MACHINE LEARNING COMS 4771, HOMEWORK 2
Assigned October 1, 2015. Due October 15, 2015 before 10am.

Please submit separate files for a) write-up, b) Matlab source files and c) figures (if you choose to include them separately from the writeup). Do not include any other files. Your write-up should be in ASCII plain text format (.txt) or Postscript (.ps) or the Adobe Portable Document Format (.pdf). Please do not submit Microsoft Office documents, LaTeX source code, or something more exotic since we will not be able to read it. LaTeX is preferred and highly recommended, but it is not mandatory. You can use any document editing software you wish, as long as the final product is in .ps or .pdf. Even if you do not use LaTeX to prepare your document, you can use LaTeX notation to mark up complicated mathematical expressions, for example, in comments in your Matlab code. See the Tutorials page for more information on LaTeX. All your code should be written in Matlab. Please submit all your source files, each function in a separate file. Clearly denote what each function does, which problem it belongs to, and what the inputs and the outputs are. Do not resubmit code or data provided to you. Do not submit code written by others. Identical submissions will be detected and both parties will get zero credit. In general, shorter code is better. Sample code is available on the Tutorials web page. Datasets are available from the Handouts web page. You may include figures directly in your write-up or include them separately as .jpg, .gif, .ps or .eps files, and refer to them by filename.

Submit your work via courseworks.columbia.edu.

Problem 1 (15 points)

VC Dimension

(A) Consider a perceptron in $\mathbb{R}^d$. How many points can it shatter or more specifically what is the VC dimension of this perceptron? Explain your answer.

(B) Prove that $VC(H) \leq \log_2 |H|$ where H is a hypothesis space. (A hypothesis on a set of n points, defines which of two classes can each point belong to. A hypothesis space is a family of all possible hypotheses).

(C) Consider the instance space $X = \{1, 2, ..., 999\}$. Let $\mathcal{C}$ be a concept class consisting of 10 concepts, c_0 through c_9. A number $n \in X$ is an element of c_i if the normal decimal representation of n contains the digit i. So, for example, the number ”778” is an element of c_7 and c_8. What is the VC dimension of $\mathcal{C}$? Justify your answer.
(Hint: To prove the VC dimension of $\mathcal{C}$ is k, you should justify the following claims. 1. There exist k elements in X that $\mathcal{C}$ can shatter. 2. No $k+1$ or more elements in X that $\mathcal{C}$ can shatter.)

Problem 2 (10 points)

Kernels

(A) Consider any Mercer kernel defined by $k(x, \tilde{x}) = \boldsymbol{\phi}(x)^\top \boldsymbol{\phi}(\tilde{x})$. We are given a sample $S = \{x_1, x_2, ..., x_n\}$ of n inputs. We can form the Kernel (Gram) matrix $\mathbf{K}$ as an $n \times n$ matrix of kernel evaluations between all pairs of examples i.e., $\mathbf{K}_{i,j} = k(x_i, x_j)$. **Mercer's Theorem**

states that a symmetric function $k(.,.)$ is a kernel iff for any finite sample S the kernel matrix $\mathbf{K}$ is positive semi-definite. Recall that a matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ is positive semi-definite iff $\mathbf{c}^\top \mathbf{K}\mathbf{c} \geq 0$ for all real-valued vectors $\mathbf{c} \in \mathbb{R}^n$. Prove Mercer's theorem in one direction: for any Mercer kernel $k(.,.)$ and finite sample S, the kernel matrix $\mathbf{K}$ is positive semi-definite.

Given any two Mercer kernels $k_1(.,.)$ and $k_2(.,.)$, prove that the following are also Mercer kernels:
a) $k(x,\tilde{x}) = \alpha k_1(x,\tilde{x}) + \beta k_2(x,\tilde{x})$ for $\alpha, \beta \geq 0$
b) $k(x,\tilde{x}) = k_1(x,\tilde{x}) \times k_2(x,\tilde{x})$
c) $k(x,\tilde{x}) = f(k_1(x,\tilde{x}))$ where f is any polynomial with positive coefficients
d) $k(x,\tilde{x}) = \exp(k_1(x,\tilde{x}))$

(B) For the kernel $K(x,y) = \exp(-\frac{1}{2} \parallel x - y \parallel^2) = \varphi(x) \cdot \varphi(y)$, write an explicit formula for φ.

Problem 3 (10 points)

In this problem, you will derive the kernel regression $\hat{y} = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$.

Recall the linear regression: $\hat{y} = \mathbf{w}^\top \mathbf{x}$ with regularized cost function:

$$L(\mathbf{w}) = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda \mathbf{w}^\top \mathbf{w}$$

We introduce new variables, $\xi_i = \mathbf{w}^\top \mathbf{x}_i - y_i$, and rewrite the objective function as following.

$$\min_{\xi,\mathbf{w}} \sum_{i=1}^{n} \xi_i^2 + \lambda \mathbf{w}^\top \mathbf{w}$$

$$\text{subject to: } \mathbf{w}^\top \mathbf{x}_i - y_i - \xi_i = 0, \quad i = 1, ..., n$$

Similar derivation of SVM you learned in class, you can derive the solution to the kernel regression in the following two steps.

a) Derive the dual using Lagrangian, and compute the solution to the dual.

b) Derive the solution to the kernel regression using the result of part a.

Problem 4 (15 points)

You will build an SVM to classify data and use cross-validation to find the best SVM kernel and regularization value. Try different polynomials and RBF kernels (varying polynomial order from 1 to 5) and varying sigma in the RBF kernel. Also, try different values of C in the SVM. First, extract the support vector machine Matlab code from Steve Gunn's software package here:

```
http://www.isis.ecs.soton.ac.uk/resources/svminfo/
In svc.m replace [alpha lambda how] = qp(...);
with             [alpha lambda how] = quadprog(H,c,[],[],A,b,vlb,vub,x0);
```

Clearly denote the various components and the function calls or scripts that execute your Matlab functions. Note, to save the current figure in Matlab as a postscript file you can type:

```
print -depsc filename.eps
```

To test your SVM, you will build a simple object recognition system. Download the data file hw2-2015-dataset.mat from the Handouts web page. This loads a matrix X of size 100 x 4 and a vector Y of size 100 x 1. The X matrix consists of 100 data points of dimensionality 4 that belong to one of two classes. Their actual labels are given in the Y vector. Train your SVM on half of the examples and test on the other half (or other random subsets of the examples if you see fit). Show performance of the SVM for linear, polynomial and RBF kernels with different settings of the polynomial order, sigma and C value. Try to hunt for a good setting of these parameters to obtain high recognition accuracy.