PCSE 595 Spring 2021 Assignment 3
Deep Learning

Due Monday April 11th

Congratulations! After completing this project, you now have the skills to create machine learning models for pretty much any task you want. First, complete part one to ensure you know the fundamentals of Keras, and to ensure you have Keras set up properly. Next, to prove your machine learning mastery, by completing part (2). You should choose a dataset that interests you, and put these skills and/or the project experience on your resume. I chose datasets/tasks that should spark interest in potential employers. Lastly, complete part (3) to practice your ability in communicating your machine learning knowledge to others. This will be important for your career, and your job interviews.

This project may be completed individually or as part of a 2-person team. **If you choose to work as a team, please send me via email your team by Friday March 25th**

Expected deliverables (submitted via Scholar)

Everyone must submit the completed iris.py.
Only one submission per team is required for parts 2 and 3

1. completed iris.py
    a. assume the data is in the same director as iris.py. Your code should run with a single click
2. Code and data for part (2). Which consists of one of the following:
    a. Code and data as a a .zip file – your code should load data, train and generate results for each classifier with a single click.
    b. A link to google colab project. Your colab project should load data, train, and generate results for each classifier with a single click. Ensure you share your colab project with me and put the data in a folder that you share with me so that I can run your code
3. Report (a .docx) as described in (3)

Project Details

Part 1: A basic classifier (iris dataset)
Individually complete iris.py. The script should load, train, and predict with a single click.

a) Download the iris dataset (https://archive.ics.uci.edu/ml/datasets/Iris)

b) Load the iris dataset and split it into test and training sets
c) Code a 3-layer densely connected neural network using keras. You are free to make whatever other design choices you want.
d) Train the network on the training data
    a) You should get pretty good performance. If you don't try adjusting the hyperparameters
e) Predict on the test data, and output test set precision, recall, and f1.
f) Perform classification and prediction on the iris dataset using a 3-layer densely connected NN using Keras.

Part 2: A more complex task
Individually or as a team select one of the datasets below and perform classification:

a) You must implement a majority class classifier
b) You must implement a "simple" baseline classifier. This is some simple method that you think of (not machine learning) to classify each sample. This is dataset dependent.
c) You must implement a machine learning classifier that uses the data directly
d) If part of a group: You must perform some additional experiment on the data or develop an additional model. This is dataset dependent.

Use your knowledge of machine learning to select which evaluation metrics you report.
You should get "good" results, which means you almost certainly have to perform a hyperparameter search for each method

Part 3: Report
Individually or as a team write a report. You can use the report from assignment 1 as a guide. The report should include the following sections:

a) Your name, and if applicable your team mate's name
b) Introduction: Description of the problem
c) Dataset: Summary of the dataset
d) Methods:
    a. Description your simple baseline
    b. Description your machine learning method
    c. If part of a group: Description of your additional experiment/additional method
e) Results:
    a. Summarize the results of each method in a table
    b. Compares each method using statistical significance tests
f) Conclusions: Make a conclusion on what the best system is
g) Future Work: Explain what your next steps would be in improving the model and how you would incorporate those changes.

You do not need to write about the iris dataset at all. This report is about part (2) only.

Important: A longer report is not better. Just make sure you include all the necessary information. I expect reports to be somewhere between 2-5 pages

Part 4: Team contribution statement
Only if you are part of a group, submit via email a statement of your contributions and a statement of your teammate's contributions.

- This is a few sentences describing your contributions to the project and the contributions of your team member. You may, but don't need to discuss your statement with your team mate. All statements will be kept confidential. (This is your chance to complain if your teammate slacked off).

Dataset Options

Activity Recognition Dataset
https://archive.ics.uci.edu/ml/datasets/Activity+recognition+with+healthy+older+people+using+a+batteryless+wearable+sensor

The goal of this task is to identify the activity being performed based on wearable motion sensor data. This is important for detecting the movements of older adults who may live alone and be prone to falls.

