Gradient Descent with Logistic Regression

In the last module, we defined the MLE solution for Logistic Regression as

$$\begin{aligned}\mathbf{w}_{MLE} &= \arg\max_{\mathbf{w}} -\sum_{i=1}^{n} \log\left(1+e^{-y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)}\right)\\ &= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \log\left(1+e^{-y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)}\right)\\ &= \arg\min_{\mathbf{w}} -\sum_{i=1}^{n} \log\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)\end{aligned}$$

where $\log \mathrm{P}(\mathbf{y}|\mathbf{X};\mathbf{w}) = -\sum_{i=1}^{n} \log\left(1+e^{-y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)}\right)$ is the log likelihood. Consequently, our goal is to find the MLE solution by *maximizing the log likelihood*. Equivalently, we can find the MLE solution by *minimizing the negative log likelihood*

$$NLL = \sum_{i=1}^{n} \log\left(1+e^{-y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)}\right) = -\sum_{i=1}^{n} \log\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)$$

Unlike convex functions you might have seen before, we cannot analytically calculate the global minima negative log likelihood even though we know that a global minima exists. Therefore, we will iteratively minimize the negative log likelihood using Gradient Descent.

Gradient of Negative Log Likelihood with respect to $\mathbf{w}$

Recall from the previous page that a gradient descent step of size α is $\mathbf{w} \leftarrow \mathbf{w} - \alpha g(\mathbf{w})$, where $g(\mathbf{w})$ is the gradient of the loss function we wish to minimize. Therefore, here we show how to compute the gradient, i.e. the first derivative, of NLL with respect to $\mathbf{w}$.

$$\begin{aligned}g(\mathbf{w}) = \frac{\partial NLL}{\partial \mathbf{w}} &= -\frac{\partial\left[\sum_{i=1}^{n} \log\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)\right]}{\partial \mathbf{w}}\\ &= -\sum_{i=1}^{n} \frac{\partial\left[\log\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)\right]}{\partial \mathbf{w}}\\ &= -\sum_{i=1}^{n} \frac{1}{\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)} \cdot \frac{\partial\left[\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)\right]}{\partial \mathbf{w}}\\ &= -\sum_{i=1}^{n} \frac{\sigma'\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)}{\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)} \cdot y_i \mathbf{x}_i\end{aligned}$$

Since $\sigma'(z) = \sigma(z)(1-\sigma(z))$ and $1-\sigma(z) = \sigma(-z)$, the gradient is:

$$
\begin{aligned}
\frac{\partial NLL}{\partial \mathbf{w}} &= -\sum_{i=1}^{n} \frac{\sigma'\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)}{\sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)} \cdot y_i \mathbf{x}_i \\
&= -\sum_{i=1}^{n} \left[1 - \sigma\left(y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right)\right] \cdot y_i \mathbf{x}_i \\
&= -\sum_{i=1}^{n} \sigma\left(-y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)\right) \cdot y_i \mathbf{x}_i
\end{aligned}
$$

In the final project, you will use this expression for the gradient – almost as is – to implement a spam email classifier with logistic regression!