150AML: Advanced Topics in Machine Learning **Spring 2008**
Course Topic: Computational Learning Theory
Department of Computer Science, Tufts University
Lecture 18 **Apr 1**
Instructor: Roni Khardon **Scribe: Roni**

More Kernels and Their Properties

These notes are slightly edited from previous scribe notes (in Spring 2006) taken by Mashhood Ishaque.

1 Kernels and Kernel Methods

In the previous lecture we introduced the idea of kernels and gave the Boolean kernels and dual perceptron algorithm that works with kernels. Here we introduce some more common kernels and kernel methods.

We say that $k(x, y)$ is a kernel function iff there is a feature map ϕ such that for all x, y,

$$k(x,y) = \phi(\vec{x}) \cdot \phi(\vec{y})$$

Any learning algorithm that only depends on the inner product of examples, and therefore, can be run kernels, is called kernel method.

Nearest Neighbors: We next show that k-nearest neighbor (kNN) is also a kernel method. kNN classifies new example by finding k closest examples in the sample and taking a majority vote on the label. So all we need is to find distances between examples. We have

$$\begin{aligned} ||\phi(\vec{x}) - \phi(\vec{y})||^2 &= \sum (\phi_i(x) - \phi_i(y))^2 \\ &= \sum \phi_i(x)^2 + \sum \phi_i(y)^2 - 2\sum \phi_i(x)\phi_i(y) \\ &= k(x,x) + k(y,y) - 2k(x,y) \end{aligned}$$

So indeed the distance can be calculated using 3 calls to the kernel function.

The polynomial kernel: Let $x, y \in R^n$. Define $k(x, y) = (\langle x, y\rangle + c)^d$ for $c, d \in R$. This corresponds to a feature map ϕ including polynomials of the original variables.

For example if $d = 2$, then:

$k(\vec{x}, \vec{y}) = (\sum x_i y_i + c)^2$
$= (\sum x_i y_i + c)(\sum x_j y_j + c)$
$= \sum_i \sum_j x_i y_i \cdot x_j y_j + 2c \sum_i x_i y_i + c^2$
$= \sum_i \sum_j x_i x_j \cdot y_i y_j + \sum_i \left(\sqrt{2c}\; x_i\right)\left(\sqrt{2c}\; y_i\right) + c^2$

ϕ has n^2 entries from $\sum_i \sum_j$ $\Longrightarrow$ (feature is $x_i x_j$)
$+$ n entries from $\sum_i$ $\Longrightarrow$ (feature is $\sqrt{2c}\; x_i$)
$+$ 1 $\Longrightarrow$ (feature is c)

The Gaussian kernel: is defined as $k(x,y) = e^{-||\vec{x}-\vec{y}||^2/\sigma}$. By using Taylor's expansion $e^a = 1 + a + \ldots + \frac{1}{k!}a^k$ one can see that $e^{\vec{x}\cdot\vec{y}}$ is a kernel with (an infinite set of) features corresponding to polynomial terms. Then we can normalize by σ and divide the corresponding features by $e^{||x||}$ and $e^{||y||}$ to get the Gaussian kernel.

2 Linear Algebra

A quick review was given using slides. See slide copies. The main result we need is as follows:

Any symmetric matrix K with real valued entries can be written in the form $K = PDP^T$ where $P = (\vec{V_1}, \vec{V_2}, ..., \vec{V_m})$, $\vec{V_i}$ are eigen vectors of K that form an orthonormal basis (so we also have $P^T = P^{-1}$) and where D is a diagonal matrix with $D_{i,i} = \lambda_i$ being the corresponding eigen values. A square matrix A is positive semi-definite (PSD) iff for all vectors c we have $c^T A c = \sum_i \sum_j c_i c_j A_{i,j} \geq 0$. It is well known that a matrix is positive semi-definite iff all the eigen values are non-negative.

3 Mercer's Theorem

The sample $S = x^1, x^2, ..., x^m$ includes m examples. The Kernel (Gram) matrix K is an $m \times m$ matrix including inner products between all pairs of examples i.e., $K_{i,j} = k(x^i, x^j)$. K is symmetric since $k(x,y) = k(y,x) = \phi(x)\ .\ \phi(y)$

Mercer's Theorem: A symmetric function $k(.,.)$ is a kernel iff for any finite sample S the kernel matrix for S is positive semi-definite.

