Assignment 1 (Knowledge Discovery and Data Mining)

Number of problems/points: Ten problems for total of 100 points
Out: January 18, 2018
Due: January 25, 2018 in class (hard copy).

Homework Policies (applicable for all 9664 assignments):
1. You are required to do the homework problems in order to pass.
2. Understandability of the solution is as desired as correctness.
3. Penalty for late homework assignments submissions is 20% per day. So, do it on time.
4. Solutions are expected to be your own work. Group work is not allowed unless explicitly approved for a particular problem. If you obtained a hint with help (e.g., through library work, discussion with another person, etc.) acknowledge your source, and write up the solution on your own. Plagiarism and other anti-intellectual behavior will be dealt with severely.

Attend Jan. 19 CIS colloquium at 11:00am-12:15pm in SERC 306. Then answer problems 1-3.
To answer remaining problems read section 1 in the textbook if needed.

Problem 1: (10 points)
How is theory-guided data science different from alternative data mining approaches.

Problem 2: (10 points)
Summarize briefly one of the applications presented by the speaker.

Problem 3: (10 points)
Summarize new content included in the 2nd edition of the “introduction to data mining” textbook by Tan P.N., Steinbach M., Kumar V. and Karpatne A. vs. the 1st edition by Tan P.N., Steinbach M., and Kumar V.

Problem 4 (10 points). An analyst sets up a sensor network in order to measure the temperature of different locations over a period. What is the data type of the data collected?

Problem 5 (10 points). The same analyst as discussed in problem (4) above finds another database from a different source
containing pressure readings. She decides to create a single database containing her own readings and the pressure readings. What is the process of creating such a single database called?

**Problem 6 (10 points).** An analyst processes Web logs in order to create records with the ordering information for Web page accesses from different users. What is the type of this data?

**Problem 7 (10 points).** Consider a data object corresponding to a set of nucleotides arranged in a certain order. What is this type of data?

**Problem 8 (10 points).** It is desired to partition customers into similar groups on the basis of their demographic profile. Which data mining problem is best suited to this task?

**Problem 9 (10 points).** Suppose in problem (8), the merchant already knows for some of the customers whether or not they have bought widgets. Which data mining problem would be suited to the task of identifying groups among the remaining customers, who might buy widgets in the future?

**Problem 10 (10 points).** Suppose in problem (9), the merchant also has information for other items bought by the customers (beyond widgets). Which data mining problem would be best suited to finding sets of items which are often bought together with widgets?