­­Class Project – Apply Machine Learning Model(s) to a Dataset of your Choosing

Find a dataset of your own choosing to perform some analysis on. The dataset must be:

1. Of an appropriate size. Needs to be large enough for the algorithm to learn some patterns that solve a particular problem. However, for grading purposes, not too big – can be emailed
2. Must not be proprietary data, either publically available or private data shared with the permission of the owner for use in this class.
3. Allow you to solve one or more prediction problems using machine learning algorithms

Once the data has been curated, determine 1-2 problems you can use the data to solve using a machine-learning algorithm. Use R, Python, Java or Weka, build a supervised or unsupervised model. If a programming language is used, please provide code, if weka is used, please copy the text from the classifier output pane for each important experiment. Then provide a write-up, detailing:

1. The choice of dataset, and the problem you are trying to solve
2. Your feature selection approach – this can be very simple, using correlation measures or related, or can use some R or python feature seletion algorithm, or some of the weka attribute selection algorithms (Select Attributes tab)
3. Your final choice of machine learning model and why. Evaluate a 2-3 different models, explain your approach to testing and evaluating the models, and which was best. Also explain your choice of evaluation metric.
4. For you final choice of model, what hyper parameters did you choose to tune or try? Note that some may be binary, which is fine also. What settings worked best? You don’t need to tune all of them, one or two is fine. Also briefly explain briefly why you chose those hyper parameters over any others, if there we more than 2 hyper parameters (i.e. why are those the most important parameters to tune for the model?)

**Deadlines**

**10.28.2014** – Submit a proposal describing the problem and the dataset, and include the dataset you intend to use. This requires no modeling work, only finding an interesting problem to tackle that’s of reasonable difficulty (i.e. not unreasonably hard).

**10.30.2014** – Discuss with the students in class if any assignments are not quite a good fit for this project. Distance learners, I will arrange a call if needed.

**11.13.2014 –** Submit final project write up, including code \ weka output. If you are having problems, please contact me for assistance several days in advance of this date.