Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 3
Due: March 27, 2007

**SVMs**

1. Download and install the `libsvm` software library from:

   `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

2. Download the `yeast` data set. The data is already divided into ten folds. The classification problem consists of predicting the label of the last column (`CYT` vs. `non-CYT`).

3. Normalize input vectors. Use SVMs combined with polynomial kernels to solve this classification problem. For each value of the polynomial degree, $d = 1, 2, 3, 4$, plot the average error plus or minus one standard deviation as a function of $C$ (let the other parameters of polynomial kernels in `libsvm`, $\gamma$ and $c$, be equal to their default values 1). Report the best value of the trade-off constant $C$ using ten-fold cross validation.

4. Let $(C^*, d^*)$ be the best pair found previously. Fix $C$ to be $C^*$. Plot the ten-fold cross-validation training and test errors for the hypotheses obtained as a function of $d$. Plot the average number of support vectors obtained as a function of $d$.

5. How many of the support vectors lie on the margin hyperplanes? Find a way to compute the *soft* margin $\rho$ and give its value for each $d$.

6. Derive the dual formulation of the SVM problem for a PDS kernel $K$ when the error term is $\frac{C}{2}\sum_{i=1}^{m}\xi_i^2$ instead. Prove that there exists a PDS kernel $K'$ such that this problem becomes equivalent to the dual optimization problem for SVMs in the *separable case*.

7. Show how to obtain the solution for this error term using `libsvm` without writing any new code. Find the best value of $d$ and $C$ using this error term and compare your results with what you found before.

**Kernels**

1. Let $H$ be a Hilbert space with the corresponding dot product $\langle \cdot, \cdot \rangle$. Show that the kernel $K$ defined over $H \times H$ by

$$K(x, y) = 1 - \langle x, y \rangle \tag{1}$$

is negative definite.

2. For any $p > 0$, let $K_p$ be the kernel defined over $\mathbb{R}_+ \times \mathbb{R}_+$ by

$$K_p(x, y) = e^{-(x+y)^p}. \tag{2}$$

Show that $K_p$ is positive definite symmetric (PDS) iff $p \leq 1$ (*hint*: you can use the fact that if $K$ is NDS, then for any $0 < \alpha \leq 1$, $K^\alpha$ is also NDS).

3. To prevent fraud, a credit-card company decides to contact Professor Villebanque and provides him with a random list of several thousand fraudulent and non-fraudulent *events*. Events can be of very different nature, they may be for example transactions of various amounts, changes of address or card-holder information, or a request for a new card.

Professor Villebanque decides to use support vector machines combined with an appropriate kernel to help predict fraudulent events accurately. It is difficult for Professor Villebanque to define relevant features for such a diverse set of events. However, the risk department of the company has created a complicated method to estimate a probability $\Pr[U]$ for any event $U$. Thus, Professor Villebanque decides to make use of that information and comes up with the following kernel defined over all pairs of events $(U, V)$:

$$K(U, V) = \Pr[U \wedge V] - \Pr[U] \Pr[V]. \tag{3}$$

Help Professor Villebanque by showing that his kernel is positive definite symmetric, which guarantees the convergence of the algorithm.