Decision Trees

Trees form a building block for gradient boosting and for random forest, both important algorithms.

Read chapter 9 for ESL, pages 305-313.

Explain to yourself the following, then discuss with your supervisor.

How might one use a classification tree to return class probabilities?

What are the three measures used for building classification trees? What are their relative merits? What’s used for regression trees?

How can the Gini index be interpreted?

What are the disadvantages of decision trees?

What does pruning aim to solve? How is it done?

What might be a problem with classification decision trees when a categorical feature with many possible values? Suggest a solution to the problem.

Read the following link, which also gives a glance into random forests, discussed next.

<https://jakevdp.github.io/PythonDataScienceHandbook/05.08-random-forests.html>

How well do you expect decision trees to learn the parity problem? Demonstrate this in code. Also argue that pre-pruning makes learning XOR difficult.

Predict the sales in the attached Car-seats data using the other variables in a decision tree. Use a validation set to tune the hyper parameters. What are the best hyper-parameters? What is the best test error? Which variables are actually used? Plot the tree and investigate the dependence of the hyper parameters on the result.