Logistic Regression MLE

Assume $y_i \in \{+1, -1\}$ and we model $\mathrm{P}(y_i|\mathbf{x}_i;\mathbf{w})$ as:

$$\mathrm{P}(y_i|\mathbf{x}_i;\mathbf{w}) = \frac{1}{1+e^{-y_i(\mathbf{w}^\top \mathbf{x}_i)}}$$

(Note: To make life easier, we assume that we absorb the parameter b into $\mathbf{w}$ by adding a constant dimension to $\mathbf{x}$)

To derive the MLE solution for logistic regression, we need to first construct the log-likelihood, namely, $\mathrm{P}(\mathbf{y}|\mathbf{X};\mathbf{w})$, where $\mathbf{y} = [y_1, \ldots, y_n]^\top$ and $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ (stacking the features horizontally so that each column is a different feature and each row is a data point). Using our assumption that our data points are independent and identically distributed, we know that the likelihood is:

$$\mathrm{P}(\mathbf{y}|\mathbf{X};\mathbf{w}) = \prod_{i=1}^n \mathrm{P}(y_i|\mathbf{x}_i;\mathbf{w}) = \prod_{i=1}^n \frac{1}{1+e^{-y_i\left(\mathbf{w}^\top \mathbf{x}_i\right)}}$$

Consequently, the log-likelihood becomes:

$$\log \mathrm{P}(\mathbf{y}|\mathbf{X};\mathbf{w}) = \sum_{i=1}^n -\log\left(1 + e^{-y_i(\mathbf{w}^\top \mathbf{x}_i)}\right)$$

So, our MLE solution for logistic regression is:

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

Unlike linear regression, logistic regression has no closed form solution (i.e. there is not an analytical expression for the optimal weight vector). Fortunately, the objective function is convex, so we can make use of gradient descent to find the optimal solution. (We will cover gradient descent in the next module.)

☆ **Key Points**

As part of our modeling, we assume an expression for the probability of the observed labels given the data and parameter vector $\mathbf{w}$.

We apply MLE and minimize the log-likelihood of the data.

The result is an expression that can be optimized using gradient descent.