EECS 281B / STAT 241B: Advanced Topics in Statistical Learning **Spring 2009**

Lecture 4 — February 2

Lecturer: Martin Wainwright *Scribe: Luqman Hodgkinson*

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

Announcements

- GSI: Junming Yin (junming@eecs.berkeley.edu); his offfice hours are posted on the webpage.
- HW #1 due Mon. Feb. 9th in class (no late assignments).

Today

- kernels and their uses
- reproducing kernel Hilbert space

Recap

So far we have looked at two simple non-parametric classifiers based on linear discriminant functions.

Methods

1. perceptron
2. max-margin SVM

- Given samples $\{x^{(i)}, y^{(i)}\}$, these are based on functions $f_\theta(x) = \langle \theta, x \rangle$ with $\theta = \sum_{i=1}^n \alpha_i y^{(i)} x^{(i)}$
- The indices $\{i : \alpha_i \neq 0\}$ index the subset of samples contributing to the classifier
- In the perceptron, α_i is related to the number of times point i is misclassified, because we increase α_i by one each time

For these methods $f_\theta(x) = \langle \theta, x \rangle = \sum_{i=1}^n \alpha_i y^{(i)} \left\langle x, x^{(i)} \right\rangle$, so we never need to represent θ explicitly. We just need weights and a representation of the inner products. This is true both for the linear perceptron algorithm and when solving the max-margin optimization problem. For the linear MM classifier, in the dual we can find α directly and never need to compute θ explicitly. This suggests a route to more powerful classifiers via the "kernel trick."

4.1 The "Kernel Trick"

Idea: Find a feature map, i.e. a function $\Phi : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{X}$ is the feature space and $\mathcal{Z}$ is a much higher dimensional space that can even be infinite dimensional. Linear classifiers in the higher-dimensional space $\mathcal{Z}$ can be very non-linear in the original feature space.

Eg. $\Phi((x_1, ..., x_d)) = (x_1, ..., x_d, ..., x_i x_j, ..., x_i x_j x_k, ..., x_d^3)$. Here $x \in \mathbb{R}^d$ and $\Phi(x) \in \mathbb{R}^{\binom{d}{1}+\binom{d}{2}+\binom{d}{3}}$ where the exponent is of order d^3. This can be generalized to map to tuples of all monomials of degree $\leq k$. In the example above, $k = 3$. Working in these high-dimensional spaces can be very difficult. If $d = 1000$, for $k = 3$, the dimension of $\mathcal{Z}$ is on the order of 10^9. For $k = 10$, the dimension of $\mathcal{Z}$ is on the order of 10^{30}, though $k = 10$ is most likely too high and may fit noise in the data.

The "kernel trick" is a computational solution to working in these high dimensions. If we had an efficient way of computing inner products, we would never need to work in the higher dimensional space, just in the space of inner products.

Suppose there were a function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\mathbb{K}(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{Z}} \forall x, y \in \mathcal{X}$. Then we could kernelize, i.e. replace every inner product with this, which is nice because then we would not pay the price of working in the high-dimensional space.

4.2 Hilbert Spaces

4.2.1 Definitions

A Hilbert space $\mathcal{H}$ is an inner product space (i.e. a set on elements closed under addition and scalar multiplication, and for an inner product $\langle f, g \rangle_{\mathcal{H}}$ is defined $\forall f, g \in \mathcal{H}$) that is complete with respect to norm $|| \cdot ||_{\mathcal{H}}$ where $||f||_{\mathcal{H}} := \sqrt{\langle f, f \rangle_{\mathcal{H}}}$.

Complete means that any Cauchy sequence $\{f_n\} \subseteq \mathcal{H}$ converges to some $f^* \in \mathcal{H}$.

A Cauchy sequence $\{f_n\}$ is a sequence such that $\forall \epsilon > 0, \exists N(\epsilon)$ such that $\forall n, m \geq N(\epsilon)$, $||f_n - f_m||_{\mathcal{H}} < \epsilon$.

4.2.2 Examples

Example (a): $\mathbb{R}^d$ with Euclidean inner product is a finite-dimensional Hilbert space

Example (b): $\ell^2(\mathbb{N}) = \{(a_1, a_2, a_3, ...) \mid \sum_{i=1}^{\infty} a_i^2 < +\infty\}$, i.e., the set of all sequences that are square summable, with $\langle a, b \rangle = \sum_{i=1}^{\infty} a_i b_i$, is an infinite-dimensional Hilbert space.

