Practice 9A: Decision Trees

Supervised Learning - 15 December 2021

1. Train and fine-tune a Decision Tree for the moons dataset.
   (a) Generate a moons dataset using `make_moons(n_samples=10000, noise=0.4)`.
   (b) Split it into a training set and a test set using `train_test_split()`.
   (c) 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`.
   (d) Train it on the full training set using these hyperparameters, and measure your model's performance on the test set.
2. Grow a forest.
   (a) Continuing the previous exercise, generate $1,000$ subsets of the training set, each containing 100 instances selected randomly.
   Hint: you can use Scikit- Learn's `ShuffleSplit` class for this.
   (b) Train one Decision Tree on each subset, using the best hyperparameter values found above. Evaluate these $1,000$ Decision Trees on the test set.
   (c) For each test set instance, generate the predictions of the 1,000 Decision Trees, and keep only the most frequent prediction (you can use SciPy's `mode()` function for this). This gives you majority-vote predictions over the test set.
   (d) Evaluate these predictions on the test set. Which are the performances compared to the single model trained in point 1.