CS281B/Stat241B: Advanced Topics in Learning & Decision Making

The Representer Theorem

Lecturer: Michael I. Jordan *Scribes: Xiaofeng Ren*

1 Addendum on the Gaussian kernel

As covered in a previous lecture, the *One-Class* SVM Classification aims at *novelty/outlier detection* in high dimensional spaces. The strategy there is to find a hyperplane in the feature space s.t. the observed data is separated away from the origin as far as possible. Here is an intuition:

The most commonly used kernel in practice is the Gaussian kernel, $K(x, x') = e^{-\frac{1}{2}\|x-x'\|}$. We can think about the feature map:

$$\Phi(x) : x \mapsto K(\cdot, x)$$

as a map from a point x to a Gaussian bump around x.

Now, let us look more closely at the functions $K(\cdot, x)$ in the feature space. What is the "length" of a "vector" $\Phi(x)$ for an arbitrary x?

$$\|\Phi(x)\|^2_{\mathcal{H}} = \langle K(\cdot, x), K(\cdot, x)\rangle_{\mathcal{H}} = K(x, x) = 1$$

by the reproducing property of $\mathcal{H}$.

And, what is the angle between any two vectors $\Phi(x)$ and $\Phi(x')$?

$$\cos(\alpha) = \langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}} = \langle K(\cdot, x), K(\cdot, x')\rangle_{\mathcal{H}} = K(x, x') \geq 0$$

The two facts above show that in the feature space, all the data points lie on the unit sphere within a single quadrant. This justifies the approach in One-Class SVM, which finds a hyperplane which separates the data away from the origin (see Figure 1).

2 The Representer Theorem

In the general case, the primal problem P is:

$$\min_{f \in \mathcal{H}} \{C(f, \{x_i, y_i\}) + \Omega(\|f\|_{\mathcal{H}})\}$$

where $\{x_i, y_i\}, i = 1, \cdots, m$ are the training data. If the problem satisfies the following conditions:

1. the loss function C is pointwise; i.e.,

$$C(f, \{x_i, y_i\}) = C(\{x_i, y_i, f(x_i)\})$$

   which only depends on $\{f(x_i)\}$, the values of f at the data points.

2. $\Omega(\cdot)$ is monotonically increasing.

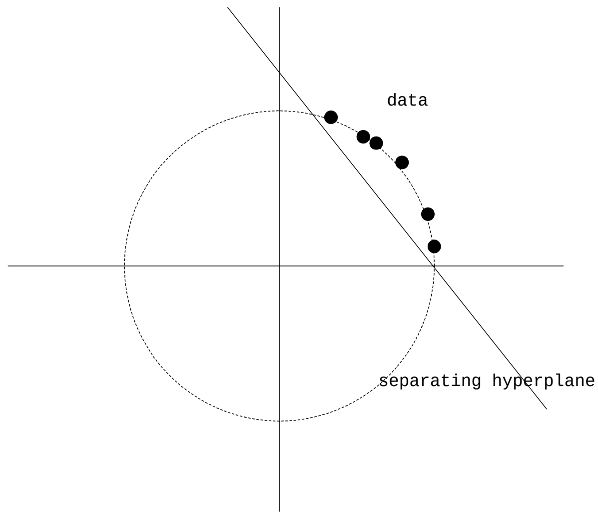

Figure 1: One-Class SVM

then the Representer Theorem (Kimeldorf & Wahba, 1971) states that every minimizer of P admits a representation of the form

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i K(\cdot, x_i)$$

I.e., the optimal f^* is a linear combination of (a finite set of) functions given by the data $\{x_i\}$.

This is a powerful result. It shows that although we search for the optimal solution in an infinite-dimensional space (in fact, if we don't have the regularization term, then we have infinitely many solutions which exactly go through all the data), adding the regularization term reduces the problem to finite-dimensional.