Example (c): Say $\mathcal{C} = \{f_i, i = 1, ..., d\}$ is some class of functions. E.g., maybe a bunch of separate estimators; if we take a weighted sum of them, it is an aggregation technique. The set

$$\text{span}(\mathcal{C}) = \{\sum_{i=1}^{d} a_i f_i \mid (a_1, ..., a_d) \in \mathbb{R}^d\}$$

is closed under addition and multiplication by scalars. We can define an inner product with $G \in \mathbb{R}^{d\times d}$, any symmetric positive definite matrix, by $\langle f_i, f_j\rangle := G_{ij}\ \forall i,j = 1,...,d$, and, thus, $\langle \sum a_i f_i, \sum b_j f_j\rangle = \sum_{i,j} a_i b_j G_{ij}$. This is a finite-dimensional Hilbert space, but a rich one, as the base functions can be quite complicated.

Example (d): $L^2[0,1] = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 f^2(t)\ \mathrm{d}t < +\infty\}$, with $\langle f, g\rangle = \int_0^1 f(t)g(t)\ \mathrm{d}t$, so that $\int_0^1 f^2(t)\ \mathrm{d}t < +\infty$ is equivalent to $||f||_{\mathcal{H}}^2 < +\infty$. This is too rich, meaning it is not a reproducing kernel Hilbert space (RKHS). Everything (a)-(c) is a RKHS so we can find kernels we like.

Example (e): $H = \{f : [0,1] \to \mathbb{R} \mid f(0) = 0, f$ diff'ble almost everywhere with $f' \in L^2[0,1]\}$. f can have isolated bad points, say a corner point, but not pathologically many. This is a Sobolev space (see Adams, 1975). The inner product is interesting, defined as $\langle f, g\rangle_H = \int_0^1 f'(t)g'(t)\ \mathrm{d}t$.

Examples (a)-(c) and (e) are RKHS. Example (d) is not. Exercise: think about this.

4.3 Reproducing Kernel Hilbert Spaces

We will focus on Hilbert spaces of functions $f : \mathcal{X} \to \mathbb{R}$.

A reproducing kernel Hilbert space (RKHS) is a Hilbert space in which $\forall x \in \mathcal{X}, \exists R_x \in \mathcal{H}$ such that $\langle R_x, f\rangle_{\mathcal{H}} = f(x), \forall f \in \mathcal{H}$. That is, function evaluation has a linear representation in the Hilbert space. R_x is the representer of evaluation for x. Note that this is just one of many definitions of RKHS.

Example (d) (revisited): Can we find g such that $\int_0^1 f(t)g(t)\ \mathrm{d}t = f(x)$. We are tempted to think of δ-functions to put infinite mass at x. In (d), δ-functions are not in $\mathcal{H}$.

4.4 Positive Semi-Definite (PSD) Kernels

A function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semidefinite (PSD) kernel if $\forall n \in \mathbb{N}(n = 1, 2, ...)$ and $\{x^{(i)}, i = 1, ..., n\} \subset \mathcal{X}$, the matrix $K \in \mathbb{R}^{n\times n}$ with $K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)})$ is symmetric and positive semi-definite. A matrix $K \in \mathbb{R}^{n\times n}$ is positive semi-definite if $\forall a \in \mathbb{R}^n, a^T K a \geq 0$.

4.5 Relationship Between RKHS and PSD Kernels

Theorem 4.1. *To every RKHS, there is a unique PSD kernel $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Conversely, given a PSD kernel, we can construct a RKHS such that $R_x(\cdot) = \mathbb{K}(\cdot, x)$ is the representer of evaluation.*

Note that $\mathbb{K}(\cdot, x) = R_x(\cdot)$ belongs to the Hilbert space in this correspondence between kernels and the reproduction property.

Given a kernel, we can build rich classes of functions. The feature set can be a subset of $\mathbb{R}$, but not necessarily. We can have kernels on combinatorial objects such as graphs, subsequences, or all subsets of a set. These are combinatorial kernels that are not continuous. Discrete kernels on combinatorial objects are very important in fields such as computational biology.

