CENG 499 – Introduction to Machine Learning

Spring 2017 – Homework 2

Tool Practice : K-Means, Decision Trees and ANNs

Selim Temizer

Feedback : Between May 22nd and May 26th, 2017
Due date : May 28th, 2017 (Submission through COW by 23:55)

Part 1. (60 points) K-Means Practice on Weka

In this part of the second homework assignment, we will cluster a numerical data set using k-means algorithm. The data set consists of 2000 points in $\mathbb{R}^5$ and it is provided to you as a text file in ***ARFF*** format (which is basically a human readable CSV, comma separated value file, with some short header information added such that it can easily be imported by the Weka tool). An important decision when using k-means algorithm is to choose a reasonable k value (especially if no expert advice is available to guide the selection). In order to complete this part, follow the steps below:

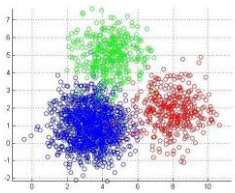
1. Download and install Weka on your computer
2. Run Weka and choose the Explorer application
3. From the Preprocess tab, open the ***ARFF*** file
4. From the Cluster tab, choose ***SimpleKMeans*** (also: initializationMethod = Farthest first)
5. Run the algorithm 15 times (start with $k = 1$ and increment k on each run)
6. Record the error for each run in the provided Excel file (it will automatically create you a plot)
7. By inspecting the plot, determine the k value using Elbow method
8. Run the algorithm again by using the k value you picked in step 7
9. Fill in the second table in the Excel file based on the results the Weka tool gives you in step 8
10. Rename the Excel file as illustrated below:

e1234567_SelimTemizer.xlsx

This might be the first time some of you will be using the Weka tool. Therefore, feel free to ask any questions you might have about this part of the assignment on the course newsgroup (including how to install Weka, how to set parameters of ***SimpleKMeans***, etc.). Also feel free to answer asked questions if you already know or have figured out how to work efficiently with Weka.

Important Technical Note : The Elbow method will help you with choosing a reasonable k value for this data set, but it may not be useful for any arbitrary data set in general

Bonus (undisclosed amount) : Use PCA to bring down the dimensionality to 3 or 2, so that the resulting ***clustered*** data could be plotted. Plot the data, and send along the plot, your brief explanations about how you applied PCA, and any source code (if you created any code).



Part 2. (20 points) Decision Tree Practice on AISpace Decision Tree Tool

Think of an ***original*** and ***realistic*** decision problem with at least 5 discrete attributes that you (or a fellow student/friend in general) face frequently. Collect at least 20 labelled training instances and save them in the input format accepted by AISpace tools. Then solve this problem using ***AISpace decision tree tool***. Save the solution in appropriate output file format (that I can open again in AISpace decision tree tool). Submit your input and output files.

Part 3. (20 points) Neural Network Practice on AISpace Neural Network Tool

Think of another ***original*** and ***realistic*** decision problem with at least 5 attributes that you (or a fellow student/friend in general) face frequently. Collect at least 20 labelled training instances and save them in the input format accepted by AISpace tools. Then solve this problem using ***AISpace neural network tool***. Save the solution again in appropriate output file format (that I can open again in AISpace NN tool). Submit your input and output files.

What to submit? (Use ***only ASCII characters*** when naming your files and folders)

1. Create a separate directory for each part (***Part1*** to ***Part3***). In Part1, just put your excel file. For Part1, if you have bonus work, also put it in Part1 directory. In Part2 and Part3, put your input and output files. Make sure that the ANN output also contains the network structure.
2. If you have any documentation for any part (like a readme file, or a WORD or PDF file possibly containing some bonus plots) put it in the respective directory.

Zip the 3 directories, (tar also works, but I prefer Windows zip format if possible), name the compressed file as <ID>_<FullNameSurname> (with the correct extension of .zip or .tar) and submit it through COW. For example:

e1234567_SelimTemizer.zip

There are a number of design decisions and possible opportunities for visual illustrations and creative extensions for this homework. For example, you may create test data in addition to training data, and have the AISpace tools plot the error on both training and test sets, and maybe to try to figure out the ideal time (in terms of epochs) to stop training the ANN, etc. You may also save and send any plots (with small explanations about them) that you think would be useful in analyzing the learning processes. There will be bonuses awarded for all types of extra effort that is documented appropriately.

Late submissions will not be accepted, therefore, try to have at least a working baseline system submitted on COW by the deadline. Good luck.