CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Multiple Kernels and Reproducing Kernel Hilbert Spaces

Lecturer: Michael I. Jordan *Scribes: Blaine Nelson and Sierra Boyd*

1 Convex Programming

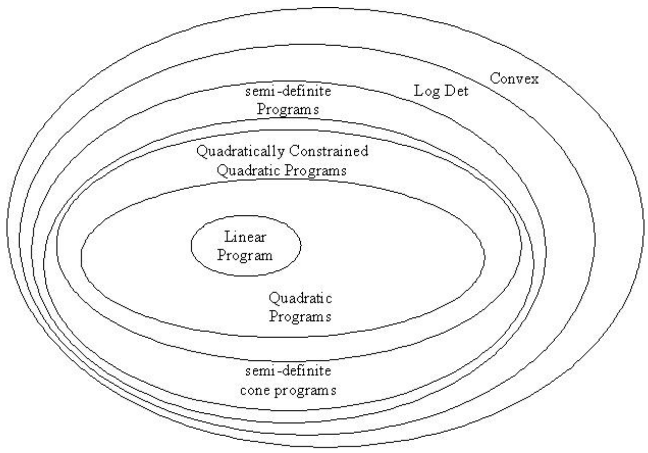

Figure 1: Diagram of Convex Programming

2 Multiple Kernels

We continue our discussion on multiple kernels from the previous lecture. Recall,

$$\begin{aligned}
&\mathrm{K} = \sum_i \mu_i K_i \\
&\mathrm{K} \succeq 0 \\
&\mathrm{tr}\,(\mathrm{K}) \leq c
\end{aligned}$$

Given data pairs $\{(x_i, y_i)\}_{i=1}^{n}$ ($x_i \in \chi$ and $y_i \in \{-1, 1\}$), let $G\,(\mathrm{K}) = \mathrm{diag}\,(Y) \cdot K \cdot \mathrm{diag}\,(Y)$ where

$$\mathrm{diag}\,(Y) = \begin{pmatrix} y_1 & & & \\ & y_2 & \mathbf{0} & \\ & \mathbf{0} & \ddots & \\ & & & y_n \end{pmatrix}$$

Thus, the (i,j)-th element of $G\,(\mathrm{K})$ is given by $G\,(\mathrm{K})_{ij} = y_i y_j \mathrm{K}\,(x_i, x_j)$

The hard-margin SVM associated with with $G\,(\mathrm{K})$ is represented by the following program:

$$\begin{array}{cc}
\max & 2\alpha^T e - \alpha^T G\,(\mathrm{K})\,\alpha \\
s.t. & \alpha^T y = 0 \quad \alpha_i \geq 0
\end{array}$$

$$e = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Moreover, the soft-margin SVM can be represented by a similar program where the last constraint is boxed $0 \leq \alpha_i \leq C$.

How can we optimize over α and μ?

Theorem 1

Multiple Kernel Optimization Theorem: *The problem of optimizing*

$$\begin{array}{ll} \max\limits_{\alpha} & 2\alpha^T e - \alpha^T \left(G\left(\mathrm{K}\right) + \tau I\right)\alpha \\ s.t. & \alpha^T y \geq 0 \quad 0 \leq \alpha_i \leq C \end{array}$$

with respect to $\{\mu_i\}$ *where* τ *is a regularization parameter reduces to:*

$$\min_{\{\mu_i\},t,\lambda,\nu,\delta} t$$

$$\begin{array}{ll} & \mathrm{tr}\left(\mathrm{K}\right) = 0 \quad \mathrm{K} = \sum\limits_i \mu_i K_i \\ s.t. & \left(\begin{array}{cc} G\left(\mathrm{K}\right) + \lambda\mathbf{I} & e + \nu - \delta + \lambda y \\ \left(e + \nu - \delta + \lambda y\right)^T & t - 2C\delta^T e \end{array}\right) \succeq 0 \\ & \nu \geq 0 \quad \delta \geq 0 \end{array}$$

which is a Semi-Definite Program (SDP).

Proof. Optimizing with respect to $\{\mu_i\}$ is equivalent to the following program:

$$\begin{array}{l} \min\limits_{\{\mu_i\}} \max\limits_{\alpha} \quad 2\alpha^T e - \alpha^T \left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right)\alpha \\ \equiv \min\limits_{\{\mu_i\},t} (t) \geq \max\limits_{\alpha} \left(2\alpha^T e - \alpha^T \left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right)\alpha\right) \end{array}$$

Taking the dual of this gives the Lagrangian,

$$L\left(\alpha,\nu,\lambda,\delta\right) = 2\alpha^T e - \alpha^T \left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right)\alpha + 2\nu^T\alpha + 2\lambda y^T\alpha + 2\delta^T\left(Ce - \alpha\right)$$

Now, we maximize over α to find the dual problem. This is done by setting $\frac{\partial L}{\partial \alpha}$ to 0. Thus we find that $\alpha = \left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right)^{-1}\left(e + \nu - \delta + \lambda y\right)$. This yields the following dual problem:

$$\min_{\nu\geq 0,\delta\geq 0,\lambda} \quad \underbrace{\left(e + \nu - \delta + \lambda y\right)^T \left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right)^{-1}\left(e + \nu - \delta + \lambda y\right) + 2C\delta^T e}_{\beta}$$

Hence the primal statement is less than or equal to t if and only if $\exists \nu, \delta, \lambda$ such that $\beta \leq t$.

