Machine Learning in the Real World

Worth: 20% of total grade [20 marks]
Due Date: Monday September 1st 2017, 11:59AM

Part I (6 marks):

There are two data files given: obscuredX and obscuredY (each with the .dta). One of them is the real dataset and one of them is a randomized dataset where the class attribute has been randomly shuffled. Can you determine which is which? Most of the marks will be for explaining how you made this determination. Please try to determine which is which using at least two methods and discuss why you think one method is easier or better than the other.

Part II (14 marks):

You have now graduated (Yea!). And are the sole employees of ``Data Are Us'', a snazzy new startup company guaranteed to make you Millionaires, if you could just attract your first paying customers! You have just been giving your big break! A unit of a big Company called ``Containers Are Us'' has given you a dataset on contaminated containers. They want to know what you can find in their database. You will need to use either obscuredX.dta or obscuredY.dta from above.

The goal is to find something that will convince this company to sign you up to do all its analyses and put you on your path to fame and fortune.

You can use any algorithms you can find on the net. Most people will either use R, Weka, or Brute. I suggest not to use brute unless you are familiar with Unix. Don't wait until the last moment. The hardest part of this assignment will be preparing the dataset. You will probably *NOT* be able to run on the whole dataset, you will have to throw parts out. (Depending on the tool you use you might be able to run the whole dataset….try it and see) Can you run on part of the dataset to bootstrap you to find the most important parts of the dataset? What question are you going to try to answer? The main variable you are trying to predict is the first one. Some of these codes represent contaminated containers and 1 represents clean containers. Which ones do you think your customer is the most interested in?

This assignment has been left intentionally vague. All grades will be based on the write up! Remember you are trying to convince this company to give you their money and not your rival company ``Data-dredging Are Us''. So convince me! This should look like a business pitch including all your data findings. Make sure you explain what algorithms you used, our company is old and conservative and we don't like black magic.

Also if you don't tell me you did something, I will assume you didn't! So make sure you tell me all the wonderful things you did in term of methodology and exploration experiments.

Please remember this assignment is meant to be fun, fun, fun! But it is only worth 20% of your grade, don't kill yourself over it. Are we having fun yet?

Marking is based on the following material:
A marketing report describing your data analysis and why this Company should give you its lucrative contract. There is a 4-6 page limit to the size of the report.

The datasets and Brute can be found at
http://www.cs.auckland.ac.nz/courses/compsci760s2c/assignments/2017/Pat/

The assignment must be submitted through the department dropbox:
https://adb.auckland.ac.nz