CS 432/536 Data Mining **Assignment 3** **Spring 2018**
Total Marks: TBA **Due Date: TBA**

1 Question 1: Classification and Clustering

Use Python or R for this question.

1.1 Datasets

You will be using the following dataset, download and study them in detail: Iris Dataset

1.2 Classification

The Iris dataset contains 3 classes of 50 instances each, where each class refers to a type of an Iris plant. The central goal here is to design a model which makes good classifications for new data, in other words one which exhibits good generalization.

Note: You need to do data analysis and show your findings with the help of different graphs. You also need to show your evaluation with the help of graphs.

1.3 Evaluation

You will need to report performance metrics (confusion matrix, accuracy, precision, recall) for the dataset using:

(a) 80:20 train-test split

(b) 10-fold cross validation

2 Question 2: Classification Tree

Construct a classification tree, of the Iris dataset, up to depth 3 (root level, intermediate level, and leaf level) using information gain as splitting criterion. Label the leaves with the probability of each class. Show your working.

3 Question 3: Theoretical

Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.