Project for the AI, 2018 Fall

1) Due on the end (Sunday, midnight) of the 18th week, past the new year;
2) All reports are supposed email to our TAs;
3) You can also contact the TAs if you have any questions or problems using our group WeChat ;
4) Reports are supposed to be five pages, single space, in a format like a conference paper, consisting of abstract, method, result, and discussions. Introduction is not necessary;
5) Both Figures and Tables should have legends;
6) Two students are working on the same project, but turning in different reports finished by each of yourself.

The data can be download from: ftp://public.sjtu.edu.cn ,
username is boyuan
password is cs2017

The data is in Gene_Chip_Data.zip
1) There should be a number of files, the biggest being the data to be used, which is p x n, where p are the features (or the individual genes), and the n are different observations (genechip experiments);
2) There should be 23,000 genes (p), and ~6000 observations (n), so a large p small n problems.
3) There are two other files which are used to describe individual genes, and experiments.
4) Note that the experiments are described by some text, (not structural), thus you will have to parse them according to your needs. For instances, to select a more specific label.
5) There are some experiments which were performed using normal cells.
6) Most of the cells are different kinds of cancer cells, which you will have to decide yourself as to what and how you are going to select and train your model.

The project is to be divided into two parts.
1) Using classical approach, such as PCA, SVM or Logistic regression (together with L1 norm, for instance), to reduce dimension and train a classifier;
2) Using deep learning to do the same

For deep learning, you will need to do some pre-learning by selecting a more appropriate model via some kind of encoding-coding with a fully-connected multilayer structure (ideally four layers). Note it is important to use L2 at the input layer, then using L1 to control the model complexity. Then using the model to further learn a deep structure by supervised regression.

Since the features do not have any locality or direct relationship, no convolutions are necessary.
It will be desirable if we could compare the results between the classical approach vs. the deep learning.

lease also feel free to contact me if have any questions by schedule an appointment via WeChat (yuanbo1960).
I will try my best to meet with you as soon as I can.

Prof. Bo Yuan, Ph.D