COM S 573: Home work 5 Spring 2014

Write your name on each page. Maximum score is 30 points, due date is **Friday, April 25, 2014** . Please hand in the solutions (CLEAN version) on the due date in class (**hard copy**). Also paste the results of your R code and the code itself into your homework. Make sure your homework is stapled! **This home work is NOT mandatory. But if you decide to do it, I will remove your worst home work grade! So the total number of home works that will determine your final grade remains 4. Nothing changes for the students not doing this home work.**

1. Suppose you have the following data

Observation	X_1	X_2	Class
1	3	1	+1
2	3	-1	+1
3	6	1	+1
4	6	-1	+1
5	1	0	-1
6	0	1	-1
7	0	-1	-1
8	-1	0	-1

   (a) [6 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).

   (b) [1 point] Calculate the margin.

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

2. In this problem (based on Problem 7 on p. 371), you will use support vector classification in order to predict whether a given car gets high or low gas mileage based on the *Auto* data set. You can download the data set at `http://www-bcf.usc.edu/~gareth/ISL/data.html`. It can be helpful to do the lab starting on p. 359 of the textbook first.

   (a) [1 point] Create a class variable that takes on a "1" for cars with gas mileage above the median, and a "-1" for cars with gas mileage below the median.

   (b) [5 points] Fit a support vector classifier with linear kernel to the data with various values of cost (i.e., the parameter C from class), in order to predict whether a car gets high or low gas mileage. Report the cross-validation errors associated with different values of this parameter. Comment on your results.

(c) [5 points] Now repeat (b), this time using SVMs with radial basis and polynomial kernels, with different values of gamma, degree and cost. Comment on your results.

In class, I said the gaussian kernel had the following form: $K(X_i, X_j) = \exp(-\|X_i - X_j\|_2^2/h^2)$. *However, in this software package* `e1071`*, the radial basis function has the form* $K(X_i, X_j) = \exp(-\gamma\|X_i - X_j\|_2^2)$. *Consequently,* $\gamma = 1/h^2$.

(d) [1 point] Which kernel function do you prefer and why?

3. (a) [6 points] Make an R program that solves the LS-SVM for nonlinear regression given the parameters γ and bandwidth h for the following gaussian kernel (do not implemenent the cross-validation procedure)

$$K(X_i, X_j) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|X_i - X_j\|_2^2}{2h^2}\right).$$

(b) [3 points] Try your program on the following function (i.e., experiment with some values of γ and h)

```
X <- seq(0, 1, length.out = 200)
y <- (sin(2*pi*(x-0.5)))^2 + rnorm(200, 0, 0.2)
```

Describe the effect when the two tuning parameters γ and h change. Finally, try the values $\gamma = 9.4365$ and $h = 0.16006$. Does this seems a good fit to you?