Homework 6, due February 21st, 11:59pm

February 14, 2024

1. Implement Logitboost using univariate (based on a single feature, with intercept) linear regressors as weak learners. At each boosting iteration choose the weak learner that obtains the largest reduction in the loss function on the training set $D = \{(\mathbf{x}_i, y_i), i = 1, ..., N\}$, with $y_i \in \{0, 1\}$:

$$L = \sum_{i=1}^{N} \ln(1 + \exp[-\tilde{y}_i h(\mathbf{x}_i)]) \tag{1}$$

where $\tilde{y}_i = 2y_i - 1$ take values ± 1 and $h(\mathbf{x}) = h_1(\mathbf{x}) + ... + h_k(\mathbf{x})$ is the boosted classifier. Please note that the Logitboost algorithm from the slides uses $y_i \in \{0, 1\}$ and the loss uses $\tilde{y}_i \in \{-1, 1\}$.

a) Using the `arcene` data, train a Logitboost classifier on the training set, with $k \in \{10, 30, 100, 300, 600\}$ boosting iterations. Plot the training loss vs iteration number for $k = 600$. Report in a table the misclassification errors on the training and test set for the models obtained for all these k. Plot the misclassification errors on the training and test set vs k. Also plot the train and test ROC curves of the obtained model with 300 features. (3 points)

b) Repeat point a) on the `dexter` dataset. (3 points)

c) Repeat point a) on the `Gisette` dataset. (3 points)