Regression Trees

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Regression

Goal: Estimate response from features

Optimal estimator: Conditional mean

Problem: Intractable to compute due to curse of dimensionality

Linear regression

Response y is approximated as an linear (affine) function of the features x

$$y \approx \sum_{i=1}^{d} \beta[i]x[i] + \alpha$$

Assumption: Response increases or decreases proportionally to each feature (if we fix other features)

Example

Response: Temperature in Manhattan (Kansas)

Features:

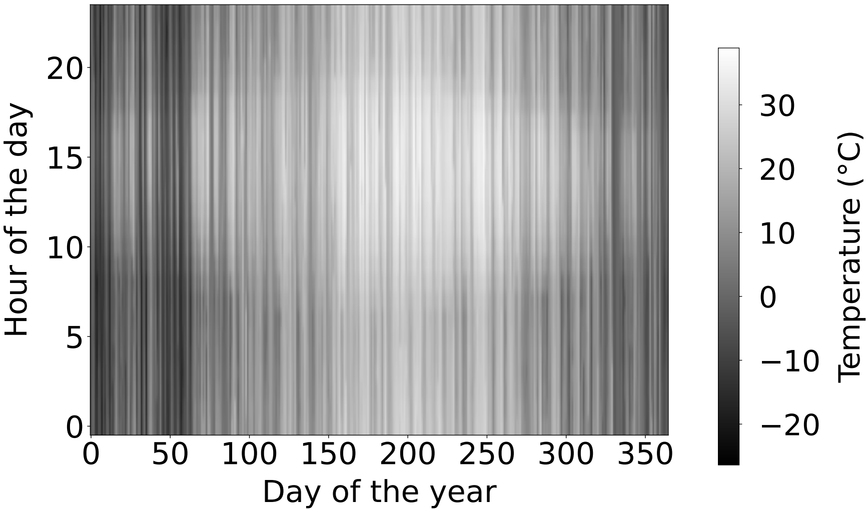
(1) Hour of the day (0-23)

(2) Day of the year (1-365)

