**502571-3 2nd Trimester 2022/2023 Lab#4**

# Topics: KNN Section: 2233

**Attached file: glass.arff**

# Due Date: Tuesday 31 January 2023 – 9:00 am

# \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**﻿Learning objectives**

In this lab you are going to: IBk, ie Weka implementation of k-Nearest Neighbours

**﻿Datasets:** description

The glass dataset comes from the U.S. Forensic Science Service contains data on six types of glass.

Glass is described by its refractive index and the chemical elements that it contains; the the aim is

to classify different types of glass based on these features. This dataset is taken from the UCI datasets, which have been collected by the University of California at Irvine and are freely available on the Web. They are often used as a benchmark for comparing data mining algorithms.

The ACE (Angeotensin--‐Converting Enzyme) dataset contains protein-­‐ligand binding data. Ligands exhibiting strong binding affinity towards a certain protein being considered as “active” with respect to it. If it is not known about the binding affinity of a ligand towards the protein, such ligand is conventionally considered as “nonactive. The goal of classification models is to be able to predict whether a new ligand will exhibit strong binding activity toward certain protein biotargets. We are interested in it because this dataset is highly unbalanced.

**﻿Task 1:** Warming-up

Load the glass dataset (glass.arf) into the explorer interface.

Q1 : How many attributes are there in the data set ?

**10 attributes**

Q2 : What are their names ?

RI, Na, Mg, Ai, Si, K , Ca, Ba, Fe

Q3 : What are the classes attributes (names) ?

Build wind float, build wind non-float, vehic wind float, vehic wind non-float, containers, tableware, headlamps

**﻿Task 2:** IBk (K-NN)

﻿Now, in the classifier frame, click Choose, then select the IBk method from the lazy submenu. The lazy submenu contains a group of methods, in which the training phase is almost omitted – it actually amounts to memorizing all instances from the training set. Instead of it, all main calculations are delayed to the test phase. That is why such methods are sometimes called lazy, instance-­‐based and memory-­‐based. The price for this “laziness” is however rather high – computations at the test phase are very intensive, and that is why such methods work very slowly during prediction, especially for big training sets. So, the abbreviation IBk means that this is an Instance-­‐Based method based on k-­‐neighbours. **The default value of k is 1**.

Run the classification algorithm lazy IBk (weka.classifiers.lazy.IBk ). Use cross-­‐validation to test its performance, leaving the number of folds at the default value of 10. Recall that you can examine the classifier options in the Generic Object Editor window that pops up when you click the text beside the Choose button. As we said, the default value of the kNN field is 1: This sets the number of neighboring instances to use when classifying

Q1: What is the performance of IBk ? (the number of correctly and incorrectly classified instances)

Correctly Classified Instances 151 70.5607 %

Incorrectly Classified Instances 63 29.4393%

Q2: Do you think that it is possible to enchance the performance of IBk for this dataset ? Try to justify your answer.

One of the ways to enhance the performance is remove any irrelevant or redundant.

There is another way is experiment with different values of k to find the optimal value for the specific dataset.

**﻿**

**Task 3:** Change the K – value

﻿Run IBk again, but increase the number of neighboring instances to k = 5 by entering this value in the KNN field. Run the classifier and create the model.

Q1: What is the performance of IBk with five neighboring instances (k = 5)? Why? look at

The number of correctly and incorrectly classified instances. Any comments about the performance of the 2 models?

Correctly Classified Instances ١٥٠ ٧٠.٠٩٣٥ %

Incorrectly Classified Instances ٦٤ ٢٩.٩٠٦٥ %

It seems to me that it is the same result or can be said that is very similar