The dataset contains data recorded from wearable motion sensors of healthy older adults. The participants were asked to perform the following activities: 1) walk to the chair, 2) sit in the chair, 3) get off the chair, 4) walk to the bed, 5) lay on the bed. The data contains information about the time, acceleration from 3 axes, signal strength, phase, and frequency. The goal is to classify each time-stamped data as 1: sit on bed, 2: sit on chair 3: lying 4: ambulating (walking)

Data was record in two different rooms with two different antennae set ups. Each folder corresponds to data recorded in each room. The gender of the participant is identified with an M or F at the end of the file name. Data in each file contains sequential recordings from the sensor.

This dataset is interesting because:

1) it is sequential, so do you have to take the class of the previous sample into account for classifying the next sample? What kind of architectures will work well for this?
2) Are there differences between males and females? Should you create two different classifiers?
3) Are there differences between sensor configurations? Do you need to create a classifier per sensor configuration, or does a classifier trained on one configuration translate to the other?

4) Are classes confused – should you create a pipelined classification approach? i.e. walking vs. sitting vs. laying then determine what they are sitting on? Other pipelined approach?
5) other – class imbalances, etc?

Data is clean and no missing values however, it is noted that recordings are sparse and noisy due to the use of a passive sensor (the sensor is batteryless).

Drug Review Dataset

https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

The goal of this task is to determine a user's satisfaction of a drug from its review.

The dataset consists of reviews and user ratings from Drugs.com. The drugs can be categorized as: contraception, depression, pain, anxiety, and diabetes drugs. The data contains the following attributes: 1) drugName (categorical): name of drug 2) condition (categorical): name of condition 3) review (text): patient review 4) rating (numerical): 10 star patient rating 5) date (date): date of review entry 6) usefulCount (numerical): number of users who found review useful.

The authors describe their method as an n-gram model which featurizes text based on n-gram counts of reviews, then use a logistic regression classifier. They achieve high results. However, the authors turn this into a 3-class classification problem (positive, neutral, negative) by bucketing the ratings.

This dataset is interesting because:

1) its NLP
2) Can a classifier work across drug types?
3) Do you need to apply any text cleaning to the data
4) What language model should you use (if any) – the authors use an n-gram approach
5) Does a classification method work well, would regression work better?

Influenza Outbreak Data

https://archive.ics.uci.edu/ml/datasets/Influenza+outbreak+event+prediction+via+Twitter+data

The goal of this task is to predict if an influenza outbreak will occur from simulated Tweets. The data contains a list of keyword counts and the location (state) of the simulated user. Tweets come from all 50 states. Example keywords include: "swine", "thera", and "mate"

data is a .mat file, which can be loaded into python using scipy.io.loadmat.

Interesting questions/ideas about the dataset:

1) Does proximity to an outbreak help predict another outbreak? How could you incorporate that into a model?
2) I wonder if state population density, or other factors can help in prediction? Could those features be incorporated?
3) The data is just keyword counts. Can you transform this into something more informative? Maybe something based on knowledge about the words, or cosine similarity to “Flue”, “sick”, etc… does that improve results?

Font Recognition Dataset

https://archive.ics.uci.edu/ml/datasets/Character+Font+Images

The goal of this task is to recognize letters from different computer fonts and handwritten images.

The data set consists of images from 153 character fonts. Some fonts were scanned from a variety of devices: hand scanners, desktop scanners or cameras. Other fonts were computer generated. The .zip file contains .csv, comma delimited files, one for each font. Each .csv file has a header row with the data set attribute names.

Interesting things about this dataset:

1. There are a lot of classes, how will that effect learning
2. Does converting the input data into an 2-D matrix increase performance (presumably you could use a CNN to classify the matrix?)
3. How will the classifier perform on a font it hasn’t seen before (e.g. hold a font out from training and see how it does)
4. How does performance differ for handwritten vs computer fonts