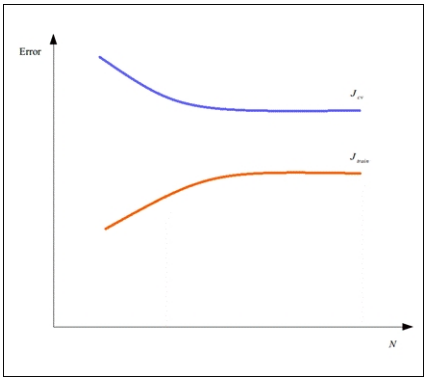
BIL470/570 Machine Learning
HW - 2
Assigned: 26.02.2019
Due: 10.3.2019
Submission: Will be announced on piazza.
Regulations: Late submissions are not allowed. Plagiarism is strictly forbidden, all that take part will be punished according to the regulations of the university.

In this assignment you are going to experiment the usage of random forests, neural networks, support vector machines and learning curves. First download the dataset we provide from the Resources section on piazza. It is a breast cancer classification dataset with $n = 569$ and $p = 30$. The first attribute (column) is the ID, (you can ignore it), the second one is class (M : malignant, B : benign) and the following 30 columns are the attributes.

A learning curve is the plot of Error vs training set size (N). It looks like the following (for the ideal case) :



where the curve labeled with J_{CV} (the blue curve) is the true error and the other one, J_{train} is the training error. Practically, we can not compute the exact true error, therefore we are going to use the cross validation (CV) error. This is the mean error (accuracy in our case) of 10 fold CV. So during performing CV for a particular sample size, compute both the training error and the test error for each fold and at the end average them. Training error average and test error average correspond to two points on J_{train} and J_{CV} respectively.

In this assignment you are going to plot the learning curves of the methods given above. You have to use phyton scikit-learn library for the models.

First, optimize the parameters. For the neural network, choose the best architecture that you could find. Parameter optimization should be done by grid search; Cross validation on training data to find the best combination of parameters. Then, based on the chosen parameters, draw the learning curves. Use matplotlib to draw the curves.

This plot can be used to infer about the bias and variance of a given method. (Think how?) You will compare the bias and variance of the methods in terms of the learning curves. Write a report of at most two pages that include the plots and your comments on the results. Which methods have high/low bias and variance, why? Were you expecting such a curve before, etc.

You will submit your code and report as a Jupyter notebook. Use and follow piazza for questions, comments and updates.