Assignment 3: Fundamentals (Part 2)

Start Assignment

Due Tuesday by 5:59pm **Points** 6 **Submitting** a text entry box or a file upload

Question 3 (1 point):

In section 3.1.1 (page 59) of the Deep Learning with Python textbook, the first couple of hidden layers of a model are defined as follows:

```
from keras import models
from keras import layers

model = models.Sequential()
model.add(layers.Dense(32, input_shape=(784,)))
model.add(layers.Dense(32))
```

Suppose we "finished" the model by adding:

```
model.add(layers.Dense(10, activation = "softmax")
```

In reviewing the documentation for the Keras Dense() layer [https://keras.io/api/layers/core_layers/dense/], we see 'If you don't specify anything, no activation is applied (ie. "linear" activation: a(x) = x).'

Would we expect this "deep" model (a model with more than one hidden layer) to beat (be more accurate than) a multinomial logistic regression model; i.e. would it beat a model defined as ...

```
from tensorflow.keras import layers, models
model = models.Sequential()
model.add(layers.Dense(10, activation = "softmax", input_shape = (784,)))
```

Why or why not?

Model 3 (5 points) ...

This week's exercise involves a multi-label text classification problem. Given a Standard Generalized Markup (.sgm) text file, we want to predict whether a Reuters new article contains the following topics:
"earn": indicates the news article contains the corporate earnings topic
"acq": indicates the news article contains the corporate acquisitions topic

"money-fx": indicates the news article contains the foreign exchange markets topic (sometimes folks call this "forex")

For each news article, we are predicting 3 probabilities:
1) the probability that the article contains the "earn" topic
2) the probability that the article contains the "acq" topic
3) the probability that the article contains the "money-fx" topic

For example, the reuters_trn/reuters_trn_01017.sgm file indicates that news article contains both "earn" and "acq" topics, as it discusses both ...
1) Corning Glass selling stock in an Argentine glass manufacturer to a European group and
2) the expected impact on net income

Please navigate to the following URL to accept the invitation for this Kaggle task:
https://www.kaggle.com/t/627fbdd1b252422ebe6ba6d97a612eeb

If you need hours added to your Azure lab VM, please send email to sadm_rudy514@uw.edu; and remember to shut down your VM after you are done using it.

Activate your conda environment:

```
conda activate py37_tensorflow
```

For this exercise, you will need to install 3 things [you will only need to install these once for the course]:
1) spaCy: the syntactic parser using the C-based extension for python (Cython)

```
pip install spacy
```

2) English language metadata for spaCy: language: English; type: core; source: web; size: small

```
python -m spacy download en_core_web_sm
```

3) The beautiful soup Standard Generalized Markup (SGM) language parser

```
pip install bs4
```

Download the 3 scripts to execute:

```
wget https://www.cross-entropy.net/ML530/reuters-vocabulary.py.txt
wget https://www.cross-entropy.net/ML530/reuters-vectors.py.txt
wget https://www.cross-entropy.net/ML530/reuters-train.py.txt
```

Execute the reuters-vocabulary.py.txt script to generate the vocabulary [consider changing vocabularySize]:

```
python reuters-vocabulary.py.txt
```

Execute the reuters-vectors.py.txt script to use the vocabulary to preprocess the news articles:

```
python reuters-vectors.py.txt
```

Execute the reuters-train.py.txt script to search for the best multi-layer perceptron [Dense() layers] model and create predictions for the test set:

```
python reuters-train.py.txt
```

Upload your predictions to Kaggle:

```
kaggle competitions submit ml530-2021-sp-reuters -f predictions.csv -m "reuters submission"
kaggle competitions leaderboard ml530-2021-sp-reuters -s
```

To receive credit, your submission should beat the random baseline for the evaluation measure. In this case, we're using the mean Area Under the Receiver Operating Characteristic (ROC) curve for the 3 topics. The area under the curve for a random multi-label classifier would be 0.5