**Instructions Lab 2\_1**

In the last lab you learnt to do some basic visualization of the data. Today you will work on a simple idea. Let us say we have some data available as a table. Each sample (row) has a feature vector of length D (i.e. a D dimensional feature space) and a class label. Assume for the time being that we have only two possible classes. The assignment for today is that you need to find that particular feature which gives the best separation between the two classes.

To achieve this, consider the breast cancer data that we used in the previous lab.

There are 9 features and each feature can take values from 1 – 10.

Let us say that we choose the i-th feature f(i) (1 <= i <= 9).

For the chosen feature take a value, say V (1 <= V <= 10).

Now consider all the samples in the table.

For the l-th sample, if f(l, i) < V then put the sample in Set1(i, V) else put in Set2(i, V).

Repeat for all samples.

Now measure the clean-ness of Set1( ) and Set2( ). We need some definition of clean-ness. So, for this assignment, we will use the following definition:

One way of defining clean-ness of a set is as the frequency of its most common constituent. For example, if a set consists of 70% of items in class “2” and 30% in class “4” then its clean-ness is 70%.

Now define the clean-ness of a tuple (i, V) as:

Clean (i , V) = Clean-ness of Set1(i, V) + Clean-ness of Set2(i, V)

Measure Clean(i,V) for all possible values of (i, V) and report the point with the highest clean-ness value.

The following example will help to make things a little more clear:

Say we have the following samples:

1000025,5,1,1,1,2,1,3,1,1,2

1002945,5,4,4,5,7,10,3,2,1,2

1017122,8,10,10,8,7,10,9,7,1,4

1018099,1,1,1,1,2,10,3,1,1,2

1018561,2,1,2,1,2,1,3,1,1,2

1041801,5,3,3,3,2,3,4,4,1,4

1043999,1,1,1,1,2,3,3,1,1,2

1044572,8,7,5,10,7,9,5,5,4,4

1047630,7,4,6,4,6,1,4,3,1,4

1048672,4,1,1,1,2,1,2,1,1,2

1049815,4,1,1,1,2,1,3,1,1,2

1050670,10,7,7,6,4,10,4,1,2,4

1050718,6,1,1,1,2,1,3,1,1,2

1054590,7,3,2,10,5,10,5,4,4,4

1054593,10,5,5,3,6,7,7,10,1,4

1056784,3,1,1,1,2,1,2,1,1,2

1057013,8,4,5,1,2,6,7,3,1,4

Say, we choose the second feature (i.e. 3rd column) and choose V = 5.

Below is the same data but with Set1 or Set2 written against it. The data in red font has to be compared with V = 5.

1000025,5,1,1,1,2,1,3,1,1,2 – Set1

1002945,5,4,4,5,7,10,3,2,1,2 – Set1

1017122,8,10,10,8,7,10,9,7,1,4 – Set2

1018099,1,1,1,1,2,10,3,1,1,2 – Set1

1018561,2,1,2,1,2,1,3,1,1,2 – Set1

1041801,5,3,3,3,2,3,4,4,1,4 -Set1

1043999,1,1,1,1,2,3,3,1,1,2 – Set1

1044572,8,7,5,10,7,9,5,5,4,4 – Set2

1047630,7,4,6,4,6,1,4,3,1,4 – Set1

1048672,4,1,1,1,2,1,2,1,1,2 – Set1

1049815,4,1,1,1,2,1,3,1,1,2 – Set1

1050670,10,7,7,6,4,10,4,1,2,4 – Set2

1050718,6,1,1,1,2,1,3,1,1,2 – Set1

1054590,7,3,2,10,5,10,5,4,4,4 – Set1

1054593,10,5,5,3,6,7,7,10,1,4 – Set2

1056784,3,1,1,1,2,1,2,1,1,2 – Set1

1057013,8,4,5,1,2,6,7,3,1,4 – Set1

Now consider the elements of Set1 for the chosen values of i and V.

1000025,5,1,1,1,2,1,3,1,1,2 – Set1

1002945,5,4,4,5,7,10,3,2,1,2 – Set1

1018099,1,1,1,1,2,10,3,1,1,2 – Set1

1018561,2,1,2,1,2,1,3,1,1,2 – Set1

1041801,5,3,3,3,2,3,4,4,1,4 -Set1

1043999,1,1,1,1,2,3,3,1,1,2 – Set1

1047630,7,4,6,4,6,1,4,3,1,4 – Set1

1048672,4,1,1,1,2,1,2,1,1,2 – Set1

1049815,4,1,1,1,2,1,3,1,1,2 – Set1

1050718,6,1,1,1,2,1,3,1,1,2 – Set1

1054590,7,3,2,10,5,10,5,4,4,4 – Set1

1056784,3,1,1,1,2,1,2,1,1,2 – Set1

1057013,8,4,5,1,2,6,7,3,1,4 – Set1

Out of the 13 elements, 9 elements belong to class “2” and the remaining belong to class “4”. Thus, clean-ness of Set1(2, 5) = 9/13.

Verify manually that clean-ness of Set2(2, 5) = 1.

Note: Do not use any library except (Pandas) for this assignment. Pandas to be used only for reading the data.

**Show your work, even if it is partial, during the lab hours. What you show during the lab will contribute toward your final grade.**