

# Assignment 3

April 16, 2021

## Question 1: 10 points

Train an SVM classifier on the MNIST dataset. Since SVM classifiers are binary classifiers, you will need to use one-versus-the-rest to classify all 10 digits. You may want to tune the hyperparameters using small validation sets to speed up the process. What accuracy can you reach?

## Question 2: 10 points

Train an SVM regressor on the California housing dataset.

## Question 3: 20 points

Train and fine-tune a Decision Tree for the moons dataset by following these steps:

- Use `make_moons(n_samples=10000, noise=0.4)` to generate a moons dataset.
- Use `train_test_split()` to split the dataset into a training set and a test set.
- Use grid search with cross-validation (with the help of the `GridSearchCV` class) to find good hyperparameter values for a `DecisionTreeClassifier`. Hint: try various values for `max_leaf_nodes`.
- Train it on the full training set using these hyperparameters, and measure your model's performance on the test set. You should get roughly 85% to 87% accuracy.

## Question 4: 20 points

Load the MNIST data and 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, and 10,000 for

testing). Then train various classifiers, such as a Random Forest classifier, an Extra-Trees classifier, and an SVM classifier. Next, try to combine them into an ensemble that outperforms each individual classifier on the validation set, using soft or hard voting. Once you have found one, try it on the test set. How much better does it perform compared to the individual classifiers?

## Requirements

These are the general rules for submitting the homework not only for this assignment.

1. Submit jupyter notebook(s) in ipynb and html formats: html format allows the grader to easily read it in canvas while ipynb allows to execute it if something is not clear from html.
2. The notebook should be well formatted:
  - Use markdown to break it into sections, bullet points, etc.
    - See `11_n3.ipynb` for example how to use it.
    - If you double click on markdown cell, you can see the code.
    - To show the formatted text, execute the corresponding markdown cell.
    - Start the notebook with a markdown title cell of the form:  
'# Assignment X, Jane Doe, date'.
    - When answering a particular question from the assignment, start the answer with markdown cell  
'## Question Y'  
that indicates what question you are answering.
3. Name your notebook (and the corresponding html file) according to the following template: `Assignment_X_Jane_Doe.ipynb`
4. If you are submitting several notebooks, append to the name '`_partZ`' and explain in canvas what each notebook is about.
5. There should not be any failed cells in the notebook. Every cell should work and have the results of the execution.
6. The notebook should contain only what is needed to answer a particular question and nothing else. More is as bad as less.
7. All the plots, if needed for the assignment, should be of production quality with readable labels, titles, etc. Also, try to find the most appropriate type of plot to clearly demonstrate your statement.
8. Due date: 04/23/2021