CSL407 Machine Learning
Homework 2

Due on 29/8/2014, 11.59pm

Instructions: Upload to your moodle account one zip file containing the following. Please do not submit hardcopy of your solutions. In case moodle is not accessible email the zip file to the instructor at ckn@iitrpr.ac.in. Late submission is not allowed without prior approval of the instructor. You are expected to follow the honor code of the course while doing this homework.

1. A neatly formatted PDF document with your answers for each of the questions in the homework. You can use latex, MS word or any other software to create the PDF.
2. Include a separate folder named as 'code' containing the scripts for the homework along with the necessary data files. Name the scripts using the problem number.
3. Include a README file explaining how to execute the scripts.
4. Name the ZIP file using the following convention rollnumber_hwnumber.zip

---

In this homework you will be implementing linear discriminants and logistic regression. You will also solve few other problems that will help to improve your understanding of these topics.

1. (6 points)You will experiment with 1-dimensional data that facilitates intuitive visualization of linear discriminant analysis. Generate data from three classes, each from a Gaussian (Normal) distribution with means 2, 3, and 4, respectively, and standard deviation of 0.2. Generate 10 samples from each class. Create test samples from [1,5] using a step size of 0.1 starting from 1. Plot two figures containing the following. In the first figure plot
   (a) color coded training data (use different colors for each class)
   (b) the three curves for corresponding to the likelihood $P(X = x|Y = k)$ for $k = 1, 2$ and 3. Use the mean and standard deviation estimated from the training dataset for each class for computing the likelihood.
   (c) the three curves corresponding to the posterior probability $P(Y = k|X = x)$ for $k = 1, 2$ and 3.

   In the second figure plot
   (a) color coded training data
   (b) the three discriminant functions for the test data
   (c) the class predicted by the linear discriminants for the test dataset.
2. (5 points) Logistic regression is also prone to overfitting similar to linear regression. One approach to prevent overfitting is to use a regularization term. Write the objective function that includes a regularization term of $||w||^2$. Derive the weight update equation for the regularized objective function for gradient descent approach.

3. (25 points)In this exercise you will experiment with regularized logistic regression and linear discriminants to predict whether credit card can be issued to an individual. As the research manager of the bank you have characterized each individual using two attributes x_1 and x_2. From these attributes, you would like to determine whether the credit card application of an individual should be accepted or rejected. To learn the models, you have a dataset of past credit card applications made by individuals and their outcomes. This is available as credit.mat or credit.txt in the zip file.

   - Load the dataset into the workspace. Plot the dataset using different colors for the two classes. The Matlab function scatter will be useful in this context.
   - Clearly the data is not linearly separable. Logistic regression models only linear decision boundaries and therefore will not perform well on this dataset. One way to fit data better is to create more features for each data point. Implement the function `featuretransform(X, degree)` that takes the data and highest degree of polynomial terms of the input attributes x_1 and x_2 to create higher order polynomials of the input attributes. For example, if degree = 4, then the transformed data point will contain 15 attributes

$$\begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1x_2 \\ x_2^2 \\ . \\ . \\ . \\ x_1x_2^3 \\ x_2^4 \end{bmatrix}$$

     We hope that this new feature will help us to model the data better.
   - You will use a default Matlab optimization solver - `fminunc` (function for unconstrained minimization) for iterating through the gradient descent algorithm for regularized logistic regression. The input arguments to the function are
     - a function that computes the value of the objective function and the gradient of the objective function for a set of data points and $w_0, w_1 \ldots, w_D$. Let us define this function as`[objval gradval] = objgradcompute(w, X, Y)`. The input to the function are the weights, input data and target values. It computes the value of objective function $J(w)$ - `objval` and a vector of values corresponding the partial derivatives of $J(w)$ with respect to $w_0, w_1, \ldots, w_D$ evaluated on the dataset - `gradval`.
     - initial value for the weights $w_0, w_1 \ldots, w_D$. Use values in the range [-0.1, 0.1] to initialize the weight vector.
     - a variable that specifies other parameters required to perform the minimization such as the maximum number of iterations. This variable is usually referred to as `options`. The parameter values for `options` can be set as follows
       `options = optimset('GradObj', 'on', 'MaxIter', 100)`
       Setting `'GradObj', 'on'` tells `fminunc` that our function `objgradcompute` returns both the objective function value and the gradient values. The maximum number of iterations is set to 100. You can change this to a higher value.

     The call to the function `fminunc` will look like
     `[w objval] = fminunc(@(w)(objgradcompute(w, X, Y)), initial_w, options);` This will perform the required number of iterations and return the final values for the weight vector and the objective function value for this weight vector.
   - Implement the function `plotdecisionboundary(w, X, Y)` that plots the non-linear decision boundary that separates the two classes as learnt by the classifier. Create a uniform grid

of values with a constant step size for the two input attributes. For each combination of the two attributes compute the value of the logistic regression function. Use the contour function to plot the curve that corresponds to regression value 0.5. Visualization will be more intuitive if the contour is overlaid on the color coded data points.

- Vary the value of the regularization parameter λ, and observe the changes in the decision boundary. Include in the report one figure each depicting under fitting and over fitting along with the value of λ.
- Implement a function `lindiscriminant(X, Y)` that computes the discriminants for two classes given the input data and target variables. The function should also compute the decision boundary using the discriminant functions and overlays it on the plot of the training data.

4. (4 points) Suppose we collect data for a group of students in a machine learning class with variables x_1 = hours studied, x_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, w_0= -8, w_1= 0.05, w_2 = 1.

   (a) Estimate the probability that a student who studies for 5h and has an undergrad GPA of 7.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 60% chance of getting an A in the class?