ALGONQUIN COLLEGE

CST8390 - LAB
BUSINESS INTELLIGENCE & DATA ANALYTICS

Week 4

LAB 4 - K Nearest Neighbor (kNN)

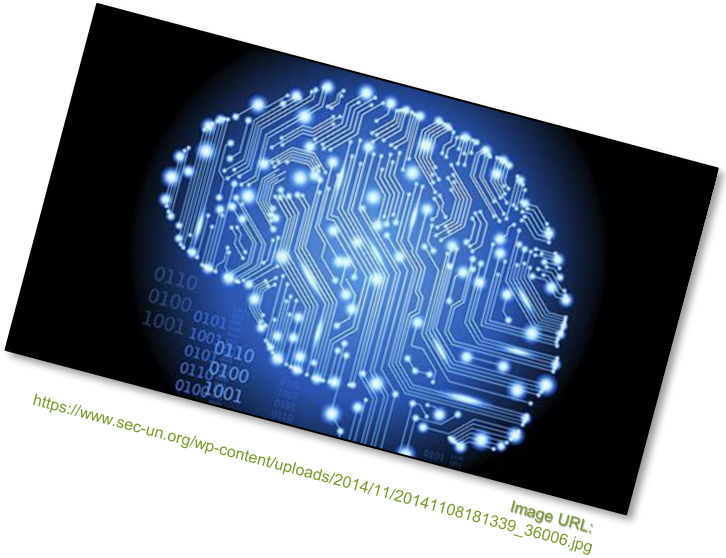
Lab 4 – kNN

PART I

- Reviewing Algorithm

PART II

- Steps
- Results



CST8390 - Lab
Business intelligence & data analytics

Lab 4 – kNN

Part I – Reviewing Algorithm

kNN: Nearest-Neighbor

- **Idea**: Discussion of nearest-neighbor learning;
- Often very accurate
- Assumes all attributes are equally important
- Statisticians have used k-NN since the early 1950s
- kD-trees can become inefficient when the number of attributes is too large.
- Complexity depends on **depth** of the tree, given by the **logarithm** of number of nodes for a balanced tree

Image URL: https://images.idgesg.net/images/article/2018/10/analytics_binary_code_network_digital-transformation-100777428-large.jpg

kNN: Decision Tree

- Amount of backtracking required depends on quality of tree (“square” vs. “skinny” nodes)
- How to build a **good tree**? Need to find good split point and split direction
  - Possible split direction: direction with greatest variance
  - Possible split point: median value along that direction
  - Using value closest to mean (rather than median).
- Can apply this split selection strategy recursively.

Image URL: https://images.idgesg.net/images/article/2018/10/analytics_binary_code_network_digital-transformation-100777428-large.jpg

kNN

- Resources from Lecture:
  - http://sens.tistory.com/277
  - http://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
  - https://www.youtube.com/watch?v=SQOdBjjA2y8
  - https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/

Remember: In Weka, kNN is called by: **Weka > Classify > Choose** and then, *weka > classifiers > lazy > IBk*

Image URL: https://images.idgesg.net/images/article/2018/10/analytics_binary_code_network_digital-transformation-100777428-large.jpg

Weka Introduction

Demo

Step-by-step (A)

I. BASIC OPERATIONS

1. Get “Wine” dataset from https://archive.ics.uci.edu/ml/datasets.html (or https://archive.ics.uci.edu/ml/index.php) and save it as a **CSV** file.