Example (c) (revisited):

$\mathcal{H} = \text{span}\{f_1, ..., f_d\}$, i.e. all linear combinations of the functions $f_i : \mathcal{X} \to \mathbb{R}$ with inner products $G_{ij} = \langle f_i, f_j \rangle$, where G is any symmetric positive definite $d \times d$ matrix. One can prove that G satisfies the properties of an inner product. Let $\{e_1, e_2, ..., e_d\}$ be any orthonormal basis of $\mathcal{H}$. Since $\mathcal{H}$ is a vector space of functions, we can always do Gram-Schmidt to get an orthonormal basis. This assumes that $f_1, ..., f_d$ are linearly independent. Define $\mathbb{K}(y, x) = \sum_{i=1}^d e_i(x)e_i(y)$. For any $x \in \mathcal{X}, \mathbb{K}(\cdot, x) = \sum_{i=1}^d e_i(x)e_i(\cdot) \in \mathcal{H}$ as $e_i(x)$ is a constant. We need to check that it reproduces. As the $\{e_i\}$ are orthonormal,

$$\langle e_i, e_j \rangle_{\mathcal{H}} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Take any $f = \sum_{i=1}^d a_i e_i \in \mathcal{H}$ for $a \in \mathbb{R}^d$. Then

$$\begin{aligned} \langle f, \mathbb{K}(\cdot, x) \rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^d a_i e_i, \sum_{i=1}^d e_j(x) e_j \right\rangle_{\mathcal{H}} \\ &= \sum_{i,j} a_i e_j(x) \langle e_i, e_j \rangle_{\mathcal{H}} \\ &= \sum_{i=1}^d a_i e_i(x) \\ &= f(x) \end{aligned}$$

Therefore, this particular Hilbert space is a RKHS. This proof applies more generally to generic finite-dimensional Hilbert spaces. Any finite-dimensional Hilbert space is reproducing, and the same line of argument can be used to construct the associated kernel.

Example (e) (revisited): Sobolev.

$\mathcal{H} = \{f : [0, 1] \to \mathbb{R} \mid f \text{ diff'ble almost everywhere, } \int_0^1 f'^2(t)\, \mathrm{d}t < +\infty, f(0) = 0\}$. Even though L^2 is not a RKHS, this Hilbert space is a RKHS.

In particular, let us define the kernel as $\mathbb{K}(x, y) := \min(x, y)$. Then $R_y(\cdot) = \mathbb{K}(\cdot, y) = \min(\cdot, y)$, which is sketched in Figure 4.1. We need to check that $R_y(\cdot) \in \mathcal{H}$. Clearly $R_y(0) = 0$. What about the diff'ble property? The derivative is

$$R'_y(x) = \begin{cases} 1 & \text{for } 0 \le x < y \\ 0 & \text{for } x > y \end{cases}$$

Thus, $R_y(\cdot)$ is not diff'ble everywhere, but it has only isolated weirdness at the point $x = y$. We can integrate the square of the derivative and check that it is finite, so $R_y(\cdot)$ does belong to the space.

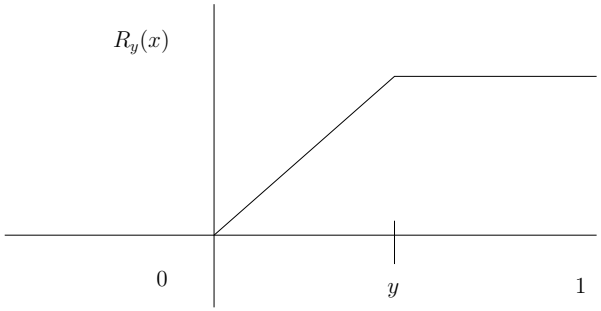

Figure 4.1. The function $R_y(x)$ for Example (e) (revisited): Sobolev.

To check the reproducing property:

$$\begin{aligned}
\langle R_y, f \rangle &= \int_0^1 R_y'(t) f'(t)\, \mathrm{d}t \\
&= \int_0^y f'(t)\, \mathrm{d}t \\
&= f(y) - f(0) \qquad \text{(by the Fundamental Theorem of Calculus)} \\
&= f(y)
\end{aligned}$$

Thus $\mathcal{H}$ is a RKHS.

To complete the correspondence between the Sobolev space $\mathcal{H}$ from Example (e) and the PSD kernel guaranteed by Theorem 4.1, one needs to show that $\mathbb{K}(x, y) = \min(x, y)$ is a PSD kernel function, i.e. for any k points, $x_1, ..., x_k$, the matrix $K \in \mathbb{R}^{k \times k}$ with $K_{i,j} = \min(x_i, x_j)$ is positive semi-definite. Exercise: Prove this by demonstrating a Gaussian process with $\mathbb{K}(x, y) = \min(x, y)$ as its covariance function.