Group Assignment 1

Multi-Output Neural Network
on Large Data

Learning objective:

Estimating a multi-output feed-forward neural network on data that is larger than what can be held in a typical computer's RAM.

Data:

All data is contained in *meta_Clothing_Shoes_and_Jewelry.json.gz* and is due to *Jianmo Ni, Jiacheng Li, Julian McAuley. Empirical Methods in Natural Language Processing (EMNLP), 2019.*
For this assignment we are using two columns: title and category. A record in the data is one product sold on Amazon.

Reading data row by row (set to read in the first 10 rows):

```
import numpy as np
import gzip
import os
os.chdir('/Users/mballin2/Desktop')

def parse(path):
  g = gzip.open(path, 'rb')
  for l in g:
    yield eval(l)


i = 0
df = {}
for d in parse('meta_Clothing_Shoes_and_Jewelry.json.gz'):
    i += 1
    X = np.array(d['title'])
    print('X (title):\n')
    print(X)
    Y = np.array(d['category'])
    print('\nY (category):\n')
    print(Y)
    if i == 10:
        break
```

Goal

The data comes from Amazon, and the category is specified by the seller of the product. The seller selects categories from a list, but there is also an ‘other’ category. When the ‘other’ category is selected the seller has to create the category. This makes it a challenging task to detect incorrect categories.

The goal in this assignment is to predict which categories a product belongs to based on its title. To clean up the categories assigned to a product we will then retain the top five categories (i.e., the five categories with the highest probability). If the model is accurate, the result will be that categories will be correct (in case a wrong category was selected from the list), and that the categories that were created but not useful (e.g., those that have typographical errors, or are inappropriate) will be filtered out.

Deliverables:

Code (.py file)

-A neural network implemented in TensorFlow that learns incrementally (record by record). The inputs are the title and the outputs (multiple) are the categories.

-All Python code to prepare the presentation

Presentation (pdf or PowerPoint)

-Performance of the model

-Useful insights based on variable importance and partial plots

-Details on how much data was used and training times