_{1k}x_k \\ b_2 + w_{21}x_1 + w_{22}x_2 + \cdots + w_{2k}x_k \\ \vdots \\ b_n + w_{n1}x_1 + w_{n2}x_2 + \cdots + w_{nk}x_k \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

$$\hat{y} = \frac{e^z}{\sum_{i=1}^k e^{z_i}} = \begin{bmatrix} \frac{e_1^z}{\sum_{i=1}^k e^{z_i}} \\ \frac{e_2^z}{\sum_{i=1}^k e^{z_i}} \\ \vdots \\ \frac{e_k^z}{\sum_{i=1}^k e^{z_i}} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_k \end{bmatrix}$$

Loss Function

$$L(\theta) = -\sum_{i=1}^{k} \delta(i,y) log\hat{y}_i$$

$$and \;\; \delta(i,j) = \begin{cases} 1 & if \;\; i = j \\ 0 & if \;\; i \neq j \end{cases}$$

Derivative

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{\partial(-\sum_{i=1}^k \delta(i,y) log\hat{y}_i)}{\partial \hat{y}_i} = -\frac{\delta(i,y)}{\hat{y}_i}$$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1-\hat{y}_j) & if \;\; i = j \\ -\hat{y}_i\hat{y}_j & if \;\; i \neq j \end{cases} => \frac{\partial \hat{y}_i}{\partial z_j} = \hat{y}_i(\delta(i,j) - \hat{y}_j)$$

$$\frac{\partial L}{\partial z_i} = \frac{\partial(-\sum_{j=1}^k \delta(j,y) log\hat{y}_j)}{\partial z_i} = -\frac{\delta(i,y)}{\hat{y}_i}\hat{y}_i(1-\hat{y}_i) - \sum_{j\neq i}^{k} \frac{\delta(j,y)}{\hat{y}_j}(-\hat{y}_j\hat{y}_i))$$

$$= -\delta(i,y)(1-\hat{y}_i) + \sum_{j\neq i}^{k} \delta(j,y)\hat{y}_i = -\delta(i,y) + \sum_{j=1}^{k} \delta(j,y)\hat{y}_i = \hat{y}_i - \delta(i,y)$$

$$\frac{\partial L}{\partial w_{ij}} = x_j(\hat{y}_i - \delta(i,y))$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - \delta(i,y)$$

$$=> \frac{\partial L}{\partial \theta_i} = x(\hat{y}_i - \delta(i,y))$$

1.2 Batch gradient descent

Loss Function

$$L(\theta) = -\sum_{u=1}^{m}\sum_{i=1}^{k} \delta(i, y^{(u)}) log \hat{y}_i^{(u)}$$

Derivative

$$\frac{\partial L}{\partial \theta_i} = \frac{1}{m}\sum_{u=1}^{m} x^{(u)}(\hat{y}_i^{(u)} - \delta(i, y^{(u)}))$$

2 One-hot encoding

2.1 Stochastic gradient descent

n features – k classes – m samples

$$\theta = \begin{bmatrix} b_1 & b_2 & \cdots & b_k \\ w_{11} & w_{21} & \cdots & w_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n} & w_{21} & \cdots & w_{kn} \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_k \end{bmatrix}$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$z = \theta^T x = \begin{bmatrix} b_1 + w_{11}x_1 + w_{12}x_2 + \cdots + w_{1k}x_k \\ b_2 + w_{21}x_1 + w_{22}x_2 + \cdots + w_{2k}x_k \\ \vdots \\ b_n + w_{n1}x_1 + w_{n2}x_2 + \cdots + w_{nk}x_k \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

$$\hat{y} = \frac{e^z}{\sum_{i=1}^{k} e^{z_i}} = \begin{bmatrix} \frac{e_1^z}{\sum_{i=1}^{k} e^{z_i}} \\ \frac{e_2^z}{\sum_{i=1}^{k} e^{z_i}} \\ \vdots \\ \frac{e_k^z}{\sum_{i=1}^{k} e^{z_i}} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_k \end{bmatrix}$$

Loss Function

$$L(\theta) = -\sum_{i=1}^{k} y_i log \hat{y}_i$$

Derivative

$$\frac{\partial L}{\partial \hat{y}_i} = \frac{\partial(-\sum_{i=1}^{k} y_i log \hat{y}_i)}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

$$\frac{\partial \hat{y}_i}{\partial z_j} = \begin{cases} \hat{y}_i(1-\hat{y}_j) & if \;\; i = j \\ -\hat{y}_i \hat{y}_j & if \;\; i \neq j \end{cases}$$

$$\frac{\partial L}{\partial z_i} = \frac{\partial(-\sum_{j=1}^{k} y_j log \hat{y}_j)}{\partial z_i} = -\frac{y_i}{\hat{y}_i}\hat{y}_i(1-\hat{y}_i) - \sum_{j \neq i}^{k} \frac{y_j}{\hat{y}_j}(-\hat{y}_j \hat{y}_i))$$

$$= y_i(\hat{y}_i - 1) + \sum_{j \neq i}^{k} y_j \hat{y}_i = \hat{y}_i - y_i$$

$$\frac{\partial L}{\partial w_{ij}} = x_j(\hat{y}_i - y_i)$$

$$\frac{\partial L}{\partial b_i} = \hat{y}_i - y_i$$

$$=> \frac{\partial L}{\partial \theta_i} = x(\hat{y}_i - y_i)$$

2.2 Batch gradient descent