Final Exam

Due: 11:50am, Thursday, December 14, 2017

Note: Please solve these problems on your own and answer all of the questions with your own efforts. Discussions with other students are strongly discouraged. Similar python scripts with other students will result in zero points for all students involved. No extension of due time is permitted because of the deadline to submit the final grades to the UNCC registrar office.

Submit your plots, descriptions, and python scripts including all of your functions.

Note: For questions 2 and 3, before performing training and testing, scale petal length and petal width (stored as a column) as follows:

$$\text{new column} = \frac{\text{old column} - \text{column min}}{\text{column max} - \text{column min}}$$

so that, after scaling, each variable spans the range [0, 1].

1. (40 points) PCA and LDA

   In dataset `dataset_1.csv`, columns correspond to variables and there are two variables named `V1` and `V2`.

   (1) Plot `V2` vs `V1`. Do you see a clear separation of the raw data?

   (2) Apply your own PCA function to this dataset without scaling the two variables. Project the raw data onto your first principal component axis, i.e. the PC1 axis. Do you still see a clear separation of the data in PC1, i.e. in projections of your raw data on the PC1 axis?

   (3) Add the PC1 axis to the plot you obtained in (1).

   (4) Apply your own LDA function to this dataset and obtain W. The class information of each data point is in the `label` column.

   (5) Project your raw data onto W. Do you see a clear separation of the data in the projection onto W?

   (6) Add the W axis to your plot. At this point, your plot should contain the raw data points, the PC1 axis you obtain from the PCA analysis, and the W axis you obtain from the LDA analysis.

   (7) Compute the variance of the projections onto PC1 and PC2 axes. What is the relationship between these two variances and the eigenvalues of the covariance matrix you use for computing PC1 and PC2 axes?

(8) Compute the variance of the projections onto the W axis.

(9) What message can you get from the above PCA and LDA analyses?

2. (40 points) Artificial Neural Network (ANN) for classification

   Modify the python script you have written for doing regression using ANN in homework assignment 2 so that your modified code can perform classification analysis.

   The major difference between using ANN for regression and using ANN for classification lies in the cost function. For classification, use the cross-entropy cost function as follows:

$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(h_\Theta(x^{(i)})\right)_k + \left(1 - y_k^{(i)}\right) \log\left(1 - \left(h_\Theta(x^{(i)})\right)_k\right)\right]$$

   where y is the actual output and $h_\theta(x)$ is the predicted output.

   Your modified ANN will contain one input layer, one hidden layer, and one output layer. The input layer has two units, the hidden layer has two units, and the output layer has one unit. For this particular ANN architecture, $K = 1$.

   With the cross-entropy cost function, your calculation of

$$\frac{\partial}{\partial a^{(L)}} J(\Theta)$$

   should be modified accordingly. L is the total number of layers in the ANN.

   Another modification you need to make to the ANN python script you have written for homework 2 is to sum the partial derivative

$$\frac{\partial}{\partial \theta_{ij}^{(l)}} J(\Theta)$$

   over all training samples. Refer to slide #43 in the neural network lecture slides.

   Apply the resulting ANN python script to perform a binary classification of the virginica and versicolor flowers in the iris dataset using petal length and petal width. Specifically, you will perform a leave-one-out analysis by using one flower for testing and the remaining 99 flowers for training. If the testing result is different from the actual flower type, the error is 1.0. Otherwise, the error is 0. Perform this leave-one-out analysis 100 times and get the average error rate.

3. (20 points) Logistic regression.

Write your own python script for classification using logistic regression. You may refer to relevant code posted on the UNCC Canvas page for our class. Perform binary classification analysis of the virginica and versicolor flowers using petal length and petal width in the iris data set using your own logistic regression function. Use leave-one-out cross validation and get the average error rate, i.e. you will perform logistic regression 100 times total with each of the 100 flowers used once for testing.

Compare classification results you have obtained using ANN with the results you have obtained using logistic regression.