Prediction Theory Beginning from a Kernel

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

A more general theory

Some more general prediction theory begins with C a compact subset of $\Re^p$ and a symmetric kernel function

$$\mathcal{K} : C \times C \to \Re$$

Ultimately, we will consider as predictors for $\mathbf{x} \in C$ related to linear combinations of sections of the kernel function, $\sum_{i=1}^{N} b_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i)$ (where the $\mathbf{x}_i$ are the input vectors in the training set). To get there in a semi-rational way, and to incorporate use of a complexity penalty into the fitting, one restricts attention to kernels that have nice properties. In particular, we suppose that $\mathcal{K}$ is continuous.

A rough outline

These slides present a version of this material that is largely correct, but logically incomplete and not indicative of how a careful exposition must go. Refer to the typed course notes for a more careful and complete story that involves "Mercer's Theorem" and (L_2) "eigenfunctions" of the kernel.

Consider a linear space of functions $\mathcal{A}$ (a subset of the square integrable functions on C) consisting (roughly) of those of the form

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} b_i \mathcal{K}(\mathbf{x}, \mathbf{z}_i)$$

for countable subsets $\{\mathbf{z}_i\} \subset C$ (assuming proper convergence of this form). Define an inner product on $\mathcal{A}$ (for $f(\mathbf{x}) = \sum_{i=1}^{\infty} b_i \mathcal{K}(\mathbf{x}, \mathbf{z}_i)$ and $g(\mathbf{x}) = \sum_{i=1}^{\infty} c_i \mathcal{K}(\mathbf{x}, \mathbf{z}_i)$) by

$$\langle f, g \rangle_{\mathcal{A}} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} b_i c_j \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$$

Representer of evaluation and reproducing kernel

From the form of the inner product, with $f(\mathbf{x}) = \sum_{i=1}^{\infty} b_i \mathcal{K}(\mathbf{x}, \mathbf{z}_i)$ and a $\mathbf{z} \in C$ (it's no loss of generality to assume that $\mathbf{z}$ is some $\mathbf{z}_i$ defining f)

$$\langle f, \mathcal{K}(\cdot, \mathbf{z}) \rangle_{\mathcal{A}} = \sum_{i=1}^{\infty} b_i \sum_{l=1}^{\infty} I[\mathbf{z}_l = \mathbf{z}] \mathcal{K}(\mathbf{z}_i, \mathbf{z}) = \sum_{i=1}^{\infty} b_i \mathcal{K}(\mathbf{z}_i, \mathbf{z}) = f(\mathbf{z})$$

and $\mathcal{K}(\cdot, \mathbf{z})$ is the representer of function evaluation in $\mathcal{A}$.

The fact that then

$$\langle \mathcal{K}(\cdot, \mathbf{x}), \mathcal{K}(\cdot, \mathbf{z}) \rangle_{\mathcal{A}} = \mathcal{K}(\mathbf{x}, \mathbf{z})$$

is the reproducing kernel property.

A function optimization problem

For applying this material to the fitting of training data, for $\lambda > 0$ and a loss function $L(y, \widehat{y}) \geq 0$ define an optimization criterion

$$\underset{f \in \mathcal{A}}{\text{minimize}} \left(\sum_{i=1}^{N} L\left(y_i, f\left(\mathbf{x}_i\right)\right) + \lambda \left\| f \right\|_{\mathcal{A}}^2 \right) \quad (1)$$

As it turns out, an optimizer of this criterion must (for the training vectors $\{\mathbf{x}_i\}$) be of the form

$$\hat{f}\left(\mathbf{x}\right) = \sum_{i=1}^{N} b_i \mathcal{K}\left(\mathbf{x}, \mathbf{x}_i\right) \quad (2)$$

and the corresponding $\left\|\hat{f}\right\|_{\mathcal{A}}^2$ is then

$$\left\langle \hat{f}, \hat{f} \right\rangle_{\mathcal{A}} = \sum_{i=1}^{N} \sum_{j=1}^{N} b_i b_j \mathcal{K}\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

Rewriting the optimization criterion

The criterion (1) is thus

$$\underset{\mathbf{b}\in\Re^N}{\text{minimize}} \left(\sum_{i=1}^{N} L\left(y_i, \sum_{j=1}^{N} b_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \right) + \lambda \mathbf{b}' \left(\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \right) \mathbf{b} \right) \qquad (3)$$

Letting $\mathbf{K} = (\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j))$ and defining

$$L_N(\mathbf{Y}, \mathbf{Kb}) \equiv \sum_{i=1}^{N} L\left(y_i, \sum_{j=1}^{N} b_j \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) \right)$$

the optimization criterion (3) is thus

$$\underset{\mathbf{b}\in\Re^N}{\text{minimize}} \left(L_N(\mathbf{Y}, \mathbf{Kb}) + \lambda \mathbf{b}' \mathbf{Kb} \right)$$

Rewriting the optimization criterion cont.

