CNG 562 Introduction to Machine Learning

Homework 2

Otto Group Product Classification

From the Kaggle page:

“

…

A consistent analysis of the performance of our products is crucial. However, due to our diverse global infrastructure, many identical products get classified differently. Therefore, the quality of our product analysis depends heavily on the ability to accurately cluster similar products. The better the classification, the more insights we can generate about our product range.

“

The dataset consists of 93 features for more than 200,000 products. The objective is to build a predictive model which is able to distinguish the main product categories.

Data is already split as training data (trainData.csv) and test data (testData.csv).

For the data and/or further information, see https://www.kaggle.com/c/otto-group-product-classification-challenge.

This part of the homework should be returned as a Jupyter Notebook with the filename <SURNAME>_otto_group_product_classification.ipynb in your GITHUB account.

In your Jupyter notebook, you must at least,

1. Download the data from Kaggle,
2. Prepare the data,
3. Train a Multi-layer Perceptron (MLP),
4. Train *at least with*
   a) K-Nearest Neighbours,
   b) Support Vector Machines,
   c) Random Forest,
   d) XGBoost Methods.
5. When possible, plot ROC curves.
6. Compare the results.

Malaria Cell Images

This dataset consists of 27k+ images of two folders Infected and Uninfected.

The objective is to predict whether a cell image is infected or not.

For data and more information, please see https://www.kaggle.com/iarunava/cell-images-for-detecting-malaria.

This part of the homework should be returned as a Jupyter Notebook with the filename <SURNAME>_malaria_cell_images.ipynb in your GITHUB account.
In your Jupyter notebook, you must at least,

1. Download the data from Google Sheets,
2. Prepare the data,
3. Train *at least with*
    a) Convolutional Neural Network,
    b) Use Transfer Learning to build other model(s) on top of neural networks trained on ImageNet. For ready-to-be-used neural networks in Keras, please see https://keras.io/applications/.

Aerial Cactus Identification

From Kaggle:

"This dataset contains a large number of 32 x 32 thumbnail images containing aerial photos of a columnar cactus (Neobuxbaumia tetetzo). Kaggle has resized the images from the original dataset to make them uniform in size. The file name of an image corresponds to its id.

You must create a classifier capable of predicting whether an images contains a cactus."

The dataset is already split as training and test sets. Use this splitting as is.

For data and more information, please see https://www.kaggle.com/c/aerial-cactus-identification.

This part of the homework should be returned as a Jupyter Notebook with the filename <SURNAME>_aerial_cactus_identification.ipynb in your GITHUB account.
In your Jupyter notebook, you must at least,

4. Download the data from Google Sheets,
5. Prepare the data,
6. Train *at least with*
    a) Convolutional Neural Network,
    b) Use Transfer Learning to build other model(s) on top of neural networks trained on ImageNet. For ready-to-be-used neural networks in Keras, please see https://keras.io/applications/.