How to select best split point in Decision Trees?

How best split is decided in Decision Trees?

- Gini impurity
- Chi-Square
- Entropy / Information Gain

How best split is decided in Decision Trees?

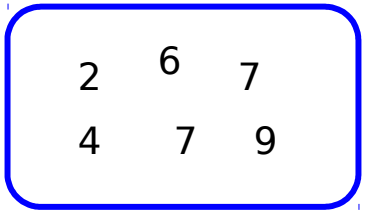
- Gini impurity
- Chi-Square
- Entropy / Information Gain
- Reduction in Variance

How best split is decided in Decision Trees?

$$\text{Variance} = \Sigma [(X - \mu)^2] / n$$

How best split is decided in Decision Trees?

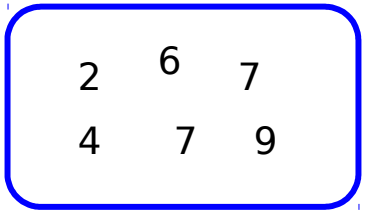
$$\text{Variance} = \Sigma [(X - \mu)^2] / n$$



Analytics Vidhya
Learn everything about analytics

How best split is decided in Decision Trees?

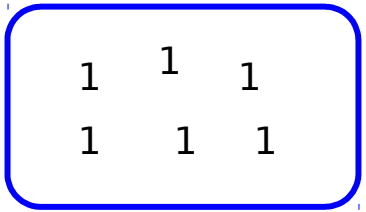
$$\text{Variance} = \Sigma [(X - \mu)^2] / n$$



Variance ~ 6

How best split is decided in Decision Trees?

$$\text{Variance} = \Sigma [(X - \mu)^2] / n$$

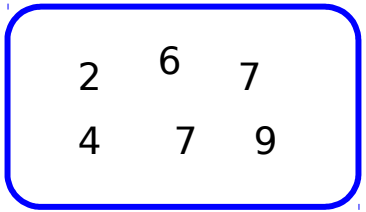


Variance =

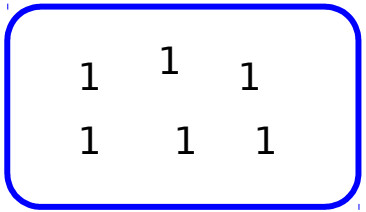
0

How best split is decided in Decision Trees?

$$\text{Variance} = \Sigma [(X - \mu)^2] / n$$



Variance ~ 6



Variance =

0

Properties of Variance

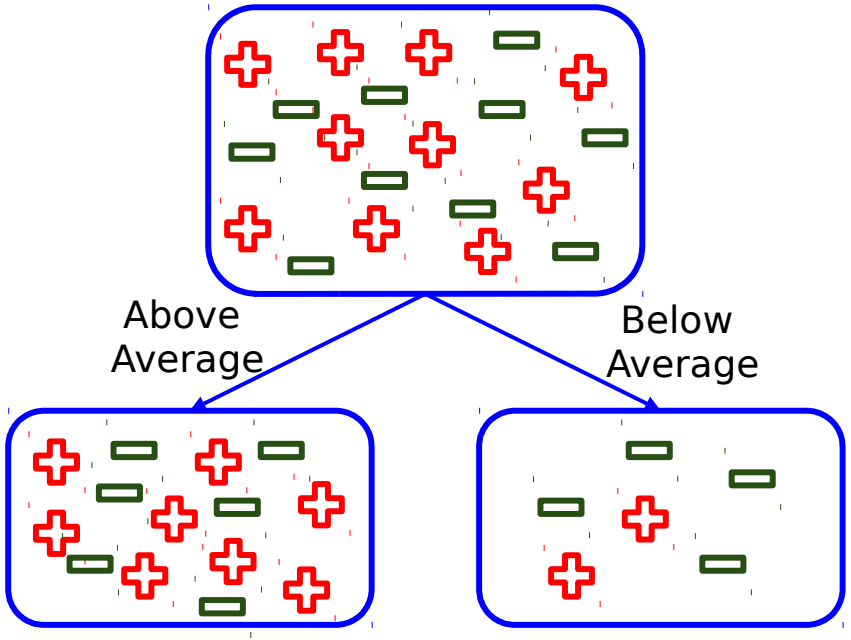
- Used when target is continuous
- Split with lower variance is selected

