Homework 7, due October 23rd, 11:59pm

October 10, 2019

1. Implement the FSA variable selection method with linear models for multi-class classification with the Vapnik loss:

$$L_D(\boldsymbol{u}, y) = \sum_{k \neq y} L(u_y - u_k), \tag{1}$$

where $L(u)$ is the Lorenz loss described in class. Use the parameters $s = 0.0001, \mu = 30, N^{iter} = 500$.

Take special care to **normalize each column** of the X matrix to have zero mean and variance 1 and to use for normalizing the test set the same mean and standard deviation that you used for normalizing the training set.

Assuming that the coefficient vector is a $p \times c$ matrix W, where p is the number of features and c is the number of classes, use the norm $\|\boldsymbol{w}_j\|, j = 1, ..., p$ of each row as the criterion to select the variables in FSA.

a) Using the `satimage` data, train a multi-class FSA classifier on the training set, starting with $\boldsymbol{\beta}^{(0)} = 0$ to select $k \in \{5, 9, 18, 27, 36\}$ features. For each k find an appropriate learning rate η to obtain a small final loss value on the training set. Plot the training loss vs iteration number for $k = 18$. (5 points)

b) Report in a table the misclassification errors on the training and test set for the models obtained for all these k. Plot the misclassification error on the training and test set vs k. (1 point)

c) Repeat points a) and b) for the dataset `covtype`, adding the misclassification errors to the table from b). (4 points).