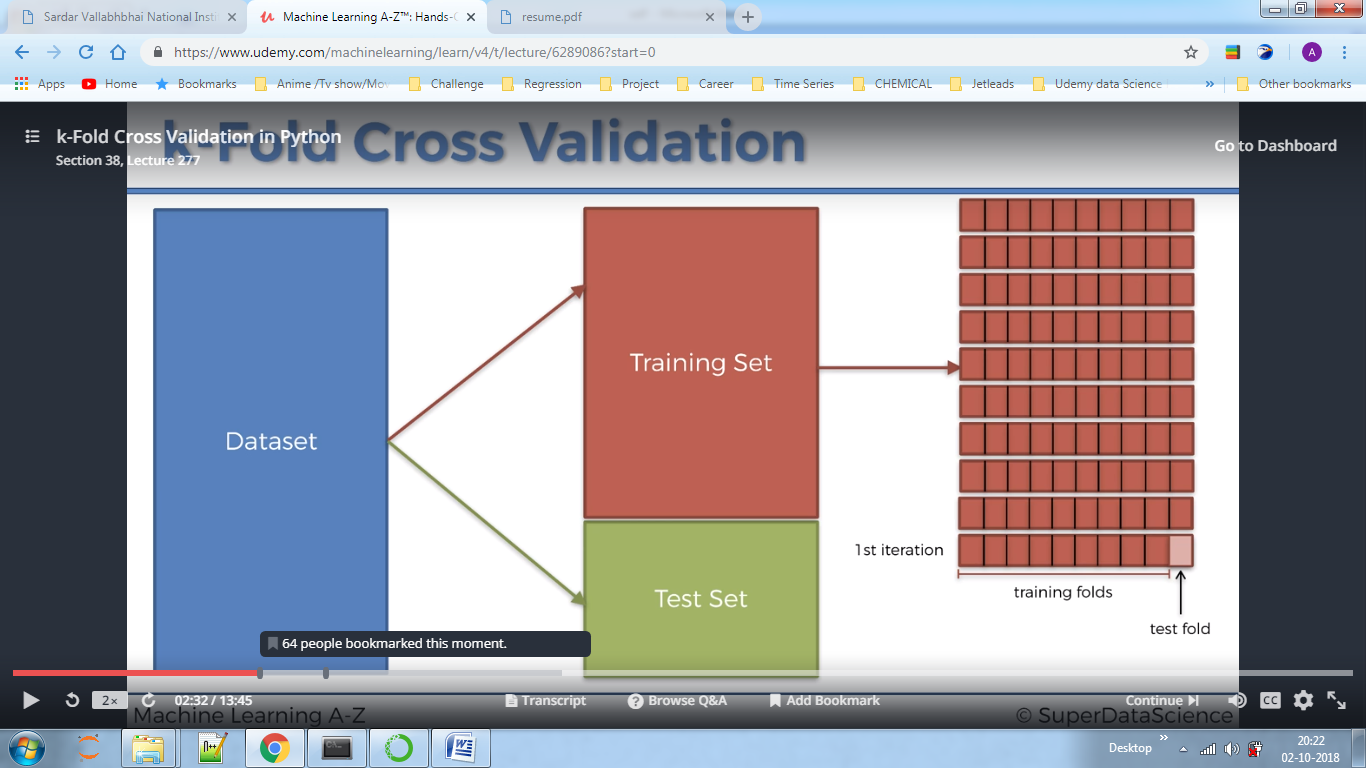
Model Selection is a way to improve our model. There are two types of parameters in a model ;1st parameter are ones which model learns by iterating itself, then comes hyper parameters which we set ; for example such as kernel=rbf(Radial Basis Function).

We use Grid Search to select best hyper-parameters for our model.

