CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

Assignment 2

Due: Thursday, April 18, 2013

This assignment consists of three problems. Two of them are "pencil and paper" problems. The third involves computation on supernova data.

1. *Classification* (30 points)

   1.1 Suppose that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$ and $X \,|\, Y = 0 \sim N(0, 1)$ and $X \,|\, Y = 1 \sim \frac{1}{2}N(-5, 1) + \frac{1}{2}N(5, 1)$.

   (a) Find expressions for the Bayes classifier and the Bayes risk.

   (b) What linear classifier minimizes the risk and what is its risk?

   1.2 Now suppose that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}$ and $X \,|\, Y = -1 \sim \text{Uniform}(-10, 5)$ and $X \,|\, Y = 1 \sim \text{Uniform}(-5, 10)$.

   (a) Find an expression for the Bayes classifier and find an expression for the Bayes risk.

   (b) Consider the linear classifier $h_\beta(x) = \text{sign}(\beta x)$ where $\beta \in \mathbb{R}$. What linear classifier β^* minimizes the risk and what is its risk?

   (c) Compute the hinge risk $R_\phi(\beta) = \mathbb{E}(1 - Y\beta X)_+$, where $(\cdot)+$ denotes positive part.

2. *Logistic regression* (30 points)

   (a) Show that if Newton's method is applied to the logistic regression log-likelihood, it leads to the reweighted least squares algorithm.

   (b) Show that if the data are perfectly linearly separable, the conditional maximum likelihood estimator for the logistic regression model does not exist. Comment on the behavior of the iteratively reweighted least squares algorithm.

   (c) Give a detailed derivation of the Newton algorithm for ridge logistic regression, using a penalty $\lambda\|\beta\|^2$. Compare it with the iteratively reweighted least squares algorithm and comment on the differences between the two.

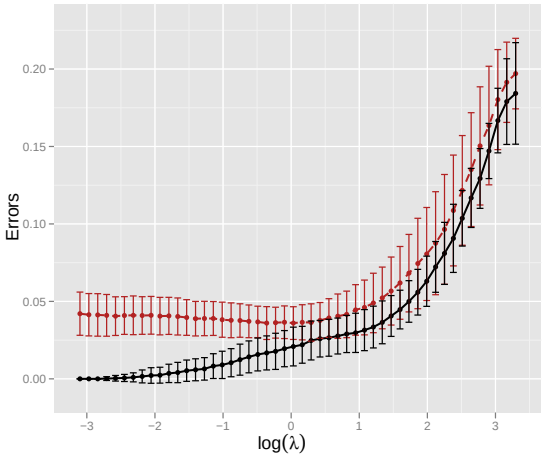
3. *Supernova classification*[1] (40 points)

Now you will implement the ridge logistic regression algorithm that you developed in the previous problem. It will be applied to the same supernovae data used to illustrate classification in class, and in the chapter notes.

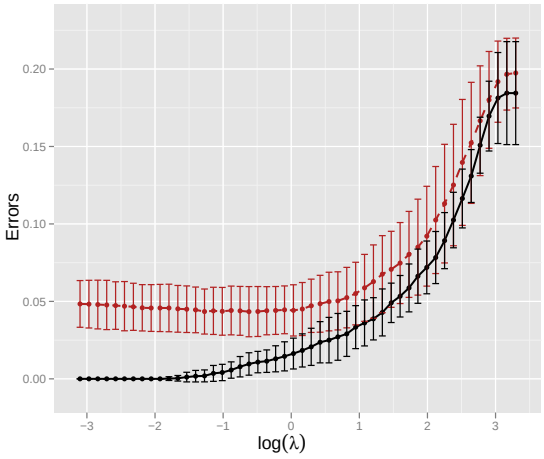
The files `supernovae.raw.dat` and `supernovae.dct.dat` contain values for 255 supernova, 206 of which are type Ia. The first file contains the time domain features, and the second contains the frequency domain features, extracted as described in the notes (`dct`="discrete cosine transform").

Your task is to implement ridge logistic regression directly, as an iterative weighted least squares algorithm, without using any R packages or Python modules that do this. Of course, you can use things like `numpy` and `scipy` for basic numerical routines.

For both the time and frequency domain data, fit models for a range of regularization parameters λ. For each λ, generate 50 random train/test splits of the data, using proportions 60%/40%. Plot the average training and test error rates, together with standard error bars ($\pm$ one standard deviation), as shown in the chapter notes for sparse logistic regression:



Sparse LR (time domain)



Sparse LR (frequency domain)

Note that for this problem you are not required to use the Amazon AWS system—you will be using this for other problems soon enough. You may implement your algorithm using R, Python, or iPython. Submit your code by placing it in the `submissions/assn2` directory on your `nfs.lsda.cs.uchicago.edu` account, naming your program according to the setup you use, `assn2_prob3.<r,py,ipynb>` as appropriate.

In your written solutions, include plots of your results. Compare to the results for ℓ_1-regularized (sparse) logistic regression. Comment on the similarities and differences, and discuss which model may be more appropriate.

[1] And speaking of exploding stars and statistics: `http://xkcd.com/1132/`