HW 1 - ML

TASK 1

HP 1

Likelihood function:

For logistic regression, the probability $p(y_n|x_n)$ can be given by the sigmoid function:

$$p(y_n = 1|x_n) = \frac{1}{1+e^{-w^T x_n}}$$

$$p(y_n = 0|x_n) = 1 - \frac{1}{1+e^{-w^T x_n}}$$

$$= \frac{e^{-w^T x_n}}{1+e^{-w^T x_n}}$$

To consolidate 2 equations,

$$p(y_n|x_n) = \frac{1}{1+e^{-y_n w^T x_n}}$$

Given the data, $L(w) = \prod_{n=1}^{N} p(y_n|x_n)$

$$L(w) = \prod_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}}$$

Log-likelihood function

Take natural log on both sides

$$\ln(L(w)) = \sum_{n=1}^{N} \ln\left(\frac{1}{1+e^{-x_n w^T y_n}}\right)$$

$$= \sum_{n=1}^{N} -\ln\left(1+e^{-y_n w^T x_n}\right)$$

To maximize the likelihood,

maximize log-likelihood $\propto \dfrac{1}{-(\text{log likelihood})}$

1.) Cross-Entropy Error

$E_{in}(w)$ for logistic regression is:

$$E_{in}(w) = -\frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{1+e^{-y_n w^T x_n}}\right)$$

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1+e^{-y_n w^T x_n}\right)$$

Now comparing the expressions, we see that minimizing the cross-Entropy error $E_{in}(w)$ is equivalent to maximizing the log-likelihood $\ln(L(w))$ as both involve minimizing / maximizing the same summation term.

→ Thus, we've shown that selecting the hypothesis h that maximizes the likelihood is equivalent to minimizing the cross-entropy error for logistic regression.

HP 2

To derive the gradient of the in-Sample error w.r.t weight vector w, we Start with the defination of the cross-entropy error for logistic regression

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln\left(1 + e^{-y_n w^T x_n}\right)$$

For the Gradient decent algorithm, we need to determine gradient of this error ~~with~~ w.r.t weight vector w.

The gradient will be a vector, and the component corresponding to the ith vector w_i can be found by differentiating E_{in} wrt w_i & then can generalize it for the entire weight vector.

For simplicity, let's consider the term inside the summation:

$$e_n(w) = \ln\left(1 + e^{-y_n w^T x_n}\right)$$

we need to compute $\frac{\partial e_n(w)}{\partial w_i}$

using chain rule & for logarithm,

$$\frac{\partial e_n(w)}{\partial w_i} = \frac{1}{1+e^{-y_n w^T x_n}} \times \frac{\partial}{\partial w_i}\left(1+e^{-y_n w^T x_n}\right)$$

$$= \frac{1}{1+e^{-y_n w^T x_n}} \times \left(-y_n x_{n_i} e^{-y_n w^T x_n}\right)$$

$$= \frac{-y_n x_{n_i} e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}}$$

using the result, gradient component for w_i is:

$$\frac{\partial E_{in}(w)}{\partial w_i} = \frac{1}{N} \sum_{n=1}^{N} \frac{-y_n x_{n_i} e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}}$$

By generalizing this for all components of the weight vector, we get the gradient:

$$\nabla E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \frac{-y_n x_n e^{-y_n w^T x_n}}{1 + e^{-y_n w^T x_n}}$$

where $x_n \to n^{th}$ input vector.

This gradient is used in the gradient descent algorithm to iteratively update the weight vector & minimize the in-sample error.