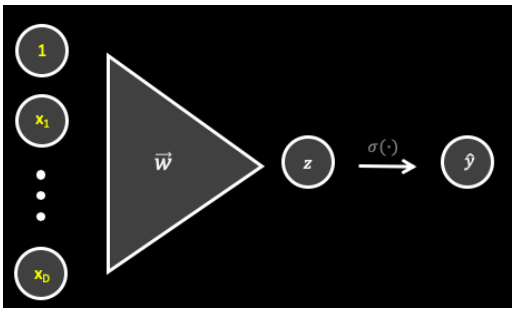
Assignment 3 Solutions

Eric Keilty

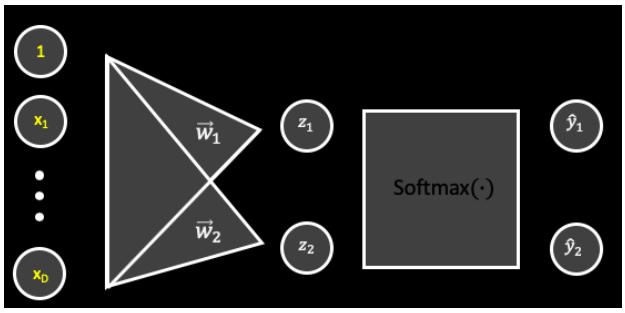
May 15, 2020

Part 1

Question



(a) Logistic Regression



(b) Multiclass Classification with $K = 2$

Can you find the relationship between $\boldsymbol{w}$ and $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$ such that each classifier makes the same prediction for any input $\boldsymbol{x} \in \mathbb{R}^D$

Answer

We want both classifiers to give the same predictions. A key insight is that if $\hat{y} = \hat{y}_2$, then both classifiers will make the same predictions. Why is this the case? The Logistic Classifier will predict a 1 if $\hat{y} > 0.5$ and a 0 otherwise. The Multiclass Classifier will predict 1 if $\hat{y}_2 > \hat{y_1}$ and a 0 otherwise. Moreover, we know that $\hat{y}_1 + \hat{y}_2 = 1$, due to the `Softmax` function. Therefore, combining both equations gives $\hat{y}_2 > 1 - \hat{y}_2 \implies \hat{y}_2 > 0.5$. Thus, if we can make $\hat{y} = \hat{y}_2$, then both classifiers will be equivalent.

$$\hat{y} = \texttt{Sigmoid}(z) = \sigma(z) = \sigma(\boldsymbol{w}^T \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}}}$$

$$\hat{y}_2 = \texttt{Softmax}(z_2) = \texttt{Softmax}(\boldsymbol{w}_2^T \boldsymbol{x}) = \frac{e^{-\boldsymbol{w}_2^T \boldsymbol{x}}}{e^{-\boldsymbol{w}_1^T \boldsymbol{x}} + e^{-\boldsymbol{w}_2^T \boldsymbol{x}}} = \frac{1}{e^{-\boldsymbol{w}_1^T \boldsymbol{x} + \boldsymbol{w}_2^T \boldsymbol{x}} + 1} = \frac{1}{1 + e^{-(\boldsymbol{w}_1 - \boldsymbol{w}_2)^T \boldsymbol{x}}}$$

Therefore, $\boxed{\boldsymbol{w} = \boldsymbol{w}_1 - \boldsymbol{w}_2}$

Part 2

Question

Determine $\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \boldsymbol{z}}$ for a multiclass regression

Answer

Derivative of Softmax

$$\hat{y}_i = \texttt{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

if $i \neq k$

$$\frac{\partial \hat{y}_i}{\partial \boldsymbol{z}_k} = \frac{0 \cdot \sum_{j=1}^{K} e^{z_j} - e^{z_i} \cdot e^{z_k}}{\left(\sum_{j=1}^{K} e^{z_j}\right)^2} = -\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \cdot \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} = -\hat{y}_i \cdot \hat{y}_k$$

if $i = k$

$$\frac{\partial \hat{y}_i}{\partial \boldsymbol{z}_k} = \frac{e^{z_i} \cdot \sum_{j=1}^{K} e^{z_j} - e^{z_i} \cdot e^{z_k}}{\left(\sum_{j=1}^{K} e^{z_j}\right)^2} = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} - \left(\frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}\right)^2 = \hat{y}_i - \hat{y}_i^2 = (1 - \hat{y}_i) \cdot \hat{y}_i$$

Full Derivative

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{CE}}}{\partial z_i} &= \frac{\partial}{\partial z_i}\left[-\sum_{k=1}^{K} y_k \log \hat{y}_k\right] = -\sum_{k=1}^{K} \frac{y_k}{\hat{y}_k} \cdot \frac{\partial \hat{y}_k}{z_i} \\
&= -\frac{y_i}{\hat{y}_i}(1 - \hat{y}_i)\hat{y}_i - \sum_{k \neq i} \frac{y_k}{\hat{y}_k} \cdot (-\hat{y}_i \hat{y}_k) \\
&= -y_i + y_i \cdot \hat{y}_i + \sum_{k \neq i} y_k \cdot \hat{y}_i \\
&= -y_i + \hat{y}_i \cdot \sum_{k=1}^{K} y_k \\
&= \hat{y}_i - y_i
\end{aligned}$$

Therefore $\boxed{\frac{\partial \mathcal{L}_{\text{CE}}}{\partial z_i} = \hat{y}_i - y_i}$, exactly the same as Sigmoid. Moreover, we can vectorize this result so that $\boxed{\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \boldsymbol{z}} = \hat{\boldsymbol{y}} - \boldsymbol{y}}$