Winter Course Statistical Learning

Trees

Marjolein Fokkema

Leiden University

Trees

A tree partitions the space spanned by predictors $X_1, , \ldots, X_p$ into areas $R_1, \ldots, R_m, \ldots, R_M$:

- R_m are *rectangular* and *non-overlapping*, each corresponding to a terminal node of the tree
- Predictive model:

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{y}_m \cdot I(x \in R_m)$$

- Predicted value $\hat{y}_m$ is a local constant:
  - Numeric response: mean or median within R_m.
  - Categorical response: Class proportions or majority class within R_m.

Trees

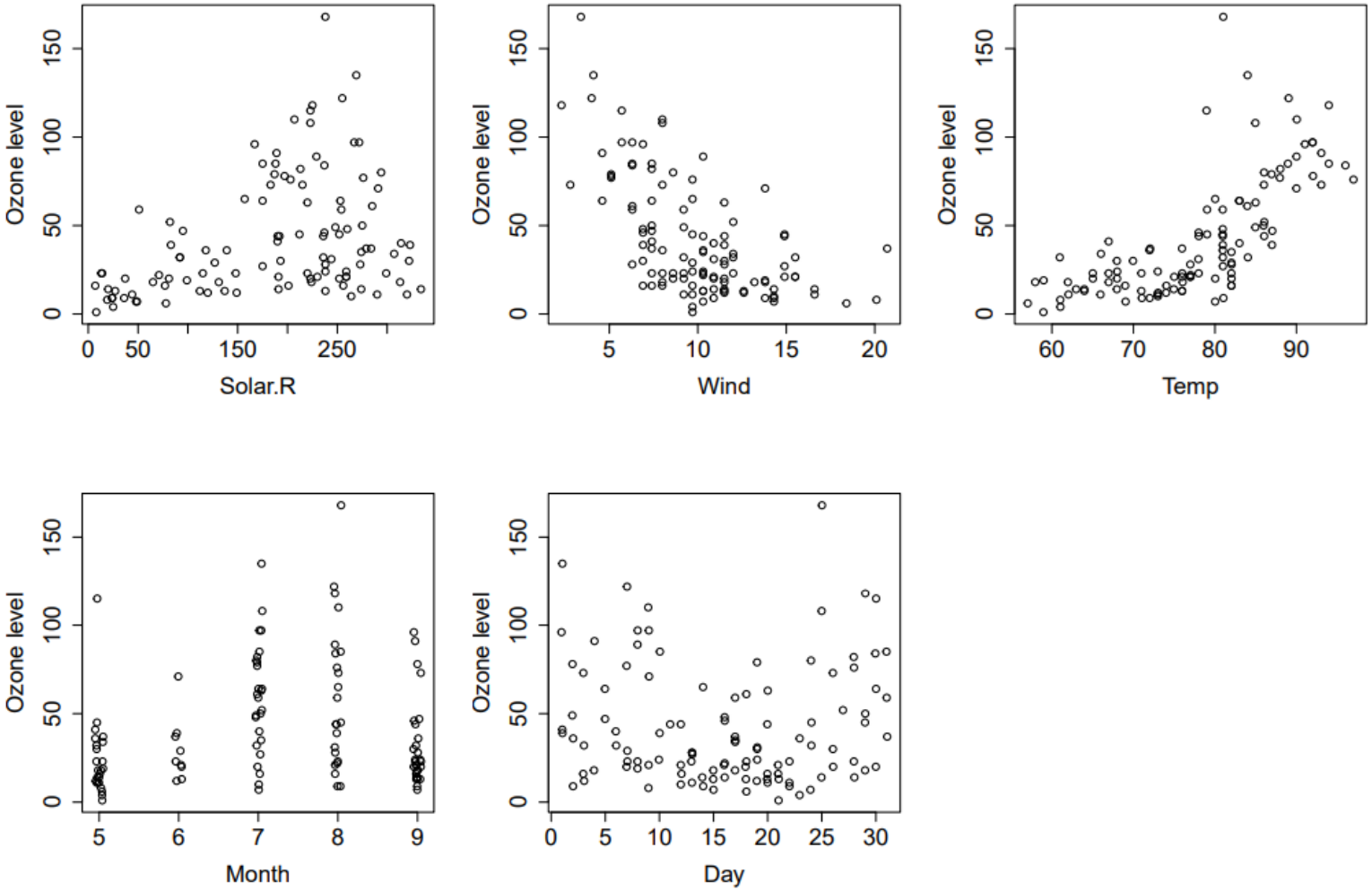
- Finding globally optimal partition: computationally infeasible.

