ELL409: Assignment 3

Maximum points: 6

Demo Schedule: 20th April 2016 (tentative)

10 April 2016

1. **Polynomial Curve Fitting** (3 points)

   Polynomial curve fitting is an example of regression. Here you will apply the concepts of linear regression for polynomial curve fitting. In regression, the objective is to learn a function that maps an input variable x to a continuous target variable y.
   For this part, you will be provided a personalised input file that contains data of the form (x_i, y_i) for $i = 1, \ldots, 100$. The relationship between x and y is of the form:

   $$y = w_0 + w_1 x + \cdots + w_M x^M + \epsilon$$

   where the noise ϵ is drawn from a Gaussian distribution with zero mean and unknown (but fixed, for a given input file) variance. M is also unknown. You can download your input data file from `http://privateweb.iitd.ac.in/~seshan/a3/<groupno>` (for e.g., `http://privateweb.iitd.ac.in/~seshan/a3/group01`) where `groupno` is your group number as listed here: `http://web.iitd.ac.in/~seshan/ell409_assignment_groups.html`. The goal is to identify the underlying polynomial (both the degree and the coefficients), as well as obtain an estimate of the noise variance.
   Specifically, the following tasks are to be accomplished:

   - To begin with, use only the first 20 data points in your file. Solve the polynomial curve fitting regression problem using error function minimisation. Define your own error function other than the sum-of-squares error. Try different error formulations and report the results.
   - Use a goodness-of-fit measure for polynomials of different order. Can you distinguish overfitting, underfitting, and the best fit?
   - Obtain an estimate for the noise variance.
   - Introduce regularisation and observe the changes. For quadratic regularisation, can you obtain an estimate of the optimal value for the regularisation parameter λ? What is your corresponding best guess for the underlying polynomial? And the noise variance?
   - Now repeat all of the above using the full data set of 100 data points. How are your results affected by adding more data? Comment on the differences.
   - What is your final estimate of the underlying polynomial? Why?

   You will be required to give a demonstration of regression, the coefficients you have obtained, and how you have done so. In addition, present visualisations of the data and results in meaningful ways.

2. **Genomic Sequence Analysis** (3 points)

   Non-coding ribonucleic acids (ncRNA) are believed to have many roles in a cell, many of which remain to be discovered. However, it is difficult to detect ncRNAs using biochemical screening methods. Recent studies have shown that computational methods can accuractely detect ncRNAs, which can be treated as supervised classification. To perform the classification, an 8-dimensional feature vector is used as input to a classifier, including the length of genomic sequence and nucleotide frequencies:

   - A feature value computed by the Dynalign algorithm[1]
   - Length of shorter sequence

[1] `http://www.ncbi.nlm.nih.gov/pubmed/11902836`

- 'A' frequencies of sequence 1
- 'U' frequencies of sequence 1
- 'C' frequencies of sequence 1
- 'A' frequencies of sequence 2
- 'U' frequencies of sequence 2
- 'C' frequencies of sequence 2

Here you will train a support vector machine (SVM) classifier to determine if a genomic sequence is an ncRNA. You can use LIBSVM - a popular open-source SVM toolbox that has been implemented in many programing languages such as C/C++, JAVA, and MATLAB - available from this webpage: `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

The training and test data sets are available here: `http://privateweb.iitd.ac.in/~seshan/a3/ncrna_data`. In each file, the data is organised as:
<label> <index1>:<value1> <index2>:<value2> ...
where each line contains a training/test example. <label> is a bipolar value (1 or -1) indicating the class label (1 indicating an ncRNA, the positive class). (Note: Test data file has all the class labels set to 0, you need to predict these class labels using your trained classifier). <index> is an integer in the range $[1, 8]$ corresponding to the 8 features listed above and <value> is a real number corresponding to a feature value which has been scaled to $[0, 1]$. If an index is omitted, it implies that the corresponding value is zero.

- Classification using linear SVM: Spilt the training data set to form validation and training data sets. Train a set of linear SVMs with different values of the regularisation parameter C using the training data set. For each value of C, train an SVM and use each trained SVM model to classify the validation data set. Plot the classification accuracy as a function of the parameter C.
- Classification using Gaussian (RBF) kernel SVM: $k(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$
  Use 5-fold cross validation to choose the best C and σ. To do so, first randomly choose 50% of the training set as the cross validation set. Next, divide the cross validation set into 5 subsets of equal size. Each subset is in turn used to validate the classifier trained on the remaining 4 subsets. So you will have 5 trained SVMs and 5 validation subsets. The cross validation accuracy is average accuracy over the 5 validation subsets. (Do not use the built-in cross validation option in LIBSVM). For both C and σ try a number of different values and be sure to try all possible of pairs of values for C and σ. Show a matrix of your cross validation results, where the entry (i, j) of the matrix corresponds to the classification accuracy on the cross validation set with i^{th} value of C and j^{th} value of σ.
  Next, use the entire training set to train an SVM classifier with the best C and σ values determined via the cross validation procedure outlined above. Finally, use the trained SVM model to classify the test data set and write the results to a file using the same format as the training data set.

You should submit the tests results together with your implementation code, and a brief report summarising all your results and your interpretation of them.