ONLINE MASTERS IN **DATA SCIENCE**

DSC 255 - MACHINE LEARNING FUNDAMENTALS

OVERFITTING IN DECISION TREE

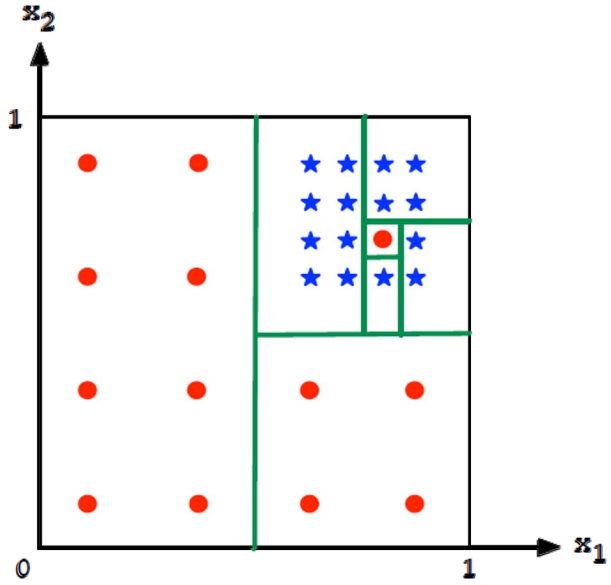
SANJOY DASGUPTA, PROFESSOR

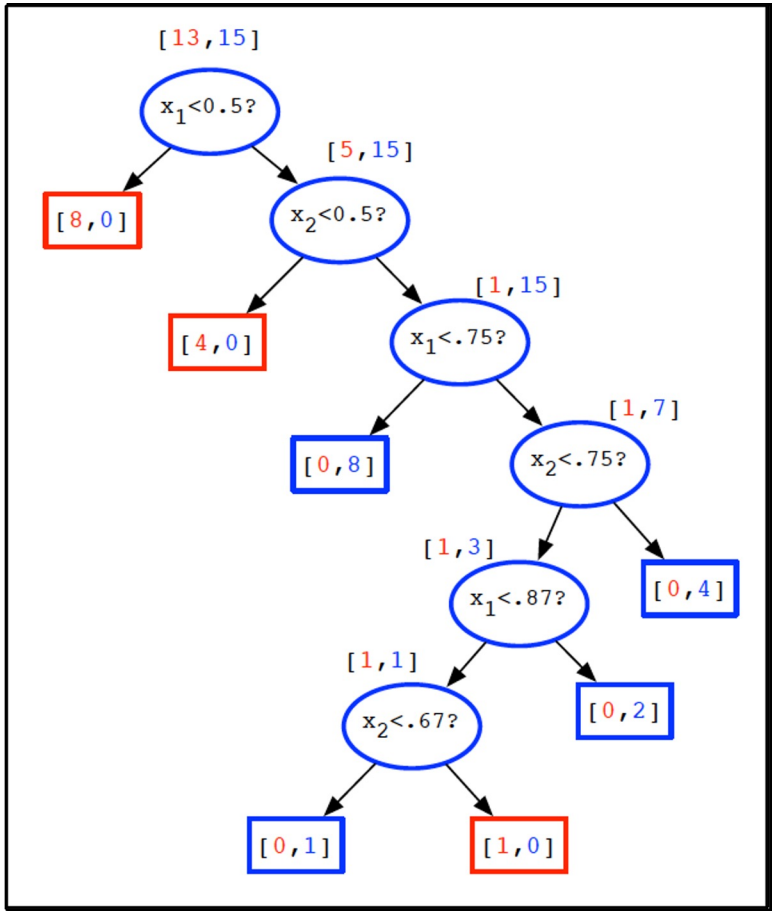
UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

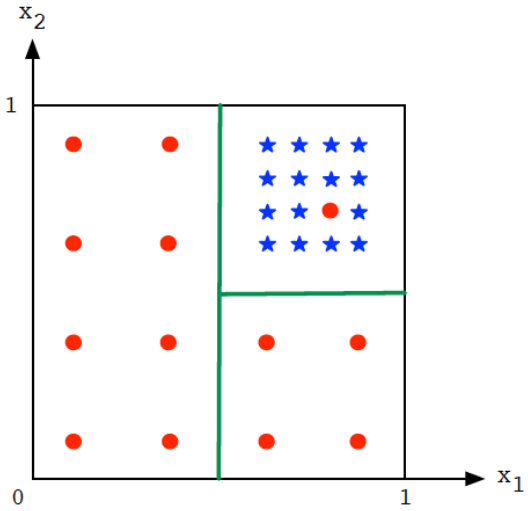
Example: Building a Decision Tree

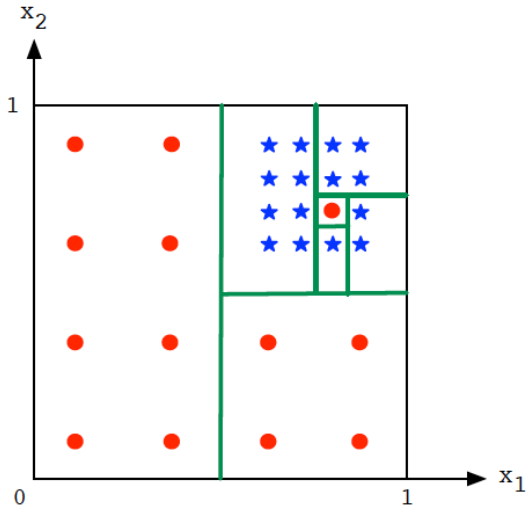




Overfitting?

