MACHINE LEARNING FROM DATA

Lab Session 9 – Ensembles

1. Goal

The goal of this session is to
- Learn to use majority voting to predict class labels by combining conceptually different machine learning classifiers
- Learn to create ensembles by bagging
- Learn to create ensembles by boosting
- Solve the MNIST classification problem using ensembles of classifiers

2. Instructions

- Download and uncompress the file Mlearn_Lab9.zip
- Answer the questions in the document Mlearn_Lab9_report_surname.pdf

3. Previous study

Read the slides corresponding to lecture 5: Evaluation, selection and combination of classifiers.

4. Moon dataset with ensembles

In the first part of the session, you will use the synthetic moon dataset with two classes and two features per class, to learn how combine classifiers using voting, bagging, Adaboost and Gradient Boosting.

Q1: What is the difference between a Random Forest Classifier and Bagging Decision trees?

5. MNIST with ensembles

In the second part you will use ensembles to try to improve the results obtained in previous Labs on the MNIST classification problem

Edit the notebook.
Add the necessary code to:

1. Load the MNIST dataset (check Lab6)
2. Split it into a training set, a validation set, and a test set (e.g., use 50,000 instances for training, 10,000 for validation -selected from the original train set with 60.000 images-, and 10,000 for testing). Use smaller sets if you have time-issues with the use of Colab.
2. Train various classifiers, such as a Random Forest classifier, an SVM classifier, a Neural Network and/or other models. You can define a list of classifiers, and then train each classifier and see the performance of each one using
```
estimators = [clf_1, clf_2, clf_3, clf_4]
for estimator in estimators:
```

```
        print("Training the", estimator)
        estimator.fit(X_train, y_train)
[estimator.score(X_val, y_val) for estimator in estimators]
```
Compute performance metrics and confusion matrices for each classifier

3. Try to combine them into an ensemble that outperforms each individual classifier on the validation set, using soft or hard voting. You can add or remove estimators from the list.

4. Once you have found one ensemble, try it on the test set. How much better does it perform compared to the individual classifiers?

Q2 Copy and analyze the results obtained with single classifiers on the test set

Q3 Copy and analyze the results obtained with the ensemble(s) on the test set, compare with the individual classifiers.

3. Optional task: train an XGBoost model on MNIST. Use the same train-val-test splits that you have used before. Compare results with the previous ensemble.

Q4 (optional task): Copy, analyze and compare results with previous approaches.