**11482 Pattern Recognition and Machine Learning  
11512 Pattern Recognition and Machine Learning PG**

**Tutorial 3 – Week 4 Discussion Questions (Lab is attached separately)**

1. Discuss what is meant by ‘curse of dimensionality’ How can this be address in learners?
2. Variance and Bias are 2 important concepts in describing the behaviour and performance of a model. They represent different sources of error that can affect a model’s ability to generalise well to new, unseen data. Explain these concepts and discuss how this needs to be considered in the building of models.
3. Discuss ‘feature selection’? Why is it necessary? Some major approaches to be covered in lectures (from Scikit-Learn) include – i. Variance threshold, ii. Univariate, iii. Recursive, iv. SelectFromModel, and v. Sequential feature selection. Discuss each and identify the constructs available in scikit=learn package.
4. Explain the differences between linear regression, and linear/non-linear classification models. What is a multiclass classification problem?
5. Accuracy of learning models may be determined by resampling/data-split techniques as below:
   1. Train and Test Sets,
   2. k-fold Cross-Validation,
   3. Leave One Out Cross-Validation (leave 1 observation out for testing for training outcomes from the rest, use MSE to evaluate through mean/variance of all tests)
   4. Repeated Random Test-Train Splits

Discuss each of the above techniques. How would you decide which technique would be best for the evaluation of your learning model?

1. Why can’t you prepare your machine learning algorithm on your training dataset and use predictions from this same dataset to evaluate performance?
2. What is over-fitting, under-fitting? How can they be addressed?
3. If you are working with training data that is full of errors and outliers, will this make it hard for the system to detect patterns? If so, what would you need to do to address this problem, so the system works properly?

Some guide for lab exercise on MNIST code - As you work through the lab code find the following

a). is the algorithm ‘supervised’ learn? Why?

b). Why do we need reshaping as in step 7: reshape(28,28)?

c). how many classes are generated?

d). what does X.shape give (step12)?

e). study the confusion matrix – step 19. Give an interpretation of the diagonal and non-zero other entries.

f). how are misclassified images (step 26) fed back to improve the classification?

g). are there any model parameters used that can give better performance?

h). Are there any hyperparameters used that can be further tuned?