Logistic Regression — MCLE

Given \( n \) pairs \((x_i, y_i)\) where \( x_i \) is \( d \times 1 \) and \( y_i \) is 0 or 1.

We aim to maximize the conditional likelihood of all \( y_i \)'s, given \( x_i \)'s.

\[
\hat{w}_{MCLE} = \underset{w}{\arg \max} \prod_{i=1}^{n} P(y_i | x_i, w)
\]

\[
\log L(w) = \log \left[ \prod_{i=1}^{n} P(y_i | x_i, w) \right] = \sum_{i=1}^{n} \log \left[ P(y_i | x_i, w) \right] = \sum_{i=1}^{n} \left[ y_i \log \left( \frac{P(y_i = 1 | x_i, w)}{P(y_i = 0 | x_i, w)} \right) \right] + (1 - y_i) \log \left[ \frac{P(y_i = 0 | x_i, w)}{P(y_i = 1 | x_i, w)} \right]
\]

Exactly one of the two terms is always zero because \( y_i = 0 \) or 1.

\[
= \sum_{i=1}^{n} \left[ y_i \log \left( \frac{e^{\mathbf{w}^T x_i}}{1 + e^{\mathbf{w}^T x_i}} \right) \right] + \log \left[ \frac{1}{1 + e^{\mathbf{w}^T x_i}} \right]
\]
\[
\sum_{i=1}^{n} \left[ y_i \left( w^T x_i \right) - \log \left( 1 + e^{w^T x_i} \right) \right]
\]

\[
\nabla_w \left[ \log l(w) \right] = \left[ \frac{\partial \log l(w)}{\partial w_0} \right.
\left. \vdots \right]
\left. \frac{\partial \log l(w)}{\partial w_n} \right]
\]

\[
\frac{\partial \log l(w)}{\partial w_j} = \sum_{i=1}^{n} \left[ y_i x_i^j - \frac{1 \times e^{w^T x_i}}{1 + e^{w^T x_i}} \times x_i^j \right] \text{ where } x_i^j \text{ is the } j\text{th component of } x_i
\]

\[
= \sum_{i=1}^{n} \left[ x_i^j \left[ y_i - \frac{e^{w^T x_i}}{1 + e^{w^T x_i}} \right] \right]
\]

This gradient can be used to perform gradient descent on \( w \).
L₂-regularized logistic regression

In addition to maximized log-conditional-likelihood, we penalize the L₂ norm of \( \omega \)

\[
\log L(\omega) = \log \left[ \prod_{i=1}^{n} P(y_i | x_i, \omega) \right] - \lambda \| \omega \|_2^2
\]

The MLE steps follow exactly the same pattern as those for plain vanilla logistic regression, except the \(-\lambda \| \omega \|_2^2\) term. Eventually, component-wise gradient is the following:

\[
\frac{\partial \log L}{\partial \omega_j} = \frac{1}{2} \sum_{i=1}^{n} \left[ x_i^j \left[ y_i - \frac{e^{x_i^j \omega}}{1 + e^{x_i^j \omega}} \right] \right] - \lambda \omega_j
\]

This gradient can be used for gradient descent to estimate \( \omega \).