CCSU
DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, April 29
2:00 – 3:00 pm in MS 101

DECISION TREES IN DATA MINING

Daniel Larose

Central Connecticut State University

We begin with the general description of a decision tree. Node "purity" is introduced, leading to a discussion of how one measures "uniformity" or "heterogeneity". Two of the main methods for measuring leaf node purity lead to the two leading algorithms for constructing decision trees, discussed in this talk, Classification and Regression Trees (CART), and the C4.5 Algorithm. The CART algorithm is walked-through, using a very small data set for classifying credit risk. The goodness of the results is assessed using the classification error measure. Next, we turn to the C4.5 algorithm. The C4.5 uses the concept of information gain or entropy reduction for selecting the optimal split. The resulting decision tree is compared with the CART model. The generation of decision rules from trees is illustrated. Then, an application and comparison of the CART and C4.5 algorithms to a real-world large data set is performed, using Clementine software. Differences between the models are discussed. The material for this presentation is drawn from Chapter Six of "Discovering Knowledge in Data: An Introduction to Data Mining" (Wiley), by Daniel Larose.

For further information:

gotchevi@ccsu.edu (860) 832-2839