COL774
Machine Learning
Nov 29, 2020

Last Class:- Decision Trees.

Entropy: $H(Y) = -\sum_y P(Y=y) \log P(Y=y)$

↳ Measure to decide which node to split on.

$MI(Y, X_j) = H(Y) - H(Y|X_j)$

↳ X_j which maximizes $MI(Y, X_j)$

Gini Index

$$\text{Gini}(Y) = 1 - \sum_y [P(Y=y)]^2$$

$$= 1 - \sum_y P(Y=y) \cdot P(Y=y)$$

Choose the Attribute X_j which leads max reduction in $\text{Gini}(Y|X_j)$

$$\arg\max_{X_j} [\text{Gini}(Y) - \text{Gini}(Y|X_j)]$$

Note:- $Y \in \{0, 1\}$ Let $P(Y=1) = P_1$ $P(Y=0) = P_0$

$$\text{Gini}(Y) = 1 - [P_1^2 + P_0^2]$$

$0 \le P_1, P_0 \le 1$

$$= 1 - [P_1^2 + (1-P_1)^2]$$

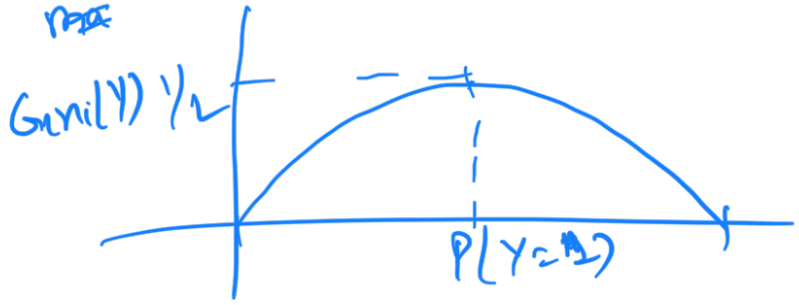
$$= 1 - [P_1^2 + 1 + P_1^2 - 2P_1]$$

$$= 2P_1 - 2P_1^2 = 2P_1(1-P_1)$$

↳ which is maximized at $P_1 = 1/2$

At extreme $\text{Gini}(Y) \to 0$

$p_0 \to 0$
or $p_1 \to 0$



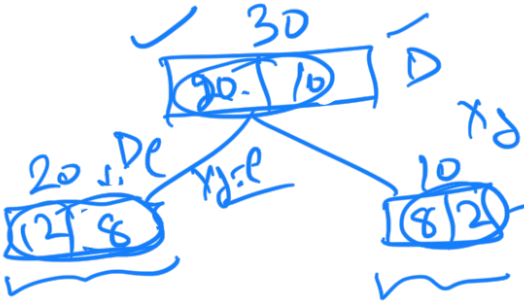
$Y \in \{0,1\}$

H(Y) vs Gini(Y)

CART → (Classification Trees) & Regression Trees

Regression Problems:-

$Y \in R$ → What is right measure to split on attribute?
↳ Entropy may not be the right measure.



For each node (after splitting)

$\hat{y}_l = \frac{1}{|De|} \sum_{i:(x^{(i)}, y^{(i)}) \in De} y^{(i)}$

Avg prediction

(Variance) J

$J_{x_j} = \frac{1}{|D|} \sum_{(x^{(i)}, y^{(i)}) \in D} [\hat{y} - y^{(i)}]^2$ $\quad \hat{y} = \frac{1}{|D|} \sum_{i=1}^{m} y^{(i)}$ $\quad (x^{(i)}, y^{(i)}) \in D$

$J_{x_j = l} = \frac{1}{|De|} \sum_{i:(x^{(i)}, y^{(i)}) \in De} (\hat{y}_l - y^{(i)})^2$

$\frac{|De|}{|D|}$

$J_{x_j \text{ split}} = \sum_{l} P(x_j = l) \cdot [J_{x_j = l}]$

Look for that attribute which leads to ... reduction in error

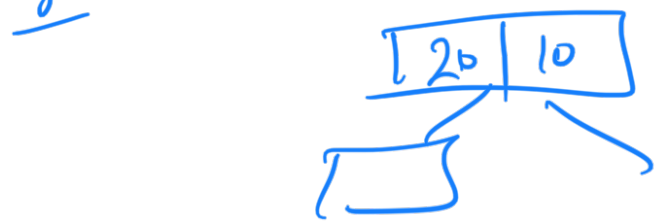
$J_{x_j} - J_{x_j, split}$

Choose

Stopping Criterion: when do you stop?
↳ stop when $|D_e| \leq$ min_threshold

Finally: splitting on cont Attributes.

x_j: continuous valued (Humidity)
$x_j \in [0, 100]$



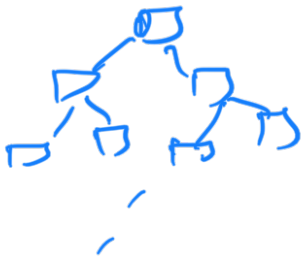
Consider a split at x_j = diff value for which we have a training data point

