CS 6140: Machine Learning	**Spring 2015**
College of Computer and Information Science	
Northeastern University	
Assignment 3	**February, 24**
Instructor: Bilal Ahmed	**Due Date: Mar, 02**

1 Ridge Regression

Consider the Ridge regression problem:

$$\hat{\mathbf{w}}_{ridge} = \arg\min_{\mathrm{w}} \left\{ \sum_{i=1}^{N} (y_i - w_0 - \sum_{j=1}^{m} w_j x_{ij})^2 + \lambda \sum_{j=1}^{m} w_j^2 \right\} \;\; ; \;\lambda \geq 0 \tag{1}$$

if we are working with centered data, then the above problem can be rewritten as:

$$\hat{\mathbf{w}}^c_{ridge} = \arg\min_{\mathrm{w}^c} \left\{ \sum_{i=1}^{N} (y_i - w_0^c - \sum_{j=1}^{m} w_j^c (x_{ij} - \bar{x}_j))^2 + \lambda \sum_{j=1}^{m} (w_j^c)^2 \right\} \;\; ; \;\lambda \geq 0 \tag{2}$$

1. Can you give the correspondence between w_i and w_i^c for all $i \in \{0, 1, \ldots, m\}$?
2. Show that the optimal value for w_0^c for centered data is the mean of the output values i.e., $\bar{y} = \sum_{i=1}^{N} y_i$.
3. Why can't we have a negative λ ?

2 Programming Ridge Regression

In this question you will implement ridge regression using the formulation shown in Equation **??**. To this end remember to center your data and do not include the constant feature in your design matrix. We will be working with the Sinusoid dataset from Assignment-2 (*do not use the validation set*). Estimate the training and test RMSE using ten-fold cross validation for the following settings:

1. Add four new feature to the dataset consisting of higher powers of $x^p,\ p \in \{1, 2, \ldots, 5\}$. Use 51 values of λ in $[0, 10]$, so that $\lambda \in \{0, 0.2, 0.4, \ldots, 10\}$. Plot train and test RMSE vs *lambda* on separate plots.
2. Add nine new feature to the dataset consisting of higher powers of $x^p,\ p \in \{1, 2, \ldots, 9\}$. Use 51 values of λ in $[0, 10]$, so that $\lambda \in \{0, 0.2, 0.4, \ldots, 10\}$. Plot train and test RMSE vs *lambda* on separate plots.

NOTE: The quantity that you will be plotting is the average RMSE across the ten folds for the test/train sets, for each value of λ.

2.1 Interpretation:

Do you see a difference in the behavior of regularization of the two synthetic datasets we created in the two problems? Explain your results in detail, keeping in mind that a zero value for λ corresponds to the solution without regularization i.e., linear regression.

3 Maximum Likelihood For Univariate Normal

Consider N samples $\{x_1, x_2 \dots, x_N\}$, generated from a univariate normal distribution:

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{3}$$

formulate the log-likelihood for the N samples, and derive the maximum likelihood estimate for the mean of the distribution μ_{ML}.

3.1 Extra-credit:

Derive the maximum likelihood estimate for the standard deviation σ_{ML}.