Regression and Neural Networks

Due date: November 8, 2017 (Wednesday) – in-class presentations. November 9, 2017 (Thursday) – final submission of all deliverables.

Teams

This is a team assignment. Teams are determined by the instructors. You will get your team assignment during the Monday, October 23 class.

Data

For this assignment you will be using Kaggle’s Filipino Family Income and Expenditure dataset available at https://www.kaggle.com/grosvenpaul/family-income-and-expenditure.

The dataset contains information about over 41,500 Filipino families. Specifically, the dataset captures the household income of each respondent family, annual expenses of the family on various categories of food, and non-food items, value of owned possessions, as well as a variety of demographic and household characteristics. The CSV file contains self-explanatory columns, and all values in the dataset are also self-explanatory.

For the purposes of this lab you need to identify three core sets of features in the datasets.

The first column of the CSV represents the total household income. In this lab, this column will play a role of your dependent variable.

A variety of columns (largely containing information about household expenses by category, and value of goods/property owned) are continuous numeric attributes in the dataset. Some of your analysis will take place exclusively on these columns.

The remaining columns in the dataset are either categorical (nominal) variables, or ordinal variables. For the purpose of this assignment we assume
that variables asking for a count (e.g., number of children in the household, 
or number of bedrooms) are ordinal. Some of your analysis will take place exclusively on these columns.

Note: There are three attributes that are continuous variables, but may 
have more in common with the other ordinal/categorical attributes. These 
attributes are Household Head Age (column AA), House Floor Area (col-
umn AO), and House Age (column AP). For the purpose of these assignment 
treat these variables as continuous/numeric variables, but put them in the 
same group as all your ordinal and nominal variables. From now on, when 
we refer to regression tasks performed on numeric attributes, we mean all 
numeric attributes except these three.

Tasks

You have to answer a number of research questions regarding this dataset. 
These questions refer both to the accuracy of prediction using specific cat-
egories of attributes in the dataset, as well as to the accuracy of prediction 
using different regression techniques we discussed in class. Such a multitude 
of questions allows you to conduct a comprehensive analysis of this dataset.

Dependent Variable. As stated above, in this assignment you will be 
always predicting the Total Household Income of the households (column 
A).

Feature sets. All analyses described below need to be performed on the 
following combinations of feature sets:

1. **Numeric Only.** Only numeric attributes as described in the section 
above. Your final models can contain a subset of the numeric features. 
You can choose any technique you know to select appropriate features, 
but for this feature set, you may not introduce any new features.

2. **Categorical only.** This set includes all nominal and ordinal variables 
as well as the three numeric variables as described in the section above. 
Your final models can contain a subset of the numeric features. You 
can choose any technique you know to select appropriate features, but 
for this feature set, you may not introduce any new features.

3. **All features.** All attributes provided in the survey. Your final models 
can contain a subset of the numeric features. You can choose any 
technique you know to select appropriate features, but for this feature 
set, you may not introduce any new features.

4. **Any features.** You are allowed to use all features provided in the 
survey, as well as engineer any additional features from the ones avail-
able to you. You are allowed to use any model selection methods to 
filter out features.

2
Analytical Methods. You will compare the following types of predictive models:

1. **Linear Regression.** This part of the project is similar to your first project - you can reuse any code that can be reused for this purpose (this includes any model selection and appropriate feature engineering code you may have).

2. **Regression Feed-Forward Neural Networks.** As a minimum, you shall implement multi-layer feed-forward neural networks for regression. You are given complete control over the implementation, as long as you are not using any pre-existing neural network or machine learning packages to emulate your neural networks. You can, as always, use pandas and numpy in support of your implementation, and you can use scikit-learn's utility functionality, but not implementations of machine learning techniques, including neural nets.

When implementing your neural networks, please note the following:

1. You can choose which activation function or functions to implement.

2. Your implementation of the neural networks must allow you to construct networks with different number of hidden layers.

3. Number of nodes in each layer of the network should be another parameter of your implementation.

4. You are expected to construct standard feed-forward neural networks where each layer has full connectivity to the next layer. However, teams that want to implement additionally other architectures of neural networks (e.g., convolutional neural networks, with a few convolutional layers preceding fully connected layers), you are allowed to do so as long as you know what you are doing. This is not required, but is allowed as an option.

**Evaluation.** Please make sure to use appropriate evaluation techniques for testing the quality of your models. You must select an appropriate error function (and make sure you are measuring the same error in both implementations). Additionally, you must select appropriate accuracy measures and an appropriate procedure for evaluating the accuracy of your models.

**Research Question 1: Linear Regression vs. Neural Networks.**
Our key question for this assignment is *whether the use of regression neural networks provides a significant improvement over linear regression models in predicting the total household income values.*

---

1If you do not know what “convolutional neural networks” is, fear not. We will cover them in class later in the course. There is no requirement to implement them now.
To answer this question, you must compare the performance of the best linear regression models you can find using the four attribute sets presented above with the performance of the best regression neural network models you can build using your neural network implementation.

For linear regression models, you can use any model selection techniques (and where allowed, feature engineering techniques) to improve your prediction accuracy.

For neural network models, you can use variation of the neural network parameters: namely, the number of hidden layers, the sizes of respective layers, the activation function (and any additional parameters you deem necessary) to discover the networks with the best accuracy (i.e., to tune your "hyperparameters").

**Research Question 2: Which sets of attributes do the best job?**

We are specifically interested in

1. whether purely numeric or purely categorical sets of attributes are better at predicting the total household income,
2. whether combining numeric and categorical attributes together produces a significant lift the accuracy of total household income prediction,
3. whether any additional feature engineering appears to have any specific positive effects on the prediction accuracy,
4. whether Linear Regression and Regression Neural Network models behave in similar ways with respect to your answers to the previous questions? That is (for example): does feature engineering affect the accuracy of linear regression predictors in the same way it affects the accuracy of regression neural network predictors.

Perform all necessary analyses to gain insight into the answers to these two research questions.

**Deliverables.**

As with all other projects, there are three sets of deliverables.

**Code.** All your code must be delivered by the final project deadline (November 9). Submission instructions (kaggle, or handin) will be provided for you shortly.

**Presentation.** Each team will present its results in class on Wednesday, November 8. The intent is for each team to have a no more than 15 min. presentation. Please make sure your presentations do not run over time.
**Report.** Your report shall be formatted as an academic paper and include a title, names of the authors, a short abstract, an introduction, a description of your implementation of Neural Networks, experimental design, results and analysis. The report shall be more or less self-contained — a reader not familiar with DATA 401 course (but, for example, familiar with the dataset itself\(^2\)) should be able to read your report and understand what you have achieved.

Final versions of the reports are due the final deadline (November 9). Hardcopy reports are preferred, but please submit an electronic copy of your report together with your final code submission.

**Good Luck!**

---

\(^2\)You should provide a brief description of the dataset, but this condition absolves you from spending a lot of efforts on it.