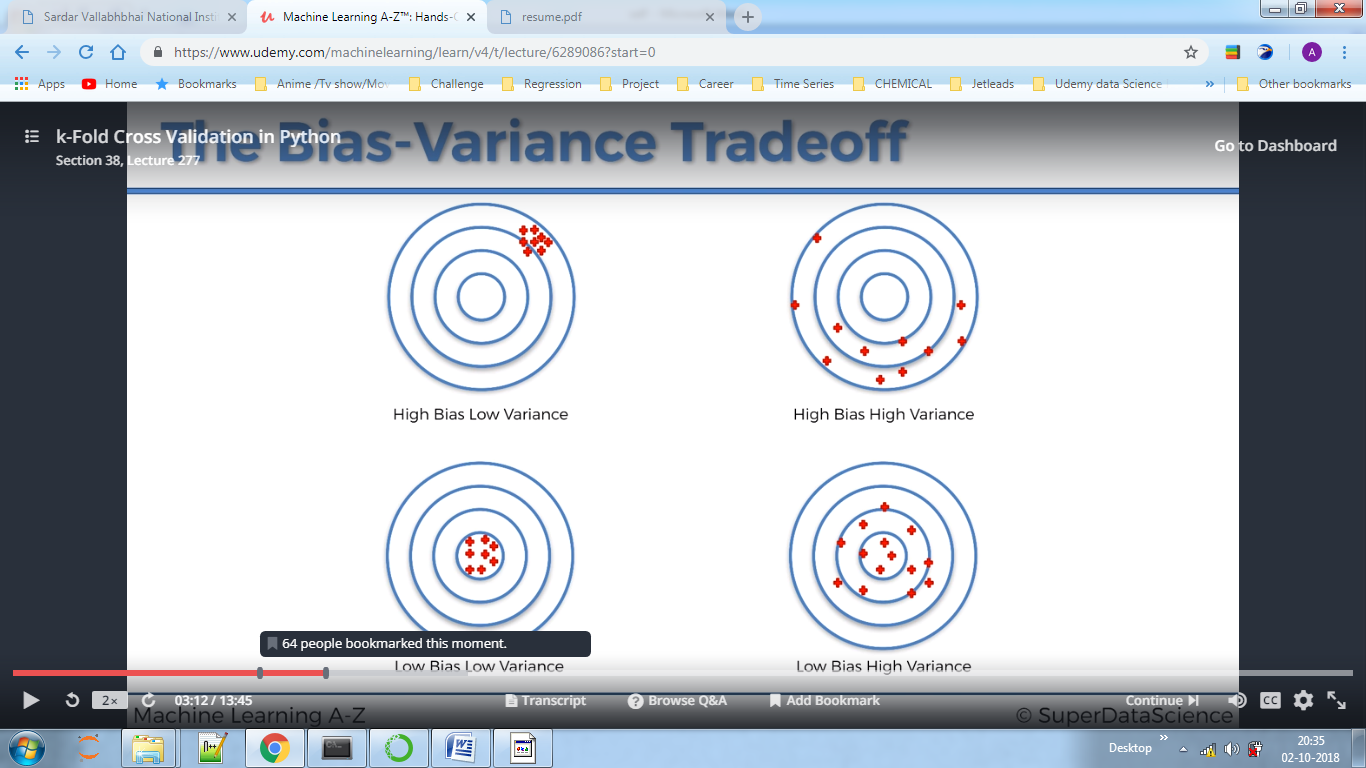
Before using Grid Search we need to use Cross-Validation.

WHY CROSS VALIDATION

Usually we divide model into test set and train set. There are many ways to divide our model into training and test set, for each training set we get different results and our accuracy is different. This difference is called variance. We avoid this using cross validation (below) and check Bias- variance tradeoff inPicture 2

