CSCI 4502/5502: Data Mining
Homework 5

Due at **12:30pm on Thursday, April 16, 2015**. Submit one file electronically at moodle: "**LastName_FirstName_Homework5.pdf**".Make sure to include your name, student id, and the Honor Code Pledge (`http://honorcode.colorado.edu/student-information/honor-code-pledge`).

1. Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. Also, suppose one-fifth of the college students are graduate students and the rest are undergraduates.
   (a) What is the probability that a student who smokes is a graduate student?
   (b) Is a randomly chosen college student more likely to be a graduate or undergraduate student?
   (c) Repeat part (b) assuming that the student is a smoker.
   (d) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.
2. Consider the following set of one-dimensional points: {93, 137, 48, 71, 162, 119}. Use the *k-means* clustering method to generate two clusters, assuming the initial centroids are 100 and 150, respectively.
3. Provide an application example for each of the three types of outliers: global outliers, contextual outliers, and collective outliers. Briefly describe the related attributes and how the specific type of outliers may be defined and detected.
4. Briefly describe one data mining tool that you have used either in this course or in other settings. What did you use this tool for? What are the key strengths and possible limitations of this tool?