Go back a few steps...

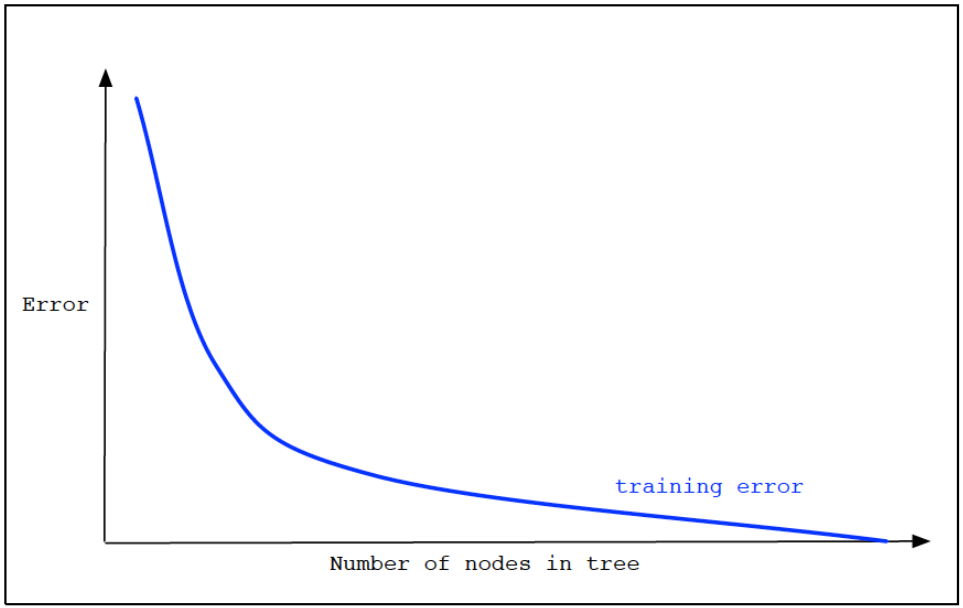




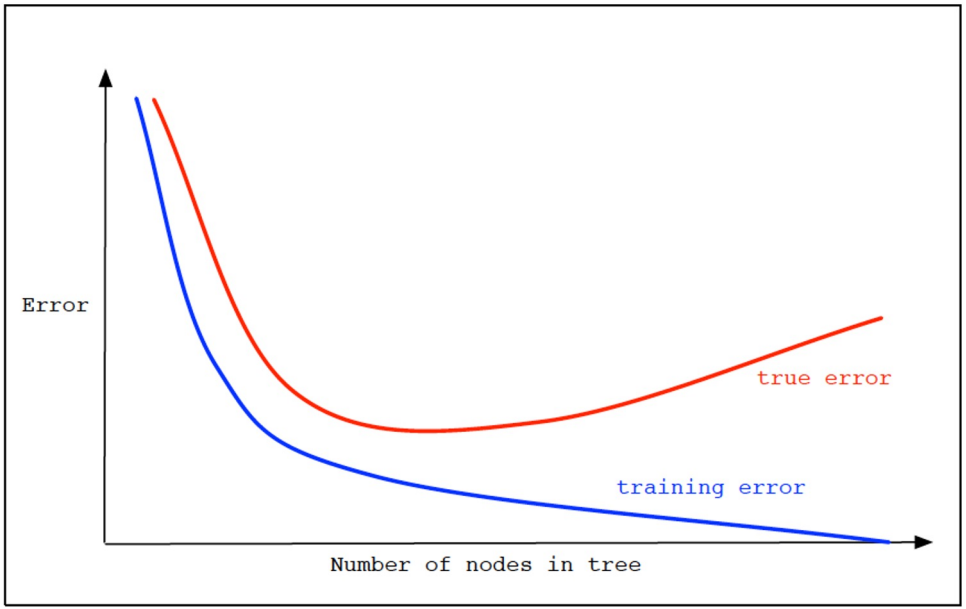
Final partition does better on training data but is more complex.
That one point might have been an outlier anyway.

We have probably ended up **overfitting** the data.

