Decision Tree Exercises - Part 2

Summary

In this exercises (DT Part 2), we explore the behavior of deceision trees when datasets are variated. Additionally we will train and evaluate a decision tree classifier for a wine dataset.

Exercise 1

Input features are often normalized before training ML models. How do you think decision trees would be affected if a dataset was differently:

- centered?
- scaled?
- rotated?

Use the functions syn1 or syn2 from Exercise_DT_setup.jpynb to create a dataset.
Train decision trees with max_depth=2 for the original data, differently centered, scaled, and rotated data.

- Does the result match your expectation?
- How does it compare to other algorithms you have used on this course?

Exercise 2

The structure of a tree is sensitive to the exact training data. Here we simulate the effect of random sampling by regenerating a synthetic dataset with various random seeds.

- Use a known random seed for generating the syn2 data. Then train a decision tree on it, use for example min_samples_leaf=5 to regularize
- Regenerate the syn2 dataset with a different random seed, to simulate the effect of a different random sample from the same underlying distribution
- Train the decision tree again.
- How different or similar are the trees?
- How do you imagine this would be if we had 100 input features instead of 2?
- Flip the class assignment of one or a few data points (to simulate mislabeled data) and train the decision tree again . What is the effect? (Try e.g. using seed=42 when generating the dataset and flipping class assignment of the last training instance)
- What does all this mean in terms of model varinace? and in terms of usefulness of the decision rules?

Exercise 3

Here we train a decision tree classifier for a wine dataset for illustration, but you are of course welcome to also try on any other data set you want.

The wine dataset is a very small dataset: Only 178 instances with 13 numeric attributes. There are three classes (Class_0, Class_1, Class_2). The data is the result of a chemical analysis of wines grown in the same region in Italy by three different cultivators. There are thirteen different measurements taken for different constituents found in the three types of wine. For details see

https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-recognition-dataset

- Load the wine dataset, and split it into training and test set (The dataset can be loaded with load_wine()).
- Use GridSearchCV to find best parameters. Try e.g. max_leaf_node from 2 to 10, and min_samples_split from 2 to 4.
- What is the depth and number of leaves of the best tree? (Hint: Use the GridSearchCV attribute best*estimator*, and the get_depth() and get_n_leaves() methods of the best tree)
- Look at the reported feature importances, which features are important? (Hint: Use the feature*importances* attribute of the best tree)
- Plot the tree, does the tree match the important features above?
- Check the performance of your final classifier on the test set.