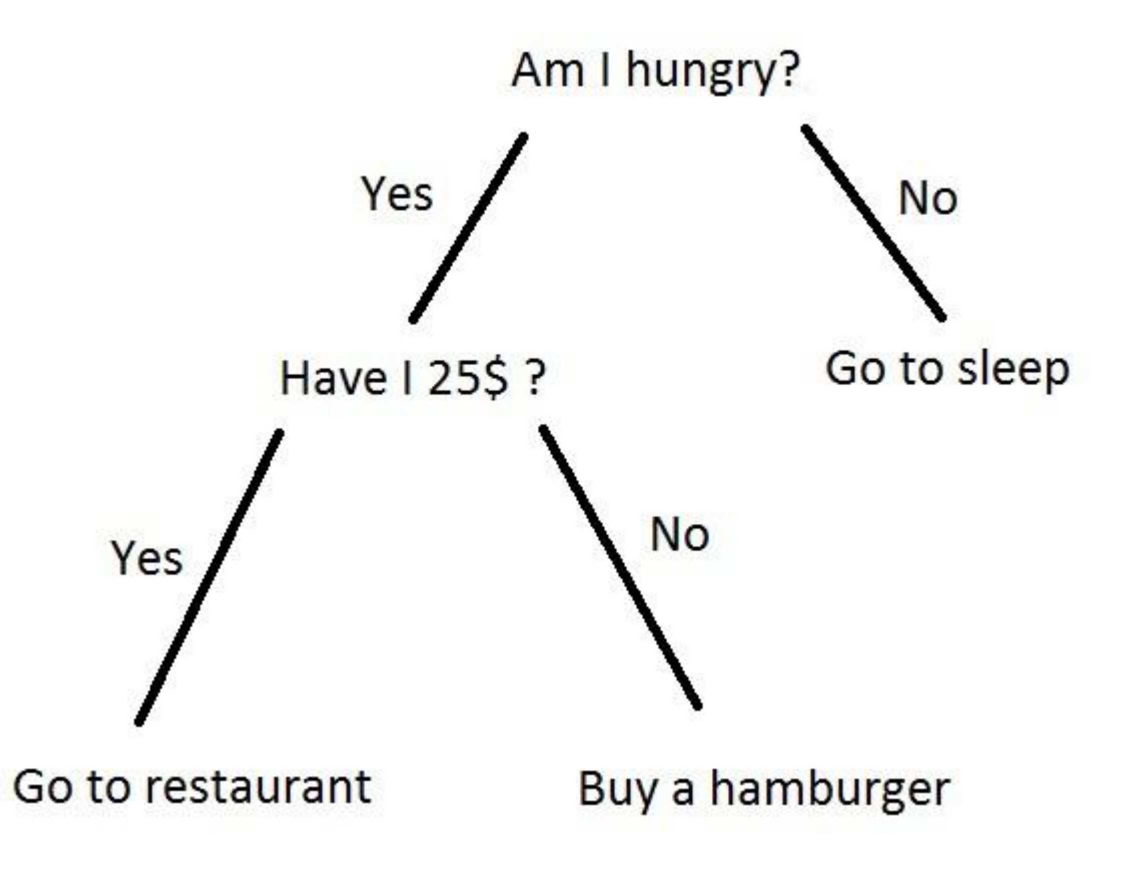
Data Science Course

Lecture 15

Decision Trees

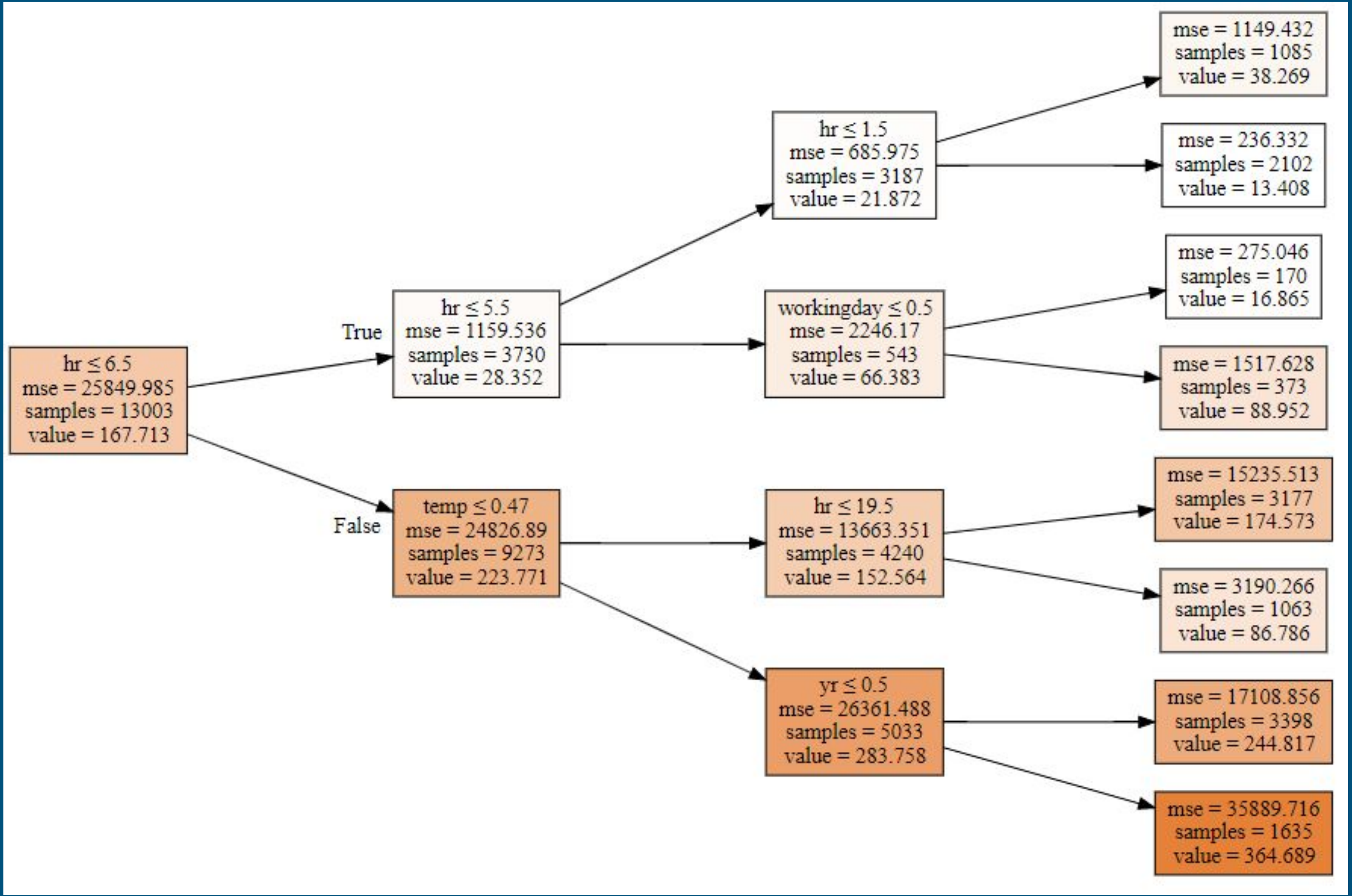
Decision Trees

- Classification and Regression Trees or CART.
- Nodes, edges, leafs.
- Sequence of splits.

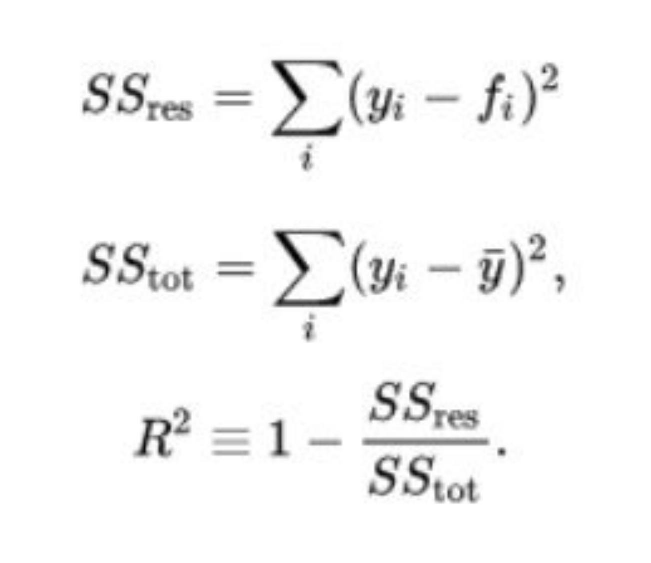

Am I hungry?
Yes
No
Have I 25$?
Go to sleep
Yes
No
Go to restaurant
Buy a hamburger

Training

- Optimal split: iterate over the predictors and their possible values.
- Target average of resulting sub-groups.
- Calculate MSE.
- Weighted average.


hr ≤ 6.5
mse = 25849.985
samples = 13003
value = 167.713
True
False
hr ≤ 5.5
mse = 1159.536
samples = 3730
value = 28.352
temp ≤ 0.47
mse = 24826.89
samples = 9273
value = 223.771
hr ≤ 1.5
mse = 685.975
samples = 3187
value = 21.872
workingday ≤ 0.5
mse = 2246.17
samples = 543
value = 66.383
hr ≤ 19.5
mse = 13663.351
samples = 4240
value = 152.564
yr ≤ 0.5
mse = 26361.488
samples = 5033
value = 283.758
mse = 1149.432
samples = 1085
value = 38.269
mse = 236.332
samples = 2102
value = 13.408
mse = 275.046
samples = 170
value = 16.865
mse = 1517.628
samples = 373
value = 88.952
mse = 15235.513
samples = 3177
value = 174.573
mse = 3190.266
samples = 1063
value = 86.786
mse = 17108.856
samples = 3398
value = 244.817
mse = 35889.716
samples = 1635
value = 364.689

Evaluation - R²



$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

Balance

Pros:

- Simple to understand, interpret, visualize.
- Can handle numerical and categorical data.
- Regression and classification.