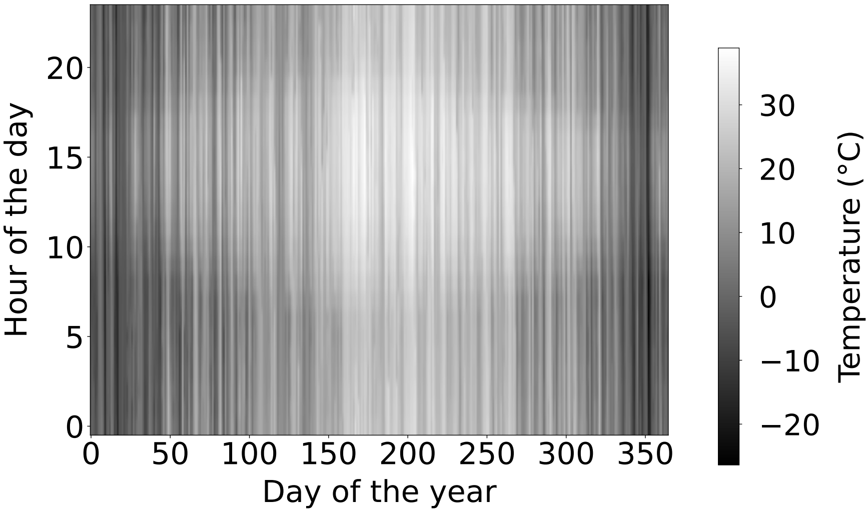
Training data: 2015

Test data: 2016

Training data

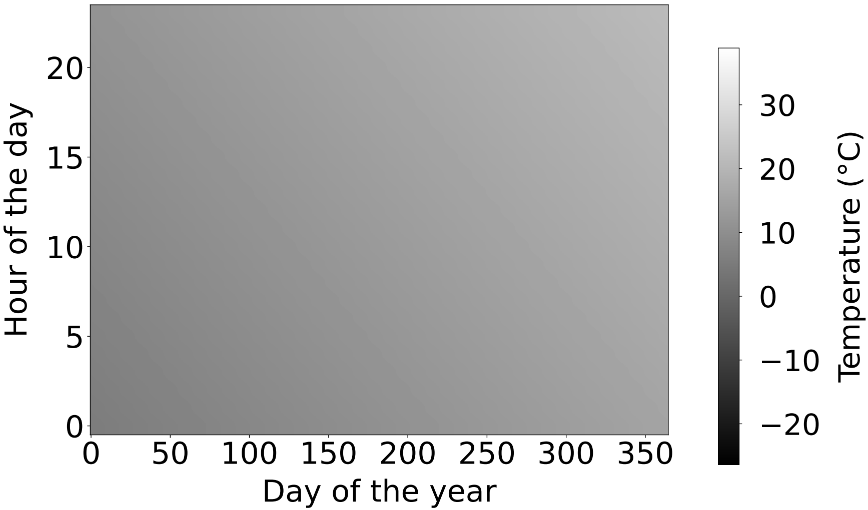


Test data



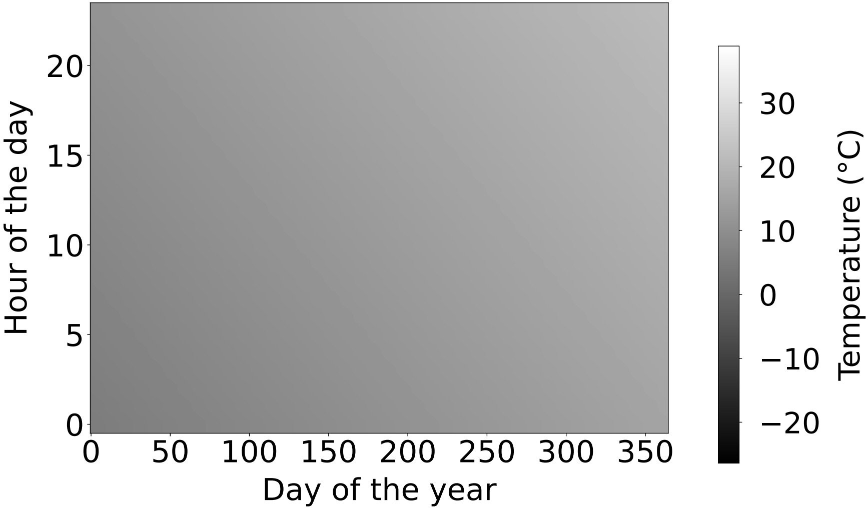
Hour of the day
0
5
10
15
20
0
50
100
150
200
250
300
350
Day of the year
Temperature (°C)
30
20
10
0
−10
−20

Linear model: 0.25 hour + 0.03 day + 5.85



Response increases or decreases proportionally to each feature (if we fix other features)

Linear model: 0.25 hour $+\ 0.03$ day $+\ 5.85$



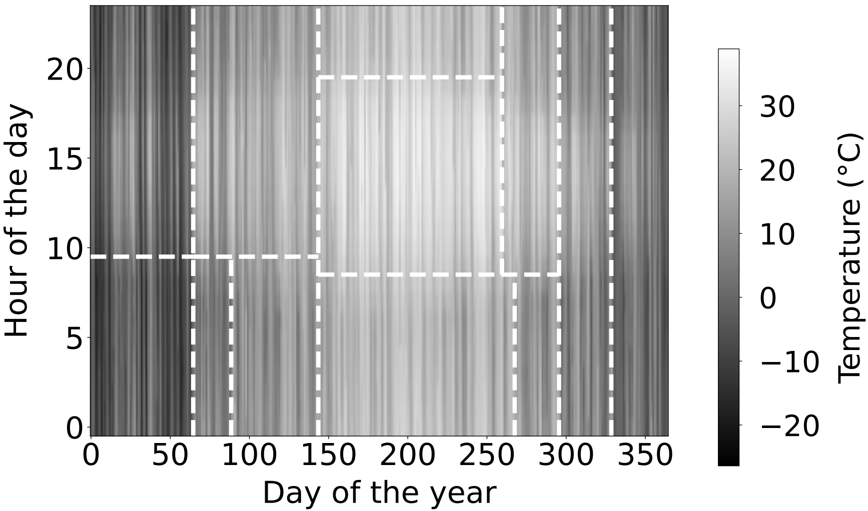
Training error: 10.8°C Test error: 11.0°C

Challenge

How to learn nonlinear model?

