Statistics 535 Projects

The Million Songs dataset

Autumn Quarter 2015 – Instructor: Marina Meila. TA: Lina Lin

DEPARTMENT OF STATISTICS

Summary:

The students of STAT 535 each implemented two predictors for the Million Songs data set with modification. That is, the binary class labels were manually constructed based on song year. These predictors were tested on a sample of 10,000 unlabeled events. Here we present the prediction results.

Data:

The dataset is a subset of the Million Songs dataset. It contains about 460,000 examples divided into 2 classes. We created the class labels ourselves, as the original problem is regression; the two classes have approximate probabilities 0.36 and 0.64. Class labels are given as +1 or -1. There are 90 features in total, 12 which describe timbre average and 78 which describe timber covariance.

The test set consists of 10,000 random events. The class probabilities in the test set are very close to that in the training set. The training set and test set are disjoint.

Methods:

Each student implemented or tested two classification methods. The first was chosen from the list below such that each method was covered by at least three people. The second was the student's choice, either from this list or an alternative method.

1. Bagged decision tree: an ensemble of decision trees obtained by either randomizing the construction of the trees, or by resampling the training set.
2. Neural network: a multilayer neural network, with 2 or more layers
3. Boosting: a boosted weak classifier of your choice
4. K-NN: K-nearest neighbors
5. Naive Bayes
6. SVM - RBF/other kernel: SVM with the Gaussian kernel or kernel of choice
7. Logistic regression
8. Generative model
9. CART: classification and regression tree.

The unlabeled test data were made available on Monday December 7th. The students had approximately a little under 72 hours to determine the classes of the songs in the test set, using the predictors trained earlier this quarter.

Figure 1(below): L_{01} loss on the test set for the predictor implemented.