Introduction to Machine Learning	**Fall Semester, 2013/14**

Homework 4: Dec 1st, 2013

Due: Dec 15th (See the submission guidelines in the course web site)

Theory Questions

1. Importance SVM:
   Assume that each training data point (x_i, y_i) is also associated with an 'importance' $v_i \in [0, 1]$.
   (a) Rewrite the primal SVM (in slide 40 of the class handouts) such that the miss-classification (or margin violation) is scaled by sample importance.
   (b) Derive the corresponding dual problem. What role does the sample importance play in the dual?

2. The number of Support Vectors:
   In $\Re^d$ Let e_i^+ be a unit vector with 1 in the i^{th} position: $(0, \dots, 0, 1, 0, \dots, 0)$ and let e_i^- similarly be $(0, \dots, 0, -1, 0, \dots, 0)$. Consider a sample set $S = \cup_{i=1}^{d} \{(e_i^+, 1), (e_i^-, -1)\}$.
   (a) Write and solve explicitly the primal SVM optimization problem for the case $d = 2$. What are the optimal separating w and b? How many Support Vectors (that is, of margin 1)?
   (b) In the general case (arbitrary dimension d), what are the optimal separating w and b? How many Support Vectors?

3. Proper Kernels:
   Given $a > 0$, $K1, K2$ proper kernels, f a real valued function, and p a polynomial with positive coefficients. For each of the functions K below, prove whether it is necessarily a proper kernel or give a counter-example:
   (a) $K(x, z) = K1(x, z) - K2(x, z)$
   (b) $K(x, z) = \frac{K1(x,z)}{K2(x,z)}$
   (c) $K(x, z) = aK1(x, z)$
   (d) $K(x, z) = -aK1(x, z)$
   (e) $K(x, z) = f(x)f(z)$
   (f) $K(x, z) = p(K1(x, z))$

Programming Assignment

For this excercise use the libsvm software library from http://www.csie.ntu.edu.tw/∼cjlin/libsvm/. Download the yeast data set from http://archive.ics.uci.edu/ml/datasets/Yeast. Our classication problem consists of predicting the label of the last column (CYT vs. non-CYT).

1. Normalize input vectors such that all attributes are in $[-1, 1]$. Use SVMs combined with polynomial kernels to solve this classication problem. For each value of the polynomial degree, $d = 1, 2, 3, 4$ plot the average error plus or minus one standard deviation (based on the number of test samples t, the standard deviation of the error estimate r_t is $\sqrt{\frac{r_t(1-r_t)}{t}}$) as a function of C (let the other parameters of polynomial kernels in libsvm be equal to their default values 1). Report the best value of the trade-off constant C using ten-fold cross validation. (Hint: use a logarithmic scale for C, that is, find the best $\log C$).

2. 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. How many of the support vectors lie on the margin hyperplanes? Compute the soft margin and give its value for each d.