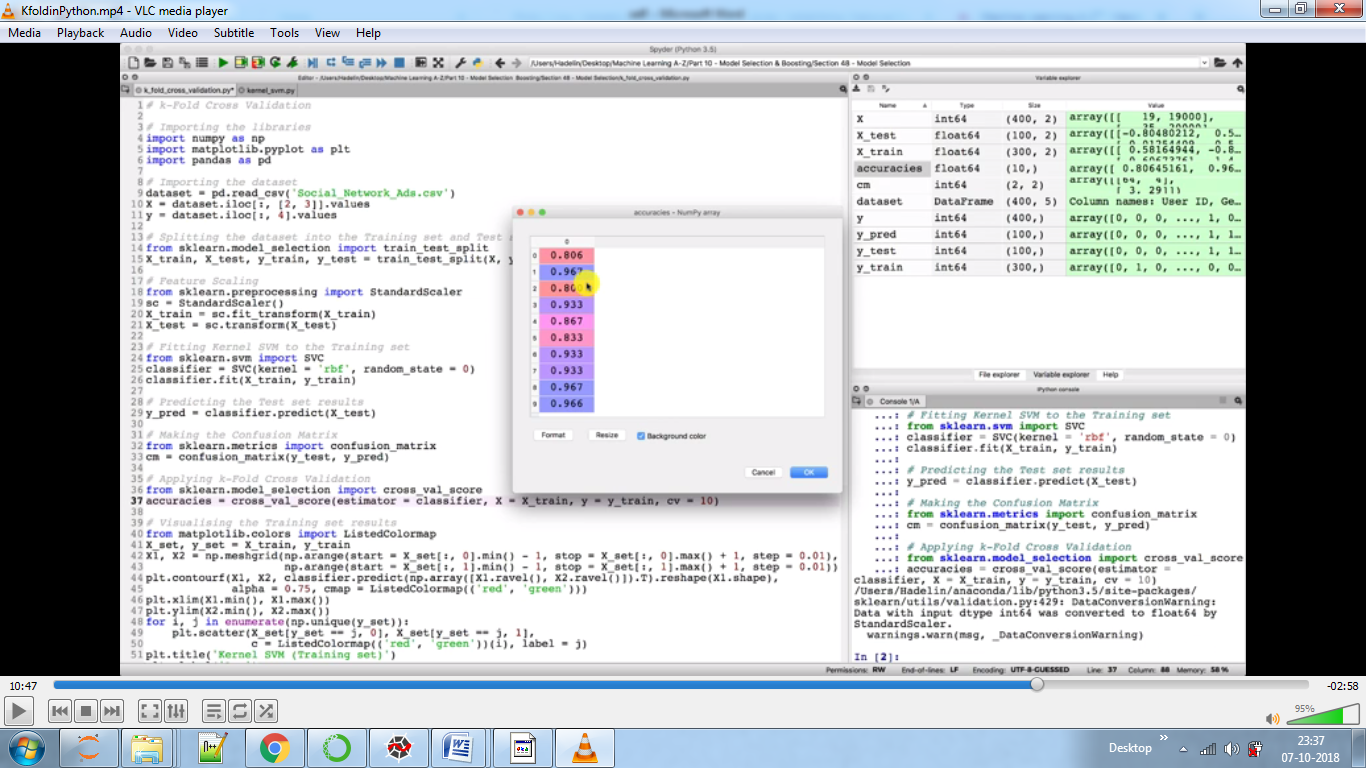


Picture 2



Below we can see 10 accuracies of each of the 10 models.

Here we just proofed that each model having different training set gives different accuracy. We take mean of all these values and conclude that mean as the average accuracy of the model.



GRID SEARCH

It allows us to decide which model to choose –linear(e.g. SVM) or non-linear(e.g. Kernel SVM).

VERY IMPORTANT-In import GridSearchCV section we set parameter = a list of dictionaries.

HOW TO SET UP DICTIONARY?

Here as we used Kernel SVM so we will see which parameters we can optimise for that model by using documentation section. Here we choose C(penalty parameter),Kernel(linear or non-linear),wont choose degree as we don’t want to see which of the non-linear, we are only dealing here for linear or non-linear, we will also choose gamma.

Now we build dictionary where key will be these three and we will provide list which contains elements to check for that parameter using our knowledge. Say as we know for C if we will increase it very much, over fitting might be the issue. We choose here C value to be 1,10,100,1000 only. For gamma we choose 0.5,0.1.0.01,0.001,0.0001.Check code on how to write this dictionary.

Setting only two keys (check code) will tell us whether the problem is linear or non-linear.

Here njobs parameter is used if the data is large.

After running we found that the optimum parameters were C=1,gamma=0.5 and kernel=rbf,now we try to optimise this more by changing numerical parameter like gamma. Instead of setting it to 0.5,0.1.0.01,0.001,0.0001 we set it to 0.2,0.3,0.4,0.5,0.6,0.7,0.8 as we found that optimum value is near 0.5 previously.

By running again the Grid search we found 0.7 to be accurate value for gamma.