SVMs Part 4: Other Related Issues

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Multi-class classification: one vs all (OVA)

Several other issues related to SVMs are discussed in HTF and Izenman. One is the matter of multi-class problems. There are both heuristic and optimality-based methods related to SVMs in the literature for cases where $\mathcal{G} = \{1, 2, \ldots, K\}$.

A heuristic one-against-all strategy is the following. Invent 2-class problems (K of them), the kth based on

$$y_{ki} = \begin{cases} 1 & \text{if } y_i = k \\ -1 & \text{otherwise} \end{cases}$$

Then for (a single) C^* and $k = 1, 2, \ldots, K$ solve the (possibly linearly non-separable) 2-class optimization problems to produce functions $\hat{g}_k(\mathbf{x})$ (underlying one-versus-all classifiers $\hat{f}_k(\mathbf{x}) = \text{sign}(\hat{g}_k(\mathbf{x}))$). A possible overall classifier is then

$$\hat{f}(\mathbf{x}) = \arg\max_{k \in \mathcal{G}} \hat{g}_k(\mathbf{x})$$

Multi-class classification: one vs one (OVO)

A second heuristic strategy is to develop a voting scheme based on pair-wise comparisons. That is, one might invent $\binom{K}{2}$ problems of classifying class l versus class m for $l < m$, choose a single C^* and solve the (possibly linearly non-separable) 2-class optimization problems to produce functions $\hat{g}_{lm}(\mathbf{x})$ and corresponding classifiers $\hat{f}_{lm}(\mathbf{x}) = \text{sign}(\hat{g}_{lm}(\mathbf{x}))$. For $m > l$ define $\hat{f}_{ml}(\mathbf{x}) = -\hat{f}_{lm}(\mathbf{x})$ and define an overall classifier by

$$\hat{f}(\mathbf{x}) = \arg\max_{k \in \mathcal{G}} \left(\sum_{m \neq k} \hat{f}_{km}(\mathbf{x}) \right)$$

or, equivalently

$$\hat{f}(\mathbf{x}) = \arg\max_{k \in \mathcal{G}} \left(\sum_{m \neq k} I\left[\hat{f}_{km}(\mathbf{x}) = 1 \right] \right)$$

Direct consideration of K classes

In addition to these fairly ad hoc methods of trying to extend 2-class SVM technology to K-class problems, there are developments that directly address the problem (from an overall optimization point of view). Pages 391-397 of Izenman provide a nice summary of a 2004 paper of Lee, Lin, and Wabha in this direction.

Other related prediction problems

Another type of question related to the support vector material is the extent to which similar methods might be relevant in regression-type prediction problems. As a matter of fact, there are loss functions alternative to squared error or absolute error that lead naturally to the use of the kind of technology needed to produce the SVM classifiers. That is, one might consider so called "ϵ insensitive" losses for prediction like

$$L_1^{\epsilon}(y,\widehat{y}) = \max(0, |y-\widehat{y}| - \epsilon)$$

or

$$L_2^{\epsilon}(y,\widehat{y}) = \max\left(0, (y-\widehat{y})^2 - \epsilon\right)$$

and be led to the kind of optimization methods employed in the SVM classification context. See Izenman pages 398-401 in this regard.