

Exercises

§14.1-14.4,14.7: Decision Trees

Python: There is an optional starter script that supports Python-based problem(s) on this sheet. Use it as scaffolding: read the comments, complete the lines marked **TODO**, and compare the output with the mathematics requested in the problem(s).

Problem 1 A regression leaf contains responses 2, 4, 5, 9. Derive the constant prediction that minimizes the leaf SSE. Compute the minimum SSE.

Problem 2 For $x = (1, 2, 3, 4, 5, 6)$ and $y = (1, 1.5, 2, 8, 8.5, 9)$, evaluate all midpoint thresholds 1.5, 2.5, 3.5, 4.5, 5.5. Find the left mean, right mean, and child SSE for each. Which threshold is selected?

Problem 3 Consider the tree: if $x_1 \leq 2$, predict 0; otherwise, if $x_2 \leq 1$, predict 3; otherwise predict 7. Find the prediction for (1, 5), (3, 0), and (3, 2). Explain how this tree represents an interaction.

Problem 4 Compute Gini impurity and base-2 entropy for: (a) a pure binary node; (b) proportions (1/2, 1/2); (c) proportions (3/4, 1/4). Which cases are most and least pure?

Problem 5 For $x = 1, \dots, 8$ and labels $(0, 0, 0, 1, 1, 1, 0, 1)$, compute the weighted child Gini and Gini gain for thresholds $s = 3.5$ and $s = 4.5$. Which split is preferred?

Problem 6 Tree A has training loss 10 and 2 leaves. Tree B has training loss 4 and 5 leaves. Using $R_\alpha(T) = R(T) + \alpha|T_{\text{leaves}}|$, find the values of α for which each tree is preferred. Find the tie value.

Problem 7 Explain why ordinary decision trees usually do not require standardization of a numerical feature before threshold splitting. Prove that the split $x \leq s$ is equivalent to a split on $z = ax + b$ when $a > 0$. What changes if the transformation reverses order?

Problem 8 Implement the regression-stump search from the lesson in Python. Report the best threshold, left mean, right mean, and child SSE. Then write a function that predicts new observations using the fitted stump.