# CSci 5521–002 Homework 0 Due Sun 9/16/2018

This homework is intended for you to test your preparation for this class. We will record your answers and mark that you have handed this in, but it will not be fully graded. It will also test that you have have figured out the software environment needed for this class:

- you have installed Matlab and can write simple programs therein;
- you can access the Canvas environment and submit assignments.

**Problem 1:** Linear regression learns a linear function of feature variables $\mathbf{X}$ to fit the responses $y$. In this problem, you will derive the closed-form solution for linear regression formulations.

1. The standard linear regression can be formulated as solving a least square problem

$$\underset{\mathbf{w}}{\text{minimize}} \quad \phi(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = \langle \mathbf{X}\mathbf{w} - \mathbf{y}, \mathbf{X}\mathbf{w} - \mathbf{y} \rangle$$

   where $\mathbf{X} \in \mathbb{R}^{n \times m}$ $(n \geq m)$ represents the feature matrix, $\mathbf{y} \in \mathbb{R}^{n \times 1}$ represents the response vector and $\mathbf{w} \in \mathbb{R}^{m \times 1}$ is the vector variable of the linear coefficients. Here the $i-j$-th element of $\mathbf{X}$, denoted $x_{ij}$, is the $j$-th attribute value for the $i$-th data sample (observation) and $y_i$ is the true response for the $i$-th data sample. This is a convex objective function of $w$. Derive the optimal $w$ by setting the gradient of the function wrt $w$ to zero to minimize the objective function. To find the gradient, you can use the following formula

$$\begin{aligned} \phi(\mathbf{w} + \boldsymbol{\delta}) &= [\mathbf{X}(\mathbf{w} + \boldsymbol{\delta})]^T \mathbf{X}(\mathbf{w} + \boldsymbol{\delta}) - 2 [\mathbf{X}(\mathbf{w} + \boldsymbol{\delta})]^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \\ &= \phi(\mathbf{w}) + 2 [\mathbf{X}\boldsymbol{\delta}]^T [\mathbf{X}\mathbf{w} - \mathbf{y}] + (\mathbf{X}\boldsymbol{\delta})^T \mathbf{X}\boldsymbol{\delta}, \end{aligned}$$

   and note that $\mathbf{w}$ must be determined so that $\phi(\mathbf{w} + \boldsymbol{\delta}) \geq \phi(\mathbf{w})$ for any possible vector $\boldsymbol{\delta}$ (why?). Here $\mathbf{X}^T$ denotes the transpose of $\mathbf{X}$.

2. In practice, a L2-norm regularizer is often introduced with the least squares, called Ridge Regression, to overcome ill-posed problems where the hessian matrix is not positive definite. The objective function of ridge regression is defined as

$$\underset{\mathbf{w}}{\text{minimize}} \qquad \tilde{\phi}(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2 = \left\| \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda} I \end{pmatrix} \mathbf{w} - \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} \right\|^2$$

   where $\lambda > 0$ and $I$ is an $m \times m$ identity matrix. This objective function is strictly convex. Derive the solution of the ridge regression problem to find the optimal $\mathbf{w}$.

**Problem 2:** Consider a coin with probability of heads equal to $\Pr(\text{H}) = p$ and probability of tails $\Pr(\text{T}) = 1 - p$. You toss it 5 times and get outcomes H,H,T,T,H.

1. What is the probability of observing the sequence H,H,T,T,H in five tosses. Also give the formula for the natural logarithm of this probability. Your formulas should be a function of $p$.

2. You have a box containing exactly 2 coins, one fair with $p = 1/2$ and one biased with $p = 2/3$. You choose one of these two coins at random with equal probability, toss it 5 times and get the outcome H,H,T,T,H.

   (a) Give the *joint* probability that the coin chosen was the fair coin ($p = 1/2$) *and* the outcome was H,H,T,T,H.

   (b) Give the *joint* probability that the coin chosen was the biased coin ($p = 2/3$) *and* the outcome was H,H,T,T,H.

3. What should the bias $p = \Pr(\text{H})$ be to maximize the probability of observing H,H,T,T,H, and what is the corresponding probability of observing H,H,T,T,H (i.e., what is the maximum likelihood estimate for $p$), assuming $p$ were unknown? Show the derivation. Hint: maximize the *log* of the function.