Whichever split leads to max. reduction in entropy (variance)
↳ choose that split

Random Forests:-

Collection of Decision Trees

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ →



Grow Multiple trees on small variations of the data

"Variance"

$D^{(1)}, D^{(2)} \dots D^{(T)}$

$|D^{(t)}| = m$

$M^{(1)}, M^{(2)} \dots M^{(T)}$

Bootstrapped Samples:-

$\forall t \in \{1 .. T\}$ $\{$

$D^{(t)} = \{\}$

$\forall i = 1$ to m $\{$

sample with replacement $\leftarrow (x^{(a)}, y^{(a)}) \sim \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$

$D^{(t)} = D^{(t)} \cup \{x^{(a)}, y^{(a)}\}$

$\}$

$\}$

$\forall i: (x^{(i)}, y^{(i)}) \sim$ Dist. $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$

$i = 1$ to m

Bagging:- Learning "ensemble" of models on Bootstrapped samples

Decision Trees | Random Forests (collection of decision trees)

↳ Create a T trees, one for every bootstrapped sample of data

→ Random Forest

① Create Bootstrapped samples $D^{(1)} .. D^{(T)}$

② Grow T trees on $D^{(1)} .. D^{(T)}$; respectively referred to as $M^{(1)} .. M^{(T)}$ } Models

③ At test time:-

Let $p^{(t)}(x) = M^{(t)}(x)$

Performs corresponding to t^{th} model

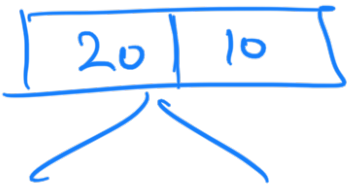
$p^{avg}(x) = \frac{1}{T} \sum_{t=1}^{T} p^{(t)}(x)$

Prediction of ensemble of trees

Bagging:- "Bootstrap Aggregating"

↳ Useful when the original model is high variance.

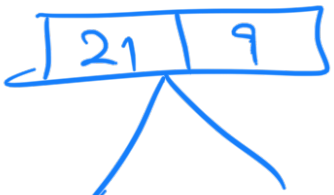
$\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$



↳ X_j:- Best Attr. to split on

↓ Grow further

$\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$



X_j' :- Best attr. to split on

↓ Grow further

$X_j \neq X_j'$

"unstable" learners.

⇒ A high amount of variance in model that we are going to learn ("overfitting"

↳ "Bagging" :- Reduces this variance by learning an ensemble of models of Bootstrapped samples

Random Forests

Specify:

Use "Average" prediction $\sum$

① Learn $M^{(1)} \ldots M^{(T)}$ models on Bootstrapped samples.

② While growing each tree MH split on a subset of randomly chosen attributes.

$X_j^* = \underset{X_j \in \{X_1 \ldots X_m\}}{\text{argmax}} \; MI(Y, X_j)$

"Robust" to noise

choose a subset of attributes

$X_j^* = \underset{X_F}{\operatorname{argmax}}\ MI(Y, X_j)$

$F^{\#}$:- # of attributes to split on

$\Leftarrow X_F \subset \{X_1 \ldots X_n\}$

$\Rightarrow$ Helps grow "Different" trees.

"Variance" :- Overfitting

③ Out of Bag error :-

↳ $M^{(t)}$:- $x^{(t)}$:- examples Not chosen as part of bootstrapped set $D^{(t)}$

Validation set that post here $\Leftarrow \left[x^{v(t)} :- \{(x^{(j)}, y^{(j)})\} \text{ s.t } (x^{(j)}, y^{(j)}) \notin D^{(t)} \right.$

$J^{(t)} = \frac{1}{|x^{v(t)}|} \sum_{(x^{(j)}, y^{(j)}) \sim x^{v(t)}} 1\{ h^{(t)}(x^{(j)}) \neq y^{(j)} \}$

↓ out of bag error

A Total out of bag error $= \frac{1}{T} \sum_{t=1}^{T} J^{(t)}$

Gradient Boosted Decision trees :-

"Weak learners" :- slightly

Boosting

Let's learn random ensemble of learners

$M_1^1 \dots M^T$

$M^{(1)} \dots M^{(t+1)}$

$M^{(t)}$ is learned by focusing on errors made by earlier model

$$M^{(t)} = \underset{M}{\operatorname{argmin}} \sum_{i=1}^{m} J\left[y^{(i)}, M^{\{1 \dots t\}}(x^{(i)})\right]$$

Search in space of $M^{(t)}$'s

squared loss

Minimizing the loss wrt

$$J\left[y^{(i)} - M^{\{1 \dots t-1\}}(x^{(i)})\right]$$

$$M^{\{1 \dots t\}}(x^{(i)}) = \sum_{t'=1}^{t} M_{t'}(x^{(i)})$$

Functional Gradient Boosting