But, from the Schur Complement Theorem, we know that,

$$\left(\begin{array}{cc} A & B \\ C & D \end{array}\right) \succeq 0 \quad \Leftrightarrow \quad \begin{array}{c} A \succeq 0 \\ D - CA^{-1}B \succeq 0 \end{array}$$

But $D - CA^{-1}B$ has the exact same form as $-\beta$. Moreover, from the problem statement, $\left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right) \succeq 0$. Thus, $\left(e + \nu - \delta + \lambda y\right)^T \left(G\left(\mathrm{K}\right) + \tau\mathbf{I}\right)^{-1}\left(e + \nu - \delta + \lambda y\right) + 2C\delta^T e \preceq 0$ if and only if,

$$\left(\begin{array}{cc} G\left(\mathrm{K}\right) + \tau I & e + \nu - \delta + \lambda y \\ \left(e + \nu - \delta + \lambda y\right)^T & t - 2C\delta^T e \end{array}\right) \succeq 0$$

Thus, this condition implies that $\beta \leq t$, which is equivalent to the original min-max problem.

$\square$

3 Hilbert Space

A Hilbert space is essentially an Euclidean space, but a Euclidean space that may be infinite-dimensional. It is a vector space (i.e., is closed under addition and scalar multiplication, obeys the distributive and associative laws, etc.). It is also endowed with an inner product $\langle \cdot, \cdot \rangle$; a bilinear form obeying the following conditions:

$$\begin{aligned} \langle x+y, z \rangle &= \langle x, z \rangle + \langle y, z \rangle \\ \langle \alpha x, y \rangle &= \alpha \langle x, y \rangle \\ \langle x, y \rangle &= \langle y, x \rangle \\ \langle x, x \rangle &\geq 0 \\ \langle x, x \rangle = 0 &\implies x = 0 \end{aligned}$$

From $\langle \cdot, \cdot \rangle$ we get a norm $\| \cdot \|$ via $\| x \| \triangleq \langle x, x \rangle^{1/2}$. This norm allows us to define notions of convergence. Adding all limit points of Cauchy sequences (the distance between elements x_i and x_{i+1} of the sequence becomes small as $i \to \infty$) to our space yields a Hilbert space—a *complete* inner product space.

Theorem 2

Cauchy-Schwartz:

$$\langle x, y \rangle \leq \| x \| \| y \|$$

is easily proved for any Hilbert space.

3.1 Examples

- R^n: $\langle x, y \rangle = x^T y$
- L_2: $\langle x, y \rangle = \int x(t) y(t) dt$
- l_2: $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$

4 Reproducing Kernel Hilbert Spaces

The Hilbert space L_2 is too "big" for our purposes, containing too many non-smooth functions. One approach to obtaining restricted, smooth spaces is the Reproducing Kernel Hilbert Space (RKHS) approach. A RKHS is "smaller" than a general Hilbert space.

Given a kernel $k(x, x')$, we will construct a Hilbert space such that k defines an inner product in that space. First define the *Gram matrix*. Given points $x_1, x_2, x_3, ... x_n$, define:

$$K_{ij} = k(x_i, x_j)$$

We say that the kernel k is *positive definite* (p.d.) if its Gram matrix is positive definite for all $x_1, x_2, ..., x_n$. Similarly, we define positive semidefinite (p.s.d.), negative definite, and negative semidefinite by the properties of the Gram matrix.

The Cauchy-Schwartz inequality holds for p.s.d. kernels:

$$k(x_1, x_2)^2 \leq k(x_1, x_1) k(x_2, x_2)$$

Proof: Form a Gram matrix of the two points x_1 and x_2:

$$K = \begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) \\ k(x_2,x_1) & k(x_2,x_2) \end{pmatrix}$$

For K to be positive semidefinite as a matrix, the determinant of K must be nonnegative:

$$\Longrightarrow k(x_1,x_1)k(x_2,x_2) - k(x_2,x_1)k(x_1,x_2) \geq 0,$$

which implies Cauchy-Schwartz.

Define the following *reproducing kernel map* for a kernel $k(\cdot,\cdot)$:

$$\Phi : x \longrightarrow k(\cdot,x).$$

I.e., to each point x in the original space we associate a function $k(\cdot,x)$.

Example: Gaussian kernel. Each point x maps to a Gaussian centered at that point as depicted in Fig. 2. Intuitively this captures the similarity of x to all other points.

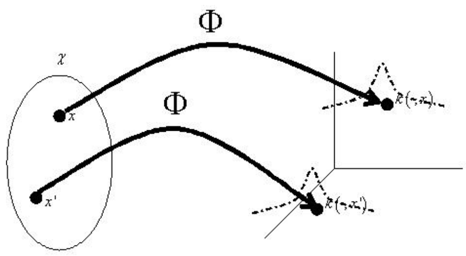

Figure 2: The mapping of an input space to a gaussian feature space.

We now construct a vector space containing all linear combinations of the functions $k(\cdot,x)$:

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

for arbitrary $\{x_i \in \chi\}_{i=1}^m$. This will be used to construct our RKHS H.

We now define an inner product. Let $f(\cdot) = \sum_i \alpha_i k(\cdot, x_i)$ and $g(\cdot) = \sum_j \beta_j k\left(\cdot, x'_j\right)$ such that $f, g \in H$, and define:

$$\langle f, g \rangle \triangleq \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k\left(x_i, x'_j\right)$$

We need to verify that this in fact defines an inner product. Symmetry is obvious: $\langle f, g \rangle = \langle g, f \rangle$. Linearity is easy to show. In the next lecture we will establish the key property: $\langle f, f \rangle = 0 \Longrightarrow f = 0$.

Kernels are analogs of Dirac delta functions. Consider L_2 (which is not a RKHS). We have:

$$f(x) = \int f(t)\delta(t, x)dt,$$

where $\delta(t, x)$ is the Dirac delta function. The Dirac delta function is the representer of evaluation for L_2, but of course it is not itself in L_2. (Which is consistent with the fact that L_2 is not a RKHS).