EECS 281B / STAT 241B: Advanced Topics in Statistical LearningSpring 2009

Lecture 5 — February 4

Lecturer: Martin Wainwright *Scribe: Jie Tang*

Note: These lecture notes are still rough, and have only have been mildly proofread.

5.1 Announcements

- HW1 due Monday Feb 9th (no late assignments)
- Minor clarifications on webpage

5.2 Today

- More on RKHS
- Representer Theorem
- Examples: max-margin classifier, general hard-margin SVM, ridge regression

5.3 Recap

Last time we introduced the concept of a reproducing kernel hilbert space (RKHS). We discussed the links between positive definite kernel functions and Hilbert spaces, and gave without proof a theorem linking every RKHS with a kernel function $\mathbb{K}$.

A *reproducing kernel Hilbert space* (RKHS) is a Hilbert space H of functions $\{f : X \rightarrow \mathbb{R}\}$ s.t. $\forall x \in X, \exists R_x \in Hs.t. < R_x, f >= f(x) \forall f \in H$, i.e. there exists a function R_x called the representer of evaluation for anye $x \in X$ s.t. we can evaluate f at x by taking the inner product in H with the representer of x.

Examples of RKHS include finite dimensional spaces and Sobolev spaces (see last lecture)

5.4 Correspondence between RKHS and PSD kernels

Theorem 5.1. *To any RKHS there exists a positive semidefinite kernel function. Conversely, given any PSD kernel we can construct a RKHS s.t.* $R_x(.) = k(., x)$.

Note: The kernel function is unique, though it requires more work to prove this.

Proof: Given a RKHS, define a kernel function via $K(x,y) :=< R_x, R_y >_H \, \forall x, y \in X$. (By the def'n of an RKHS, there is a representer $R_x \forall x \in X$). This is a valid kernel function only if it is positive semidefinite Given any $x_1, ..., x_n \in X$, and any $a \in \mathbb{R}^n$, we need to show $a^T K a \geq 0$ where $K_{ij} = K(x_i, x_j)$

Expanding out the quadratic form, we have

$$\begin{aligned} a^T K a &= \sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) && (5.1) \\ &= \sum_{i,j=1}^{n} a_i a_j < R_{x_i}, R_{x_j} >_H && (5.2) \\ &= || \sum_{i=1}^{n} a_i R_{x_i} ||_H^2 \geq 0 && (5.3) \\ & && (5.4) \end{aligned}$$

Conversely, say we have a PSD kernel K. We must construct a space s.t. $K(.,x)$ is the representer, which has reproducing property, and is a Hilbert space (complete).

Define a linear space of functions:

$$L = \text{span}\{K(.,x) | x \in X\} = \left\{ \sum_{i=1}^{n} a_i K(., x_i) | n \in N, a \in \mathbb{R}^n, \{x_1, ..., x_n\} \subset X \right\} \tag{5.5}$$

Define an inner product on this space:

$$< \sum_{i=1}^{n} a_i K(., x_i), \sum_{j=1}^{m} b_j K(., y_j) >= \sum_{i,j} a_i b_j K(x_i, x_j) \tag{5.6}$$

Because the matrix K is positive semidefinite, this expression must be ≥ 0, and this inner product is positive semidefinite.

Next, we must check that $K(.,x)$ has the representer property:

$$< K(.,x), f(.) >=< K(.,x), \sum_{j=1}^{m} b_j K(., y_i) >= \sum_{j=1}^{m} b_j K(x, y_i) = f(x) \tag{5.7}$$

Finally, we must make sure the space H is complete, i.e. that it is actually a Hilbert space)

Let f_n be a Cauchy sequence in L i.e. $\forall \epsilon > 0, \exists N(\epsilon) s.t. \forall n, m \geq N(\epsilon)$

$$||f_n - f_m||_H < \epsilon$$

For any $x \in X, f_n, f_m$, we can check that the Cauchy sequence converges to some optimum f^*.

$$\begin{aligned} |f_n(x) - f_m(x)| &= | < K(.,x), f_n - f_m >_H | \\ &\leq ||K(.,x)||_H ||f_n - f_m||_H \end{aligned}$$

where the last line above follows by Cauchy Schwarz. Since $||f_n - f_m|| < \epsilon$, we have that $|f_n(x) - f_m(x)| < \epsilon$ and $f_n(x) \to f^*(x)\ \forall x \in X$

(This is special property of the representer of evaluation: in general functions such an optimization can converge to a limit without converging pointwise).

To ensure that our Hilbert space is complete, we can add all limits $f_n \to f*$ to our space. Our final result is:

$$H = \overline{\{\text{span} K(.,x) | x \in X\}} \tag{5.8}$$

where $\overline{X}$ denotes the closure of the space X.

□

5.5 Applications of RKHS

How do we use the RKHS machinery to develop classifiers or other statistical estimators?

5.5.1 Max-margin Classifiers

Consider a generalization of hard-margin SVMs. Recall that to train an SVM we solve the following quadratic program.

$$\begin{aligned} \min \quad & \frac{1}{2}||\theta||_2^2 \\ \text{s.t. } y^i < \theta, x^i > \quad \geq \quad & 1\ \forall i = 1, ..., n \end{aligned}$$

Assuming linearly separable training data, this q. program maximizes the margin between the decision boundary and the nearest training examples. These are called max-margin classifiers. Maximizing the margin is appealing because it minimizes the expected risk of classifying unknown data.

