Regularized Least Squares Classification with a Gaussian Regularization Prior

Jason Rennie
jrennie@csail.mit.edu

November 23, 2003

Abstract

Rifkin describes how Regularized Least Squares Classifiers can be learned efficiently via the Conjugate Gradients algorithm [1]. We extend his framework to incorporate the notion of the regularization term being a prior on the weight vector. This allows us to incorporate any prior information we have about weights for individual features. Transfer is one scenario where this may be of use.

Rifkin [1] states the Regularized Least Squares Classification problem as

$$\min_{f \in \mathcal{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_K^2. \tag{1}$$

Using the representer theorem, he shows that this is equivalent to

$$\min_{c \in \mathbb{R}^n} \frac{1}{2n} (y - Kc)^T (y - Kc) + \lambda c^T K c, \tag{2}$$

where $K_{ij} = K(x_i, x_j)$ and K is the kernel function.

We consider classification as an encoding problem. We encode the model and the data given the model. The least squares objective,

$$(y - Kc)^T (y - Kc) \tag{3}$$

is the negative log-likelihood of y for an un-normalized Normal distribution with mean Kc and identity covariance matrix, $N(Kc, I)$. The regularization term,

$$\lambda c^T K c \tag{4}$$

is the negative log-likelihood of $K^{1/2}c$ for a Normal distribution with zero mean and identity covariance matrix, $N(0, I)$. Note that $K^{1/2}c$ parameterizes a linear decision boundary in the kernel space.

Our modification is to relax the constraint that the covariance matrix of the regularization term be a scaled identity matrix. Now we consider a Normal with

zero mean and full covariance matrix, $N(0, \Sigma)$. This allows us to incorporate prior information about the decision boundary. Let $X^T = K^{1/2}$. The new least squares objective is

$$\min_{c \in \mathbb{R}^n} \frac{1}{2n}(y - XX^T c)^T (y - XX^T c) + \frac{1}{2} c^T X \Sigma^{-1} X^T c. \tag{5}$$

For our problems, which are of high dimension, we do not forsee utilizing a full covariance matrix, but rather a diagonal covariance matrix, where we only specify the variance for each feature.

To solve this problem, we set the gradient to zero. The gradient is

$$\nabla F_c = -\frac{1}{n} XX^T y + \frac{1}{n} XX^T XX^T c + X\Sigma^{-1} X^T c. \tag{6}$$

So, we get the optimal c by solving

$$XX^T XX^T c + nX\Sigma^{-1} X^T c = XX^T y. \tag{7}$$

Unfortunately, we cannot usefully further reduce the expression as was possible when the covariance matrix was the identity.

One scenario where this framework is useful is transfer. The problem of transfer is to learn to classify examples given two sources of information. The first source is the traditional source of supervision: labeled examples drawn from the distribution in question. The second source is labeled examples from somewhat similar problems. Call them reference tasks. For example, if the main task is topic classification, then the reference task examples would come from other topic classification problems. While the words used to identify topics in the main task may be very different from the words used to identify topics in the reference tasks, there will be similarities in how the language is used to identify important words.

References

[1] Ryan Rifkin. *Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning*. PhD thesis, Massachusetts Institute of Technology, 2002.