Gini Impurity in Decision Tree

Getting Started with ML

```
model.fit(X_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

Gini Impurity

- Gini is a measure of purity.
- For Decision Tree it will tell us where to split.
- It less computationally expensive and preferrable than Entropy.

Gini Index

$$Gini(Situation) = 1 - \sum_{i=1}^{c} (p_i)^2$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Gini(Situation) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini(Play\ Golf) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini(9,5)\ where\ we\ have\ 5\ NO\ and\ 9\ YES$$

$$Gini\ (9,5) = 1 - \{\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\}$$

$$Gini\ (9,5) = 0.46$$

Outlook	mp	Humidity	Windy	Play Golf
Rainy	ot	High	FALSE	No
Rainy	ot	High	TRUE	No
Overcast	ot	High	FALSE	Yes
Sunny	ild	High	FALSE	Yes
Sunny	ool	Normal	FALSE	Yes
Sunny	ool	Normal	TRUE	No
Overcast	ool	Normal	TRUE	Yes
Rainy	ild	High	FALSE	No
Rainy	ool	Normal	FALSE	Yes
Sunny	ild	Normal	FALSE	Yes
Rainy	ild	Normal	TRUE	Yes
Overcast	ild	High	TRUE	Yes
Overcast	ot	Normal	FALSE	Yes
Sunny	ild	High	TRUE	No

$$Gini(Situation) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini(Play\ Golf, Outlook) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini(Play\ Golf, Outlook) = P(Sunny) * G(3,2) + P(Overcast) * G(4,0) + P(Rainy) * G(2,3)$$

$$Gini\ (Play\ Golf, Outlook) = \{\left(\frac{5}{14}\right) * 0.48 + \left(\frac{4}{14}\right) * 0 + \left(\frac{5}{14}\right) * 0.48\}$$

$$Gini(Play\ Golf, Outlook) = .343$$

Outlook	Temp	umidity	Windy	Play Golf
Rainy	Hot	igh	FALSE	No
Rainy	Hot	igh	TRUE	No
Overcast	Hot	igh	FALSE	Yes
Sunny	Mild	igh	FALSE	Yes
Sunny	Cool	ormal	FALSE	Yes
Sunny	Cool	ormal	TRUE	No
Overcast	Cool	ormal	TRUE	Yes
Rainy	Mild	igh	FALSE	No
Rainy	Cool	ormal	FALSE	Yes
Sunny	Mild	ormal	FALSE	Yes
Rainy	Mild	ormal	TRUE	Yes
Overcast	Mild	igh	TRUE	Yes
Overcast	Hot	ormal	FALSE	Yes
Sunny	Mild	igh	TRUE	No

$$Gini(Situation) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini(Play\ Golf, Temp) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini\ (Play\ Golf, Temp) = P(Hot) * G(2,2) + P(Mild) * G(4,2) + P(Cool) * G(3,1)$$

$$Gini\ (Play\ Golf, Temp) = \{\left(\frac{4}{14}\right) * 0.5 + \left(\frac{6}{14}\right) * 0.4444444444444445 + \left(\frac{4}{14}\right) * 0.375\}$$

$$Gini(Play\ Golf, Temp) = 0.444$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Gini(Situation) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini\ (Play\ Golf, Humidity) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini\ (Play\ Golf, Humidity) = P(High) * G(3,4) + P(Normal) * G(6,1)$$

$$Gini\ (Play\ Golf, Humidity) = \{\left(\frac{7}{14}\right) * 0.489795918367347 + \left(\frac{7}{14}\right) * 0.24489795918367355\}$$

$$Gini(Play\ Golf, Humidity) = 0.3675$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Gini(Situation) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini(Play\ Golf, Windy) = 1 - \sum_{i=1}^{c} (p_i)^2$$

$$Gini\ (Play\ Golf, Windy) = P(TRUE) * G(3,3) + P(FALSE) * G(6,2)$$

$$Gini\ (Play\ Golf, Windy) = \{\left(\frac{6}{14}\right) * 0.5 + \left(\frac{8}{14}\right) * 0.375\}$$

$$Gini\ (Play\ Golf, Windy) = 0.428$$

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No

$$Gini\ (9{,}5) = 0.46$$

$$Gini(Play\ Golf, Outlook) = .343$$

$$Gini\ (Play\ Golf, Temp) = 0.444$$

$$Gini\ (Play\ Golf, Humidity) = 0.3675$$

$$Gini\ (Play\ Golf, Windy) = 0.428$$