1 Tikhonov Regularization

Developing a generalizable learning algorithm first entails minimizing empirical error on the training data set. We define the empirical risk $I_s[f]$ with a loss function V on training data $(x_i, y_i)_{i=1}^n$ as

$$I_s[f] = \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i). \tag{1}$$

Then empirical risk minimization (ERM) comprises the optimization problem of minimizing $I_s[f]$:

$$\min_{f\in\mathcal{H}} I_s[f] = \min_{f\in H} \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i). \tag{2}$$

Is this problem well-posed? Recall that a well-posed problem's solutions are:

- Existant,
- Unique, and
- Stable.

If the positive loss function V is strictly convex (no flat regions) and coercive (growing rapidly at extrema), there will exist a unique minimizer. The familiar squared loss and hinge loss functions are convex, but the 0-1 loss function is not.

In order to ensure stability, Tikhonov regularization alters the optimization problem with a positive real number, the regularized functional λ, and instead attempts to find the minimizer of $I_s[f] + \lambda\|f\|_{\mathcal{H}}^2$:

$$\min_{f\in\mathcal{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \lambda\|f\|_{\mathcal{H}}^2 \right\} \tag{3}$$

Tikhonov regularization constitutes one way to use prior information about training data to impose stability on ill-posed problems.

2 Representer Theorem

Any minimizer over the RKHS $\mathcal{H}$ of the regularized empirical functional

$$I_s[f] + \lambda\|f\|_{\mathcal{H}}^2 \tag{4}$$

can be represented by

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) \tag{5}$$

for some n-tuple $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$ provided that $\lambda > 0$. Minimizing over the Hilbert space now equates to minimizing over $\mathbb{R}^n$. This is a very nice result: we've shown that an optimization problem over a potentially infinite-dimensional space has a solution that can be expressed as a kernel expansion in terms of training set data.

One proof of the representer theorem is outlined below.

Proof: Define the linear subspace of $\mathcal{H}$,

$$\overline{\mathcal{H}} = \left\{ f \in \mathcal{H} \mid f = \sum_{i=1}^{n} \alpha_i K_{x_i}; (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\}. \tag{6}$$

This subspace $\overline{\mathcal{H}}$ is the space spanned by representers of the training set. Now let $\overline{\mathcal{H}}^{\perp}$ be a linear subspace of $\mathcal{H}$ and be orthogonal to $\overline{\mathcal{H}}$. Thus,

$$\mathcal{H} = \overline{\mathcal{H}} \oplus \overline{\mathcal{H}}^{\perp} \tag{7}$$

since $\overline{\mathcal{H}}$ is finite-dimensional, and

$$\overline{\mathcal{H}}^{\perp} = \left\{ f \in \mathcal{H} \mid \langle f, \sum_{i=1}^{n} \alpha_i K_{x_i} \rangle_{\mathcal{H}} = 0 \text{ for all } x_i \in \overline{\mathcal{H}} \right\}. \tag{8}$$

Each $f \in \mathcal{H}$ may be decomposed into a component, $\overline{f}$, along $\overline{\mathcal{H}}$ and a component, $\overline{f}^{\perp}$, along $\overline{\mathcal{H}}^{\perp}$:

$$f = \overline{f} + \overline{f}^{\perp}. \tag{9}$$

Then the empirical risk appears as

$$I_s[f] = \frac{1}{n} \sum_{i=1}^{n} V(\overline{f}(x_i) + \overline{f}^{\perp}(x_i), y_i). \tag{10}$$

By the reproducing property, the $\overline{f}^{\perp}$ term will be nullified in computing the inner product with the representer K_{x_i}. We then see that

$$I_s[f] = I_s[\overline{f}] + I_s[\overline{f}^{\perp}] = I_s[\overline{f}]. \tag{11}$$

Also, because of orthogonality,

$$\|\overline{f} + \overline{f}^{\perp}\| = \|\overline{f}\| + \|\overline{f}^{\perp}\|. \tag{12}$$

Now minimizing the regularized empirical risk over $\mathcal{H}$,

$$\min_{f \in \mathcal{H}} \left\{ I_s[f] + \lambda \|f\|_{\mathcal{H}}^2 \right\} = \min_{f \in \mathcal{H}} \left\{ I_s[\overline{f}] + \lambda(\|\overline{f}\|_{\mathcal{H}}^2 + \|\overline{f}^{\perp}\|_{\mathcal{H}}^2) \right\} \tag{13}$$

Since

$$\lambda(\|\overline{f}\|_{\mathcal{H}}^2 + \|\overline{f}^{\perp}\|_{\mathcal{H}}^2) \geq \lambda \|\overline{f}\|_{\mathcal{H}}^2, \tag{14}$$

the resulting minimizer must have $\|\overline{f}^{\perp}\|_{\mathcal{H}}^2 = 0$ and belong to subspace $\overline{\mathcal{H}}$.

□