COM S 474: Homework 3 Spring 2017

Write your name on each page. Maximum score is 30 points, due date is **Friday, March 24, 2017** . Please hand in the solutions (CLEAN version) on Friday, March 24, 2017 in class (hard copy). **Staple all the pages together**. NO credit will be given if work is not shown!! A good advice: if you don't know the answer, make an appointment with the TA or instructor. Do not simply google and copy/paste the answer. That is not a good practice and will not help you during the exam!

1. Consider the following data points with corresponding class labels (see Table 1).

Observation	X_1	X_2	Class
1	2	2	+1
2	2	-2	+1
3	-2	-2	+1
4	-2	2	+1
5	1	1	-1
6	1	-1	-1
7	-1	-1	-1
8	-1	1	-1

Table 1: Data points with corresponding classlabels

Consider the following mapping function φ

$$\varphi\begin{pmatrix}X_1\\X_2\end{pmatrix} = \begin{cases} \begin{pmatrix}4 - X_2 + |X_1 - X_2|\\4 - X_1 + |X_1 - X_2|\end{pmatrix} & \text{if } \sqrt{X_1^2 + X_2^2} > 2 \\ \begin{pmatrix}X_1\\X_2\end{pmatrix} & \text{otherwise.} \end{cases}$$

(a) [8 points] Find the equation of the hyperplane (in terms of w) WITHOUT solving a quadratic programming (QP) problem. Make a sketch of the problem (i.e., plot the data, unique hyperplane and corresponding dashed lines. Look at the notes on SVM posted on BlackBoard).

(b) [1 point] Calculate the margin.

(c) [2 points] Find the α's of the SVM for classification (again WITHOUT solving a QP problem).

(d) [4 points] Perform SVM with Matlab or R with RBF kernel (hyper parameter included) and plot the final result? Is this result different from the one you obtained by hand?

2. [5 points] Verify by means of simulation that each bootstrap sample (replicate) will contain $1 - 1/e \approx 63.2\%$ of the original sample. Consider the following experiment: generate an increasing number of standard normal random variables i.e., from $n = 1$ to $n = 10000$. For each of the generated data sets draw one bootstrap sample of the same size. Calculate, for each bootstrap sample, what fraction of the original sample is contained in the bootstrap sample. Make a plot of sample size vs. fraction of original data set. Also provide your code.

3. [10 points] Compare SVM and LS-SVM (and decide on the kernel function) on the ozon level detection data set on Blackboard (see `https://archive.ics.uci.edu/ml/datasets/Ozone+Level+Detection`). This data set contains 72 measurement variables and was measured between 1/1/1998 and 12/31/2004. All missing values have been removed. The goal is to detect detect whether there was too much ozon (class label 0) or a normal day (class label 1). The class label is the last variable. Set up the simulation and clearly describe what you are doing and why. Finally, state, according to your findings (boxplots, ROC curves, etc.), the best classifier for this problem.