Data is in Wine.data and info is in data.names

Add **attribute** names as the first row in the CSV file

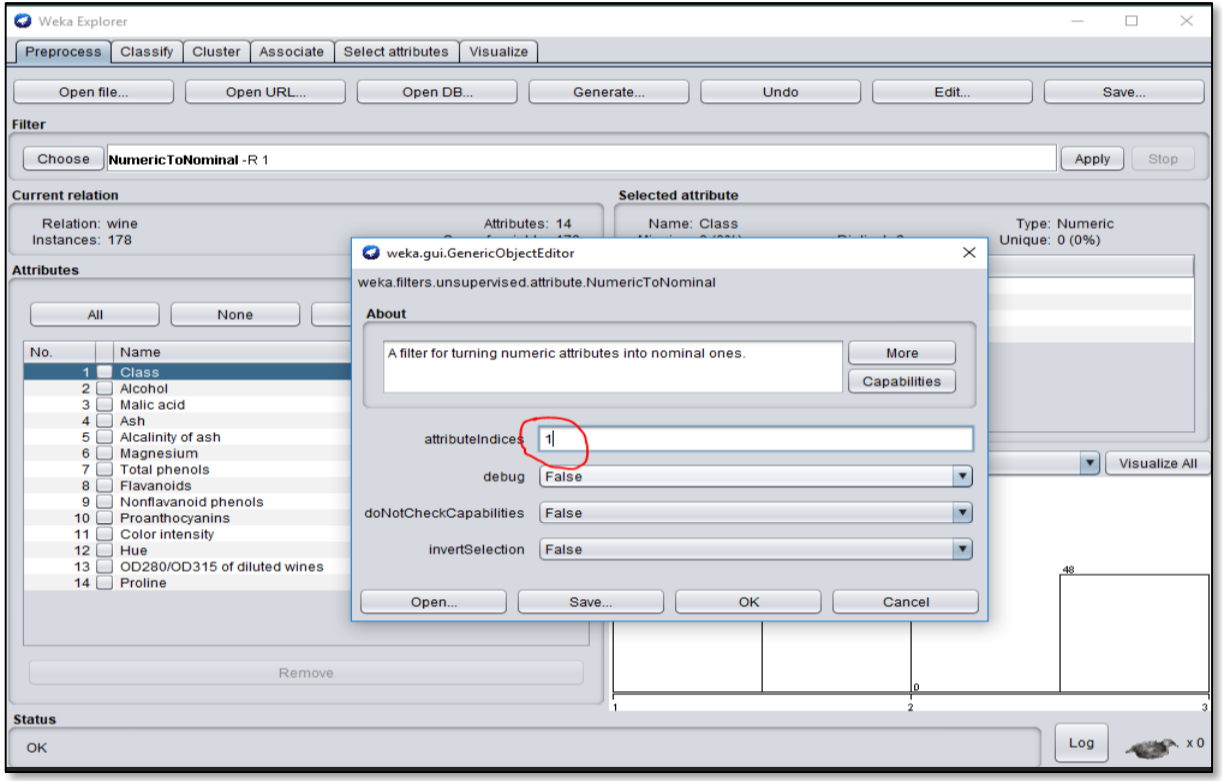
2. Explore and learn about the **relevance** of various attributes of the dataset.

Step-by-step (B)

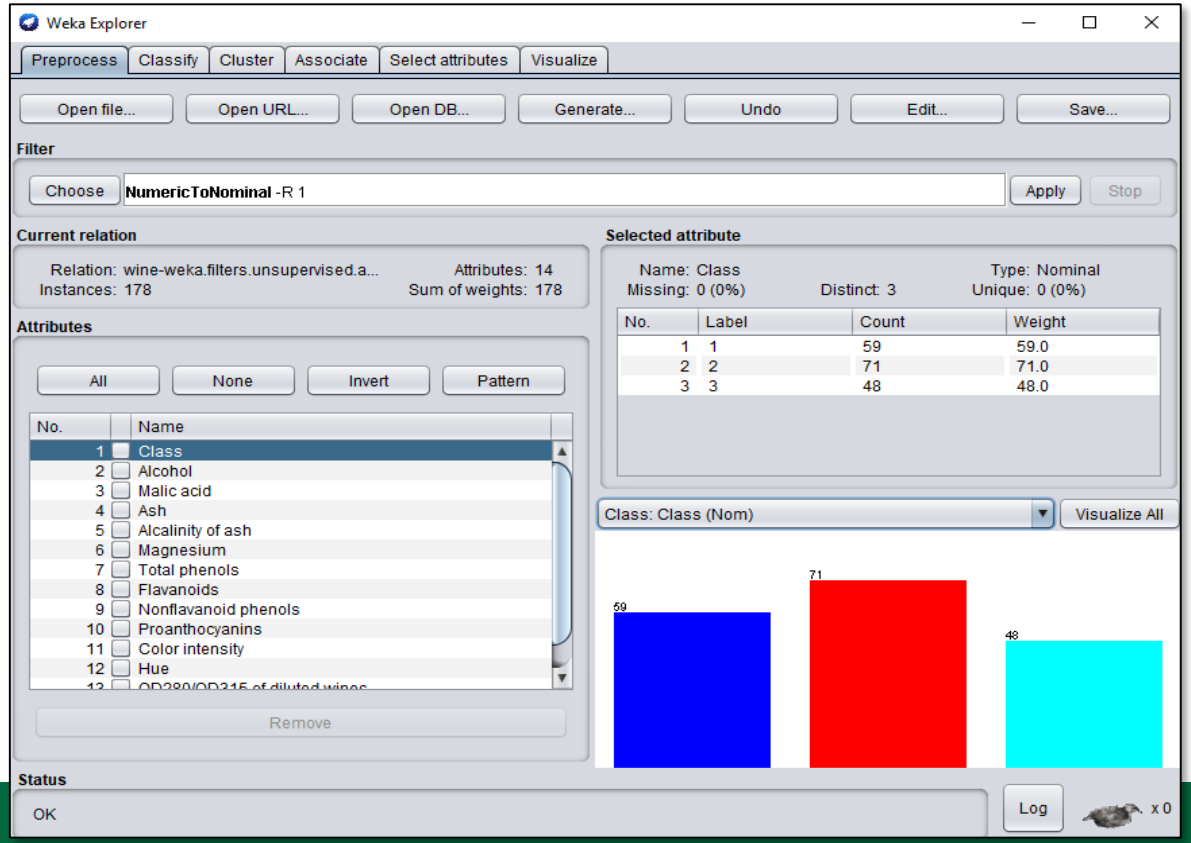
3. Load the file to **Weka**.

4. Check how various attributes are converted in **Weka**. Class is considered as **numeric** instead of **nominal**. Apply filter NumericToNominal to convert class datatype to nominal. When you apply filter, you need to specify the index of the attribute you need to apply the filter.

Step-by-step (C)

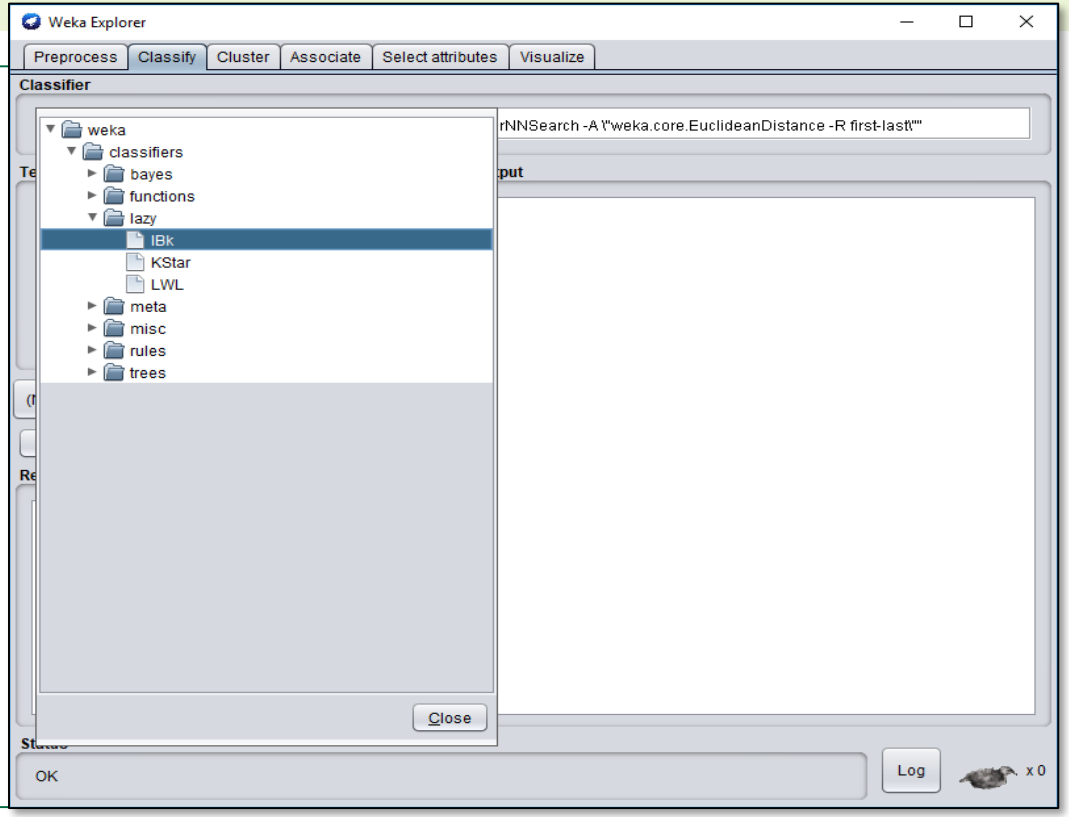


Step-by-step (D)



Step-by-step (E)

5. Now, we need to perform classification using kNN method. For that, click on "Classify" tab. For this lab, we use kNN. For that, choose IBk which is **Instance Based k Nearest Neighbors** from Lazy in the tree view.



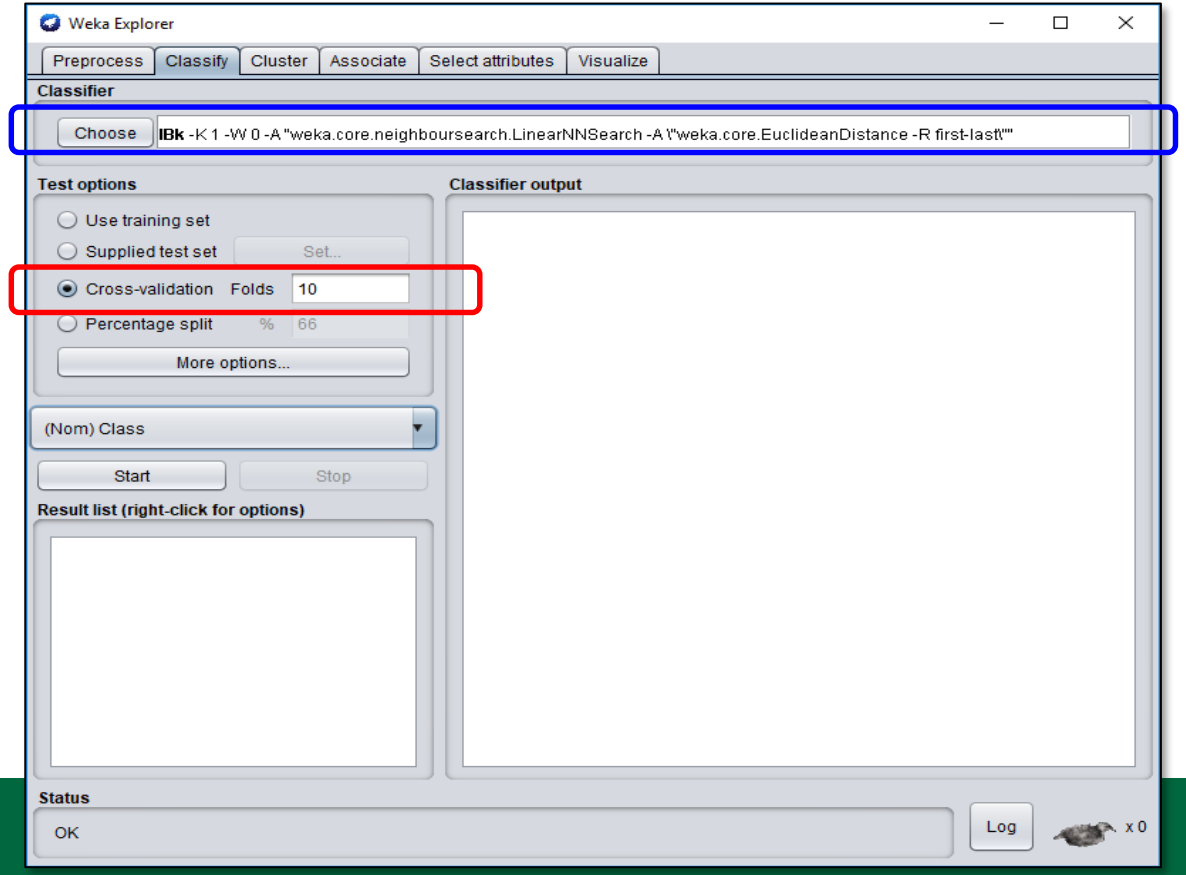
Step-by-step (F)

6. As mentioned earlier, our first attribute is the class label. We need to set that now in the classify panel.

IBk –K 1 –W 0 –A “weka.core.neighboursearch.LinearNNSearch ….”

- This is the **parameter list** for the algorithm.
- Click on this text to set the value of k.
- Set k as 3.
- Close the window.
- Now, set the cross-validation to 10 Folds if it’s not already there.
- Now click “Start” to run the algorithm.

Step-by-step (G)



Step-by-step (H)

7. There should be a lot of text in the right-hand side of the window with the results of the algorithm. Find the line that says "Correctly classified instances".

   (a) What is the **percentage** of correctly classified items?
   (b) What are the **True Positive (TP)** rates of each class?
   (c) Look at the **confusion matrix**, which class is incorrectly classified?

8. Now click on the "Choose" button to modify the number of neighbours that are used in the **kNN** search to 5.

   (a) What is the **percentage** of correctly classified items?
   (b) What are the **True Positive (TP)** rates of each class?
   (c) Look at the **confusion matrix**, which class is incorrectly classified?

Step-by-step (I)

9. Run the algorithm **several times**, always increasing the value of N by two, and always an **odd number**: 1, 7, 9, 11, 13. Each of your tests will be in the window of the lower left. Fill in the following table.

K	percentage of correctly classified instances
1	
7	
9	
11	
13	

Which class is being mis-classified? _____

Step-by-step (J)

10. Repeat **step 9** with "Percentage Split" of 70. Fill in the following table.

K	percentage of correctly classified instances
1	
3	
5	
7	
9	
11	
13	

Step-by-step (K)

REMEMBER:

- Show your answers to the lab professor when you are done (in **Weka** and document).
- This lab has 5 marks so ensure that you have all your answers filled in.

Note: Due Date: Week 5 in corresponding lab sessions.

What about my analysis?

- The previous questions can help you to do your own analysis:

FOR YOUR ANALYSIS:

- *What is the purpose of "confusion matrix"? What is its importance?*
- *Explain with your own words the **kNN** method.*

https://www.marketingdirecto.com/wp-content/uploads/2018/01/ciencia-datos.jpg

Image URL:

Open questions…

- Before we finish, do you have any doubt / question?

Image URL: https://cdn-az.allevents.in/events3/banners/26b150363d5757da8578fa6a1481585a368f12d4247adfe95837e6ea6c5ab2af-rimg-w1200-h549-gmir.jpg?v=1569691155

See you…

- Remember:
  - Labs require practice and it is ok committing errors and learning with them.
  - Do not forget to show your results…
  - Any questions, let me know…

sousap@algonquincollege.com

Image URL: https://thumbs.gfycat.com/MaleFrigidBull-size_restricted.gif

Thank you for your attention!