- To turn it into feasible task:
  - Splits are recursive, each based on only a single splitting variable and value.
  - Two-way splits only.
  - Splits are found in a *greedy* fashion: In each node, split minimizing current error is implemented (no looking ahead).
  - R_m are rectangular and non-overlapping.

- Resulting structure easy to visualize as a decision tree.

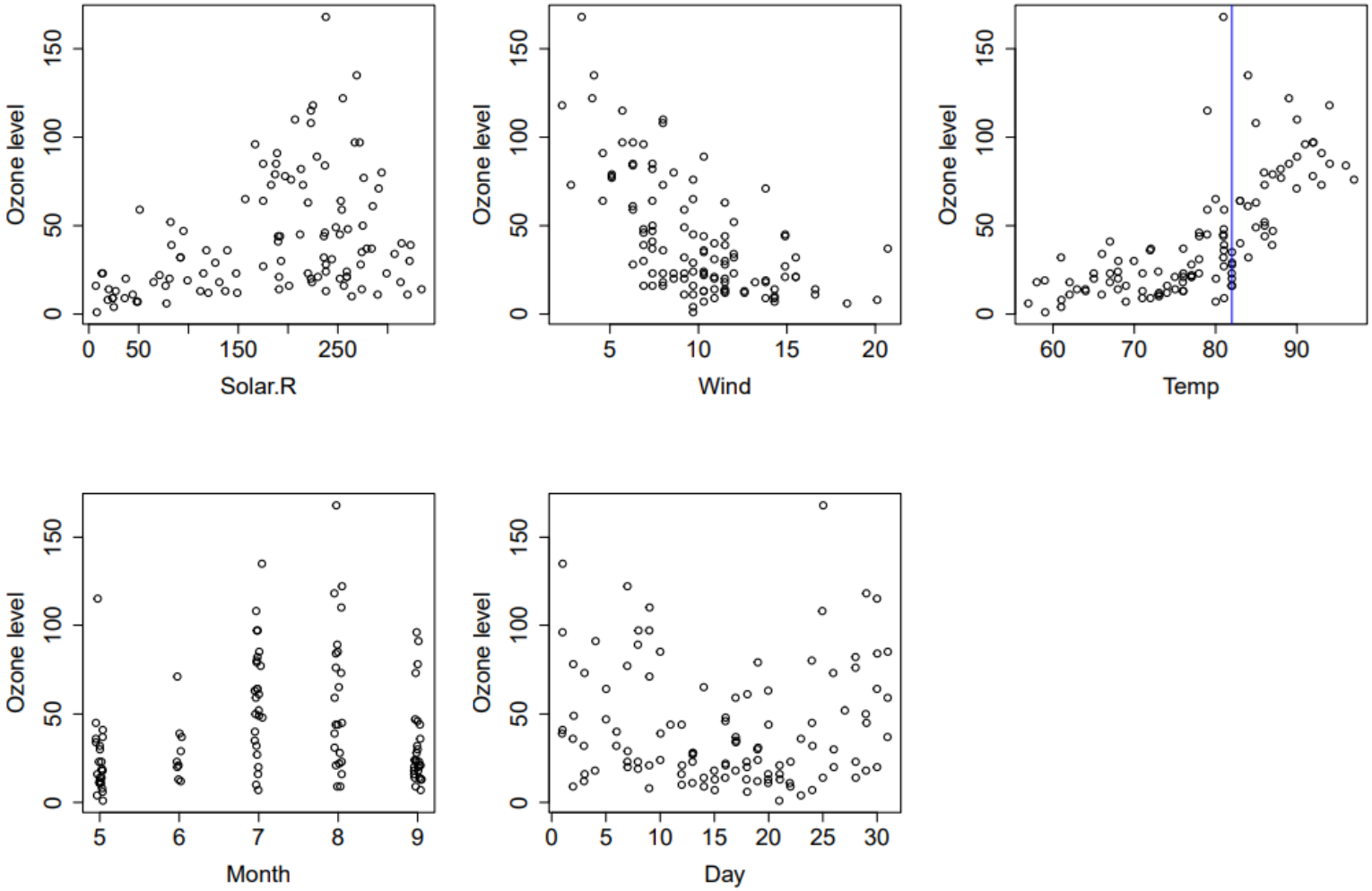
Exhaustive search for splits

Example dataset:



