**20CS6037 Machine Learning**

**Fall Semester 2015**

**Assignment 3**

**Assigned 10/6/2015**

**Due 10/14/2015 on BB 11:59PM**

In this assignment you are to implement in Matlab the following concept learning algorithms :

* FindS
* ListE (list-then-eliminate)
* CE (Candidate\_Elimintation.

You will apply each of these algorithms on the Iris dataset.

However, your program should be general enough so that it can be applied to other labeled data sets in which a data point is of the form

[attr1 attr2 attr3 …. attrn class]

A complete description of the data set is available at <http://en.wikipedia.org/wiki/Iris_flower_data_set>.

The dataset can be downloaded from <http://archive.ics.uci.edu/ml/datasets/Iris>.

The Matlab command **load fisheriris** will load it into your workspace.

Again, the iris data set describes three types of iris flowers: Setosa, Versicolor, and Vriginica.

**You will want to learn each class separately:** for example, suppose we want to learn Setosa. Then each example of one of the two other classes is a negative example for Setosa.

**Example of data from the iris dataset**

**Setosa**

5.1 3.5 1.4 0.2 **1**

4.9 3.0 1.4 0.2 **1**

4.7 3.2 1.3 0.2 **1**

4.6 3.1 1.5 0.2 **1**

5.0 3.6 1.4 0.2 **1**

**Versicolor**

5.6 3.0 4.1 1.3 2

5.5 2.5 4.0 1.3 2

6.1 3.0 4.6 1.4 2

5.8 2.6 4.0 1.2 2

5.0 2.3 3.3 1.0 2

**Virginica**

6.3 3.3 6.0 2.5 3

5.8 2.7 5.1 1.9 3

7.1 3.0 5.9 2.1 3

4.9 2.5 4.5 1.7 3

7.3 2.9 6.3 1.8 3

6.7 2.5 5.8 1.8 3

You will need to select at random a subset of data from each class and use that for training. Then you will use the remaining data points for test purposes.

The attributes are real-valued, so you will need to discretize them, that is, to divide the range of an attribute into intervals (and eventually label them).

One way to discretize the data for this assignment is as follows:

1. For each attribute, collect all the values of that attribute that appear in training examples, order them.
2. Identify uninterrupted runs of values that correspond to one class only; (If a value appears in more than one class, consider for it the class to which it is most often associated with.)

These runs become now the attribute values: that means that, for example, for sepal length, instead of values such as 5.2 you will have as value an interval [5.0, 5.7].

In the generalization hierarchy, allow for intermediate values:

* Union of intervals to generalize two different intervals to obtain more general value.
* Intersection of intervals to obtain more specific attribute value

Matlab functions you may need: **unique, intersect, setdiff, setxor, ismember**

**What to turn in**

Your assignment should be uploaded on Blackboard.

Please turn in one file only! This file should be a Matlab script (.m file) with the name **NameAss2.m. Each team member must upload their own file. This file must contain at the top the names of all team members.**

In addition to the Matlab code (which I should be able to run) this file must contain example runs included as comments.