W3WI DS304.1 Applied Machine Learning Fundamentals

Derivation of the Gradient for Softmax Regression

$$\frac{\partial}{\partial \Theta_{ij}} \mathcal{J}(\boldsymbol{\Theta}) = -\frac{\partial}{\partial \Theta_{ij}} \left(\sum_{k=1}^{K} y_k \log\big(g_k(\boldsymbol{z})\big) \right)$$

$$= -\sum_{k=1}^{K} y_k \cdot \frac{\partial}{\partial \Theta_{ij}} \log\big(g_k(\boldsymbol{z})\big)$$

[**Apply chain rule**]

$$= -\sum_{k=1}^{K} y_k \cdot \frac{\partial \log\big(g_k(\boldsymbol{z})\big)}{\partial g_k(\boldsymbol{z})} \cdot \frac{\partial g_k(\boldsymbol{z})}{\partial z_j} \cdot \frac{\partial z_j}{\partial \Theta_{ij}}$$

[**Derivative of** $\log$ **(first factor produced by chain rule)**]

$$= -\sum_{k=1}^{K} y_k \cdot \frac{1}{g_k(\boldsymbol{z})} \cdot \frac{\partial g_k(\boldsymbol{z})}{\partial z_j} \cdot \frac{\partial z_j}{\partial \Theta_{ij}}$$

[**Separate cases** $k = j$ **and** $k \neq j$]

$$= \overbrace{-y_j \cdot \frac{1}{g_j(\boldsymbol{z})} \cdot \frac{\partial g_j(\boldsymbol{z})}{\partial z_j} \cdot \frac{\partial z_j}{\partial \Theta_{ij}}}^{k=j} - \sum_{\substack{k=1 \\ k\neq j}}^{K} \overbrace{y_k \cdot \frac{1}{g_k(\boldsymbol{z})} \cdot \frac{\partial g_k(\boldsymbol{z})}{\partial z_j} \cdot \frac{\partial z_j}{\partial \Theta_{ij}}}^{k\neq j}$$

$$= \left(-y_j \cdot \frac{1}{g_j(\boldsymbol{z})} \cdot \frac{\partial g_j(\boldsymbol{z})}{\partial z_j} - \sum_{\substack{k=1 \\ k\neq j}}^{K} y_k \cdot \frac{1}{g_k(\boldsymbol{z})} \cdot \frac{\partial g_k(\boldsymbol{z})}{\partial z_j} \right) \cdot \frac{\partial z_j}{\partial \Theta_{ij}}$$

[**Derivative of the softmax function**]

$$= \left(-y_j \cdot \frac{1}{g_j(\boldsymbol{z})} \cdot g_j(\boldsymbol{z}) \cdot (1 - g_j(\boldsymbol{z})) + \sum_{\substack{k=1 \\ k\neq j}}^{K} y_k \cdot \frac{1}{g_k(\boldsymbol{z})} \cdot g_k(\boldsymbol{z}) \cdot g_j(\boldsymbol{z}) \right) \cdot \frac{\partial z_j}{\partial \Theta_{ij}}$$

[**Cancel terms**]

$$= \left(-y_j + y_j \cdot g_j(\boldsymbol{z}) + \sum_{\substack{k=1 \\ k\neq j}}^{K} y_k \cdot g_j(\boldsymbol{z}) \right) \cdot \frac{\partial z_j}{\partial \Theta_{ij}}$$

[**Put the two cases** $k = j$ **and** $k \neq j$ **back together**]

$$= \left(-y_j + \sum_{k=1}^{K} y_k \cdot g_j(\boldsymbol{z}) \right) \cdot x_i$$

[$g_j(\boldsymbol{z})$ **does not depend on index** k**. Therefore, we can pull it out of the sum**]

$$= \left(-y_j + g_j(\boldsymbol{z}) \cdot \sum_{k=1}^{K} y_k\right) \cdot x_i$$

[$\boldsymbol{y}$ **is a one-hot vector, therefore the sum of its components is equal to 1**]

$$= (-y_j + g_j(\boldsymbol{z})) \cdot x_i$$
$$= (g_j(\boldsymbol{z}) - y_j) \cdot x_i$$