CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Introduction to Kernels

Lecturer: Michael I. Jordan *Scribe: Thinh Nguyen*

1 Review on SVM

Recall that the solution to the optimization problem:

$$\begin{aligned} \min \quad & \frac{1}{2}w^T w \\ s.t \quad & y_i(x_i^T w + b) \geq 1, \quad \alpha \geq 0 \end{aligned} \tag{1}$$

is:

$$w = \sum_{i:\alpha_i > 0} \alpha_i y_i x_i \tag{2}$$

Therefore, the fitted hyperplane is:

$$f(x) = \sum_{S.V} \alpha_i y_i x_i^T x + b \tag{3}$$

2 Kernel Trick

The idea is to replace $x^T y$ with a kernel $K(x,y) = \Phi^T(x)\Phi(y)$ for some $\Phi(.)$
A few examples of kernels:

$$K(x,y) = e^{-\frac{1}{2}\|x-y\|^2}$$
$$K(x,y) = (x^T y)^n$$
$$K(x,y) = Dirichlet$$

To derive the Dirichlet kernel, consider a function $f(x)$ written in a Fourier basis:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{N} a_k \cos(kx) + \sum_{k=1}^{N} b_k \sin(kx) \tag{4}$$

Define $C = (\frac{a_0}{2}, a_1, ..., b_1, ...b_n)^T$ and $\Phi(x) = (1, \cos x, \cos 2x, ..., \sin x, \sin 2x, ...)^T$ then $f(x) = C^T\Phi(x)$
Now note that

$$\Phi^T(x)\Phi(y) \quad = \quad 1 + \sum_{k=1}^{N} (\cos kx \cos ky + \sin kx \sin ky)$$

$$
\begin{aligned}
&= \sum_{k=0}^{N} \cos k(x-y) \\
&= Re\left[\sum_{k=0}^{N} e^{ik(x-y)}\right] \\
&= Re\left[\frac{1-e^{i(N+1)(x-y)}}{1-e^{-i(x-y)}}\right] \\
&= \frac{\sin((N+1)(x-y)/2)}{2\sin((x-y)/2))}, \qquad (5)
\end{aligned}
$$

which is the Dirichlet kernel.

There are two main tools for interpreting a kernel in terms of a mapping $Phi(x)$ to a high-dimensional "feature space":

1. Reproducing kernel Hilbert space
2. Mercer's theorem

We will first briefly discuss Mercer's theorem, and will return to these tools in later lectures.

3 Mercer's Theorem

Letting $K(x,y)$ be symmetric, i.e., $(K(x,y) = K(y,x)$, there exists an expansion:

$$K(x,y) = \Phi^T(x)\Phi(y)$$

if and only if:

$$\int K(x,y)g(x)g(y)dxdy \geq 0$$

for all g $\in L^2$.

There is analogy between Mercer's theorem and eigenvector expansions in the discrete case. For the discrete case, recall that if A is a symmetric positive definite matrix then:

$$A = \sum_{i}^{m} \lambda_i x_i x_i^T$$

$$Ax_j = \sum_{i}^{m} \lambda_i x_i x_i^T x_j = \lambda_i x_i$$

$$A = \sum_{i}^{m} (\sqrt{\lambda_i} x_i \sqrt{\lambda_i} x_i^T)$$

Note that $\lambda_i \geq 0$ and x_i are orthogonal vectors.

Consider the integral equation : $f(x) = \int K(x,y)g(y)dy$. $K(x,y)$ plays the role of the matrix A in the discrete case. A function $\Phi(x)$ is an eigenfunction of $K(x,y)$ if:

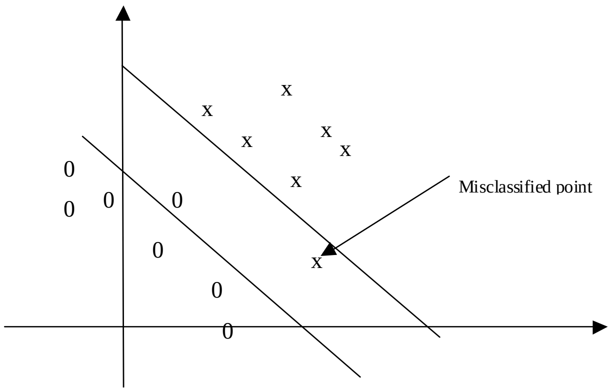
$$\lambda\Phi(x) = \int K(x,y)\Phi(y)dy.$$

We can expand the kernel in terms of its eigenfunctions:

$$K(x,y) = \sum_i \frac{\Phi_i(x)\Phi_i(y)}{\lambda_i},$$

and also expand arbitrary functions in this eigenbasis: $g(x) = \sum_i \alpha_i \Phi_i(x)$.

4 Non-separable Classification



Introduce slack variables:

$$\begin{aligned} y_i(w^T x_i + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0 \end{aligned}$$

for an error to occur in training set, $\xi_i \geq 1 \Rightarrow \sum_i \xi_i$ is the upperbound on number of errors.

Primal problem:

$$\begin{array}{ll} \min & \frac{1}{2} w^T w + \frac{C}{m} \sum_{i=1}^{m} \xi_i \\ s.t. & y_i(x_i^T w + b) - 1 + \xi_i \geq 0 \end{array}$$

(continue in next lecture)