Term Project Phase II

(100 points)

Due Date:

MassHousing	**4:00 PM Thursday, November 15, 2016**
Crime Forecasting	**4:00 PM Thursday, November 15, 2016** **(The team meet with Prof. Melissa Morabito on 11/09/2016)**
PhD-Student projects	**4:00 PM Thursday, November 15, 2016** PhD-Student Projects: You are expected to present the 2nd phrase of your project using neural networks. A major milestone should be achieved during this phrase.

Submission Requirements

1. One submission per team. Name your file with teamleadlastname_firstname_team#_ph1. For example, team 1 of lead John Smith should name their file as Smith_John_team1_ph1.zip.
2. Prepare 7-minute PPT slides to discuss about your project in class on the Nov 15th, 2016 including your project demo, the design of the neural networks, and experimental results.
3. Two files should be submitted under your Blackboard account: 1) submit a single zipped file of all the software programs you developed for this assignment through your Blackboard account. 2) Submit a separate PDF file of your 7-minute PPT slides through your Blackboard account.
4. Submit the paper copy including the PPT slides and program source code. Paper copy should be bound firmly together as one pack (for example, staple, but not limited to, at the left corner). 5 points will be deducted for unbounded homework.
5. No hard copies or soft copies results in 0 points.

MassHousing Project

The objective of this phase is to build a revised neural network that can better predict the financial risk of various real estate developers that work closely with MassHousing.

The two teams for the project will be designated as sub-teams.

Team 1:

- Ashrita Muthukumaraswami (Team Lead)
- Shruthi Manjeshwar Jeevananda
- Pavithra B Venkatachalam

Team 4:

- Peter Babokhov (Team Lead)
- Achuth Kamath Miyyar
- Alesia Razumova

The tasks below have been prioritized by what should be done for this phase, and what can be deferred to the subsequent phase, in the interest of time.

The general experiment and the team-specific tasks are the immediate priorities for Phase II. The tasks that can be deferred can be found under the category “To Be Determined.”

General experimental design for both Team 1 and Team 4

- Design a three-layer neural network. Give an explanation, in Layman’s terms, of the type of neural network being used. The input layer of the neural network will consist of all 51 columns from the MassHousing data. The Financial Rating Letter Grade column will make up the five-neuron output layer, one neuron for each letter grade (A, B, C, D, F). This column will serve as the benchmark for prediction.

- The data from the input layer will be normalized with min/(max - min) scores (for Team 1) and with max/min scores, z-scores, and frequency values for categorical data (for Team 4).

- For phase II, the data will be split into three parts, training, test, and validation. The training data will be all the data from the statement dates before 2010. The test data will be all the data from the statement date of 2010. Lastly, he validation data will be selected at random from the training data. Each team will decide which portion of the training data they want to use for validation.

- The neural network will be built using Python, specifically the network2.py code provided by MNIST.

Designated to Team 1

- Determine the best weights for the seven neurons in the hidden layer. The weights of the seven neurons (from the input neurons) will be sent to Max Ward for review.
- Experiment with larger values for the hidden layer (25, 50, and 75 neurons). The analysis from the different numbers will be used to determine the best number of neurons for the hidden layer.
- For analysis of the original seven weights to be sent to Max, and the larger values for the hidden layer, make use of precision and recall.
- Precision tells you how many of the selected items are relevant (usefulness), and recall/sensitivity tells you how many of the relevant items were selected (completeness). That is, find out how many true positives, true negatives, false positives, and false negatives are obtained in the experiments. Give an explanation in Layman's terms of these findings.
- The input data for Team 1 will be normalized with min/(max - min) scores.

Designated to Team 4

- Apply log transformation on the data before normalizing it, to reduce the amount of negative values.
- Experiment with normalizing the values. For numeric values, use min/(max - min) scores, but also experiment with z-scores or any other suitable method. Report the method that gives a higher accuracy.
- For categorical values (such as Loan Closed Date, Statement Date, rm_key, etc.), replace each categorical value with its a frequency value.
- Determine which of the five letter grades from the Financial Rating Letter Grade column, the output for the neural network, have been predicted with greatest accuracy. In addition, analyze which of the letter grades are easier to predict with greater accuracy. Use precision and recall here as well.

To Be Determined

- Evaluation of the true positive, true negative, false positive, and false negative data from the hidden layer and letter grade experiments. Find whether these sets have any identifiable qualities, and whether they are alike in any way. In the interest of time, this task will be deferred to the subsequent phase.
- Visualize the data of the experiments and results (scatter plots, treemaps, decision trees, etc.). Due to time constraints, this task is not top priority, and will be deferred to the next phase.

- Experimentation with a lower learning rate, or different numbers of epochs can be deferred to the next phase.

- Experimentation with non-sigmoidal neural networks, in order to handle non-normalized data without getting any runtime errors in the Python code.

- Use the features learned from the neural network as input features into a classifier algorithm, such as support vector machine (SVM) or random forest.

Real-Time Crime Forecasting Project

1 Requirements Goal

Project goal: Using CNN to solve the Real-Time Crime Forecasting problem.

Due Date: Tuesday, November 9th, 2016 **Software tools:** Theano, Lasagne, Tensorflow ...
Programming language: python

Evaluation (Prediction Accuracy Index (PAI)):

$$\frac{\left(\frac{n}{N}\right)*100}{\left(\frac{a}{A}\right)*100} = \frac{HitRate}{AreaPercentage} = \text{Prediction Accuracy Index} \quad (1)$$

where n is the number of crimes in areas where crimes are predicted to occur (e.g. hotspots), N the number of crimes in study area, a the area (e.g. km^2) of areas where crimes are predicted to occur (e.g. area of hotspots), and A the area (e.g. km^2) of the study area.

Simple Code of CNN: http://craterdetect.cs.umb.edu:443/notebooks/examples/Lasagne

2 Assignment 1

Assignment goal: Change the sample code of CNN and apply it to solve the Real-Time Crime Forecasting problem.

Requirement: Using at least 2 convolutional layers, 1 fully connected layer.

Simple Code of CNN: http://craterdetect.cs.umb.edu:443/notebooks/examples/Lasagne

3 Assignment 2

Assignment goal: Using extra features to test your model on different Calls-for-Service(Burglary, Motor Vehicle Theft, and Street Crime), and compare the results.

Data:

Calls-for-service data of Portland: http://www.nij.gov/funding/Pages/fy16-crime-forecasting-challenge.aspx

Census Data: http://www.census.gov/2010census/popmap/index.php

Factfinder Data: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk **Extra Features Suggestion:** Number of crimes, Poverty- public assistance, Unemployment, Number of young people (under 18), Weather.

4 Assignment 3

Nihar A report of "long term and short term prediction" from Jerry Ratcliffe's talk.

Asma A report of the work that Predpol and Rutgers (two other teams working on this project) are doing on crime prediction.

Predpol: http://www.predpol.com/how-predpol-works/ Rutgers: http://www.rutgerscps.org/rtm.html

Yong A report of the paper "Convolutional LSTM Network: A Machine Learning Approach for Pre- cipitation Nowcasting".

4.1 Data collection

A report includes following information:

1. What kind of data we can get? e.g. population of each census tract.
2. How to get it? e.g. the architectural and relationship of the data tables.
3. Write sample code for data loading.

Akshay Census data collection http://www.census.gov/2010census/popmap/index.php

Sindhu Factfinder data collection

Factfinder: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

Shwetha Portland city data collection

Portland city: https://www.portlandoregon.gov/28130