One way to prove the result is to use a Fourier basis, but this is complicated, involving calculus of variations. Here we give a simple, coordinate-free proof:

Without loss of generality, we assume that the second term in P has the form $\bar{\Omega}(\|f\|^2_{\mathcal{H}})$. Since this is just a monotonic transform, and in the original problem we allow for all monotonic functions Ω, we don't lose any generality.

Consider the linear subspace $\mathcal{H}_{\mathcal{D}}$ of $\mathcal{H}$ spanned by the functions $K(\cdot, x_i)$, $i = 1, \cdots, m$. Every f in the Hilbert space $\mathcal{H}$ has a unique decomposition, a component in the subspace and a component orthogonal to it:

$$f(\cdot) = f_{\|}(\cdot) + f_{\perp}(\cdot) = \sum_{i=1}^{m} \alpha_i K(\cdot, x_i) + f_{\perp}(\cdot)$$

where $f_{\perp}$ is perpendicular to the subspace $\mathcal{H}_{\mathcal{D}}$, i.e., $\langle f_{\perp}, K(\cdot, x_i)\rangle_{\mathcal{H}} = 0$ for all $i = 1, \cdots, m$.

Use the reproducing property,

$$f(x_j) = \langle f(\cdot), K(\cdot, x_j)\rangle = \sum_i \alpha_i \langle K(\cdot, x_i), K(\cdot, x_j)\rangle + \langle f_{\perp}(\cdot), K(\cdot, x_j)\rangle$$

The second term vanishes, so

$$f(x_j) = \sum_i \alpha_i K(x_j, x_i)$$

I.e., the values of f at the data points only depend on the coefficients $\{\alpha_i\}$ and not the perpendicular component $f_\perp$.

Why is this fact important? Because the loss function C is pointwise, so the first term only depends on the values of f at the data points. We can establish equivalence classes for the functions in $\mathcal{H}$ s.t. f and f' are equivalent if and only if $f(x_j) = f'(x_j)$ for all the data x_j.

The first term in P is the same for all the functions within each equivalence class. For the second term,

$$\Omega(\|f\|_\mathcal{H}) = \bar{\Omega}(\|f\|^2_\mathcal{H}) = \bar{\Omega}(\|\sum_i \alpha_i K(\cdot, x_i)\|^2_\mathcal{H} + \|f_\perp\|^2_\mathcal{H})$$

Ω is monotonic, so the minimizer of P within each equivalence class, i.e., for α_i's fixed, is the one which satisfies $\|f_\perp\| = 0$.

The global minimizer f^* of the primal problem P belongs to some equivalence class and it must be the minimizer within that class. Hence it satisfies $\|f^*_\perp\| = 0$, i.e., $f^* = \sum_i \alpha_i K(\cdot, x_i)$.

3 Another Way to Understand SVM: L_1 Regularization

The framework above does not provide us any intuition why, in SVM, some or most of the α_i's are zero. The regularization in the primal problem of SVM, $w^T w$, is L_2-like. As we have seen in the example of linear regression, L_1 regularization tends to make some of the parameters zero, while L_2 regularization usually does not.

One motivation comes from *basis-pursuit denoising* (Chen & Donoho), which studies the following cost:

$$J(\alpha) = \frac{1}{2}\|f(\cdot) - \sum_{i-1}^{N} \alpha_i \varphi_i(\cdot)\|^2_{L_2} + \lambda\|\alpha\|_{L_1}$$

The L_2 norm in the first term can not be calculated exactly. Instead, we approximate it by averaging over the data points,

$$J(\alpha) \approx \frac{1}{2N}\sum_{n=1}^{N}\left(y_n - \sum_{i=1}^{N}\alpha_i\varphi_i(x_n)\right)^2 + \lambda\sum_{i=1}^{N}|\alpha_i|$$

