Lab III - Concept Drift Detection

module III, ML Model Deployment & Maintenance

DT4031, *Applied Machine Learning*
Jens Lundström

Introduction & Assignment

In this lab you will investigate a dataset representing streaming data to be used for binary classification. You have two partitions of the data **A** & **B**. **A** is a dataset containing 1000 samples (sequential in time) for ***training & evaluation*** and **B** is another sequential 1000 samples representing labeled ***production data*** .

The first task is to create a new Notebook and to load the two datasets, select a classifier (reuse code from earlier labs), train a model and evaluate it (evaluate generalization error) – using the dataset **A**. Question: How well is the classifier performing?

The second task is to test the model on the dataset **B** (*production* dataset). Questions: How well is the trained classifier performing on the *production* dataset? Is it better or worse than on the generalization error of dataset **A**?

As a third task, try to measure how the features are changing over time (concept drift). This could be done by applying the Kolmogorov-Smirnov Test (*K-S Test*, read more about it further down) to see if the distribution of the features remains constant or if they are changing over time (index of the production dataset). Your reference distribution can be the 250 first values of dataset **A, called** ***FEATURE_SAMPLES_DS_A*** . First try to do a K-S Test between *FEATURE_SAMPLES_DS_A* and for indeces 0 to 250 of the **production dataset B** . Are they having the same distribution? (can we reject the null hypothesis?). Secondly, try to do a K-S Test between *FEATURE_SAMPLES_DS_A* and for indeces 250 to 500 of the **production dataset B** . Are they having the same distribution? (can we reject the null hypothesis?). Try other windows in the production dataset B, can you estimate where the concept drift starts (at which index in dataset **B**)? (*optional question*: which type of concept shift occurs: abrupt, gradual or incremental? Try to guess)

A fourth task (optional), try to retrain the model with parts of dataset B occurring <u>after</u> the start of the concept drift (which could be estimated in the third task). How is the generalization error now? Is it improved?

Lab report instructions

The lab solution shall be documented in the form of a written report. As a data scientist it is your task to explain what you have done in a detail such that the solution is reproducible, the reasoning behind the decisions, the outcome (results) and the conclusions you have drawn from such results. It is suggested to have the following lab report structure: Introduction & Motivation; Methods & Solutions; Results; Conclusions & Discussion; Appendix. Keep the number of pages in the lab report limited, no more than five pages (excluding *Appendix* which should contain Notebook code and figures not directly linked to the result).

Lab assistance

These functions are suggested to be used when solving the lab.

numpy.loadtxt(…)
sklearn.tree .DecisionTreeClassifier (…)
classifier_model .fit(…)
classifier_model .score(…)
sklearn.model_selection.train_test_split (…)

```
from scipy import stats
stats.ks_2samp(FEATURE_SAMPLES_DS_A, FEATURE_SAMPLES_DS_B)
```

Feel free to reach out to Kunru or Jens if you have questions regarding the lab.

Regarding K-S Test (two-sample K-S test)

The K-S Test is a statistical hypothesis test used when testing if two samples are of equal distributions. If the null hypothesis is that the two samples are drawn from the same distributions. One of your task in this lab is to see if the null hypothesis can be rejected.

You can find more information about the K-S test on the (currently, 2022-02-18) excellent Wikipedia article: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Moreover, see documentation for the function ***ks_2samp*** (suggested to be used in the lab) here: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html