# CSC 215-01 Artificial Intelligence (Spring 2019)

# Mini-Project 2: Yelp Business Rating Prediction using Tensorflow

Name: Megha Mathpal

Sac ID: 219695880

**Due at 4:00 pm, Wednesday, February 27, 2019**

## Problem statement:

In this project we aim to predict a business's stars rating using the reviews of that business and review count based on neural network implementation in Tensorflow. We focus on comparison of regression model (last project) to the best RMSE of Tensorflow regression neural network model as well as comparison of classification model with each previously obtained classification.

## Methodology:

- Implementation of tensorflow models by understanding the tensorflow funtions and
- Used multiple layers and many combinations of optimizers and activation functions.
- Used loop to get better result and to avoid the local optimum

## Experimental Results and Analysis:

Experiment done on final data set having stars, review count, text and applying one hot encoding on star rating and taking input as review count.
I have applied various activation on all models and changed number of neurons and tried to apply different optimizers as well.
The results are shown is table below.

**Regression without tips:**

| optimizer | Activation and layers | Number of neuron count | MSE | RMSE |
|---|---|---|---|---|
| Adam | (relu,relu) | (20,25) | 0.05 | 0.23 |
| **Adam** | **(sigmoid,sigmoid)** | **(50,20)** | **0.04** | **0.22** |
| **Adam** | **(tanh, tanh)** | **(100,60)** | **0.04** | **0.22** |
| SGD | (relu,relu,relu) | (100,60,30) | 0.70 | 0.26 |
| SGD | (sigmoid,sigmoid) | (100,50) | 0.74 | 0.27 |
| SGD | (tanh, tanh) | (100,55) | 0.05 | 0.24 |

**regression with tips:**

| | | | | |
|---|---|---|---|---|
| Adam | (sigmoid,relu,tanh) | (50,30,10) | 0.04 | 0.21 |

**Classification without tips:**

| Optimizer | Activation and layers | Number of neuron count | F1 score |
|---|---|---|---|
| Adam | (relu,relu, relu) | () | 0.74 |
| Adam | (tanh,tanh,tanh) | | 0.73 |
| Adam | (relu,sigmoid, sigmoid, sigmoid) | (100,80,20,10) | 0.74 |
| **Adam** | **(sigmoid,relu,tanh)** | **(50,30,10)** | **0.75** |
| SGD | (relu,relu,relu,relu) | | 0.13 |
| SGD | (tanh,tanh,tanh) | (50,40,10) | 0.16 |
| SGD | (sigmoid, sigmoid, sigmoid) | (50,30,10) | 0.13 |
| SGD | (sigmoid,relu,tan) | (70,40.,20) | 0.13 |

**Classification with tips:**

| | | | |
|---|---|---|---|
| Adam | (sigmoid,sigmoid) | (50,20) | 0.76 |

**Scikit models :**

| **Models** | **With Tips** | **Without Tips** |
|---|---|---|
| Linear regression (RMSE) | 0.26 | 0.55 |
| KNN(F1 score) | 0.39 | 0.47 |
| Logistic Regerssion(F1 score) | 0.52 | 0.37 |
| MNB (F1 score) | 0.35 | 0.37 |
| SVM (F1 score) | 0.68 | 0.70 |

## Task done:

- Converting json to csv
- Data preprocessing (merge, groupby)
- Cleaning data (removing missing values, numbers)
- MinMax, vectorizing(tfidVectorizer), feature normalization
- Implemented all previous project models (with tips dataframe)
- Implementing tensor flow model
- Used earlystopping and modelcheckpoint

- While using tenserflow i have used EarlyStopping and ModelCheckpoint while training neural networks.
- Tuned these hyperparameters when training neural networks using Tensorflow:
  **Activation:** relu, sigmoid, tanh
  **Number of layers and neuron count for each layer**
  **Optimizer:** adam and sgd.

## Project Reflection:

I first implemented all the models with the tips data included with scikit learn. Then implemented tensorflow model for regression and classification without the tips data. Then for comparison I picked the best of all and implemented with the tips data. Because of the huge data set even on google colab I was facing problems. The ram was getting crashed midway.

## Extra Feature:

I have added a new file to data called tip.json.

Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions. I have used the following fields:

```
{
    // string, text of the tip
    "text": "Secret menu - fried chicken sando is da bombbbbbb Their zapatos are good too.",


    // integer, how many compliments it has
    "compliment_count": 172,

    // string, 22 character business id, maps to business in business.json
    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",


}
```

When we add a new feature or data to data frame it might help us to predict the result better:

I picked the best regression and classification tensorflow model and to compare the results obtained from both. I implemented the tensorflow model with the tip data included just for the best combination of optimizer and activation resulted from the data set without the tips data.