BITS F464 - Machine Learning
II Semester 2019-2020
Assignment #2
Weightage: 15%
Due Data: 29th May, 2020

---

1. Active Learning
   (a) Convert a supervised learning problem (multiclass classification – having more than 3 classes) into an Active Learning problem by randomly removing class labels of data points (retain only 10% of labelled points). The removed labelled points will work as a human oracle!!
   (b) Use stream-based and pool-based scenarios to label additional 10%, 20%, 30%, 40% data points.
      i. use uncertainty sampling (Least confident, Margin Sampling, & Entropy) to label points. Compare the three measures of informativeness.
      ii. Use QBC (Vote Entropy & KL divergence) with at least 5 committee members, to label points. Compare the two measure of disagreement.
      iii. What is the size (number of points) of the version space? Order points to label in such a way that the version space gets reduced by maximum with each point chosen to be labelled.
      iv. Incorporate the additional labelled points (separately, the best from i & ii) into your model and compare with corresponding models trained with randomly chosen labelled points. Also, compare with stream-based scenario.
      v. From 90% of unlabelled points, randomly pick 40% of the points and use clustering (using K-means with K=number of class labels). In each cluster, randomly label 20% of the points to label remaining points in the cluster. How accurate is the cluster-based labelling? How much saving it results in if each label costs you Rs. 100 and each labelling takes one hour.
2. Implement SOM and apply it to one real life problem involving high dimensional data. Describe your SOM architecture and plot your results in a 2D hexagonal grid.

Navneet Goyal
goel@

Please note that the weightage of Assignment #1 has been increased from 10% to 15%.