One direction of the theorem is easy: if $k()$ is a kernel, and K is the kernel matrix with $K_{i,j} = k(x_j, x_j)$. Then $c^T K c = \sum_i \sum_j c_i c_j K_{i,j} = \sum_i \sum_j c_i c_j \phi(x_i)\phi(x_j) = (\sum_i c_i \phi(x_i))(\sum_j c_j \phi(x_j)) = ||(\sum_j c_j \phi(x_j))||^2 \geq 0$.

For the other direction we will prove a weaker result.

Theorem: Consider a finite input space $X = \{x^1, x^2, ..., x^m\}$ and the kernel matrix K over the entire space. If K is positive semi-definite then $k(.,.)$ is a kernel function.

Proof: By the linear algebra facts above we can write $K = PDP^T$.

Define a feature mapping into a m-dimensional space where the lth bit in feature expansion for example x^i is $\phi_l(x^i) = \sqrt{\lambda_l}\ (\vec{V_l})_i$.

The inner product is

$$\begin{aligned} \phi(\vec{x^i}) \cdot \phi(\vec{y^j}) &= \sum_{l=1}^{m} \phi_l(x^i)\ \phi_l(x^j) \\ &= \sum_{l=1}^{m} \lambda_l (V_l)_i\ (V_l)_j \end{aligned}$$

We want to show that

$$k(x^i, x^j) = \phi(\vec{x^i}) \cdot \phi(\vec{y^j})$$

Consider i, jth entry of the matrix $K = k(x^i, x^j)$. We have the following identities where the last one proves the result.

$$K_{i,j} = [PDP^T]_{i,j}$$

$$= \quad [[PD]P^T]_{i,j}$$

$$\begin{aligned} [PD] &= (\vec{V_1}, \vec{V_2}, ..., \vec{V_m})D \\ [PD]_{i,l} &= (V_l)_i \ \lambda_l \\ [[PD]P^T]_{i,j} &= \sum_{l=1}^{m} (V_l)_i \lambda_l (V_l)_j \end{aligned}$$

Note that Mercer's theorem allows us to work with a kernel function without knowing which feature map it corresponds to or its relevance to the learning problem. This has often been used in practical applications.

4 More Properties of Kernels

Consider any space X of samples and kernels $k_1(.,.)$ and $k_2(.,.)$ over X. Then $k(.,.)$ is a kernel with

(1) $k(x,y) = k_1(x,y) + k_2(x,y)$
(2) $k(x,y) = ak_1(x,y)$ where $a > 0$
(3) $k(x,y) = f(x) \cdot f(y)$ for any function f on x
(4) $k(x,y) = k_1(x,y) \ \cdot \ k_2(x,y)$
(5) $k(x,y) = \frac{k_1(x,y)}{\sqrt{k_1(x,x)}\sqrt{k_1(y,y)}}$

Proof of (1):

$$K = K_1 + K_2$$

where we add matrices component-wise

$$\forall \vec{x}, \ \vec{X}^T \ K \ \vec{X} = \vec{X}^T \ K_1 \ \vec{X} + \vec{X}^T \ K_2 \ \vec{X} \geq 0$$

Another Proof: Let

$$\phi^1(x) = (\phi^1_1(x), ..., \phi^1_{N_1}(x))$$
$$\phi^2(x) = (\phi^2_1(x), ..., \phi^2_{N_2}(x))$$

be the feature map for K_1 and K_2

Define $\phi(x)$ by concatenating the feature maps (or alternate features if the spaces are infinite)

$$\phi(x) = (\phi^1_1(x), ..., \phi^1_{N_1}(x), \phi^2_1(x), ..., \phi^2_{N_2}(x))$$

The mapping clearly satisfies $\phi(x) \cdot \phi(y) = \phi^1(x) \cdot \phi^1(y) + \phi^2(x) \cdot \phi^2(y)$.

Proof of (2):

$$k(x,y) = (\sqrt{a}\phi^1_1(x), ..., \sqrt{a}\phi^1_N(x))(\sqrt{a}\phi^1_1(y), ..., \sqrt{a}\phi^N_1(y) = ak_1(x,y)$$

Proof of (3): there is just one feature defined by $f()$

Proof of (4): multiply out the ϕ expressions for k_1 and k_2 to see that k is a kernels with the space of products of features from ϕ^1 and ϕ^2.

Proof of (5):

Let $\phi^1(x)$ be as above. Define $\phi(x)$ by $\phi_i(x) = \frac{\phi^1_i(x)}{\|\phi^1(x)\|}$. Then $k()$ calculates the inner product for $\phi()$.