Letting $\mathbf{P} = \mathbf{K}^{-}$ (a symmetric generalized inverse of $\mathbf{K}$) the criterion is

$$\underset{\mathbf{b} \in \Re^{N}}{\text{minimize}} \left(L_N \left(\mathbf{Y}, \mathbf{Kb} \right) + \lambda \mathbf{b}' \mathbf{K}' \mathbf{P} \mathbf{K} \mathbf{b} \right)$$

i.e.

$$\underset{\mathbf{v} \in C(\mathbf{K})}{\text{minimize}} \left(L_N \left(\mathbf{Y}, \mathbf{v} \right) + \lambda \mathbf{v}' \mathbf{P} \mathbf{v} \right) \tag{4}$$

That is, the function space optimization problem (1) reduces to the N-dimensional optimization problem (4). A $\mathbf{v}_\lambda \in C(\mathbf{K})$ (the column space of $\mathbf{K}$) minimizing $L_N(\mathbf{Y}, \mathbf{v}) + \lambda \mathbf{v}'\mathbf{P}\mathbf{v}$ corresponds to $\mathbf{b}_\lambda$ minimizing $L_N(\mathbf{Y}, \mathbf{Kb}) + \lambda \mathbf{b}'\mathbf{K}\mathbf{b}$ via

$$\mathbf{K}\mathbf{b}_\lambda = \mathbf{v}_\lambda \tag{5}$$

The SEL problem

For the particular special case of squared error loss, $L(y, \widehat{y}) = (y - \widehat{y})^2$, this development has a very explicit punch line. That is,

$$L_N(\mathbf{Y}, \mathbf{Kb}) + \lambda \mathbf{b}'\mathbf{Kb} = (\mathbf{Y} - \mathbf{Kb})'(\mathbf{Y} - \mathbf{Kb}) + \lambda \mathbf{b}'\mathbf{Kb}$$

Some vector calculus shows that this is minimized over choices of $\mathbf{b}$ by

$$\mathbf{b}_\lambda = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} \tag{6}$$

and corresponding fitted values are

$$\widehat{\mathbf{Y}}_\lambda = \mathbf{v}_\lambda = \mathbf{K}(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y}$$

Then using (6) under squared error loss, the solution to (1) is from (2)

$$\hat{f}_\lambda(\mathbf{x}) = \sum_{i=1}^{N} b_{\lambda i} \mathcal{K}(\mathbf{x}, \mathbf{x}_i) \tag{7}$$

The most general problem

CFZ provide a result summarizing the most general available version of this development, known as "The Representer Theorem." It says that if $\Omega : [0, \infty) \rightarrow \Re$is strictly increasing and

$$L\left(\left(\mathbf{x}_1, y_1, h\left(\mathbf{x}_1\right)\right), \ldots, \left(\mathbf{x}_N, y_N, h\left(\mathbf{x}_N\right)\right)\right) \geq 0$$

is an arbitrary loss function associated with the prediction of each y_i as $h\left(\mathbf{x}_i\right)$, then an $h \in \mathcal{A}$ minimizing

$$L\left(\left(\mathbf{x}_1, y_1, h\left(\mathbf{x}_1\right)\right), \left(\mathbf{x}_2, y_2, h\left(\mathbf{x}_2\right)\right), \ldots, \left(\mathbf{x}_N, y_N, h\left(\mathbf{x}_N\right)\right)\right) + \Omega\left(\|h\|_{\mathcal{A}}\right)$$

has a representation as

$$h\left(\mathbf{x}\right) = \sum_{i=1}^{N} \beta_i \mathcal{K}\left(\mathbf{x}, \mathbf{x}_i\right)$$

Unpenalized components

Further, if $\{\psi_1, \psi_2, \ldots, \psi_M\}$ is a set of real-valued functions and the $N \times M$ matrix $(\psi_j(\mathbf{x}_i))$is of rank M, then for $h_0 \in$span$\{\psi_1, \psi_2, \ldots, \psi_M\}$ and $h_1 \in \mathcal{A}$, an $h = h_0 + h_1$ minimizing

$$L((\mathbf{x}_1, y_1, h(\mathbf{x}_1)), (\mathbf{x}_2, y_2, h(\mathbf{x}_2)), \ldots, (\mathbf{x}_N, y_N, h(\mathbf{x}_N))) + \Omega(\|h_1\|_{\mathcal{A}})$$

has a representation as

$$h(\mathbf{x}) = \sum_{i=j}^{M} \alpha_j \psi_j(\mathbf{x}) + \sum_{i=1}^{N} \beta_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i)$$

The important generality provided above is that linear combinations of the functions $\psi_i(\mathbf{x})$ go unpenalized in fitting.

SEL implications

Then for the SEL case, take

$$\underset{N\times M}{\mathbf{\Psi}} = (\psi_j(\mathbf{x}_i))$$

and

$$\mathbf{R} = \left(\mathbf{I} - \mathbf{\Psi}\left(\mathbf{\Psi}'\mathbf{\Psi}\right)^{-1}\mathbf{\Psi}'\right)\mathbf{Y}$$

An optimizing $\boldsymbol{\alpha}$ is $\hat{\boldsymbol{\alpha}} = \left(\mathbf{\Psi}'\mathbf{\Psi}\right)^{-1}\mathbf{\Psi}'\mathbf{Y}$ where $\hat{\beta}_\lambda$ optimizes

$$(\mathbf{R} - \mathbf{K}\beta)'(\mathbf{R} - \mathbf{K}\beta) + \lambda\beta'\mathbf{K}\beta$$

and the earlier argument implies that $\hat{\beta}_\lambda = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{R}$.