Steps to calculate Variance for a split

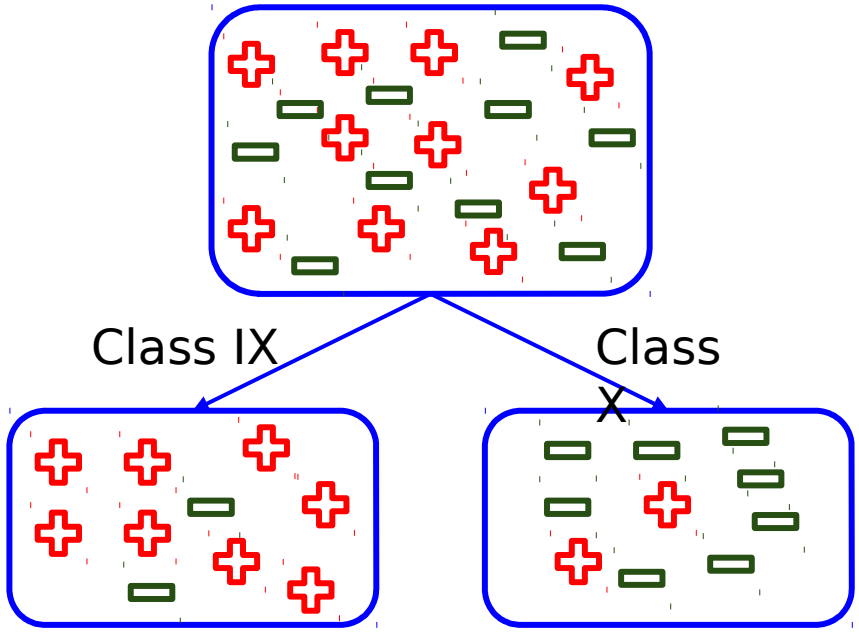
- Calculate the variance of each child node

- Variance = $\Sigma [(X - \mu)^2] / n$
- Calculate the variance of each split as weighted average variance of each child node

Steps to calculate Variance for a split



Split on Performance in Class



Split on Class

Analytics Vidhya
Learn everything about analytics

Steps to calculate Variance for a split

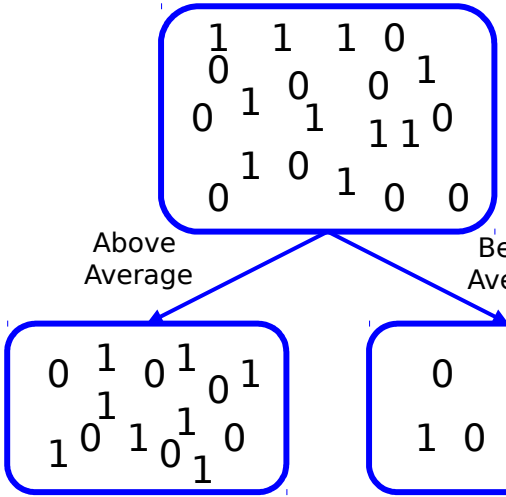
- Plays Cricket = 1
- Do not play Cricket = 0

Steps to calculate Variance for a split

Split on Performance in Class

Steps to calculate Variance for a split

- Above Average node:
  - Mean = (8*1 + 6*0) / 14 = 0.57
  - Variance = [8*(1-0.57)^2 + 6*(0-0.57)^2] / 14 = 0.245
- Below Average node:
  - Mean = (2*1 + 4*0) / 6 = 0.33
  - Variance = [2*(1-0.33)^2 + 4*(0-0.33)^2] / 6 = 0.222
- Variance: Performance in Class:
  (14/20)*0.245 + (6/20)*0.222 = 0.238



