Final Project:
The Life of a Data Scientist

Due Monday, May 1st at 11:59 PM via e-mail to any instructor

I. Overview

Welcome to the final project for the Data Science Decal! In this semi-open ended project, you will act as a data scientist by performing a typical end-to-end modeling procedure. Throughout the semester, you have learned about how to clean and manipulate data, many different modeling methods, and how to optimize said models. Now, it is your turn to do all of the above on a dataset of your choosing (well, our choosing) in order to experience the typical pipeline of raw data to meaningful results.

You will work in **groups of 3-4** students to clean and featurize data, implement, compare, and optimize **3 different** machine learning models on one of the datasets described below. **One of the models you use must be some kind of neural network**.

As a reminder, here are the models we have covered this semester:

- Linear regression
- Logistic regression
- K-nearest neighbors
- K-means
- Decision trees and random forests
- Neural networks (feed-forward, convolutional, recurrent)
- Support vector machines (We haven't covered SVMs, but you may implement them!)

II. Datasets

Here are some datasets that we believe will be reasonable and fun to work with. **Please choose ONE** of the below datasets for your project.

- Mushroom identification
  - This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family
  - Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended (this latter class was combined with the poisonous one)
  - Your job is to classify whether the mushroom is poisonous or non-poisonous

  - See `http://archive.ics.uci.edu/ml/datasets/Mushroom` for the dataset and a detailed description

- Identifying hate speech in tweets
  - A sampling of Twitter posts that have been judged based on whether they are offensive or contain hate speech, as a training set for text analysis
  - See `https://www.crowdflower.com/wp-content/uploads/2016/03/twitter-hate-speech-classifier-D` `csv` for the dataset

- UFO reports
  - NUFORC geolocated and time standardized ufo reports for close to a century of data. 80,000 plus reports
  - See `https://github.com/planetsig/ufo-reports` for the dataset and a detailed description
  - You may perform regression or classification here!

- Horses for courses
  - Daily horse racing (thoroughbred) information that has(is) being actively collected and aggregated from a variety of sources
  - Years covered are just 2016, country is irrelevant to the dataset
  - Here, we want to predict the final position of a horse in the race (i.e. 1st, 2nd, 3rd place, etc...)
  - Please see `https://www.kaggle.com/lukebyrne/horses-for-courses` for the dataset and a detailed description

- Human activity recognition with smartphones
  - The Human Activity Recognition database was built from the recordings of 30 study participants performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors
  - The objective is to classify activities into one of the six activities performed
  - Please see `https://www.kaggle.com/uciml/human-activity-recognition-with-smartphones` for the dataset and a detailed description

- Credit card fraud detection
  - The datasets contains transactions made by credit cards in September 2013 by european cardholders
  - This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions
  - Your job is to classify whether or not a transaction was fraudulent
  - The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
  - Please see `https://www.kaggle.com/dalpozz/creditcardfraud` for the dataset and a detailed description

- Polish companies bankruptcies
  - The dataset is about bankruptcy prediction of Polish companies
  - Please see `http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data` for the dataset and a detailed description
- Breast cancer diagnosis
  - Please see `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29` for the dataset and a detailed description
- Automobile MPG (regression)
  - Please see `http://archive.ics.uci.edu/ml/datasets/Auto+MPG`
- Individual household power consumption (regression)
  - Please see `http://archive.ics.uci.edu/ml/datasets/Energy+efficiency`
- Fertility
  - 100 volunteers provide a semen sample analyzed according to the WHO 2010 criteria
  - Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits
  - Please see `http://archive.ics.uci.edu/ml/datasets/Fertility`
- Amazon book reviews
  - Please see `http://archive.ics.uci.edu/ml/datasets/Amazon+book+reviews`
  - Is a review positive or negative?

III. Project outline

As discussed above, your group will choose one of the above datasets to complete your project with. Again, you must model the data with **3 different** models, and **one of them must be some kind of neural network**.

Deliverable: Please submit a write-up in pdf format that follows the following structure. It is in your best interest to dedicate time to constructing a nice write-up which **you would be proud putting your name on**. For many of you, this will be your first "big" data science project, and it is something that you might want to talk about in future interviews. We suggest LaTeX format, but this is not required. There is no minimum or maximum page requirement, but you are expected to show significant effort.

I. Problem statement and goal of analysis

Give an overview of your project and some motivation

II. Data preprocessing

What features are you using? Did you do any featurization to the raw data? If so, what did you do?

III. Data modeling

Describe the 3 models you chose to use. Describe any hyperparameter tuning here. When training and evaluating models, make sure to split your data into a training and validation set. Did you perform any cross-validation?

IV. Comparison of models

Which model performed best? Why?

V. Discussion of results

What was your best accuracy? What are the implications? What went wrong?

VI. Conclusions

Here, you should reflect on your project and discuss what you have learned. Is there anything that you wish you would have tried if time permitted? What was your biggest challenge, and how did you overcome it?

The above is a **bare minimum** of what you should turn in. You are **deeply encouraged** to go above and beyond. Remember, this is for your benefit!

IV. Grading and Tips

We will grade your write-up in terms of the challenge of your project, the professionalism of your report, the functionality of your models, and your effort. A strong effort includes going above and beyond the bare minimum by doing things such as extra featurization, cross-validation, trying more than 3 models, choosing a difficult dataset, and more. The rubric is fairly loose, and you will be graded on a scale from 0-100.

In order to pass this class, you must turn in something we deem worthy. Please check with us throughout your work if you are unsure of the quality of your project. If you did not do so well on past quizzes and assignments, this is your chance to go above and beyond to help you get a passing grade.

Remember, **scikit-learn** and **Keras** are your friends. Training the models should be the easy part of this project! The problem formulation, data processing, and model optimization should be the hard part.

THIS IS NOT AN EASY PROJECT - DO NOT PROCRASTINATE! WE ESTIMATE THAT A HIGH-QUALITY PROJECT COULD TAKE ANYWHERE BETWEEN 10 TO 20 HOURS

Good luck, and have fun! c: