Waikato Environment for Knowledge Analysis (Weka)

CIS*4500 (Fall 2019)

Weka Introduction

- Weka is a suite of machine learning software written in Java and developed at the University of Waikato, New Zealand

- Weka contains a collection of visualization tools, machine learning algorithms, and data preprocessing utilities along with graphical interfaces to access these functions (the workbench)

- Many extension packages can be added to Weka through a package manager

https://en.Wikipedia.org/wiki/Weka_(machine_learning)

Classification/Regression Tools

- Bayes (14): NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialText, etc.
- Functions (29): Logistic, LDA, SMO (SVM), etc.
- Trees (18): J48 (Decision Trees), Random Forest, etc.
- Lazy (5): IBk (K-Nearest Neighbors), etc.
- Rules (12):
- Meta (36):

https://wiki.pentaho.com/display/DATAMINING/Classifiers

Download Weka for Mac

- Weka 3 – Data Mining with Open Source Machine Learning Software:
  - Go to the website: https://www.cs.waikato.ac.nz/ml/weka/downloading.html
  - Download: weka-3-8-3-oracle-jvm.dmg (143.1 MB)
  - Open to create a folder "weka-3-8-3" and drop it to a desirable location
  - Run the package with the command: "java –jar weka.jar"

References for Weka

- Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Morgan Kaufmann, 2016.
- Weka Manual 3-8-3 (comes with the weka package)
- How To Work Through a Multi-Class Classification Project in Weka by Jason Brownlee:

  https://machinelearningmastery.com/multi-class-classification-tutorial-weka/

- Text Classification with a Complete Example by Ashraf Uddin:

  Ashraftsau.blogspot.com/2015/03/text-classification-in-weka-with.html

Why Try Different Classifiers?

- No free lunch theorem (David H. Wolpert, 1996): No single classifier works best across all possible scenarios

- Each algorithm has its own characteristics based on certain assumptions

- In practice, we often need to try multiple algorithms in order to select the best model for a particular problem