Machine Learning Project Proposal

By Adam Clark and Alexander Dean

1 DataSet

The data that will be used for this project originates from the book *An R companion to Applied Regression*, we found it on the website below and it is licensed as public domain.

The features include the following:

- Rank - Professor, Assistant Professor, Associate Professor
- Discipline - A (theoretical) or B (applied)
- Years since PHD
- Years of Service
- Sex

The label of this data is a professor's salary.

1.1 DataSet URL

The dataset was found on the following website:
https://bigml.com/user/totyb/gallery/dataset/50f303103b56354d2a000405

The description of the data on this page reads:

"The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

Fox J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition Sage."

1.2 DataSet URL

This is the dataset for the salaries of regular employees in the University of Maine System. It is in PDF form and may need to manually be converted to a csv or similar. The UMS data is publicly available on the internet under an undefined license.

Origin:
http://www.maine.edu/about-the-system/ums-data-book/
Origin:
http://www.maine.edu/about-the-system/ums-data-book/human-resources-reports/

November 2018:
http://staticweb.maine.edu/wp-content/uploads/2018/11/UMS_PUBLICINFO_MAIN_11-05-2018.pdf?0d0f03
September 2018:
http://staticweb.maine.edu/wp-content/uploads/2018/09/UMS_PUBLICINFO_MAIN_04-03-2018-Revised.pdf?0d0f03
November 2017:
http://staticweb.maine.edu/wp-content/uploads/2017/11/PUBLICINFO-November-2017.pdf?ca0c38
April 2017:
http://staticweb.maine.edu/wp-content/uploads/2017/04/PUBLICINFO-April-2017.pdf?0d0f03
November 2016:
http://staticweb.maine.edu/wp-content/uploads/2013/06/PUBLICINFO-November-2016.pdf?0d0f03
April 2016:
https://staticweb.maine.edu/wp-content/uploads/2016/04/PUBLICINFO-April-2016-1.pdf?565a1d

2 Project Idea

The idea for this project is to create a machine learning model that fits the data from the dataset noted above. This model will be used to predict the salaries of professors based on the features listed in section 1. The data will also be displayed in a series of meaningful graphs that show a range of salaries for any one feature. Though this is not related to machine learning, it is useful in predicting a salary based on one feature alone.

Once this model is in place, the salaries of USM professors will be provided to the model, and the salaries will be compared.
It will be interesting to compare these values. Similarly, it will be interesting to compare how certain features align with the averages from this dataset. For example, perhaps female professors are paid more at USM than the average female professor from the dataset. This project will focus primary on learning. It is for this reason that

additional statistical work will be done. While machine learning models act as a good way of predicting future data, more traditional statistical approaches act as a better way of displaying and visually comparing data.

3 Approach

Because the main focus of this project is on Machine Learning, the first step of the project will be to develop a machine learning model that can accurately predict professors' salaries. A linear regression model will be used as a model for this. Each feature mentioned in section 1 will be a feature in the model. The label, of course, will be the professor's salary.

As mentioned in section 2, a number of statistical graphs will be created in addition to the linear regression. These will be created using more traditional statistical approaches, but will be used as a manual comparison tool.

It is also a possibility that, given enough time, a neural network model will be used to predict the salaries of professors. This will be done by making "bins" of ranges of salaries. These bins will be the various classes the neural network uses.

Most of the work for the project will be done in Matlab due to its robust statistical and mathematical functionality. Time permitting, a secondary method, such as the Tenserflow API for Python, may be used as well. This will add an additional layer of learning to the project, for the accuracy for the accuracy of the Tenserflow model and the Matlab model can be compared.

4 Referenced Projects

4.0.1 Ref Project 1

URL: https://www.polyglotdeveloper.com/r-projects/2016-09-30-Predicting-salaries-using-linear-regression/
This project uses the R language to create various data plots & guess professor salaries based off of a similar data set to ours.

4.0.2 Ref Project 2

URL: https://www.kaggle.com/koki25ando/nba-salary-prediction-using-multiple-regression
This project aims to find the salaries of NBA players. There dataset has significantly more features than ours such as the players height and other sports related statistics. This project also uses R.