Lecture-40

Instability, Smoothness & Repeated Subtrees

Instable :- More variance. Prone to Noises

Lack of smoothness :- Since decision tree has piecewise rectangular boxes. clear cut boundary

Repeated subtrees -:- Same set of split made at different points in decision tree.

Loss $L_{KK'} : \sum_{K=1}^{K} \hat{p}_{mK}(1-\hat{p}_{mK}) = \sum_{K=1}^{K} \hat{p}_{mK}\hat{p}_{mK'}$

(K: True, K': estimated)

-:- Advantages :-

- Simple to understand and interpret.
- handle both numerical and categorical data.
- Performs well with large datasets.

disadvantages -:

- Non-Robust : A small change in the training data can result in a large change in the tree. and consequently final Predictions.
- For data including categorical variables with different number of levels, Information-gain is biased in favour of attributes with more levels.
- Decision tree can learn very complex decision boundary. Overfitting.