HW4: Data Mining - tool-based

Total points: 6

Summary: In this homework, you are going to use three UI-based tools (no coding!), to carry out data mining: **WEKA, KNIME, RapidMiner.** There are 3 questions you need to answer - linear regression, using each tool.

Description

WEKA

Start by downloading WEKA, from https://www.cs.waikato.ac.nz/ml/weka. Note - you can use an older 32-bit version, if your laptop is unable to run the latest 64-bit one. FYI WEKA is written in Java, so you need to install Java [most likely you already have it] prior to installing WEKA. WEKA is powerful and capable - you can continue using WEKA long after this course, and in the future, even consider extending it by writing plugins for it. Here are tutorials: https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/.

Here is a famous (in the ML/DM community) dataset called the 'Boston Housing Dataset'. As you can read from the description, it is a dataset that contains data regarding houses in several Boston suburbs, published in 1993. It has 506 rows (records) of data, and 14 columns (attributes). For this HW, we'll use the 'MEDV' (median home price) attribute as the "class" (the output to predict). In other words, using existing data from the other 13 columns, we want to be able to learn to predict MEDV for a new record (ie. row) that contains known values for those 13 'input' columns. Note that Zillow, TopHap, etc. routinely carry out such an analysis.

As you can see, the data is in a WEKA-native format called ARFF [https://www.cs.waikato.ac.nz/ml/weka/arff.html], which resembles, but is more descriptive than, CSV.

Q1 (2 points). Build a **linear regression** equation, to predict MEDV. Include a screenshot that shows the linear equation. How many terms are in the equation, and 'why'? In other words, discuss the resulting equation.

KNIME

Here is another dataset to use (scientists go out 'in the field' to painstakingly collect such data! ML might be able to automate some/all of it). It consists of 4177 rows of data regarding abalone shells, where each row resulted from measuring 9 parameters/features/values for each shell. The data is in text format (.arff format, for input to WEKA, like above), do take a look at it. The idea is to be able to predict the 9th value, number-of-rings, given the other 8 values, using the existing dataset to learn how to predict.

Next, download and install KNIME ("nime"), and work through the quickstart tutorial. KNIME is also UI-driven, like WEKA; additionally, it's also visual-dataflow-driven, which means we can do data mining with it, by 'connecting the boxes' (where each box reads data or does mining or writes data, etc).

Q2 (2 points). Use KNIME to perform **linear regression** [on all parameters, not a subset]. You need these nodes: AARF Reader, Linear Regression Learner. Create and connect the nodes, and execute each. What is the linear equation? Include a screenshot.

RapidMiner

Download RapidMiner Studio, and play with it for a bit - it is also dataflow-based, just like KNIME.

Q3 (2 points). Bring in the shells.arff data (in the operators list, look under Data Access -> Files -> Read), and only work with these 4 params: length,diameter,height,num_rings (use a 'Select Attributes' node, and type in a regular expression that specifies length,diameter,height,num_rings, or use the 'subset' attribute filter to pick the ones we want - search the documentation for how (additionally, this will help: https://www.youtube.com/watch?v=tQ7oDnQXhmQ). Do a **linear regression** to predict num_rings, from length,diameter,height. Question: what is the equation? Include a screenshot. Note that you need a 'Set Role' node where you would set num_rings to be a "label", before doing the regression (to let

the regression node know which attribute to predict, using the other non-label ones). The regression itself would be done using a 'Linear Regression' operator.

What to submit: a single .zip, named HW4_<yourname>.zip, with:

• screenshots, named Q1.{jpg,png} etc

• a single README.txt file, with answers for the questions (ie. regression equations)

It's highly worth knowing how to use such tools for analysis, in addition to only knowing how to do so using Python or R code **- the interface-driven tools are just as powerful**, because they encapsulate, with point-and-click UI, a variety of data-mining algorithms/code - resulting in a product that (even) non-programmers, eg. business analysts, managers etc. can use.

That said, R is a wonderful language for data analysis and plotting - you got a taste of it in HW3 :) Here is a little reference card (not mine) to tell you more. Fun fact: in R, the assignment operator can be flipped! In other words, a <- 5 and 5 -> a are both legal syntax [wow]. Fyi - extensive spatial data analysis/mining is possible using R, and this gives you a taste of it. And there's a little mo'R'e, Me Hea'R'ties - YaRrr/rrrr!!

Please upload your (.zip) submission on to D2L as usual.

ENJOY!!
