Lab 8/Homework 4

Download the KKBox's Churn Prediction training dataset from kaggle.

Classification

1. Use the first half of the training data (rows) to train a logistic regression model to make prediction, and then test on the second half of the training data. What is the classification error you obtain? What is the AUC of the testing? Make a ROC curve of it.
2. Conduct 5-fold cross-validation to estimate the classification error. Is it larger or smaller than the classification error you obtain in (1)? Which one will be closer to the true classification error of your classifier, do you think?
3. Train a logistic model using all the training dataset, and make a prediction over the kaggle testing dataset. Then make your first submission of your prediction result to kaggle! What is your kaggle ranking and score?