Lecture 8: 17 February, 2022

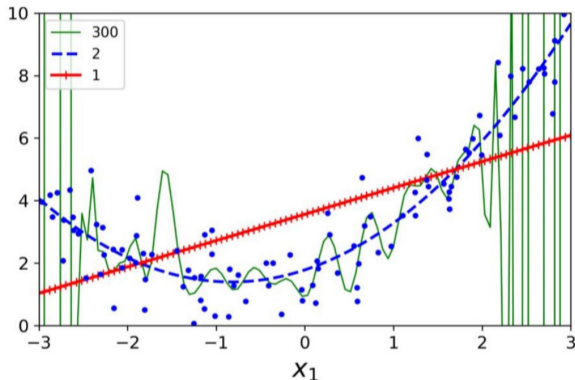Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
January–May 2022

Overfitting

- Overfitting: model too specific to training data, does not generalize well

- Regression — use regularization to penalize model complexity

- What about decision trees?

- Deep, complex trees ask too many questions

- Prefer shallow, simple trees
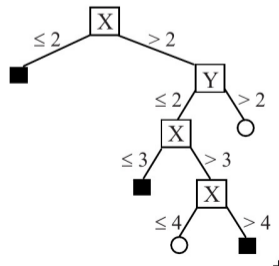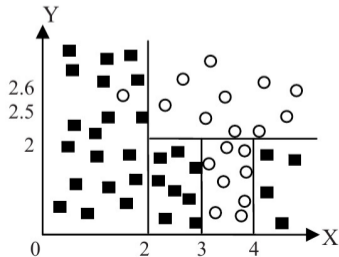
Tree pruning

- Remove leaves to improve generalization

- Top-down pruning
  - Fix a maximum depth when building the tree
  - How to decide the depth in advance?

- Bottom-up pruning
  - Build the full tree
  - Remove a leaf if the reduced tree generalizes better
  - How do we measure this?

Tree pruning

Overfitted tree
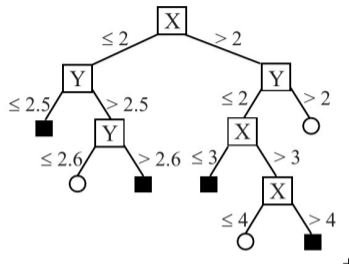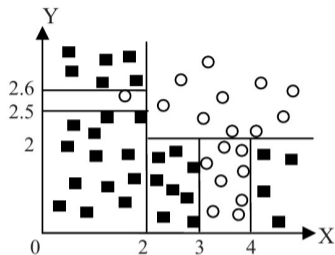
Pruned tree

Bottom up tree pruning

- Build the full tree, remove leaf if the reduced tree generalizes better

- How do we measure this?

- Check performance on a test set

- Use sampling theory [Quinlan]

- Given n coin tosses with h heads, estimate probability of heads as h/n
  - Estimate comes with a confidence interval: $h/n \pm \delta$
  - As n increases, δ reduces: 7 heads out of 10 vs 70 out of 100 vs 700 out of 1000

- Impure node, majority prediction, compute confidence interval

- Pruning leaves creates a larger impure sample one level above

- Does the confidence interval decrease (improve)?

- Predict party affiliation of US legislators based on voting pattern
    - Read the tree from left to right

- After pruning, drastically simplified tree

- Quinlan's comment on his use of sampling theory for post-pruning

Now, this description does violence to statistical notions of sampling and confidence limits, so the reasoning should be taken with a large grain of salt. Like many heuristics with questionable underpinnings, however, the estimates it produces seem frequently to yield acceptable results.

```
physician fee freeze = n:
    adoption of the budget resolution = y: democrat (151)
    adoption of the budget resolution = u: democrat (1)
    adoption of the budget resolution = n:
        education spending = n: democrat (6)
        education spending = y: democrat (9)
        education spending = u: republican (1)
physician fee freeze = y:
    synfuels corporation cutback = n: republican (97/3)
    synfuels corporation cutback = u: republican (4)
    synfuels corporation cutback = y:
        duty free exports = y: democrat (2)
        duty free exports = u: republican (1)
        duty free exports = n:
            education spending = n: democrat (5/2)
            education spending = y: republican (13/2)
            education spending = u: democrat (1)
physician fee freeze = u:
    water project cost sharing = n: democrat (0)
    water project cost sharing = y: democrat (4)
    water project cost sharing = u:
        mx missile = n: republican (0)
        mx missile = y: democrat (3/1)
        mx missile = u: republican (2)
```