CSE 255, Winter 2015: Assignment 1

Instructions

This is an **open-ended** assignment in which you are expected to write a detailed report documenting your results. Please submit your solution electronically to Dongcai Shen (doshen@cs.ucsd.edu), on or before February 23 (lecture 6). This assignment is worth **30%** of the final grade.

This assignment may be conducted **individually or in pairs**. Both members of a pair will receive the same grade, though assignments conducted in pairs should be **substantially deeper**. Make sure to mention your partner's name when submitting. Submissions should be in the form of a written report, which is expected to be between 6 and 10 pages, or roughly 3-5 thousand words, plus figure, tables, and equations. Assignments conducted individually should be on the shorter end of this spectrum, assignments conducted in pairs on the longer end.

Examples of datasets and projects that may be of interest in this assignment were discussed in Lecture 3: `http://cseweb.ucsd.edu/~jmcauley/cse255/slides/lecture3_assignment1.pdf`

Tasks

Assignments will be graded based on their coverage of the following six components. There is room for flexibility, though you should have good reasons for deviating from the basic structure below. Examples of what might be included in these sections is described in the above link. Each of the six sections below will contribute approximately 5 out of 30 percent of the grade for this assignment.

1. Identify a **dataset** to study, and perform an exploratory analysis of the data. Describe the dataset, including its basic statistics and properties, and report any interesting findings. This exploratory analysis should motivate the design of your model in the following sections. Datasets should be reasonably large (e.g. more than 50,000 samples).

2. Identify a **predictive task** that can be studied on this dataset. Describe how you will evaluate your model at this predictive task, what relevant baselines can be used for comparison, and how you will assess the validity of your model's predictions.

3. Describe **literature** related to the problem you are studying. If you are using an existing dataset, where did it come from and how was it used? What other similar datasets have been studied in the past and how? What are the state-of-the-art methods currently employed to study this type of data? Are any of them suitable for comparison?

4. Identify which **features** shall be relevant to the task at hand. Why do you expect them to be useful for prediction, and how does your exploratory analysis justify the features you have selected? What forms of pre-processing were required to select or manipulate the features?

5. Describe your **model**. Explain and justify your decision to use the model you proposed. How will you optimize it? Did you run into any issues due to scalability, overfitting, etc.? What other models did you consider for comparison? What were your unsuccessful attempts along the way? What are the strengths and weaknesses of the different models being compared?

6. Describe **Results and conclusions**. How well does your model perform compared to alternatives, and what is the significance of the results? What is the interpretation of your model's parameters? Why did the proposed model succeed why others failed (or if it failed, why did it fail)?