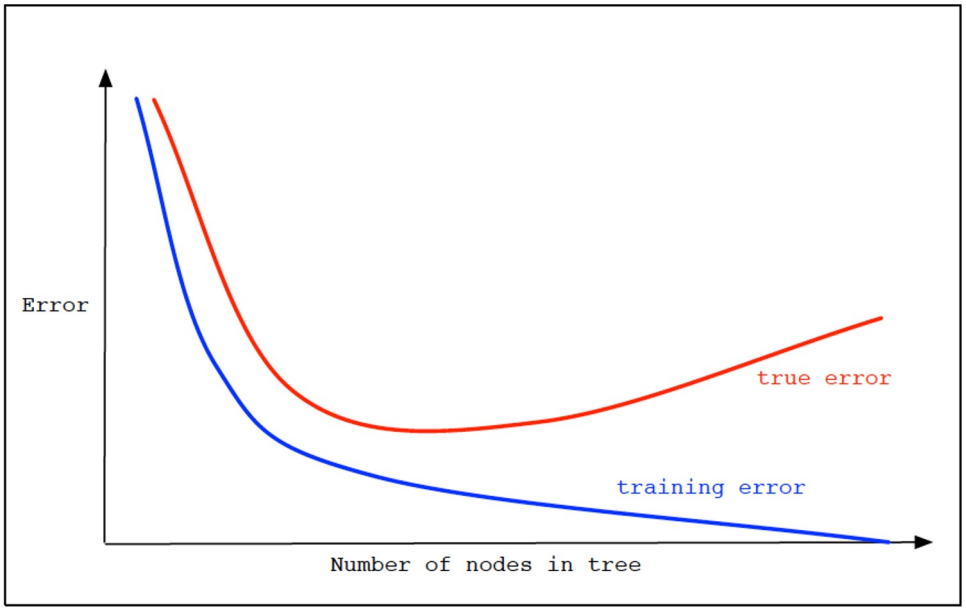
Overfitting Picture



Overfitting Picture



Overfitting Picture

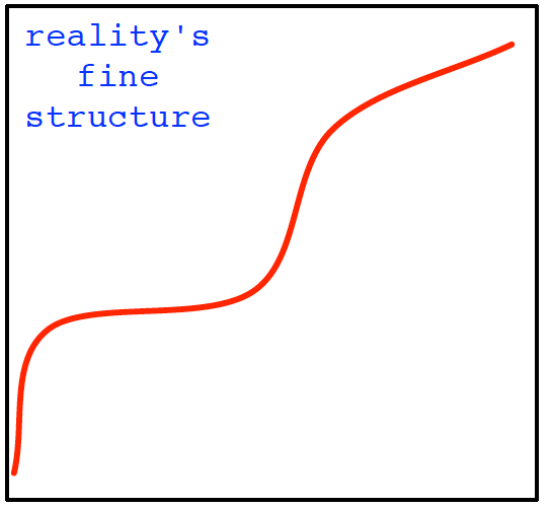


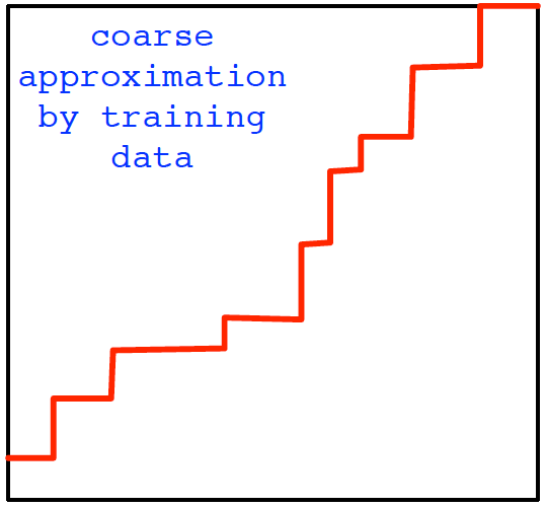
As we make our tree more and more complicated:

- training error keeps going down
- but, at some point, true error starts increasing!

Overfitting: Perspectives

- The true underlying distribution D is the one whose structure we would like to capture.
- The training data reflects the structure of D, so it helps us.
- But it also has chance structure of its own - we must avoid modeling this.





Decision Tree Issues

A very expressive family of classifiers:

- Can accommodate any type of data: real, Boolean, categorical, …
- Can accommodate any number of classes
- Can fit any data set
- Statistically consistent

Decision Tree Issues

A very expressive family of classifiers:

- Can accommodate any type of data: real, Boolean, categorical, …
- Can accommodate any number of classes
- Can fit any data set
- Statistically consistent

But this also means that there is a serious danger of overfitting.

Building a Decision Tree

- Start with a single node containing all data points
- Repeat:
  - Look at all current leaves and all possible splits
  - Choose the split with the greatest benefit

When to stop?

Building a Decision Tree

- Start with a single node containing all data points
- Repeat:
  - Look at all current leaves and all possible splits
  - Choose the split with the greatest benefit

When to stop?

- When each leaf is pure?
- When the tree is already pretty big?
- When each leaf has uncertainty below some threshold?

Building a Decision Tree

- Start with a single node containing all data points
- Repeat:
  - Look at all current leaves and all possible splits
  - Choose the split with the greatest benefit

When to stop?

- When each leaf is pure?
- When the tree is already pretty big?
- When each leaf has uncertainty below some threshold?

Common strategy: keep going until leaves are pure.
Then, shorten the tree by **pruning**, to correct for overfitting.