# CSCI 6333 Data Mining & Warehousing

**Module 5: Anomaly Detection, Avoiding False Discoveries**

**Homework Assignment Four**

**All problems are equal-weighted with 20 points each.**

1. Consider distance-based outliers with respect to two parameters and : A data point is an outlier if the number of points within a radius to is less than . Explain in detail one major weakness of this formulation.
2. Consider the density-based outliers. Describe in detail the time complexity of finding local outlier factors (LOFs) for all data points in a set of points. Assume that the dimension is and the parameter k is given.
3. Consider the connectivity-based outliers. Describe in detail the time complexity of finding connectivity outlier factors (COFs) for all data points in a set of points. Assume that the dimension is and the parameter k is given.
4. Explain in detail two different methods to use k-means clustering to detect outliers.
5. Consider the problem of determining whether a coin is a fair one, i.e. P(heads)=P(tails)=0.5, by flipping the coin 10 times. Use the binomial theorem and the basic probability to answer the following questions.
   1. A coin is flipped ten times and it comes up heads every time. What is the probability of getting 10 heads in a row and what would you conclude about whether the coin is fair?
   2. Suppose 10,000 coins are each flipped 10 times in a row and the flips of 10 coins results in all heads, can you confidently say that these coins are not fair?