Notes on Logistic Regression

Liangjie Hong

February 19, 2010

1 Basic Model

We use X_i ($|X| = N$) to represent i-th feature of a data instance. Y_l ($|Y| = M$) is the label of the l-th data instance in the corpus. We want to model the probability that given a feature vector $\overline{X_l}$ how likely the label is $Y_l = k$, namely $P(Y_l = k|\overline{X_l})$. In this section, we only consider the two-class case.

$$P(Y = 1|X) = \frac{\exp\left(\sum_{k=0}^{N} w_k X_k\right)}{1 + \exp\left(\sum_{k=0}^{N} w_k X_k\right)}$$
$$P(Y = 0|X) = \frac{1}{1 + \exp\left(\sum_{k=0}^{N} w_k X_k\right)}$$

We want to model the likelihood of the total data as follows:

$$\begin{aligned}
L(W) &= \sum_{l=1}^{M} \left[Y^l \log P(Y^l = 1|X^l, W) + (1 - Y^l) \log P(Y^l = 0|X^l, W)\right] \\
&= \sum_{l=1}^{M} \left[Y^l \log \frac{P(Y^l = 1|X^l, W)}{P(Y^l = 0|X^l, W)} + \log P(Y^l = 0|X^l, W)\right] \\
&= \sum_{l=1}^{M} \left[Y^l \log \exp(\sum_{k=0}^{N} w_k X_k^l) + \log \frac{1}{1 + \exp\left(\sum_{k=0}^{N} w_k X_k^l\right)}\right] \\
&= \sum_{l=1}^{M} \left[Y^l \left(\sum_{k=0}^{N} w_k X_k^l\right) - \log\left(1 + \exp\left(\sum_{k=0}^{N} w_k X_k^l\right)\right)\right]
\end{aligned}$$

There is no closed-form of the maximum solution of this total likelihood. We use iterative algorithms to obtain the global maximum (e.g., Gradient Descent):

$$\begin{aligned}
\frac{\partial L(W)}{\partial w_k} &= \sum_{l=1}^{M} \left[Y^l X_k^l - \frac{X_k^l \exp\left(\sum_{k=0}^{N} w_k X_k^l\right)}{1 + \exp\left(\sum_{k=0}^{N} w_k X_k^l\right)}\right] \\
&= \sum_{l=1}^{M} X_k^l \left[Y_l - \frac{\exp\left(\sum_{k=0}^{N} w_k X_k^l\right)}{1 + \exp\left(\sum_{k=0}^{N} w_k X_k^l\right)}\right]
\end{aligned}$$

2 L2 Regularization

In order to overcome the problem of over-fitting, we can incorporate a L2 regularizer into the objective function (the likelihood of the data):

$$W \leftarrow \arg\max_{W} \sum_{l}^{M} \log P(Y^l|X^l, W) - \frac{\lambda}{2}||W||^2$$

Thus, the derivative with one additional penalty term is :

$$\frac{\partial L(W)}{\partial w_k} = \sum_{l=1}^{M} X_k^l \left[Y_l - \frac{\exp\left(\sum_{k=0}^{N} w_k X_k^l\right)}{1 + \exp\left(\sum_{k=0}^{N} w_k X_k^l\right)} \right] - \lambda w_k$$

3 L1 Regularization

We also can incorporate L1 regularizer into the objective function:

$$W \leftarrow \arg\max_{W} \sum_{l}^{M} \log P(Y^l|X^l, W) - \lambda||W||_1$$

Although this objective function is convex, it is not differentiable. Though simple iterative algorithms cannot be applied, see references for more details.