This cost term does not look like SVM yet, but the L_1 regularization does force some of the α_i's to be zero. Girosi pointed out that the following modified cost term does lead to SVM:

$$J_{SVM}(\alpha) = \frac{1}{2}\|f(\cdot) - \sum_{i=1}^{N}\alpha_i K(\cdot, x_i)\|^2_\mathcal{H} + \lambda\|\alpha\|_{L_1}$$

where in the first term we use the RKHS norm instead of the L_2 term. The surprising fact is that with the RKHS norm the first term can be calculated exactly, using the reproducing property:

$$\begin{aligned} J_{SVM}(\alpha) = &\ \tfrac{1}{2}\|f\|^2_\mathcal{H} - \textstyle\sum_i \alpha_i \langle f(\cdot), K(\cdot, x_i)\rangle + \tfrac{1}{2}\sum_i\sum_j \alpha_i\alpha_j \langle K(\cdot, x_i), K(\cdot, x_j)\rangle + \lambda\|\alpha\|_{L_1} \\ = &\ \tfrac{1}{2}\|f\|^2_\mathcal{H} - \textstyle\sum_i \alpha_i f(x_i) + \tfrac{1}{2}\sum_i\sum_j \alpha_i\alpha_j K(x_i, x_j) + \lambda\sum_i|\alpha_i| \end{aligned}$$

The first term, the RKHS norm of f, is independent of the α_i's. If we assume that $y_i = f(x_i)$, i.e., noise-free, the optimization problem becomes

$$\min_f\left\{-\sum_i \alpha_i y_i + \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j K(x_i, x_j) + \lambda\sum_i|\alpha_i|\right\}$$

and this is exactly the dual problem of SVM. It has the form of a L_2-like term (in the RKHS) plus a L_1 regularizer. Notice that formally it looks very different from the primal problem.

4 Introduction to Linear Operators

4.1 Linear operators

For a given Hilbert space $\mathcal{H}$, a function $T : \mathcal{H} \to \mathcal{H}$ is called a *linear operator* if:

1. $T(f_1 + f_2) = Tf_1 + Tf_2$, for any $f_1, f_2 \in \mathcal{H}$;
2. $T(\alpha f) = \alpha(Tf)$, for any $f \in \mathcal{H}$ and α scalar.

Some examples of linear operators:

1. For a finite-dimensional Hilbert space, a linear operator $A : g \mapsto f$ can be represented as an $n \times n$ matrix, where $n = \dim \mathcal{H}$. Choose a basis $\{\phi_i\}$ for $\mathcal{H}$, if under this basis g has a representation $g = \sum_i g_i \phi_i$ and similarly $f = \sum_i f_i \phi_i$, then A has a matrix representation (a_{ij}), such that $g = Af$ iff $g_i = \sum_j a_{ij} f_j$.
2. Integral operators are linear. For example, $(Tf)(\cdot) = \int K(\cdot, x) f(x) dx$, where K is a given kernel.
3. Differential operators are linear. For example, $Tf = \frac{d}{dx} f$, $Tf = \left(a \frac{d^2}{dx^2} + b \frac{d}{dx} \right) f$.

4.2 Adjoint operators

Adjoint operators are defined as following: a linear operator T^* is adjoint to a linear operator T if $< Tf, g >=< f, T^* g >$ for all $f, g \in \mathcal{H}$.

Examples of adjoint operators:

1. In R^n, $\langle x, y \rangle = x^T y$, so we have $\langle Ax, y \rangle = \langle x, A^T y \rangle$. The adjoint operator of A is the transpose of A.
2. What is the adjoint operator of $\frac{d}{dx}$ in L_2? Since

$$\begin{aligned} \left\langle \frac{d}{dx} f, g \right\rangle &= \int \tfrac{d}{dx} f(x) g(x) dx \\ &= f(x) g(x)|_{-\infty}^{+\infty} - \int f(x) \tfrac{d}{dx} g(x) dx \\ &= \int f(x) \left[\left(-\tfrac{d}{dx} \right) g(x) \right] d(x) \end{aligned}$$

The first term vanishes because both f and g are in L_2 so the limit goes to zero. The adjoint operator of $\frac{d}{dx}$ is $-\frac{d}{dx}$.