Lecture 12: Variable Importance in Tree Models

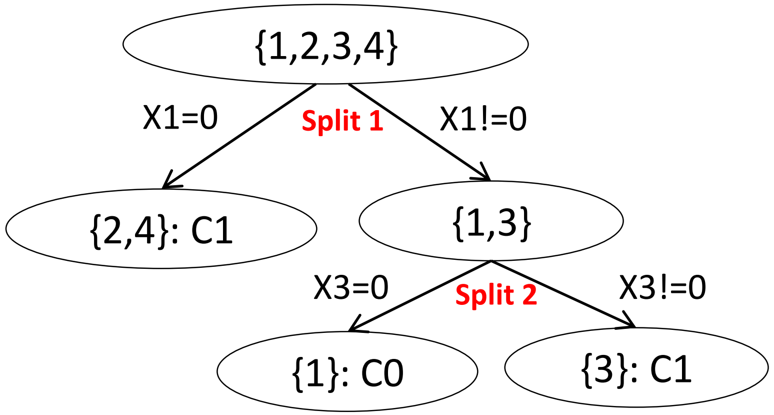
Instructor: Prof. Shuai Huang

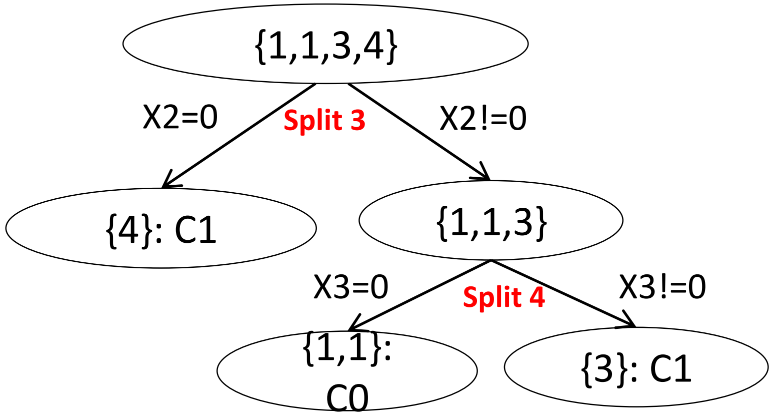
Industrial and Systems Engineering

University of Washington

Importance score in Random Forest (RF)

ID	X_1	X_2	X_3	Class
1	1	1	0	C0
2	0	0	0	C1
3	1	1	1	C1
4	0	0	1	C1



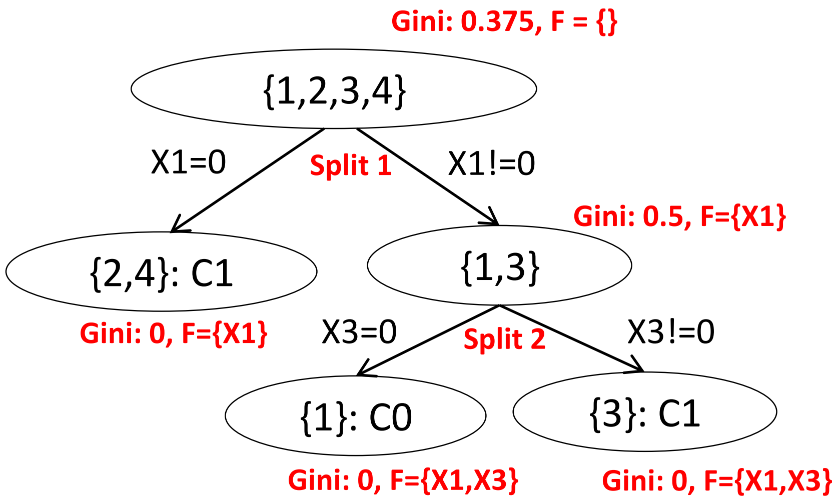


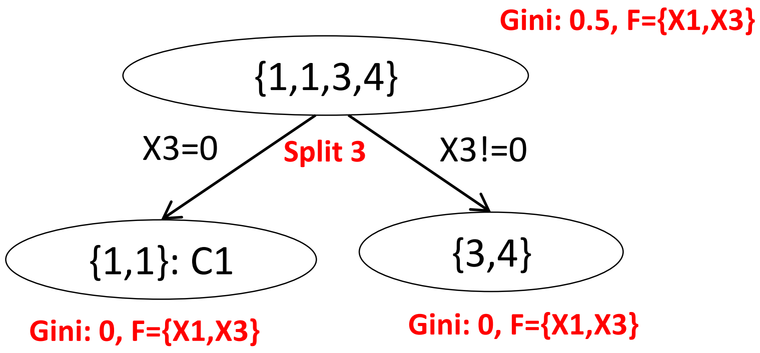
Importance score in Regularized RF (RRF)

ID	X_1	X_2	X_3	Class
1	1	1	0	C0
2	0	0	0	C1
3	1	1	1	C1
4	0	0	1	C1

The regularized impurity gain of variable X_i at a node is calculated as

$$Gain'(X_i) = \left\{ \begin{array}{ll} \lambda \cdot Gain(X_i) & X_i \notin F \\ Gain(X_i) & X_i \in F \end{array} \right\}$$





Importance score in Guided RRF (GRRF)

In GRRF, instead having one λ for all variables, each variable X_i can have its own λ_i:

$$Gain'(X_i) = \begin{Bmatrix} \lambda_i \cdot Gain(X_i) & X_i \notin F \\ Gain(X_i) & X_i \in F \end{Bmatrix},$$

where λ_i is

$$\lambda_i = (1 - \gamma)\lambda_0 + \gamma * w_i,$$

where λ_0 controls the base regularization, $w_i \in [0,1]$ is a prior of importance of each variable v_i, and $\gamma \in [0,1]$ controls the weight from the prior.