**Problem 3:** Below is the pseudo-code of perceptron algorithm for binary classification, where $(\mathbf{x}^t, r^t)$ is the $t$-th data sample: $\mathbf{x}^t$ is the vector of attribute values (real numbers) and $r^t = \pm 1$ is the class label for the $t$-th sample:

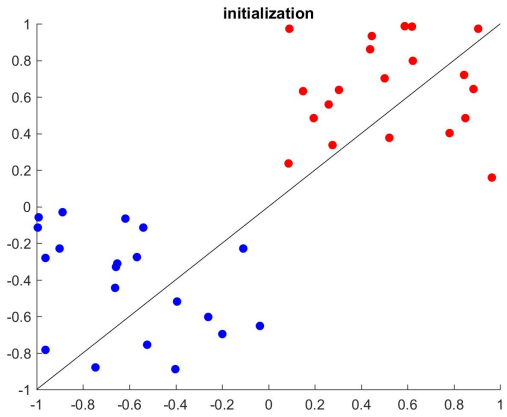
1. $\mathbf{w} = \mathbf{w}_0$.
2. **Do** Iterate until convergence
3. **For** each sample $(\mathbf{x}^t, r^t)$, $t = 1, 2, \cdots$
4. **If** $r^t \langle \mathbf{w}, \mathbf{x}^t \rangle \leq 0$
5. $\mathbf{w} = \mathbf{w} + r^t \mathbf{x}^t$

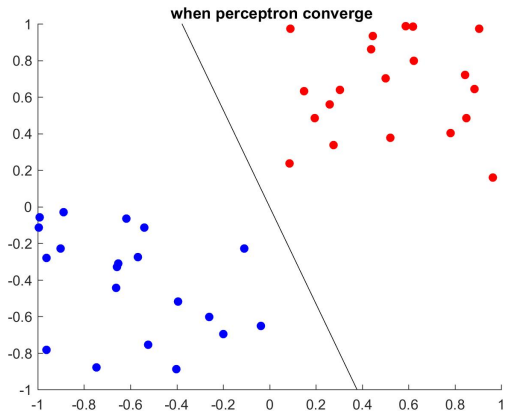
Here "convergence" means $\mathbf{w}$ does not change at all over one pass through the entire training dataset in the loop starting in step 3.

1. Implement the perceptron algorithm and test it on the provided data. To begin, do "`load data1.mat`" to load the file the data file into MATLAB. $\mathbf{X} \in \mathbb{R}^{40 \times 2}$ is the feature matrix of 40 samples in 2 dimensions and $\mathbf{r} \in \mathbb{R}^{40 \times 1}$ is the label vector ($\pm 1$). Use initial value $\mathbf{w}_0 = [1; -1]^T$. Now, run your perceptron algorithm on the given data. How many iterations does it take to converge?

2. Visualize all the samples (use 2 different colors for the 2 different classes) and plot the decision boundary defined by the initial $\mathbf{w}_0$. Plot the decision boundary defined by the $\mathbf{w}$ returned by the perceptron program.
   Hint: To visualize the samples you could use the MATLAB function call

```
scatter(X(:,1), X(:,2), 50, y, '*');
```




Type `help scatter` for more information. Plotting the boundary is equivalent to plotting the line $\mathbf{w}^T\mathbf{x} = w_1x_1 + w_2x_2 = 0$. Since all the sample points are located within the square $\{(x_1, x_2), -1 \leq x_1, x_2 \leq +1\}$, choose two points $(a, b)$ and $(c, d)$ by setting $a = -1, c = +1$ and solving for $b, d$, or else set $b = -1, d = +1$ and solving for $a, c$, and then draw the line between the two points $(a, b)$ and $(c, d)$ with the command

```
hold on; plot([a,c],[b,d]); hold off;
```

Use the `hold` function to add the line to the existing scatter plot, and `axis` to adjust the axes, if needed. Draw both the initial boundary and the final boundary on the same plot.

# Submission

- **Things to submit:**

  1. `hw0_sol.pdf`: a document contains all the derivations of Problem 1 & 2 and the plot asked by Problem 3.
  2. `MyPerceptron.m`: a MATLAB function defined with header `function [w, step]=MyPerceptro y, w0)`, where X is the feature matrix, y is a label vector ($\pm 1$) and w0 is the initial value for the parameter vector w. In the output, w is the parameter found by perceptron and step represents the number of steps the algorithm takes to converge. The function should also display the plot of samples and boundary.
  3. Zip both files into a single zipped file and name it as your *`lastname`*`.zip`.

- **Submit**: All material must be submitted electronically via Canvas. This homework will not be graded but required as a proof of satisfying the prerequisites for taking the class.