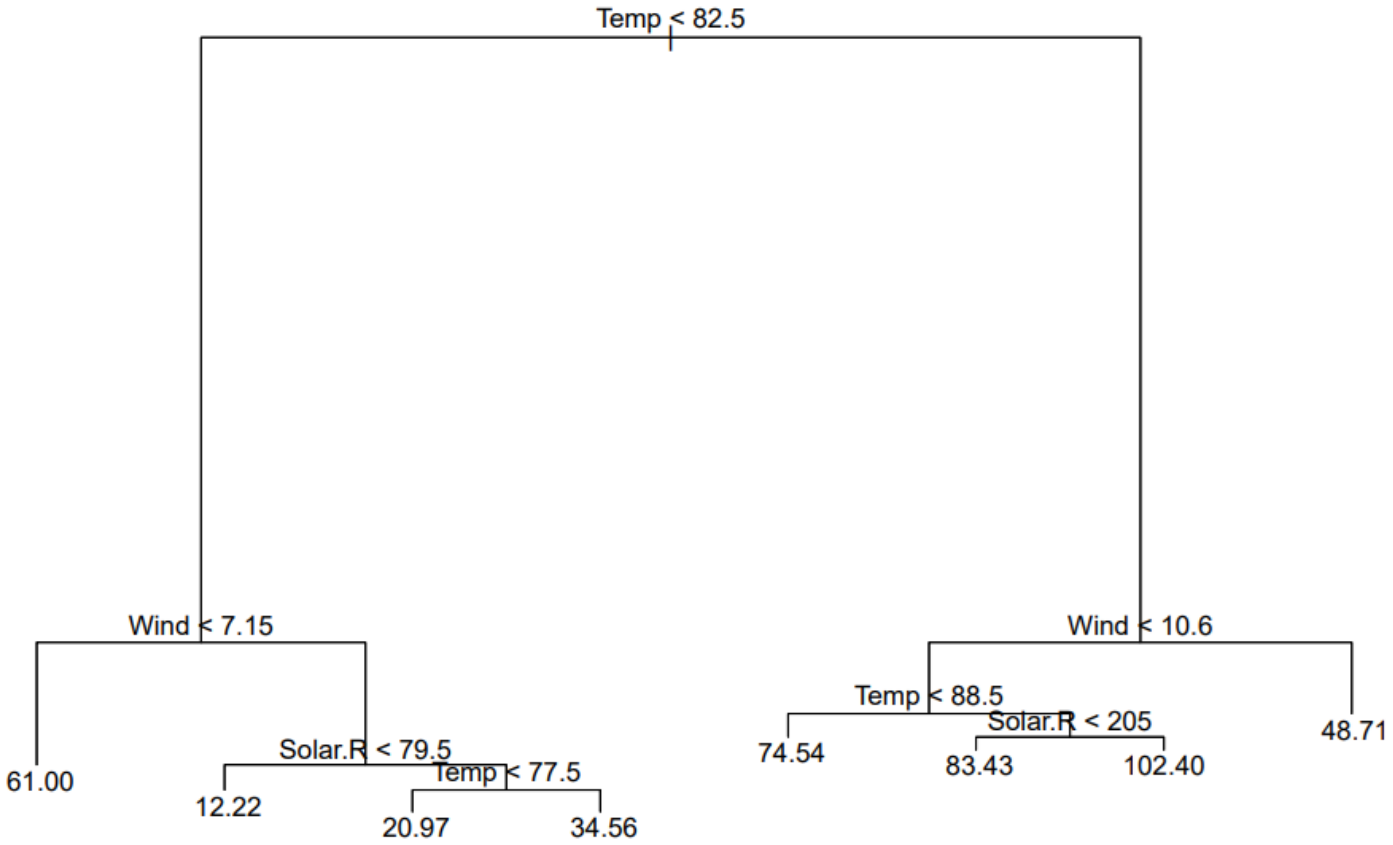
Greedy splitting

Find split that minimizes error in resulting two areas:



