Quiz02

1. Gradient of the Cross Entropy Loss

Consider the data $(\vec{x}_i, y_i)$ for $i = 1, \ldots, n$ where $\vec{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. Consider the parameter vector $\vec{w} \in \mathbb{R}^d$. For each $i \in \{1, \ldots, n\}$, define the *logistic function* $p_i \colon \mathbb{R}^d \mapsto \mathbb{R}$ given as

$$p_i(\vec{w}) = \frac{1}{1 + e^{-\vec{w}^\top \vec{x}_i}}. \tag{1}$$

(a) Find the gradient of the function $p_i(\vec{w})$.

Solution: The gradient is

$$\nabla p_i(\vec{w}) = \begin{bmatrix} \frac{\partial p_i}{\partial w_1}(\vec{w}) \\ \vdots \\ \frac{\partial p_i}{\partial w_d}(\vec{w}) \end{bmatrix}. \tag{2}$$

Here

$$\frac{\partial p_i}{\partial w_j}(\vec{w}) = \frac{(\vec{x}_i)_j e^{-\vec{w}^\top \vec{x}_i}}{\left(1 + e^{-\vec{w}^\top \vec{x}_i}\right)^2}. \tag{3}$$

Thus

$$\nabla p_i(\vec{w}) = \vec{x}_i \cdot \frac{e^{-\vec{w}^\top \vec{x}_i}}{\left(1 + e^{-\vec{w}^\top \vec{x}_i}\right)^2} \tag{4}$$

(b) For $i \in \{1, \ldots, n\}$, the *cross entropy* of $p \in [0, 1]$ against y_i is defined as

$$H_i(p) \doteq -y_i \log(p) - (1 - y_i) \log(1 - p). \tag{5}$$

Find the gradient of the function $\ell_i(\vec{w}) \doteq H_i(p_i(\vec{w}))$ with respect to $\vec{w}$.

Solution: The gradient is

$$\nabla_{\vec{w}} \ell_i(\vec{w}) = \begin{bmatrix} \frac{\partial \ell_i}{\partial w_1}(\vec{w}) \\ \vdots \\ \frac{\partial \ell_i}{\partial w_d}(\vec{w}) \end{bmatrix}. \tag{6}$$

We can use the chain rule to find each component:

$$\begin{aligned}
\frac{\partial \ell_i}{\partial w_j}(\vec{w}) &= -\left[\frac{\partial H_i}{\partial p}(p_i(\vec{w}))\right]\left[\frac{\partial p_i}{\partial w_j}(\vec{w})\right] && (7)\\
&= -\left[\frac{y_i}{p_i(\vec{w})} - \frac{1 - y_i}{1 - p_i(\vec{w})}\right]\left[\frac{(\vec{x}_i)_j e^{-\vec{w}^\top \vec{x}_i}}{\left(1 + e^{-\vec{w}^\top \vec{x}_i}\right)^2}\right] && (8)\\
&= -\left[\frac{y_i}{1/(1 + e^{-\vec{w}^\top \vec{x}_i})} - \frac{1 - y_i}{e^{-\vec{w}^\top \vec{x}_i}/(1 + e^{-\vec{w}^\top \vec{x}_i})}\right]\left[\frac{(\vec{x}_i)_j e^{-\vec{w}^\top \vec{x}_i}}{\left(1 + e^{-\vec{w}^\top \vec{x}_i}\right)^2}\right] && (9)\\
&= -\left[y_i(1 + e^{-\vec{w}^\top \vec{x}_i}) - \frac{(1 - y_i)(1 + e^{-\vec{w}^\top \vec{x}_i})}{e^{-\vec{w}^\top \vec{x}_i}}\right]\left[\frac{(\vec{x}_i)_j e^{-\vec{w}^\top \vec{x}_i}}{\left(1 + e^{-\vec{w}^\top \vec{x}_i}\right)^2}\right] && (10)\\
&= -(\vec{x}_i)_j \left[y_i \frac{e^{-\vec{w}^\top \vec{x}_i}}{1 + e^{-\vec{w}^\top \vec{x}_i}} - (1 - y_i)\frac{1}{1 + e^{-\vec{w}^\top \vec{x}_i}}\right] && (11)\\
&= -(\vec{x}_i)_j \left[y_i(1 - p_i(\vec{w})) - (1 - y_i)p_i(\vec{w})\right] && (12)\\
&= -(\vec{x}_i)_j \left[y_i - p_i(\vec{w})\right] && (13)\\
&= (\vec{x}_i)_j \left[p_i(\vec{w}) - y_i\right]. && (14)
\end{aligned}$$

Thus

$$\nabla_{\vec{w}} \ell_i(\vec{w}) = \vec{x}_i \left[p_i(\vec{w}) - y_i \right]. \tag{15}$$