Foundations of Machine Learning
Department of Computer Science, NYU
Homework assignment 3
Due: May 3, 2005

1. Problem 1: Metrics and Kernels

   Let X be a non-empty set and $L : X \times X \rightarrow \mathbb{R}$ be a negative definite symmetric kernel such that $L(x, x) = 0$ for all $x \in X$.

   (a) Show that there exists a Hilbert space H and a mapping $x \mapsto K_x$ from X to H such that:

   $$L(x, y) = ||K_x - K_y||^2.$$

   Assume that $L(x, y) = 0 \Rightarrow x = y$. Show that $\sqrt{L}$ defines a metric on X.

   (b) Use this result to prove that the kernel $K(x, y) = \exp(-|x - y|^p)$, $x, y \in \mathbb{R}$, is not positive definite for $p > 2$.

   (c) The kernel $K(x, y) = \tanh(a(x \cdot y) + b)$ was shown to be equivalent to a two-layer neural network when combined with support vector machines. Show that K is not positive definite if $a < 0$ or $b < 0$. What can you conclude about the corresponding neural network when $a < 0$ or $b < 0$?

2. Problem 2: Boosting

   This problem studies boosting-type algorithms defined with objective functions different from that of AdaBoost. We assume that the training data is given as m labeled examples $(x_1, y_1), \ldots, (x_m, y_m) \in X \times \{-1, +1\}$. Let Φ be a strictly increasing convex and differentiable function over $\mathbb{R}$ such that: $\forall x \geq 0, \Phi(x) \geq 1$ and $\forall x < 0, \Phi(x) > 0$.

   (a) Consider the loss function $L(\alpha) = \sum_{i=1}^{m} \Phi(-y_i f(x_i))$ where f is a linear combination of base classifiers: $f = \sum_{t=1}^{T} \alpha_t h_t$ as with AdaBoost. The goal is to derive a new boosting algorithm using the objective function L. Characterize the best base classifier h_u to select at each round of boosting if we use coordinate descent.

   (b) Plot the following functions (1) misclassification loss $\Phi_1(-u) = 1_{u \leq 0}$; (2) least squared loss $\Phi_2(-u) = (1 - u)^2$; (3) SVM loss $\Phi_3(-u) = \max\{0, 1 - u\}$; and (4) logistic loss $\Phi_4(-u) = \log(1 + e^{-u})$. Do they satisfy all the hypotheses of the problem?

(c) For each loss function verifying the hypotheses, derive the corresponding boosting algorithm, including the pseudocode. How does the algorithm differ from AdaBoost?