Greedy splitting

Repeat procedure in each resulting area:



Error (impurity) measures for split selection

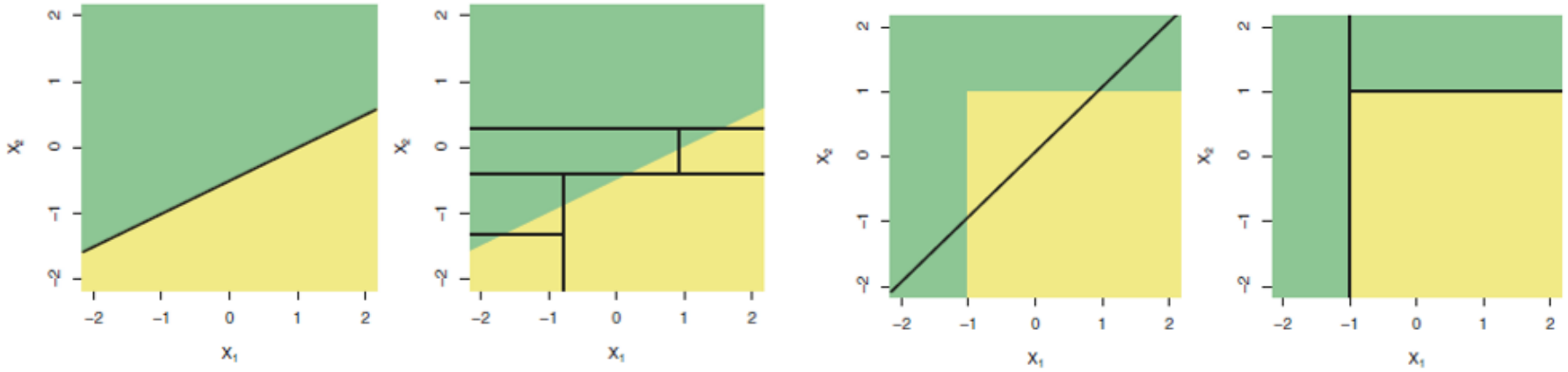
Numeric responses (m denotes terminal node, $\hat{y}$ mean or median):

- Mean squared error: $Q_m = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{y}_i)^2$
- Mean absolute error: $Q_m = \frac{1}{N_m} \sum_{x_i \in R_m} |y_i - \hat{y}_i|$

Categorical response with K classes (k denotes class, m terminal node, $\hat{p}$ observed proportion):

- Misclassification error: $Q_m = 1 - \max_k \hat{p}_{mk}$ (insensitive to changes in probabilities, do not use)
- Gini index: $Q_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$
- Cross entropy (a.k.a. deviance): $Q_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$ (cross entropy and gini index yield identical results)

Bias-variance trade-off



More splits → lower bias (more flexible), higher variance (more unstable).

As always, magnitude of bias and variance also depends on the true (but unknown) shapes of association, and sample size.

Sudden shifts in predicted values make trees more unstable (less smooth) than e.g. (generalized) linear and additive models.

Want to find optimal tree size, which minimizes sum of squared bias and variance.

Minimizing bias and variance through pruning

One way to reduce instability: Pruning

- Again, we optimize (error + penalty): $\sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$
- Find optimal value of α through k-fold CV

Variable selection bias

- Exhaustive search for splits (i.e., selecting variable and cutpoint at once) introduces variable selection bias: Variables of equal predictive power but with more possible cutpoints are more likely to be selected for splitting.

- Solution: At each split, select splitting variable first, then select optimal splitting value for that variable.

- This approach is used by conditional inference trees:

  1. Select splitting variable based on statistical association tests (conditional inference tests; Strasser and Weber, 1999)
  2. Given splitting variable, select cutpoint that minimizes error as usual.

- Additional benefit: Statistical tests provide natural stopping criterion for splitting: No need for pruning.