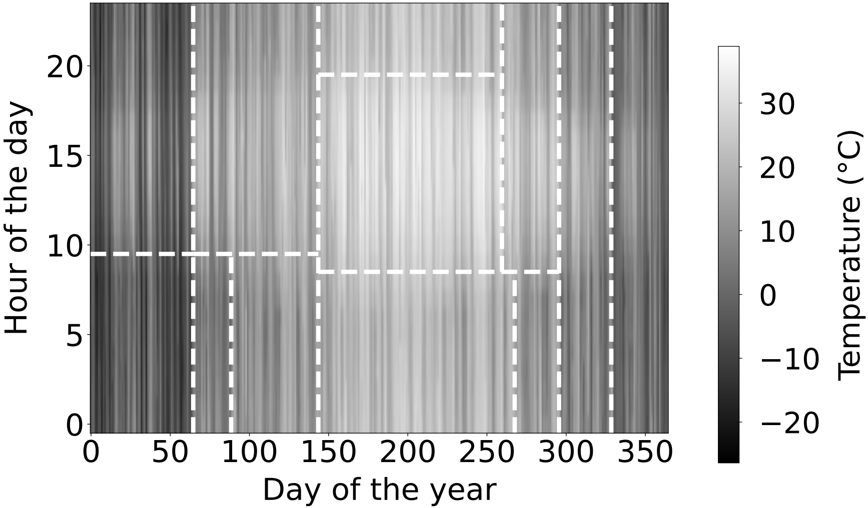
Idea

(1) Partition feature space into regions



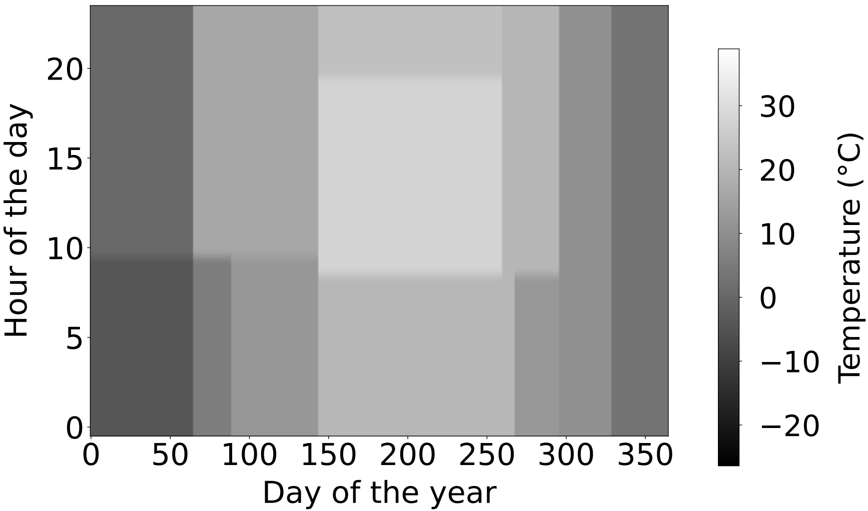
Idea

(2) Assign constant estimate to each region

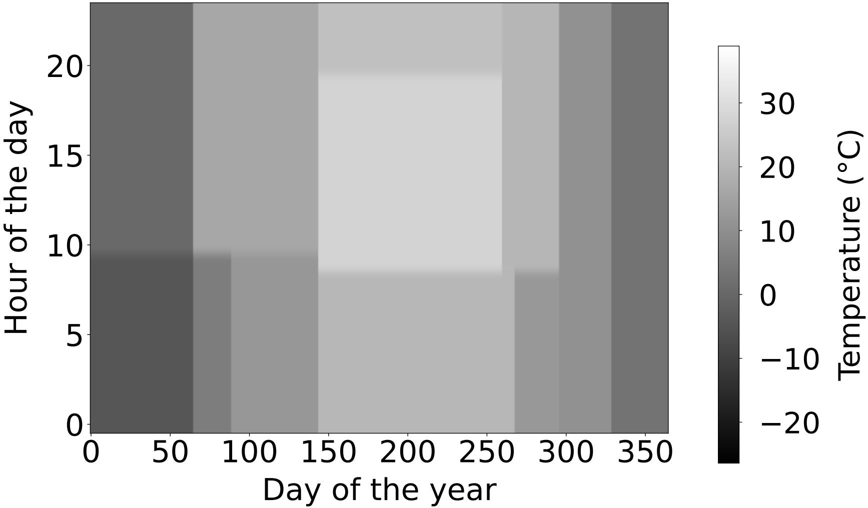


Idea

(2) Assign constant estimate to each region



Works pretty well!



Training error: 5.5°C Test error: 6.2°C

Two key questions

How to compute constant estimate?

How to choose the regions?

Constant estimate?

Consider the n_R feature-response pairs (x_i, y_i) in region R

$$\mathrm{RSS}(\alpha) := \sum_{\{i : x_i \in R\}} (y_i - \alpha)^2$$

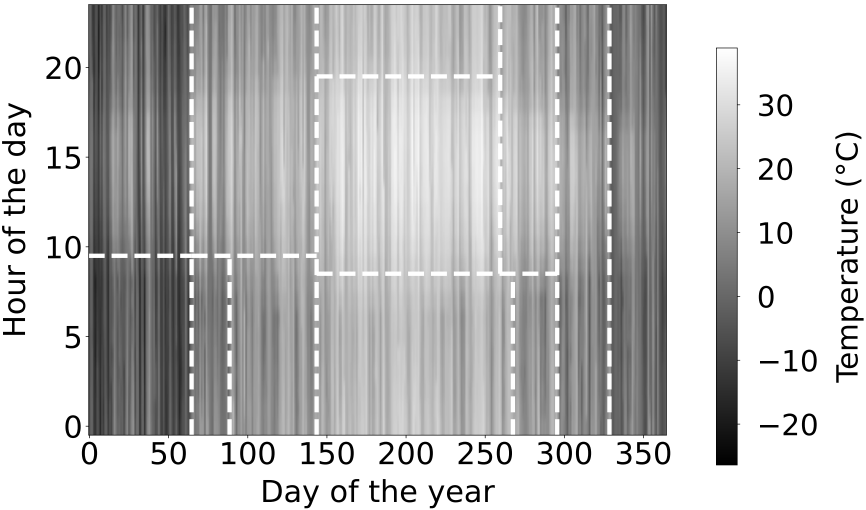
$$\frac{\mathrm{d}\,\mathrm{RSS}(\alpha)}{\mathrm{d}\alpha} = -\sum_{\{i : x_i \in R\}} (y_i - \alpha)$$