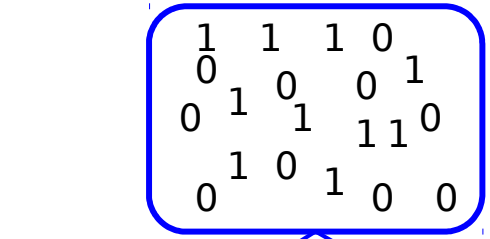
Analytics Vidhya
Learn everything about analytics

Steps to calculate Variance for a split

Split on Class

Steps to calculate Variance for a split

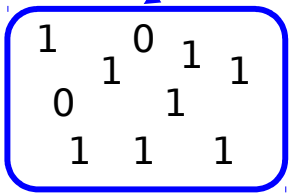
- Class IX node:
  - Mean = (8*1 + 2*0) / 10 = 0.8
  - Variance = [8*(1-0.8)^2 + 2*(0-0.8)^2] / 10 = 0.16
- Class X node:
  - Mean = (2*1 + 8*0) / 10 = 0.2
  - Variance = [2*(1-0.2)^2 + 8*(0-0.2)^2] / 10 = 0.16
- Variance: Class:
  (10/20)*0.16 + (10/20)*0.16 = 0.16



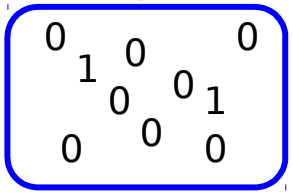
Students = 20
Play Cricket = 10
Do not play = 10
Prob. play = 0.5
Prob. Not play = 0.5

Class IX

Class X



Students = 10
Play Cricket = 8
Do not play = 2
Prob. play = 0.8
Prob. Not play = 0.2



Students = 10
Play Cricket = 2
Do not play = 8
Prob. play = 0.2
Prob. Not play = 0.8

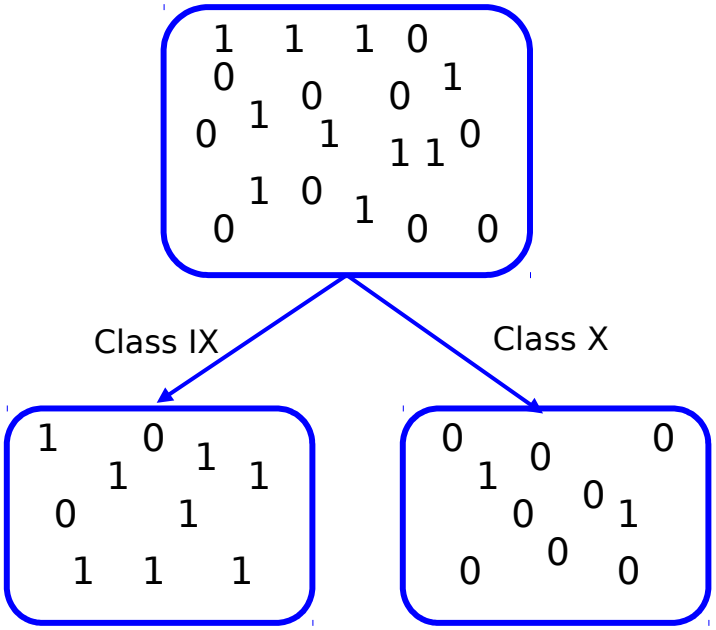
Steps to calculate Variance for a split

Split	Variance
Performance in Class	0.238
Class	0.16

Steps to calculate Variance for a split

Split	Variance
Performance in Class	0.238
Class	0.16

Steps to calculate Variance for a split



Split on Class

Thank
You!

Analytics Vidhya
Learn everything about analytics