Note: we can also introduce slack variables to account for nonseparable data; hw problem 1.5.

We generalize this algorithm by performing the optimization over an RKHS H.

$$\begin{aligned} \min \quad & \frac{1}{2}||\theta||_H^2 \\ \text{s.t. } y^i f(x^i) \quad \geq \quad & 1\ \forall i = 1, ..., n \end{aligned}$$

Our classification rule becomes $f(x^i) \geq 1$ for some function f in the RKHS associated with our kernel. The loss function for classification is given by

$$L(x^i, y^i, f(x^i)) = \sum_{i=1}^{n} I(y^i f(x^i) \geq 1)$$

Instead of just linear functions, we can now use polynomial kernels, meaning richer decision boundaries can be fit by this classifier. However, more powerful classifiers also run the risk of overfitting.

5.5.2 Lagrangian duality

In the original SVM, we can always find a solution $\theta = \sum_i \alpha_i y^i x^i$, where $\alpha \in \mathbb{R}^n$ comes from the dual problem. In the generalized SVM algorithm, the space we are optimizing over could be an infinite dimensional space. But, we can show (via the representer theorem, next time) that the solution always takes the form:

$$f(.) = \sum_{i=1}^{n} \alpha_i y^i K(., x^i)$$

Note: In many settings, the α_i are likely to be sparse, i.e. contain many zeroes. Only a few of the data points x_i with nonzero α must be involved in the computation: these are support vectors.

5.6 Representer Theorem

Theorem 5.2. *Let $\Omega : [0, \infty] \to \mathbb{R}$ be strictly increasing and let $l : (X \times Y \times \mathbb{R})^n \to \mathbb{R} \cup \{\infty\}$ be a loss function. Consider $\min_{f \in H} l((x^i, y^i, f(x^i))) + \lambda_n \Omega(||f||_H^2)$ (think of Ω as identity), $\lambda_n > 0$, and Y is $\{-1, 1\}$ in classification. Ω is a regularization operator which puts a penalty on more complicated functions in our space.*

Then any optimal solution has the form:

$$f(.) = \sum_{i=1}^{n} \alpha_i K(., x^i) \tag{5.9}$$

Intuitively, the loss function only depends on observed data points, so the optimal solution function should only depend on kernel functions centered at the observed data points.

Proof: (due to Kimeldorf & Wahba 1970s, proved for kernel ridge regression. Later, Smola and Scholkopf derive a more general version, using the same ideas.)

Any $f \in H$ can be written as

$$f = \sum_{i=1}^{n} \alpha_i k(., x^i) + f_\perp$$

where $f_\perp = V^\perp, V = \overline{\text{span}\{K(., x^i), i = 1, ..., n\}}$

We will show that the $f_\perp$ component is 0.

By the representer property of RKHS, $\forall j = 1, ..., n$

$$\begin{aligned} f(x^j) &= < K(., x^j), f >_H \\ &= \sum_{i=1}^{n} \alpha_i K(x^j, x^i) + < K(., x^j), f_\perp >_H \end{aligned}$$

Note that $K(., x^j) \in V, f_\perp \in V^\perp$, so their inner product is 0 and the second term is zero. By the Pythagorean Theorem,

$$\Omega(||f||_H^2) = \Omega(||\sum_{i=1}^{n} \alpha_i K(., x^i)||_H^2 + ||f_\perp||_H^2) \tag{5.10}$$

By monotonicity, if $||f_\perp|| \geq 0$, then the regularizer term Ω will be larger because Ω is strictly increasing. Therefore, $||f_\perp|| = 0$

□

5.7 Example: linear vs. kernel ridge regression

The standard form of the linear regression problem is $y^i =< \theta, x^i > +w^i$where $w \in \mathbb{R}$, $\theta \in \mathbb{R}^d.\theta$ is the regression vector, x^i are the covariates / predictors, and they generate a response $y^i \in \mathbb{R}$. Given samples $i = 1, ..., n$ and a loss function $l(y^i, x^i, f(x^i)) = \sum_{i=1}^{n}(y^i - < \theta, x^i >)^2$, add a regularizer term to get

$$\min_{\theta \in \mathbb{R}^d} \{L(\theta) + \frac{\lambda}{2}||\theta||_2^2\} \tag{5.11}$$

This is classic ridge regression, optimizing over in $\mathbb{R}^d$. The well-studied optimal solution in unique, closed form:

$$\hat{\theta}_{RID} = (X^T X + \lambda I)^{-1} X^T y$$

where $X = [x^1, x^2, ..., x^n]^T$, a $n \times d$ vector, and $y = [y^1, y^2, ..., y^n]^T \in \mathbb{R}^n$

With some fancy linear algebra, we can rewrite this in the following form:

$$\hat{\theta}_{RID} = X^T (X X^T + \lambda_n I_{n \times n})^{-1} y \tag{5.12}$$

The key matrix here is XX^T: this is a kernel Gram matrix $K_{ij} =< x^i, x^j >$ for the ordinary inner product. To kernelize this, we replace the inner product matrix with a PSD kernel matrix K. This allows us to optimize over functions $f \in H$, and use the RKHS norm for the regularization terms.