CCSU
DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Friday, October 7
2:00 – 3:00 PM
Maria Sanford, Room 101

INTRODUCTION TO THE K-MEANS CLUSTERING ALGORITHM

DANIEL LAROSE

CENTRAL CONNECTICUT STATE UNIVERSITY

ABSTRACT

We begin with a review of the overall clustering methodology in data mining, including the concept of distance. Good clustering algorithms seek to construct clusters of records such that the between-cluster variation is large compared to the within-cluster variation. Next, we turn to the *k*-means clustering algorithm, beginning with the definition of the steps involved in the algorithm. Cluster centroids are defined. The *k*-means algorithm is walked-through, using a tiny bivariate data set, showing graphically how the cluster centers are updated. A conjecture is made regarding the interpretation of the root mean square error for the clustering solution. An application of *k*-means clustering to a large *churn* data set is undertaken, using *SAS Enterprise Miner*. The resulting clusters are profiled. Finally, the methodology of using cluster membership for further analysis downstream is illustrated, with the clusters identified by *SAS Enterprise Miner* helping to predict churn. The presentation consists of material from Chapter 8 of *Discovering Knowledge in Data: An Introduction to Data Mining* (Larose, 2005, John Wiley and Sons). The talk includes an update on the latest happenings in the Data Mining program at CCSU.

The talk is easily accessible to anyone with an interest in the analysis of data.

AFTERMATH:
Refreshments will follow the colloquium at Castaneda's
(1590 Stanley St. – across from the administration building)

For further information:

gotchevi@ccsu.edu 860-832-2839
castanedan@ccsu.edu 860-832-2851