Home Work 1

Due Date: February 17th

Exercise

Code in any programming language the linear logistic regression algorithm with ridge-regression for binary classification using steepest descent with a fixed learning rate. You need to run your code in a subset of the NESARC database that has been recently provided.

The input variables will be: 42 (needs encoding), 48, 63 (encode), 64, 68 (normalization), 79, 3673 (needs encoding), 114 (needs encoding), 131, 136, 144, 163, 196, 217, 226, 230, 232 (unknowns), 294 (unknowns), 306-308 (merge and unknowns) and 310.

Add a input variable that it is equal to 1 for any input to encode a bias from the origin. Basically, we can to estimate $y = \text{sign}(\mathbf{w}^\top \mathbf{x} + w_0) = \text{sign}(\widetilde{\mathbf{w}}^\top \widetilde{\mathbf{x}})$, where $\widetilde{\mathbf{w}} = [\mathbf{w}^\top w_0]^\top$ and $\widetilde{\mathbf{x}} = [\mathbf{x}^\top 1]^\top$.

The variable that we want to predict is 313 (encode).

Deliverables

- The code and instructions in how to run it in a mac. The inputs to the code should be an $N \times D$ inputs matrix $\mathbf{X}$, an $N \times 1$ outputs vector $\mathbf{y}$, a nonnegative scalar ν for the regularizer weight, and a positive scalar for the step-size η.
- The weight vector when the regularizer is turned-off.
- Training and test errors. Do you think it has overfitted?
- A 200-word paragraph (or less) interpreting the weight vector. Detail what surprised you or what did you find normal in the solution.

Note: You need to code the logistic regression yourself it is part of the assignment. This is a machine-learning course, not a data-mining course. But, if I were you, I would definitely try any package out there that solves logistic regression and