$$= n_R \alpha - \sum_{\{i : x_i \in R\}} y_i$$

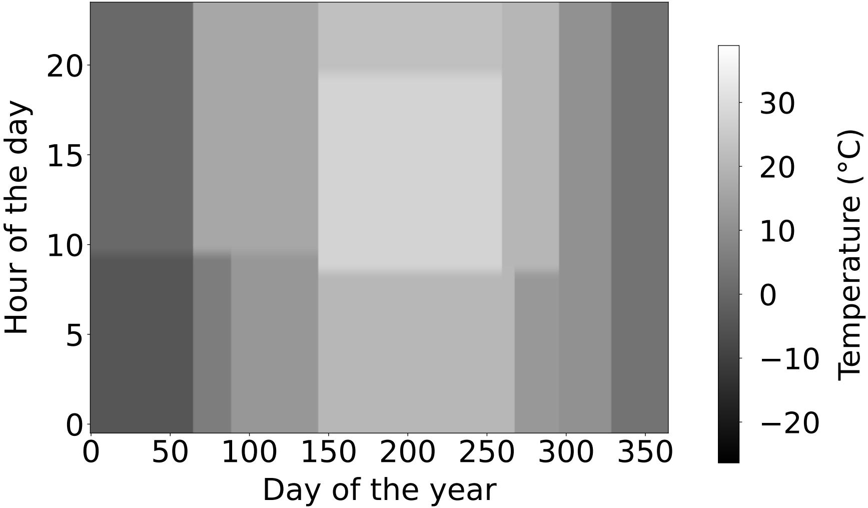
$$\frac{\mathrm{d}^2\,\mathrm{RSS}(\alpha)}{\mathrm{d}\alpha^2} = n_R$$

$$\alpha_{\min} = \frac{1}{n_R} \sum_{\{i : x_i \in R\}} y_i$$

Constant estimate?



Just average!

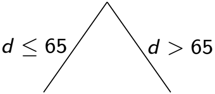


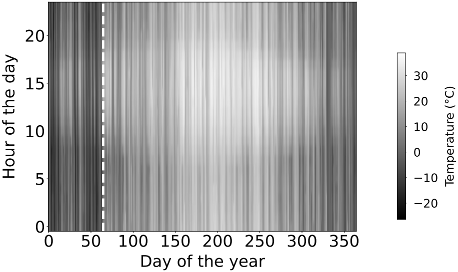
Regions?

Possible regions explode exponentially with number of features!

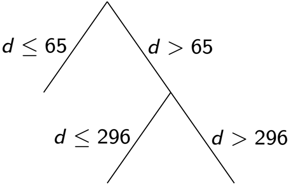
Idea: Use a binary tree to represent the regions

Tree

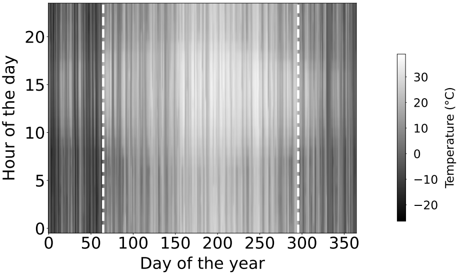

$d \leq 65$
$d > 65$


Hour of the day
Day of the year
Temperature (°C)
0
5
10
15
20
0
50
100
150
200
250
300
350
30
20
10
0
−10
−20

Tree

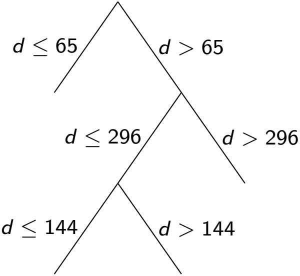


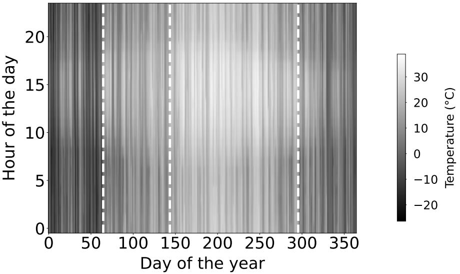
$d \leq 65$
$d > 65$
$d \leq 296$
$d > 296$



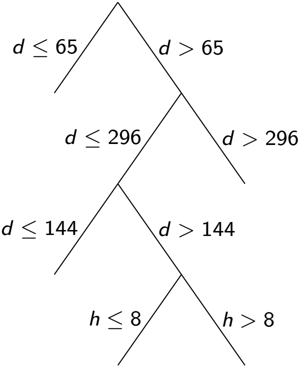
Hour of the day
0
5
10
15
20
Day of the year
0
50
100
150
200
250
300
350
Temperature (°C)
30
20
10
0
−10
−20

