Feature Selection for (Nonlinear) Regularized Least-Squares Classification

Gene Yeo & Alex Rakhlin

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: We are interested in adapting a method of feature selection used in Support Vector Machine Classification (SVM) for Regularized Least-Squares Classification (RLSC), and their performances on toy problems and large-scale problems. For large scale non-linear problems, several approximation schemes have been used to avoid directly storing the entire kernel matrix(for nonlinear RLSC) [2]. Feature selection using these different approximation schemes will be studied.

Motivation: Tikhonov regularization with a quadratic loss function and equality constraints leads to the Regularized Least-Squares Classification (RLSC) algorithm. RLSC has been shown to perform favorably compared to the Support Vector Machine (SVM) formulation, which solves an inequality constrained convex optimization problem. To avoid storing the entire kernel matrix, approximation schemes, such as bagging smaller classifiers, and low rank approximations to the kernel matrix have been implemented for non-linear RLSC. Here we are interested in developing feature selection techniques for RLSC, and in their sensitivity to the different approximation schemes.

Previous Work: Feature selection techniques have been introduced for SVMs [1]. We extend this approach to RLSC.

Approach: Before we described the feature selection problem, we state briefly the RLSC problem:

\[
\min_{f \in H} \sum_{i=1}^{\ell} (y_i - f(x_i))^2 + \lambda \|f\|^2_K. \tag{26}
\]

where \(y_i\) are the binary labels of examples, \(x_i, f(.)\) is the decision function, \(\lambda\) is the regularization parameter and \(\|f\|^2_K\) is the norm of the function in the Reproducing Kernel Hilbert Space. The solution \(f^*\) to the above regularization problem has the form

\[
f^*(x) = \sum_{i=1}^{\ell} c_i K(x, x_i), \tag{27}
\]

so our function finding problem is reduced to the \(\ell\)-dimensional problem of finding the \(c_i\), where the optimal \(c\) can be found by solving

\[
(K + \lambda I)c = y, \tag{28}
\]

where \(I\) denotes an appropriately-sized identity matrix.

The feature selection problem is: given a set of functions \(y = f(x, c)\) we want to find scaling factors \(\sigma \in \mathbb{R}^n\) such that \(x \mapsto (x \ast \sigma)\), and the parameters \(c\) of the function, \(f\) that give the minimum value of

\[
\int V(y, f((x \ast \sigma), c))dP(x, y) \tag{29}
\]

where \(P(x, y)\) is not known, \(x^* = (x_1 \sigma_1, ..., x_n \sigma_n)\), \(V(., .)\) is a loss functional. The smallest scaling factors will determine the features that can be removed. Assuming that the maximal margin is size \(M\) and the images of the training samples are within a sphere of radius \(R\), the following theorem is true.
Theorem 1 If images of training data of size $l$ belonging to a sphere of size $R$ are separable with the corresponding margin $M$, then the expectation of the error probability has the bound

$$\mathbb{E}[\text{err}] \leq \frac{1}{l} E\left\{ \frac{R^2}{M^2} \right\} = \frac{1}{l} E\{R^2W^2(e^0)\},$$

(30)

where expectation is taken over sets of training data of size $l$. Therefore, we use Equation 30 and minimize Equation 31 over $c$. This can be performed by gradient descent on Equation 31 with respect to $\sigma$.

$$R^2W^2(\sigma) = R^2(\sigma)W^2(e^0, \sigma)$$

(31)

where the maximal margin corresponds to

$$W^2(e^0, \sigma) = e^{0T}K_\sigma e^0$$

(32)

and the optimal $e^0$ is computed from Equation 28 with $K = K_\sigma$ (Equation 33).

To compute the radius $R$ for kernel $K_\sigma$, defined below as:

$$K_\sigma(x, z) = K((x * \sigma), (z * \sigma))$$

(33)

we have to compute Equation 34.

$$R^2(\sigma) = \max_\beta \sum_{i=1}^\ell \beta_i K_\sigma((x_i), (x_i)) - \sum_{i,j=1}^\ell \beta_i \beta_j K_\sigma((x_i), (x_j))$$

(34)

subject to $\sum_i \beta_i = 1, \beta_i \geq 0, i = 1, ..., \ell$.

For feature selection in nonlinear large scale problems, three low-rank approximations of the kernel matrix can be utilized:

1. Subset method: $(K_{mm} + I\lambda\ell)e_m = y$.
2. Nystrom method: $(K_{lm}K_{mm}^{-1}K_{ml} + I\lambda\ell)e = y$.
3. Rectangle method: $(K_{ml}K_{lm} + K_{mm}\lambda\ell)e_m = K_{ml}y$.

**Future Work:** We will first try feature selection on toy data before comparing the performance of RLSC-feature selection to SVM-feature selection on large scale datasets.

**Research Support:** Research at CBCL is sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032 Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

**References:**
