DSP 556 Homework 3: Decision Trees and Support Vector Machines

For this assignment, we will be focusing on Decision Trees, and Support Vector Machines. These models are available in `sklearn`. As always, remember to apply proper k-fold cross validation techniques.

We will use the `Forest Cover` dataset for classification models and the `Housing Prices` dataset for regression models. These datasets are available in the `datasets` directory on GitHub. Make sure to go over the dataset summary provided.

For each model you implement on each dataset, answer the following questions.

From the documentation on each model, what parameters could you try tweaking from the default settings to make the model:

- have better accuracy?
- train faster?
- predict faster?
- do all three?

For each model, what selections of parameters cause the model to shatter on this dataset?

Discuss how you went about parameter selection for your models. Document the effects of selecting different parameters on model accuracy and training time. Which parameters had the most noticeable impacts on model performance? Why?

Covertype

This is a classification dataset and we have already created a train-test split for you. You will investigate the various parameters of the models to make sure that they do not over-fit the data.

For this dataset, you will build the following models:

- `DecisionTreeClassifier` from `sklearn.tree`
- `LinearSVC` from `sklearn.svm`

Only once you are satisfied with your models' performance on the training data may you run your models with the test data.

Investigate the test set and comment on any differences you might observe compared to the training set. What, if any, ethical implications can you draw from this experience?

Prices

This is a regression dataset and we have already created a train-test split for you. You will investigate the various parameters of the models to make sure that they do not over-fit the data.

For this dataset, you will build the following models:

- `DecisionTreeRegressor` from `sklearn.tree`
- `SVR` from `sklearn.svm`
- `LinearSVR` from `sklearn.svm`

Discussion

Don't forget about the week 3 discussion. One important facet of a machine learning model is the notion of *interpretability* – that is, if you look at a trained model, does it provide you some insight into how it works? Or is it a black box? Please discuss the interpretability of Decision Trees and Support Vector Machines.