Tree

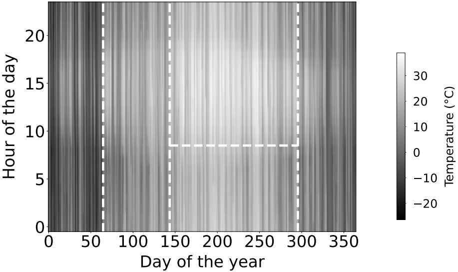

$d \leq 65$
$d > 65$
$d \leq 296$
$d > 296$
$d \leq 144$
$d > 144$


Hour of the day
0
5
10
15
20
0
50
100
150
200
250
300
350
Day of the year
Temperature (°C)
−20
−10
0
10
20
30

Tree

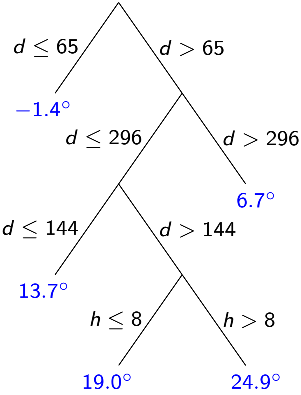


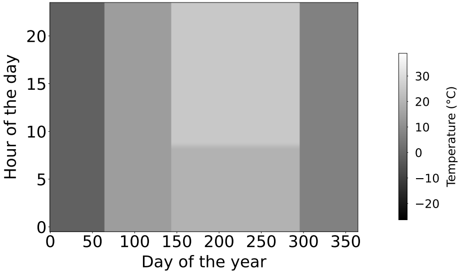
$d \leq 65$
$d > 65$
$d \leq 296$
$d > 296$
$d \leq 144$
$d > 144$
$h \leq 8$
$h > 8$



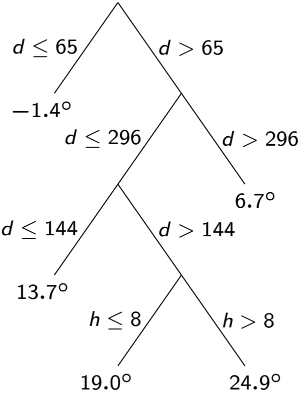
Hour of the day
0
5
10
15
20
0
50
100
150
200
250
300
350
Day of the year
Temperature (°C)
30
20
10
0
−10
−20

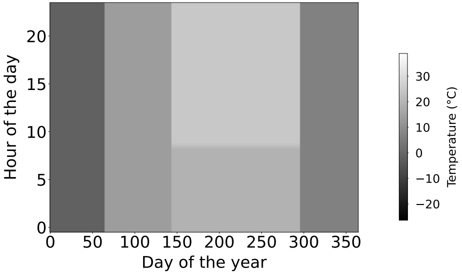
Region estimates



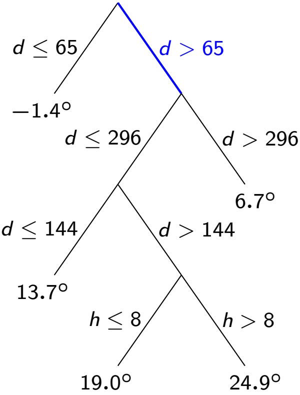


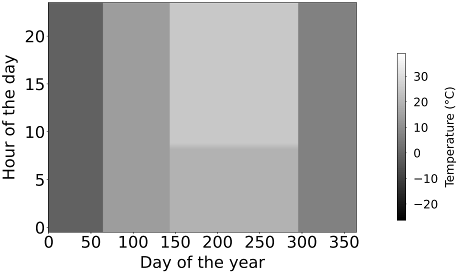
August 19 ($d := 251$) at 3 am ($h := 3$)?



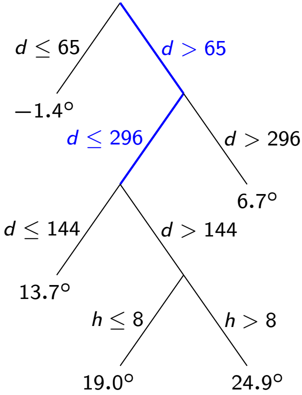


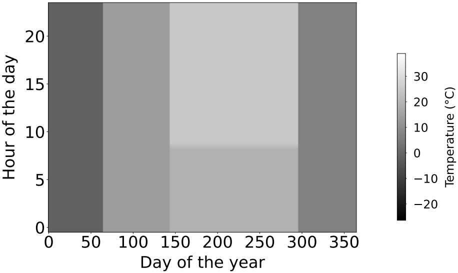
August 19 ($d := 251$) at 3 am ($h := 3$)?