Cons:

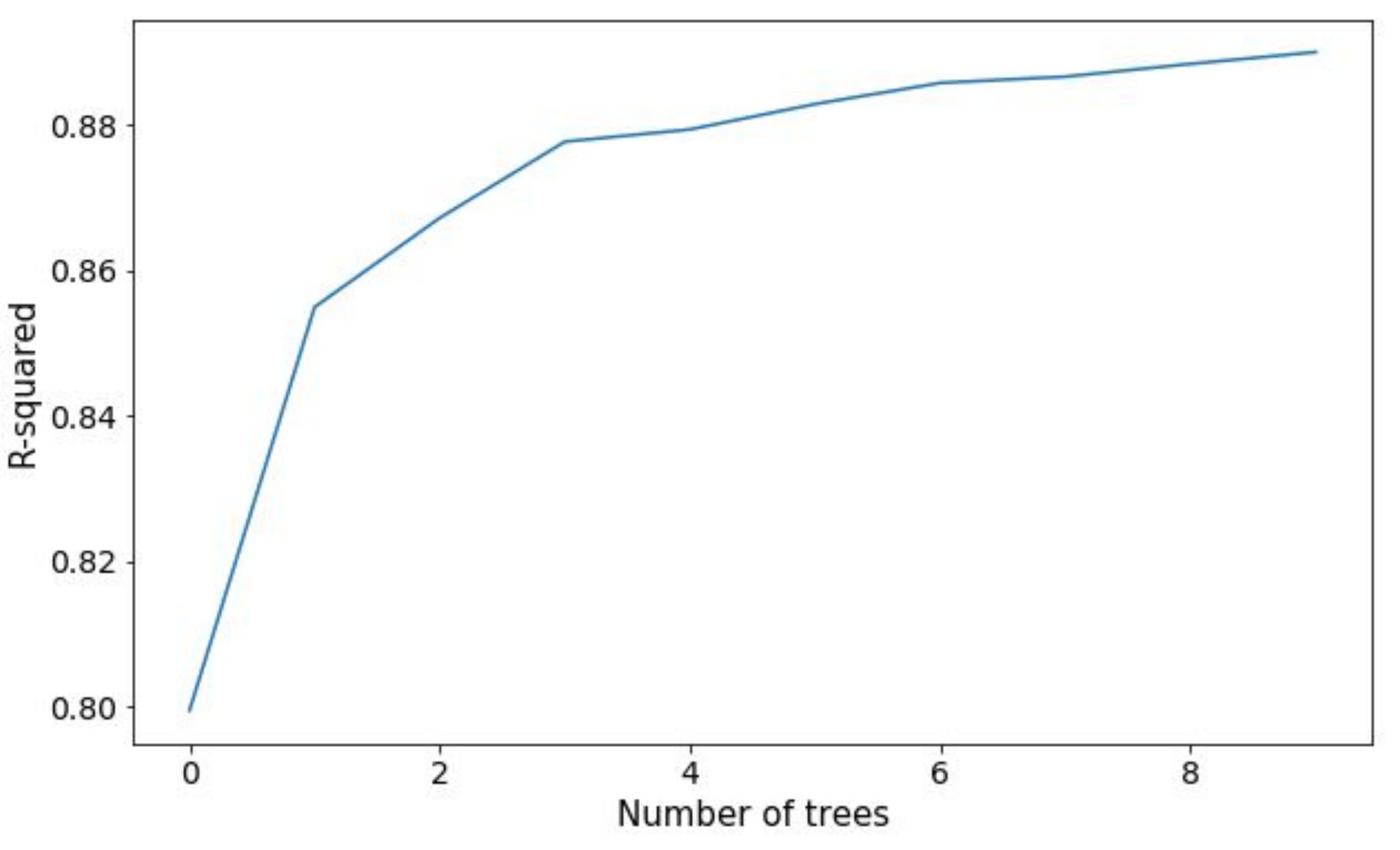
- Over-complex trees that do not generalize the data well.
- Biased trees if some classes dominate.

Bagging

- Objective: reduce the variance (robustness and accuracy).
- Multiple models not correlated.

Random Forest

- Small tweak to decorrelate the trees.


R-squared
0.80
0.82
0.84
0.86
0.88
0
2
4
6
8
Number of trees