

Homework 12, due December 1st, 11:59pm

November 19, 2020

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

$$L(\mathbf{u}, y) = \sum_{k \neq y} \ell(u_y - u_k), \quad (1)$$

where $\ell(u)$ is the logistic loss described in class. Use the parameters $\lambda = 0.0001$, $\mu = 100$, $N^{iter} = 500$.

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 $\|\mathbf{w}_j\|$, $j = 1, \dots, p$ 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 $\beta^{(0)} = 0$ 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 $k = 27$. (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) using the cross-entropy loss function:

$$L(\mathbf{u}, y) = -u_y + \ln\left(\sum_{k=1}^C \exp(u_k)\right), \quad (2)$$

adding the misclassification errors to the table from b). (4 points).