CS534 — Implementation Assignment 1 — Due 11:59PM Oct 10th, 2014

General instructions.

1. The following languages are acceptable: Java, C/C++, Matlab, Python and R.
2. You can work in team of up to 3 people. Each team will only need to submit one copy of the source code and report.
3. Your source code and report will be submitted through the TEACH site

`https://secure.engr.oregonstate.edu:8000/teach.php?type=want_auth`

Please clearly indicate your team members' information.
4. Be sure to answer all the questions in your report. You will be graded based on your code as well as the report. In particular, **the clarity and quality of the report will be worth 10 pts**. So please write your report in clear and concise manner. Clearly label your figures, legends, and tables.
5. In your report, the results should always be accompanied by discussions of the results. Do the results follow your expectation? Any surprises? What kind of explanation can you provide?

Linear regression with L_2 regularization (total points: 60 + 10 pts)

For the first part of the assignment, you need to implement linear regression with L_2 (quadratic) regularization, which learns from a set of N training examples $\{\mathbf{x}_i, y_i\}_{i=1}^N$ an weight vector $\mathbf{w}$ that optimize the following regularized Sum of Squared Error (SSE) objective:

$$\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda\|\mathbf{w}\|^2 \tag{1}$$

To optimize this objective, you need to implement the gradient descent algorithm. Because some features have very large values, you may wish to normalize the features to have zero mean and unit variance or to the range between zero and one. This will allow the gradient descent procedure to converge better.

Data: you are provided with a training set and a test set, both in csv format. Each row of the data file contains the data for one example. The first column contains the dummy feature taking the constant value of 1 for all examples and the last column stores the target y values for each example. Here is what you need to do and report:

1. Explore different learning rate for gradient descent (10 pts). Describe your strategy in determining the learning rate. What learning rate or learning rates did you observe to be good for this particular dataset? Report your observations.

2. Experiments with different λ values (20 pts): You will test the effect of the regularizer on your linear regressor. It is often the case that we don't really what the right λ value should be and we will need to consider a range of different λ values. For this project, consider at least the following values for λ: $0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100$. Feel free to explore the choices of λ in a more finer search resolution. Report the SSE (the first term in the Eq. 1) on the training data and the test data respectively for each value of λ. Your discussion of the results should clearly answer the following questions:

(a) What trend do you observe from the training SSE as we change λ value?

(b) What tread do you observe from the test SSE?

(c) Provide an explanation for the observed behaviors.

3. Selecting the best λ with 10-fold cross-validation. (30 pts). Cross-validation is a commonly used technique for selecting hyperparameters (e.g., λ). In k-fold cross validation, you split the training data into k parts. To evaluate a candidate parameter value λ, we repeat the following for k times. Each time, we train on $k-1$ parts and test on the remaining one part (this single part is called the validation set) and measure its SSE. After k times, each of the ten parts would have served as the validation set exactly once and we then sum up the k SSEs as the cross-validated SSE for the λ value. Among all candidate λ values, we select the

one that has the lowest cross-validated SSE. Implement the k-fold cross-validation functionality for selecting λ's. Apply 10-fold cross-validation to the training data and report the cross-validated SSEs for different λ values. Answer the following questions:

(a) Which λ would you choose based on the cross-validation results?

(b) Is the cross-validated SSE larger or smaller than the training SSE? Can you provide an explanation for that?

(c) What trend do you observe from the cross-validated SSEs as we change λ? Which one do you think this trend resemble more, the trend of training SSE or the trend of test SSE? Why do you think this is the case?