SVMs Part 3: Function Space Penalized Fitting

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Function space optimality argument

A more satisfying argument for use of "the kernel trick" might be based on appeal to optimality/regularization considerations. The following comes from a 2002 *Machine Learning* paper of Lin, Wahba, Zhang, and Lee.

Consider the abstract function space $\mathcal{A}$ associated with the non-negative definite kernel $\mathcal{K}$ and the optimization problem involving the "hinge loss"

$$\begin{array}{l}\text{minimize} \\ h \in \mathcal{A} \\ \text{and } \beta_0 \in \Re\end{array} \quad \sum_{i=1}^{N} \left(1 - y_i \left(\beta_0 + h\left(\mathbf{x}_i\right)\right)\right)_+ + \lambda \frac{1}{2} \|h\|_{\mathcal{A}}^2 \qquad (1)$$

The hinge loss is convenient here in making optimality arguments and there are several reasons that it is quite natural.

Expected hinge loss

Concerning the use of hinge loss:

1. Recall the development of Section 1.5.3 for the hinge loss $h_3(u) = (1-u)_+$ and the fact that for $(\mathbf{x}, y) \sim P$ a function g minimizing the expected hinge loss $\mathrm{E}(1 - yg(\mathbf{x}))_+$ is

$$g^{**}(\mathbf{x}) = \text{sign}\left(P[y = 1|\mathbf{x}] - \frac{1}{2}\right)$$

(the minimum risk classifier under 0-1 loss). So dividing the whole criterion (1) (hinge loss plus constant times squared $\mathcal{A}$ norm of h) by N, an empirical version of expected hinge loss is involved, and one can hope that an element h of $\mathcal{A}$ and value β_0 will be identified so that voting function $\beta_0 + h(\mathbf{x})$ is close to the optimal 0-1 loss classifier.

Expected hinge loss cont.

2. Further (again recalling the development in Section 1.5.3) since for all u, $I[u \leq 0] \leq (1-u)_+$ the criterion to be optimized involves an empirical version of a bound for the 0-1 loss error rate for the voting function $\beta_0 + h(\mathbf{x})$. So, again, one can hope that the voting function of a classifier with a good 0-1 loss error rate will be identified in the minimization.

Connection to SV classifier margin violations

3. Recalling the form of the SV classification optimization problem, the quantity $(1 - y_i (\mathbf{x}_i'\boldsymbol{\beta} + \beta_0))_+$ is the fraction of the margin (M) by which input $\mathbf{x}_i$ violates its cushion around the classification boundary hyperplane. (Points on the "right" side of their cushion don't get penalized at all. Ones with $(1 - y_i (\mathbf{x}_i'\boldsymbol{\beta} + \beta_0))_+ = 1$ are on the classification boundary. Ones with $(1 - y_i (\mathbf{x}_i'\boldsymbol{\beta} + \beta_0))_+ > 1$ are points mis-classified by the voting function.) The average of such terms is an average fractional (of the margin) violation of the cushion and the optimization seeks to control this. So the loss really is related to the SV classification ideas.

Simplification via the "representer theorem"

The "Representer Theorem" implies that an optimizing $h \in \mathcal{A}$ must be of the form

$$h_{\boldsymbol{\beta}}(\mathbf{x}) = \sum_{j=1}^{N} \beta_j \mathcal{K}(\mathbf{x}, \mathbf{x}_j) = \boldsymbol{\beta}' \mathbf{k}(\mathbf{x})$$

So the minimization problem is

$$\underset{\substack{\boldsymbol{\beta} \in \Re^N \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \sum_{i=1}^{N} \left(1 - y_i \left(\beta_0 + \boldsymbol{\beta}' \mathbf{k}(\mathbf{x}_i)\right)\right)_+ + \lambda \frac{1}{2} \left\| \sum_{j=1}^{N} \beta_j \mathcal{K}(\cdot, \mathbf{x}_j) \right\|_{\mathcal{A}}^2$$

that is, (for $\mathbf{K} = (\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j))$)

$$\underset{\substack{\boldsymbol{\beta} \in \Re^N \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \sum_{i=1}^{N} \left(1 - y_i \left(\beta_0 + \boldsymbol{\beta}' \mathbf{k}(\mathbf{x}_i)\right)\right)_+ + \lambda \frac{1}{2} \boldsymbol{\beta}' \mathbf{K} \boldsymbol{\beta}$$

Quadratic optimization problem

Now this is equivalent to the optimization problem

$$\underset{\substack{\boldsymbol{\beta} \in \Re^N \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \sum_{i=1}^{N} \xi_i + \lambda \frac{1}{2} \boldsymbol{\beta}' \mathbf{K} \boldsymbol{\beta} \text{ subject to } \begin{cases} y_i \left(\boldsymbol{\beta}' \mathbf{k}\left(\mathbf{x}_i\right) + \beta_0 \right) + \xi_i \geq 1 \ \forall i \\ \text{for some } \xi_i \geq 0 \end{cases}$$

which for $\underset{N \times N}{\mathbf{H}} = \left(y_i y_j \mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) \right)$ has dual problem of the form

$$\text{maximize } \mathbf{1}' \boldsymbol{\eta} - \frac{1}{2\lambda} \boldsymbol{\eta}' \mathbf{H} \boldsymbol{\eta} \text{ subject to } \mathbf{0} \leq \boldsymbol{\eta} \leq \mathbf{1} \text{ and } \boldsymbol{\eta}' \mathbf{y} = 0 \qquad (2)$$

or

$$\text{maximize } \mathbf{1}' \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}' \left(\frac{1}{\lambda^2} \mathbf{H} \right) \boldsymbol{\alpha} \text{ subject to } \mathbf{0} \leq \boldsymbol{\alpha} \leq \lambda \mathbf{1} \text{ and } \boldsymbol{\alpha}' \mathbf{y} = 0$$

Connection to the heuristic argument

That is, function space optimization problem (1) has a dual that is (*for the choice of* $C^* = \lambda$ *and kernel* $\frac{1}{\lambda^2}\mathcal{K}(\mathbf{x}, \mathbf{z})$) the same as that produced by the heuristic argument. Then, if $\boldsymbol{\eta}^{\text{opt}}$ is a solution to (2), Lin et al. say that an optimal $\boldsymbol{\beta} \in \Re^N$ is

$$\frac{1}{\lambda}\mathbf{diag}(y_1, \ldots, y_N)\,\boldsymbol{\eta}^{\text{opt}}$$

(producing coefficients to be applied to the functions $\mathcal{K}(\cdot, \mathbf{x}_i)$). On the other hand, the heuristic argument prescribes that for $\boldsymbol{\alpha}^{\text{opt}}$ the solution to its dual problem, coefficients in the vector

$$\mathbf{diag}(y_1, \ldots, y_N)\,\boldsymbol{\alpha}^{\text{opt}}$$

get applied to the functions $\frac{1}{\lambda^2}\mathcal{K}(\cdot, \mathbf{x}_i)$. Upon recognizing that $\boldsymbol{\eta}^{\text{opt}} = \frac{1}{\lambda}\boldsymbol{\alpha}^{\text{opt}}$ it is evident that for $C^* = \lambda$ and kernel $\frac{1}{\lambda^2}\mathcal{K}(\cdot, \cdot)$, the heuristic argument produces a solution to the optimization problem (1).