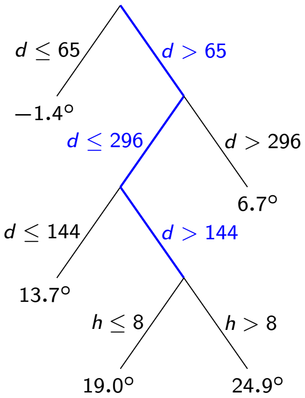


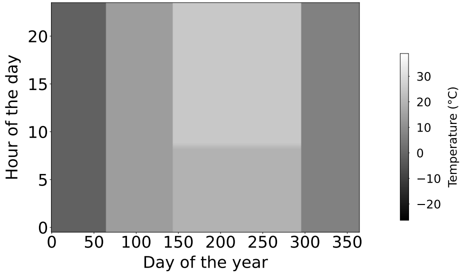
August 19 ($d := 251$) at 3 am ($h := 3$)?



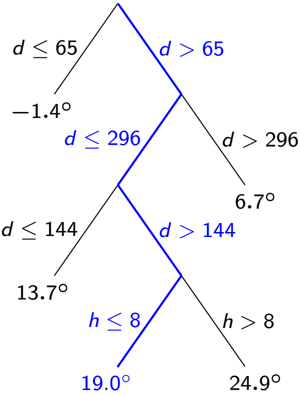


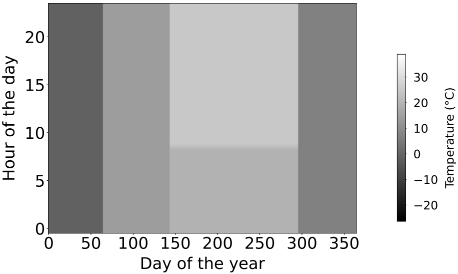
August 19 ($d := 251$) at 3 am ($h := 3$)?





August 19 ($d := 251$) at 3 am ($h := 3$)?





Interpretable!

Actual temperature (in 2023): 22°

How do we build the tree?

Idea: Choose tree with smallest training error

Tree with depth h and 2^h leaves

Number of possible bifurcations? $b := 2^h - 1$

At each bifurcation (1) d features and (2) t thresholds

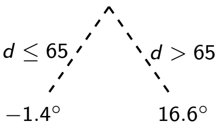
Number of possible trees? $(dt)^b$

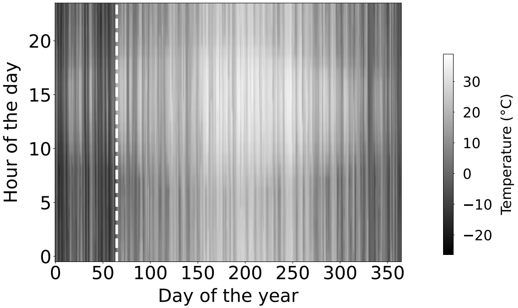
For $h := 4$, $d := 10$, $t := 100$: 10^{45} trees!

