DAP 1

Due: Midnight on Friday, April 17th

As the course progresses you will be building a data analysis project for a real-world scenario of your choosing. You will pick a project topic, a machine learning application that interests you, define problems for which you plan to apply machine learning techniques, and identify one or several datasets suitable for your topic of interest. You will then explore how best to apply machine learning algorithms to solve it: preprocess the data, conduct an exploratory data analysis, develop models using supervised learning and unsupervised learning algorithms, and refine the models for performance improvement.

DAP 1: Project Design and Datasets (50 pts)

Step 1 of your Data Analysis Project (DAP)

Your first step is (1) to pick a topic or application you would like to tackle and define a set of problems, and (2) to identify one or several datasets suitable for your project, and (3) to write a milestone plan.

Project topic and category

We suggest that you pick an application area that you will enjoy working with, since you will be stuck with it for the whole quarter. So, pick something that you can get excited and passionate about! For example, pick something you are interested in – predictive models in public health domain, social networks analysis, stock market forecasting, etc. Be brave rather than timid, and do feel free to propose ambitious things that you're excited about. (Just be sure to ask Professor for help if you're uncertain how to best get started.)

In your project design document, **below your project title**, please include the project category. The category can be one of:

- Athletics & Sensing Devices
- Audio & Music
- Finance & Commerce
- General Machine Learning
- Life Sciences
- Natural Language
- Physical Sciences
- Theory & Reinforcement Learning

(If you feel a category is missing, please let me know.)

Problem definitions

Once you have identified a topic of interest, it can be useful to look up existing research on relevant topics by searching related keywords on an academic search engine such as: http://scholar.google.com. You will find more concrete examples of possible problems you would like to work on. Please consider problems in each of the following three categories: supervised learning (i.e., regression & classification) and unsupervised learning (i.e., cluster analysis).

Datasets

Another important aspect of designing your project is to identify one or several datasets suitable for your topic of interest. There are a couple of data repositories for machine learning research (e.g., Kaggle, the UCI machine learning repository, etc.) where you can find suitable datasets for your project.

If that data needs considerable preprocessing to suit your task, or that you intend to collect the needed data yourself, keep in mind that this is only one part of the expected project work, but can often take considerable time, so pace your project accordingly.

While we don't want you to have to spend much time collecting raw data, the process of inspecting and visualizing the data, trying out different types of preprocessing, and doing error analysis is often an important part of machine learning. Hence if you choose to use pre-prepared datasets (e.g. from Kaggle, the UCI machine learning repository, etc.), we encourage you to do some data exploration and analysis to get familiar with the problem. This task will be in the 2nd step of the project (DPA2).

Milestone

The milestone will help you make sure you are on track, and should describe what you've accomplished so far, and very briefly say what else you plan to do.

- You should write it as if it's an "early draft" of what will turn into your final data analysis report so that you can re-use most of the milestone text in your final report.
- Please write the milestone (and final report) keeping in mind that the intended audience are the Professor and machine learning expertise. Thus, for example, you should not spend two pages explaining what logistic regression is.
- Your milestone should include the full names of all your team members and state the full title of your project and a category of the project. Note: We will expect your final writeup to be on the same topic as your milestone.

Design document format:

Your design document should be a PDF document and include the following information:

- The title of the project, the project category, the full names of all of your team members,
- A 300-500 word description of what you plan to do.
    - Motivation: What problem are you tackling?
    - Method: What machine learning techniques are you planning to apply or improve upon?
    - Intended experiments: What experiments are you planning to run?
    - How do you plan to evaluate your machine learning algorithm?
    - Links to one or two relevant datasets.
    - One example of prior research on the topic.

Submission Guideline

- Submit **a design document** to Canvas (one submission per group).
- Follow the convention for submission file names:

  {group#}_DAP{project phase number}.pdf (Ex) group1_DAP1. pdf

Some Tips

1. Analysis of image and sound data is beyond the scope of this course so please consider numerical and text datasets.
2. If you are looking for project ideas or if you are unsure whether your proposed application is appropriate, please come to Professor's office hours, and I would be happy to brainstorm and suggest some project ideas.