Artificial Intelligence and Machine Learning

Derivation of the Gradient for Softmax Regression

We compute the partial derivative of the cross-entropy cost function:

$$\frac{\partial}{\partial \theta_{ij}} \ell^{\text{CE}}(\boldsymbol{h}_{\Theta}(\boldsymbol{x}), \boldsymbol{y}) = \frac{\partial}{\partial \theta_{ij}} \left(- \sum_{k=1}^{K} y_k \log(\zeta_k(\boldsymbol{z})) \right)$$

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

[Apply chain rule]

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

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

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

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

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

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

[Derivative of the softmax function (see exercise sheet)]

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

[Cancel terms]

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

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

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

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

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

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

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

$$= \boxed{(\zeta_j(\boldsymbol{z}) - y_j) \cdot x_i}$$

$\square$