On L2-norm Regularization and the Gaussian Prior

Jason Rennie
jrennie@ai.mit.edu

May 8, 2003

Abstract

We show how the regularization used for classification can be seen from the MDL viewpoint as a Gaussian prior on weights. We consider the problem of transmitting classification labels; we select as our model class logistic regression with perfect precision where we specify a weight for each feature. This is unrealistic since the encoding length of any such model is infinite, but if we use a Gaussian prior on weights and ignore constant factors, we find that the encoding length objective exactly matches that of logistic regression with an L2-norm regularization penalty. Through this understanding, we see that the tradeoff parameter is the variance of the Gaussian prior. It also delineates steps for improved regularization—both decreased resolution and feature selection could be used to decrease the encoding length.

1 The Problem

Let $(x_1, \ldots, x_n)$ be a set of examples. Let $(y_1, \ldots, y_n)$, $y_i \in \{+1, -1\}$, be a set of binary lables for the examples. The problem we address is that of encoding the labels as efficiently as possible. The labels have little internal structure of their own; we use the information in the examples to help predict the labels. Compression is a natural way to judge the degree to which a system has learned. In this case, compression judges the effectiveness of using the examples to predict the labels. Note that to make use of the examples, we must encode the mechanism for extracting information, so the framework imposes a sort of natural regularization.

2 Encoding

To encode the labels, we estimate a conditional distribution using a linear classifier. In one part, we encode the weights for the linear classifier and in the second part, we encode the labels, to the extent that they have not been specified by the classifier. Define

$$p(y_i = +1|x_i; \vec{w}) = g\left(\sum_{k=0}^{l} x_{ik} w_k\right) \tag{1}$$

to be the conditional probability of label y_i being positive given example x_i. $g(z) = \frac{1}{1+e^{-z}}$ is the logistic function. x_{ik} is the value of the k^{th} feature of example i. w_k is the weight for feature k. $k = 0$ is the special "bias" feature; $x_{i0} = 1$ for all i. l is the number of non-bias features. Let $z_i = \sum_k x_{ik} w_k$. Then, if we ignore the discrete practicality of coding, the encoding length for label y_i is

$$L(y_i|x_i; \vec{w}) = -\log g(y_i z_i). \tag{2}$$

All that remains to be encoded is the weights.

To encode the weights, we assume a Gaussian prior with mean zero and variance σ^2,

$$p(w_k) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w_k^2}{2\sigma^2}\right). \tag{3}$$

But this is a density, not a probability mass function as we require. However, we are not concerned with absolute encoding lengths—relative encoding lengths are sufficient since we are comparing between models in a restricted class. Using this prior and treating it as a probability mass function, we get a (reltaive) encoding length of

$$L(w_k) = -\log p(w_k). \tag{4}$$

Now, we can write down the total encoding length. The total encoding length sums the encoding length for all labels and all weights. We do not encode the "bias" weight. The total encoding length is

$$L_{\text{tot}} = -\sum_i \log g(y_i z_i) + \sum_{k=1}^{l} \left(\frac{1}{2}\log 2\pi\sigma^2 + \frac{w_k^2}{2\sigma^2}\right). \tag{5}$$

3 Regularized Logistic Regression

Logistic regression maximizes the (log-)likelihood of the labels, where the likelihood of a label is as defined in equation 1. For more information on logistic regression, see [1]. We subtract a constant multiple of the the square of the L2-norm to regularize the weights. This gives us an objective,

$$J_{lr} = \sum_i \log g(y_i z_i) - \frac{C}{2} \sum_{k=1}^{l} w_k^2, \tag{6}$$

that we would like to maximize.

To minimize the total encoding length as defined above, we can ignore the $\frac{1}{2} \log 2\pi\sigma^2$ constant term. Hence, the encoding length objective that we wish to minimize is

$$L'_{\text{tot}} = -\sum_i \log g(y_i z_i) + \frac{1}{2\sigma^2} \sum_{k=1}^{l} w_k^2. \tag{7}$$

Reversing the sign and substituting $C = \frac{1}{\sigma^2}$ gives us the regularized logistic regression objective.

4 Conclusion

It is clear from section 3 that the L2-norm regularizer used for logistic regression (and other learning algorithms) is not arbitrary, but rather a direct result of imposing a Gaussian prior on weights. We can also see that there is much room for improvement. The L2-norm assumes unlimited precision and does not encourage feature selection. We could improve the regularization by restricting weight values to a discrete set and allowing the classifier to select out features that are not useful. These two steps would bring us closer to an encoding that is more efficient than the trivial one where each label is encoded with a single bit. These are areas for future work.

References

[1] Jason Rennie. Logistic regression. http://www.ai.mit.edu/∼jrennie/writing/lr.pdf, April 2003.