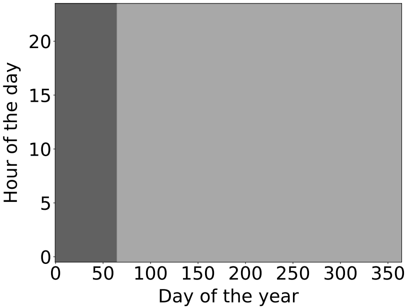
Recursive binary splitting

Add one bifurcation at a time, being greedy

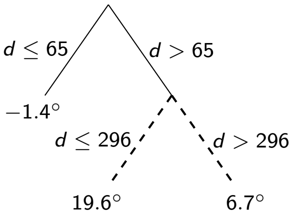
Bifurcation 1

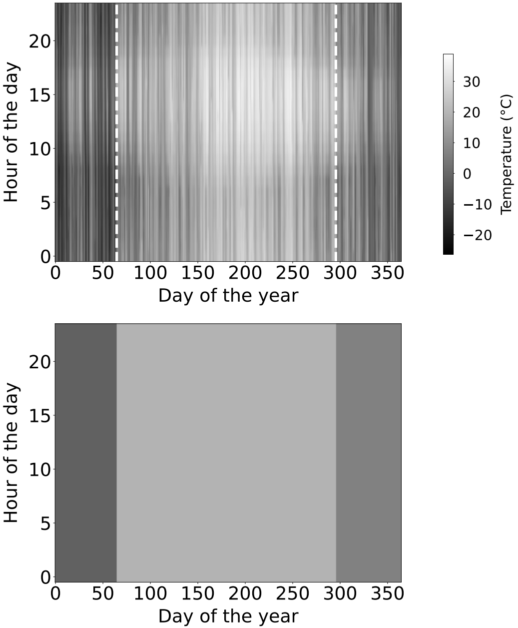




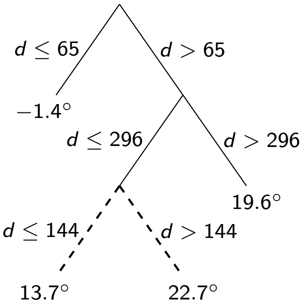


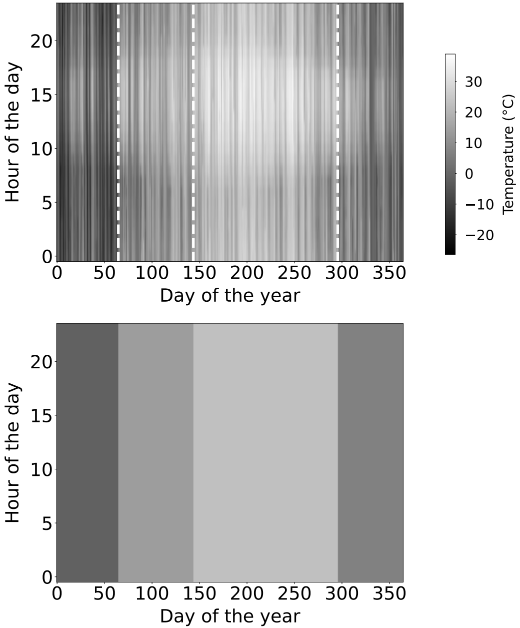
Bifurcation 2





Bifurcation 3





Residual Sum of Squares (RSS)

Data: (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n)

Regions: R_1, . . . , R_m

Estimates at each region: α_1, . . . , α_m

$$\text{Residual Sum of Squares} := \sum_{r=1}^{m} \sum_{\{i : x_i \in R_r\}} (y_i - \alpha_r)^2$$

Choosing a split

$$\text{RSS} := \sum_{r=1}^{m} \sum_{\{i : x_i \in R_r\}} (y_i - \alpha_r)^2$$

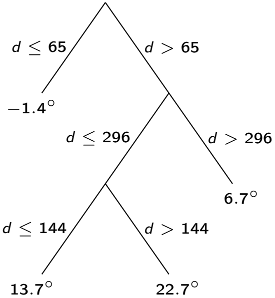
If we split region R_r into subregions A and B

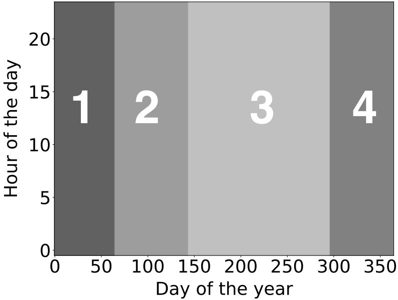
$$\triangle \text{RSS} := \sum_{\{i:x_i \in R_r\}}^{n} (y_i - \alpha_r)^2 - \sum_{\{i:x_i \in A\}}^{n} (y_i - \alpha_A)^2 - \sum_{\{i:x_i \in B\}}^{n} (y_i - \alpha_B)^2$$

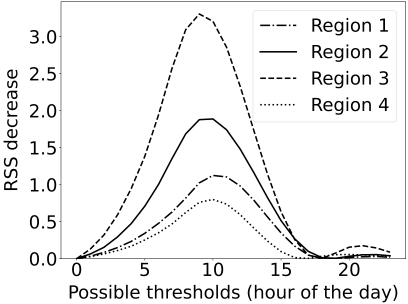
Depends on (1) region, (2) feature and (3) threshold

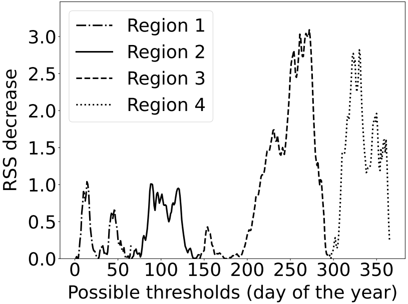
Choose split that maximizes $\triangle$ RSS

