DLI Accelerated Data Science Teaching Kit

Module 2 Lab:
Data Annotation in Active Learning

OBJECTIVE

Data annotation is to label collected data, which is necessary to build machine learning models, especially for supervised learning models in different applications such as Image Processing, Natural Language Processing, and Smart Grid.

Active learning is a category of machine learning models in which a learning algorithm can interactively query a user to label samples with the predefined label. The goal of this lab is to build a pipeline that implements data annotation in the active learning.

PREREQUISITES

You have to complete all submodule 2.4 lecture slides and install packages below.

sklearn, modAL, numpy, IPython, matplotlib

INSTRUCTIONS

- Import ***MNIST*** dataset (http://yann.lecun.com/exdb/mnist/) with *sklearn.datasets*
- Split the dataset into training and testing datasets $\mathbf{D_{training}}$ and $\mathbf{D_{testing}}$
- Select 100 training samples to build a subset $\mathbf{D_s}$ from $\mathbf{D_{training}}$, which is used to build a classifier for active learning
- Build a data pool of unlabeled data $\mathbf{D_{pool} = D_{training} - D_s}$
- Initializing a classifier ***C*** based on **Logistic Regression** with $\mathbf{D_s}$ via for active learning
- Implement data annotation on $\mathbf{D_{pool}}$ by querying users to labeled samples that are from $\mathbf{D_{pool}}$, extend $\mathbf{D_s}$ with the annotation results to build to $\mathbf{D'_s}$, and retrain ***C*** with $\mathbf{D'_s}$ and check the performance, where the performance is evaluated with accuracy