Recitation 7: November 24

Lecturer: Mariano Schain *Scribe: ym*

7.1 Summary of Lecture on Kernels

Recall that the dual program is

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j < x_i, x_j > \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \; \forall i \in [1, m] \\ & \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned}$$

Given a solution α we can recover the primal variables: The weights w are

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \ ,$$

and the bias b is

$$b = y_i - < w, x_i >= y_i - \sum_{j} \alpha_j y_j < x_j, x_i >$$

for any support vector i such that $0 < \alpha_i < C$.

Note that the solution of the optimization problem need only the values $< x_i, x_j >$ and not the actual points x_i. This is also true for the classification rule given a new point x,

$$h(x) = sign(< w, x > +b) = sign(\sum_{i1=}^{m} \alpha_i y_i < x_i, x > +b)$$

where only the inner products $< x_i, x >$ are needed.

The main idea is to replace the inner products by a kernel function. The benefit would be to enrich the hypothesis class (using non-linear kernels) and keep the computational efficiency.

7.2 Proper kernel

Assume the input space is χ. A proper kernel function maps $K : \chi \times \chi \to \mathbb{R}$. The properties of a proper kernel are:

1. Symmetry: $K(a, b) = K(b, a)$ for any $a, b \in \chi$.
2. Positive semi definite $\overline{\overline{K}}$: For any m and any $x_1, \ldots, x_m \in \chi$ we define (keeping m implicit) $\overline{\overline{K}} \triangleq K[x_1, \ldots, x_m]$, that is $\overline{\overline{K}}_{i,j} = K(x_i, x_j)$. We require that for any $c \in \mathbb{R}^m$ we have $c^t \overline{\overline{K}} c \geq 0$. (Equivalently, there exists a matrix P such that $\overline{\overline{K}} = PP^t$, or alternatively, all the eigenvalues of $\overline{\overline{K}}$ are non-negative.)

A proper kernel induces a mapping $\phi : \chi \to H$, where χ is the input space and H is the feature space, such that

$$\forall a, b \in \chi \quad K(a, b) =< \phi(a), \phi(b) >$$

This implies that K implicitly computes inner products in H. Specifically, $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$ and $\overline{\overline{K}} = PP^t$ where $P^t = [\phi(x_1) \cdots \phi(x_m)]$, i.e., $\phi(x_i)$ is the i-th column of P^t. From now on the term kernel is used to refer to a proper kernel.

Claim 7.1 $\forall x, x' \in \chi \;\; K(x, x')^2 \leq K(x, x)K(x', x')$

Proof: For any set of m points the matrix $\overline{\overline{K}}$ has a non-negative determinant (since all its eigenvalues are non-negative). Consider $K[x, x'] = \begin{pmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{pmatrix}$. Then

$$det(K[x, x']) = K(x, x)K(x', x') - K(x, x')K(x, x') = K(x, x)K(x', x') - K^2(x, x') \geq 0$$

□

7.3 Normalized Kernel

Given a proper kernel K, we define a *normalized kernel* K' as follows:

$$K'(x, x') = \begin{cases} 0 & K(x, x) = 0 \quad \text{or} \quad K(x', x') = 0 \\ \frac{K(x,x')}{\sqrt{K(x,x)K(x',x')}} & \text{otherwise} \end{cases}$$

For eaxmple, the kernel defined by

$$K(x, x') = e^{<x,x'>/\sigma^2}$$

when normalized, becomes the Gaussian kernel introduced in class:

$$K'(x,x') = e^{<x,x'>/\sigma^2} e^{-0.5\|x\|^2/\sigma^2} e^{-0.5\|x'\|^2/\sigma^2} = e^{-\|x-x'\|^2/(2\sigma^2)}$$

From Claim 7.1 we have $K(x,x')^2 \leq K(x,x)K(x',x')$, and therefore

$$|K'(x,x')| \leq 1$$

and (for $K(x,x) \neq 0$) we have $K'(x,x) = 1$.

This shows the interpretation of the kernel as similarity function: For a given $x \in \chi$, the maximal simmilarity 1 is achieved by x itself. Furthermore, when we consider the mapping ϕ' induced by a normalized kernel K' we have (using the fact that $K'(x,x) =< \phi'(x), \phi'(x) >= 1$)

$$\begin{aligned}\|\phi'(a) - \phi'(b)\|^2 &= < \phi'(a) - \phi'(b), \phi'(a) - \phi'(b) > \\ &= < \phi'(a), \phi'(a) > -2 < \phi'(a), \phi'(b) > + < \phi'(a), \phi'(a) > \\ &= 2(1 - K'(a,b))\end{aligned}$$

we can see that the distnace in feature space between the mappings of two input space vectors a and b approaches 0 as $K'(a,b)$ approaches 1.

We can show that K' is positive semi-definite by considering $\overline{\overline{K'}}$. We have,

$$c^t \overline{\overline{K'}} c = \sum_{i,j=1}^{m} \frac{c_i, c_j K(x_i, x_j)}{\sqrt{K(x_i,x_i)K(x_j,x_j)}} =< \sum_{i=1}^{m} \frac{c_i \phi'(x_i)}{\sqrt{K(x_i,x_i)}}, \sum_{j=1}^{m} \frac{c_j \phi'(x_j)}{\sqrt{K(x_j,x_j)}} >\geq 0$$

7.4 Closure property of kernels

We will consider two closure operations of proper kernels: $K_1 + K_2$ and $K_1 * K_2$. The first ($K = K_1 + K_2$) can be interpreted as an "OR", requiring that if two input vectors are similar in the feature space induced by K_1 (that is, have high value when K_1 is applied to them) OR the feature space induced by K_2 then they are similar in the feature space induced by K. Similarly, the second ($K = K_1 * K_2$ can be interpreted as an "AND", requiring that if two input vectors are similar in the feature space induced by K_1 AND are similar in the feature space induced by K_2 then they are similar in the feature space induced by K.

We show that $K_1 + K_2$ is a proper kernel by showing that for any $c \in \chi$ we have

$$c^t(\overline{\overline{K_1 + K_2}})c = c^t(\overline{\overline{K}}_1 + \overline{\overline{K}}_2)c = c^t\overline{\overline{K}}_1 c + c^t\overline{\overline{K}}_2 c \geq 0$$

To show that $K = K_1 * K_2$ is propor we consider the matrices $P_1^t = (\phi_1(x_1) \cdots \phi_1(x_m))$ and $P_2^t = (\phi_2(x_1) \cdots \phi_2(x_m))$, such that $\overline{\overline{K}}_1 = P_1 P_1^t$ and $\overline{\overline{K}}_2 = P_2 P_2^t$. Assume that the length

of ϕ_1 is ℓ_1 and the length of ϕ_2 is ℓ_2. The mapping for K is ϕ and define as $\phi(x)$ of length $\ell_1\ell_2$ where the (serialized) $(i,j)^{th}$ entry in $\phi(x)$ is $\phi_1(x)_i\phi_2(x)_j$ (where $\phi(x)_r$ is the r^{th} entry in $\phi(x)$).

By construction, we have that K as defined is a proper kernel ($\overline{\overline{K}} = PP^t$ where $P^t = (\phi(x_1)\cdots\phi(x_m))$. It remains to show that indeed $K(a,b) = K_1(a,b)K_2(a,b)$ for all $a, b \in \chi$.