Choosing a split

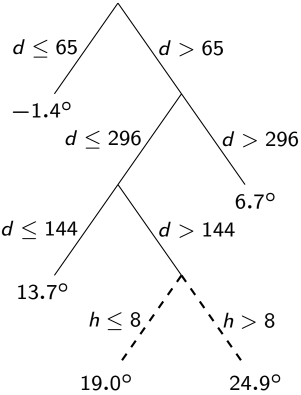


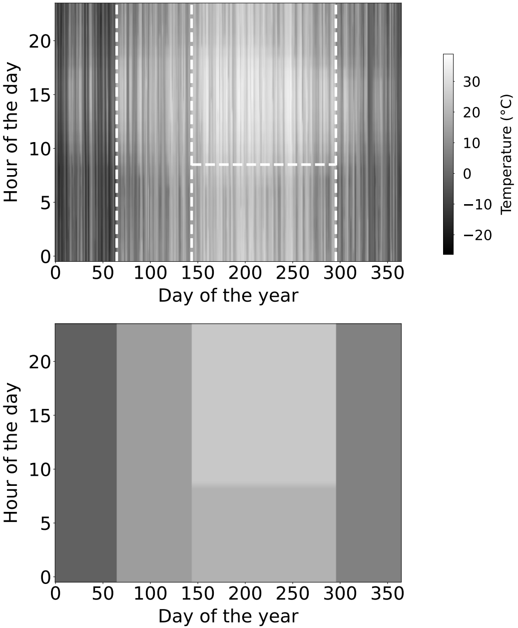




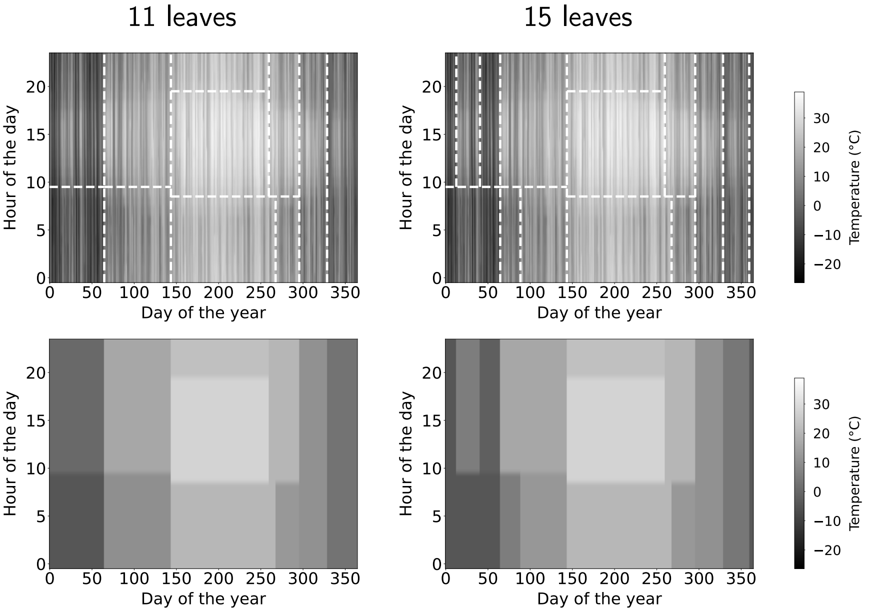


Bifurcation 4



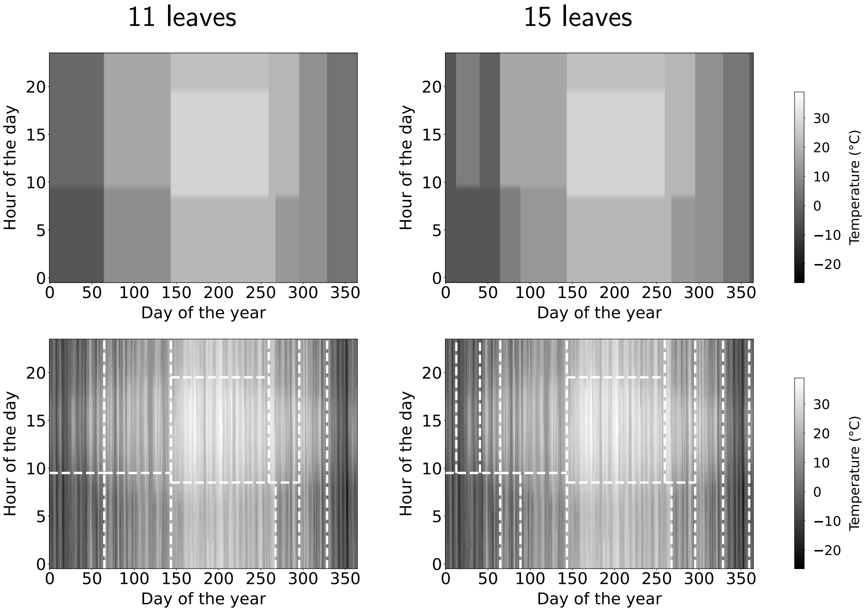


When to stop?

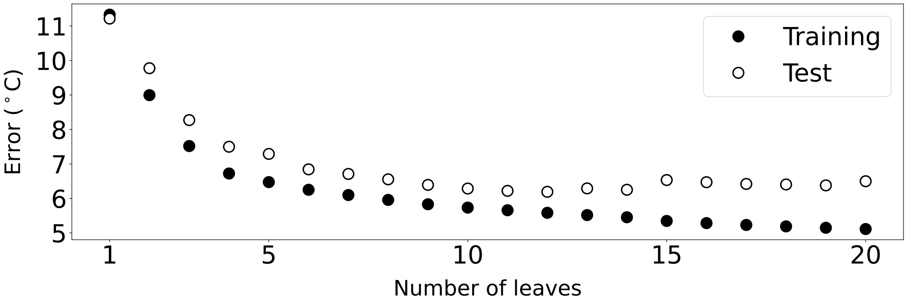


11 leaves
15 leaves
Hour of the day
Day of the year
Temperature (°C)
0
5
10
15
20
0
50
100
150
200
250
300
350
−20
−10
0
10
20
30

Test data


11 leaves
15 leaves
Hour of the day
Day of the year
Temperature (°C)

Training and test error



What have we learned?

How to build nonlinear regression models using trees

To be careful about overfitting!