Homework 6

Jean-Philippe Vert

Due March 30, 2009

We wish to construct positive definite kernels for finite sets of points in the interval $[0,1]$. Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ be two such sets of length n and m.

1. Show that the following kernel is positive definite for any $\sigma > 0$:

$$K_1(X,Y) = \sum_{x \in X} \sum_{y \in Y} \exp\left(-\frac{(x-y)^2}{2\sigma^2} \right).$$

2. To any finite set X of length n we associate the function $g_X : \mathbb{R} \to \mathbb{R}$ defined by:

$$g_X(t) = \frac{1}{n} \sum_{x \in X} \exp\left(-\frac{(x-t)^2}{2\sigma^2} \right).$$

Show that the following kernel is positive definite for any $\sigma > 0$:

$$K_2(X,Y) = \int_{\mathbb{R}} g_X(t) g_Y(t) dt \,.$$

Is there a simple relation between $K_1(X,Y)$ and $K_2(X,Y)$?

3. Let $\mathcal{P}$ be a partition of $[0,1]$. For any bin $p \in \mathcal{P}$, let $n_p(X)$ be the number of points of X which are in p. Show that the following kernels are positive definite:

$$K_3(X,Y) = \sum_{p \in \mathcal{P}} \min(n_p(X), n_p(Y)) \,,$$

$$K_4(X,Y) = \prod_{p \in \mathcal{P}} \min(n_p(X), n_p(Y)) \,.$$

4. Let T_D be a complete binary tree of depth D, that is, a directed graph such that, starting from the root, each node has two children, until the nodes in the D-th generation which have no children (nodes with no children are called *leaves*). The nodes of T_D are denoted $s \in T_D$. How many nodes are there in T_D?

5. We denote by $S(T_D)$ the set of connected subgraphs of T_D which contain the root and such that all their nodes have either 0 or 2 children. What is the size of $S(T_D)$ for $D = 10$?

6. For $0 < p < 1$, we consider the following rule to generate randomly a tree in $S(T_D)$. We start at the root, and give it two children with probability p, and no child with probability $1-p$. If it has no child, then the process stops and the tree generated is the root only. Otherwise, the same rule is applied independently to both children, which have themselves 0 or 2 children with probability $1-p$ and p. The process is repeated iteratively to all new children, until no more child is generated, or until we reach the D-th generation where nodes have no children with probability 1. For any $T \in S(T_D)$ we denote by $\pi(T)$ the probability of generating T by this process. For any real-valued function h defined over the set of nodes $s \in T_D$, propose a factorization to compute the following sum efficiently:

$$\sum_{T \in S(T_D)} \pi(T) \prod_{s \in leaves(T)} h(s) \,.$$

7. Suppose that each leaf $s \in leaves(T_D)$ is associated to a interval $p(s)$ of $[0,1]$ which together form a partition. For any node $s \in T_D$ we denote by $D(s)$ the set of leaves of T_D which are descendant of s, and we associate to s the subset $p(s) \subset [0,1]$ defined by:

$$p(s) = \bigcup_{l \in D(s)} p(l) \,.$$

For any $T \in S(T_D)$, show that the following function is a positive definite kernel:

$$K_T(X,Y) = \prod_{s \in leaves(T)} \min(n_{p(s)}(X), n_{p(s)}(Y)) \,.$$

8. Show that the following function is a positive definite kernel and propose an efficient implementation to compute it

$$K_5(X,Y) = \sum_{T \in S(T_D)} \pi(T) K_T(X,Y) \,.$$