Mining when the Stationary Distribution Assumption is Violated

So far we have assumed the stationary distribution assumption. Now we will examine how to mine when this assumption is violated and touch upon which mining techniques are sensitive to violations of this assumption. We will begin with the simplest situation: 1) that the test set is different than the training set (Zadrozny paper). We then move onto a more complicated situation 2) that the concept to mine "drifts" based on time (Fan and Street Papers) and finally, we move onto the most complex situation, 3) the test set changes adverserially based on the previous predictions of the learner (Domingos paper).

Read the papers and think about when these situations occur as they most certainly will when you apply classification mining techniques in practice. Situation 1) was motivated by mining of hospital patients records and application of the subsequent records to the entire population. Situation 2) was motivated by mining of ATM streaming traffic information. Finally, situation 3) was motivated by trying to predict various types of fraud (see below)

04/05/05 - Learning With Fixed Sample Bias
04/07/05 - Learning With Concept Drift -1
04/12/05 - Learning With Concept Drift -2

Mining On Distributed Resource Limited Platforms

So far we have assumed a single processor and a computationally “rich” platform. In these papers we explore distributed but computationally limited platforms.

04/14/05-04/19/05 – Distributed Mining on Sensor Networks - 1
04/14/05-04/19/05 – Distributed Mining on Sensor Networks - 2

Constrained Clustering

We covered clustering under constraints for non-hierarchical clustering. We can now cover some more advanced topics with clustering under constraints

04/19/05 – The Most Informative/Best Constraints To Select

Progressively Increasing the Size of the Training Data Set

Until now, we have assumed that all of the data available will be mined. However, most mining algorithms following a behavior that is colloquially known as the learning curve phenomenon that can best be thought of as the diminishing effect on improving model generalization error by adding more training data. One simple solution is to mine a random sample, but how big should the sample be? This work looks at progressively increasing the sample size until the “break-even” point is reached.
04/21/05 – **Efficient Progressive Sampling** (This lecture will be by Ashwin)

Privacy Preserving Data Mining

When we mine data sets, we do not wish any patterns found to be able identify individuals/entities in the data set. How can we ensure this does not occur?

04/27/05-05/03/05 **Privacy Preserving Data Mining**

Mining when the Stationary Distribution Assumption is Violated

04/25/05 - **Classification when the environment is adversarial**