Cs434 Assignment 1
Due: Monday Oct 13th in class

Part I. (14pts)
Through out the course, Weka will be a very useful tool for you to explore different machine learning and data mining algorithms. The purpose of this part of the assignment is to familiarize you with this software package. For this assignment, please download and install the Weka software package (version 3.4) from

http://www.cs.waikato.ac.nz/ml/weka/

If you don't have access to a computer for this purpose, please inform the instructor and special arrangement can be made to accommodate your need.

Basic information on this software package can be found in the following tutorial:

http://easynews.dl.sourceforge.net/sourceforge/weka/ExplorerGuide-3.4.pdf

Note that there are many documents available on the Weka webpage introducing different aspects of Weka. For example, another useful thing to read is the following document, which describes the input format for Weka, i.e., the "arff" (attribute relation file format) format.

http://www.cs.waikato.ac.nz/~ml/weka/arff.html

Please use Weka to explore the "iris" data set that comes with the software. To open this data set, choose "explorer" from the Weka GUI chooser, which opens a panel with several tabs. Select the "preprocess" tab and click "Open file", then click on the "data" folder, choose the "iris.arff" file.

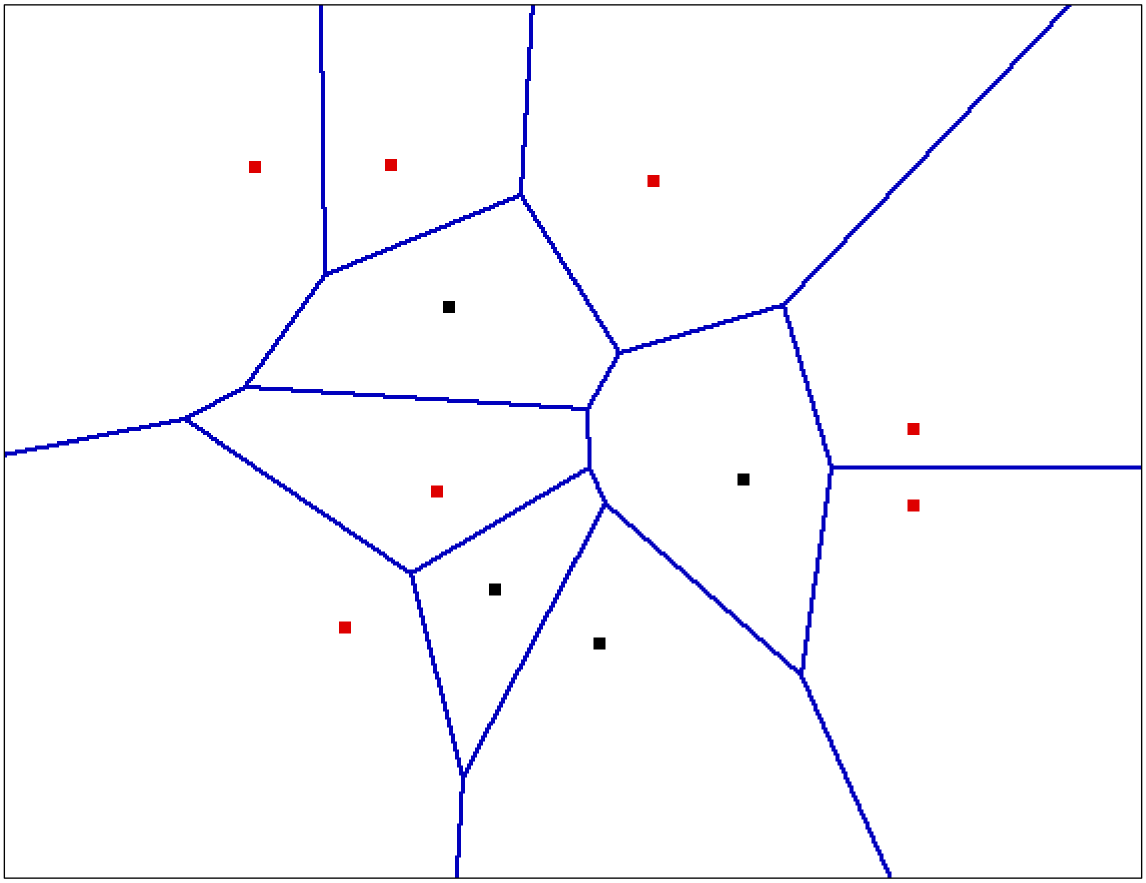
With the help of the Weka software, answer the following questions:

1. How many classes there are in this data set? (2pts)
2. How many (non-class) attributes there are? (2pts)
3. What are the mean and standard deviation for each attribute? (2pts)
4. If you were to choose only one attribute to build your classifier, which attribute should you choose? (4pts)
5. Which pair of attributes provides the best discrimination among classes? (4pts) (Suggestion: use the visualization tool to look for good separations among classes.)

Note that questions 4 and 5 are subjective; please provide your reasons for the answers. Reasons could be describe in words and/or shown through figures.

Part II

1. Below is a set of 2-d data points, with black dots representing positive class and red dots representing negative class. The blue line segments show the Voronoi diagram of these points. (14pts)
   a. What is the training error of 1-nearest neighbor? (2pts)
   b. What is the training error of 3-nearest neighbor? (4pts)
   c. Please mark out the 1-nearest neighbor decision boundary for this data set, which should be a subset of the blue line segments. (4pts)
   d. Now consider 3-nearest neighbor, **true or false**: the decision boundary of 3-NN is also formed by a subset of these blue line segments, but a different subset from the answer of (a). Explain your answer. (4 pts)



The Perceptron Algorithm

```
Let $\mathbf{w}_0 \leftarrow (0,0,0,...,0)$
$\mathbf{t} \leftarrow \mathbf{0}$
Repeat until all training examples are correctly classified
    Accept training example $i : (\mathbf{x}^i \ , y^i)$
    $u^i \leftarrow \mathbf{w} \cdot \mathbf{x}^i$
    if $y^i \cdot u^i \ <= \ 0$
        $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y^i \mathbf{x}^i$
        $t \leftarrow t + 1$
```

2. (15pts) Consider the following training set of three points (-1, -), (1, -) (4, +). Apply the above illustrated Perceptron algorithm to these three points (accepting the training examples in the given order) and learn a straight line separating positive from negative examples. To do this, you are required to (1) transform the data into the canonical representation by adding a bias attribute 1 to each example; (2) record the intermediate and final weight vectors starting from w_1, and plot each corresponding decision boundary (using the 2-d space of the canonical representation for plotting the lines). Please use a separate plot for each decision boundary.