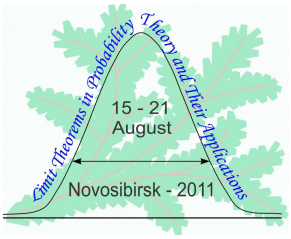

Plug-in Approach to Active Learning

Stanislav Minsker

Georgia Institute of Technology
sminsker@math.gatech.edu

Let $(X, Y) \in S \times \{-1, 1\}$ be a random couple with unknown distribution $\mathcal{P}$, where X is an observation and Y - a binary label to be predicted. The goal of a learning algorithm is to construct a *classifier* - a measurable function $f : S \mapsto \{-1, 1\}$. In practice, distribution $\mathcal{P}$ remains unknown but the learning algorithm has access to a training data set - a sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from $\mathcal{P}$.

It often happens that the cost of obtaining the training data is associated with *labeling* the observations while the pool of observations itself is almost unlimited. This suggests to measure the performance of a learning algorithm in terms of its *label complexity*, the number of labels required to obtain a classifier with desired accuracy (rather then the total number of observations which is commonly used in *passive learning* literature). Active Learning theory explores the possible advantages of this modified framework. We will present a new active learning algorithm based on nonparametric estimators of the regression function and explain main improvements over the previous work. Our investigation provides upper and lower bounds for the performance of proposed method over a broad class of underlying distributions.