Logistic Regression — MCLE

Given n pairs $(\vec{x_i}, y_i)$ where $\vec{x_i}$ 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

$$\vec{w}_{MCLE} = \arg\max_{\vec{w}} \prod_{i=1}^{n} P(y_i \mid \vec{x_i}, \vec{w})$$

$$\log l(\vec{w}) = \log \left[\prod_{i=1}^{n} P(y_i \mid \vec{x_i}, \vec{w}) \right]$$

$$= \sum_{i=1}^{n} \log \left[P(y_i \mid \vec{x_i}, \vec{w}) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log \left[P(y_i = 1 \mid \vec{x_i}, \vec{w}) \right] + (1 - y_i) \log \left[P(y_i = 0 \mid \vec{x_i}, \vec{w}) \right] \right]$$

exactly one of the two terms is always zero because $y_i = 0$ or 1

$$= \sum_{i=1}^{n} \left[y_i \left[\log \left[\frac{P(y_i = 1 \mid \vec{x_i}, \vec{w})}{P(y_i = 0 \mid \vec{x_i}, \vec{w})} \right] \right] + \log \left[P(y_i = 0 \mid \vec{x_i}, \vec{w}) \right] \right]$$

$$= \sum_{i=1}^{n} \left[y_i \left[\log \left[\frac{(e^{\vec{w}^T \vec{x_i}}) / (1 + e^{\vec{w}^T \vec{x_i}})}{(1) / (1 + e^{\vec{w}^T \vec{x_i}})} \right] \right] + \log \left[\frac{1}{1 + e^{w^T x}} \right] \right]$$

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

$$\nabla_{\vec{w}} \left[\log l \right] = \begin{bmatrix} \frac{\partial \log l(\vec{w})}{\partial w_0} \\ \vdots \\ \frac{\partial \log l(\vec{w})}{\partial w_n} \end{bmatrix}$$

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

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

This gradient can be used to perform gradient descent on $\vec{w}$.

L_2-regularized logistic regression

In addition to maximized log-conditional-likelihood, we penalize the l_2 norm of $\vec{w}$

$$\therefore \log l(\vec{w}) = \log\left[\prod_{i=1}^{n} P(y_i | \vec{x}_i, \vec{w})\right] - \lambda \|\vec{w}\|_2^2$$

The MCLE steps follow exactly the same pattern as those for plain vanilla logistic regression, except the $-\lambda\|\vec{w}\|_2^2$ term. Eventually, component-wise gradient is the following:

$$\frac{\partial \log l}{\partial w_j} = \sum_{i=1}^{n}\left[x_i^j\left[y_i - \frac{e^{\vec{w}^T\vec{x}_i}}{1+e^{\vec{w}^T\vec{x}_i}}\right]\right] - 2\lambda w_j$$

This gradient can be used for gradient descent to estimate $\vec{w}$.