CS5783: Machine Learning
Final Project

Prof. Christopher Crick

For your final project, you are to find a significant and interesting dataset and perform data analysis and prediction using four different machine learning approaches. One of these approaches must be a deep learning neural network; if you wish, you may create two such networks with very different structures and behaviors to count as two of the four algorithms. **Note:** If you have a different idea for something you would like to do as a project, come talk to me and make your case. In particular, you are *always* welcome to produce an article using machine learning in your own research, which we could then turn into a publication in a journal or conference. An example of a successful such paper from a student in a previous year can be found at `http://cs.okstate.edu/~chriscrick/Maulik-JFluidMech-19.pdf`.

This project has three deliverables:

- A four-minute class presentation on your dataset, what you ultimately plan to discover, and some initial results and visualizations. This will probably amount to three or four slides. The presentations will be given on Monday, November 18 and Wednesday, November 20. The order of presentations will be determined randomly; all of the slide PDFs must be submitted to Prof. Crick by Sunday evening. He will assemble them all into a single presentation to be delivered from his laptop, to save switching time and technical difficulties.

- A written report. Introduce your dataset and explain the value of understanding it through machine learning. Explain the approaches you took and your design decisions. The heart of the report will be composed of twenty figures, their captions and accompanying text. These figures should illustrate your discoveries and experiments. You can show changes with respect to varying hyperparameters, effects of overfitting, intermediate results, incorrectly and correctly classified examples – anything that helps tell your story.

- The code you wrote to support your experiments. Unlike in assignments, where you have been expected to provide low-level implementations of the algorithms we studied, you may use available implementations in common libraries. Anything in Tensorflow or Scipy/Sklearn is fair game. However, you may not use some random code you found floating around in Github – your code should be executable by anyone, including your professor. The code should contain a link or a description to the data or (best) a little program to download and organize it.

Here are a few places to look for data, to get you started.

- US government agency data at `http://data.gov`

- Health-care data at `https://www.healthdata.gov`
- Federal reserve economic data at `https://fred.stlouisfed.org/`
- Facebook social media data at `https://developers.facebook.com/docs/graph-api`
- Images with bounding boxes and labels at `http://image-net.org/`
- Earthquake data at `http://earthquake.usgs.gov/data/`

Obviously, the possibilities are endless. Your job is to find a dataset that can support a good machine learning story. **Note:** You should *not* use a data source that has been specifically curated for the use of machine learning students. These are often collected on websites called something like "Machine Learning Repository", such as `https://archive.ics.uci.edu/ml/index.php`. Obviously, this also includes all of the datasets that can be loaded at a keystroke via something like the Keras library. Have fun with this, and good luck!