Example 3: Gini index (CART)

- **Example 3**. Let *D* be the given training data, where there are **nine** tuples belonging to the class *buys_computer* = *yes* and the remaining **five** tuples belong to the class *buys_computer* = *no*. A (root) node *N* is created for the tuples in *D*.

Example 3: Gini index (CART)

- We first use Eq. (8.7) for the **Gini index to compute the impurity of *D***:

$$Gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459.$$

Example 3: Gini index (CART)

• To find splitting criterion for tuples in D, we need to compute **Gini index for each attribute**.

• Let's start with attribute *income* and consider **each of the possible splitting subsets**.

• Consider subset {*low*, *medium*}. This would result in **10** tuples in partition D_1 satisfying condition "*income* $\in$ {*low*, *medium*}." Remaining **four** tuples of D would be assigned to partition D_2 (i.e., *income* $\in$ {*high*}). Gini index value computed based on this partitioning is

Example 3: Gini index (CART)

TID	*age*	*income*	*student*	*credit_rating*	*Class: buy_computer*
9	youth	low	yes	fair	yes
5	senior	low	yes	fair	yes
7	middle_aged	low	yes	excellent	yes
6	senior	low	yes	excellent	no
11	youth	medium	yes	excellent	yes
10	senior	medium	yes	fair	yes
4	senior	medium	no	fair	yes
12	middle_aged	medium	no	excellent	yes
8	youth	medium	no	fair	no
14	senior	medium	no	excellent	no
13	middle_aged	high	yes	fair	yes
3	middle_aged	high	no	fair	yes
1	youth	high	no	fair	no
2	youth	high	no	excellent	no

Example 3: Gini index (CART)

- Gini index value computed based on this partitioning is

$$Gini_{income \in \{low,\, medium\}}(D) =$$

$$\frac{10}{14} Gini(D_1) + \frac{4}{14} Gini(D_2)$$

$$= \frac{10}{14}\left[1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right] + \frac{4}{14}\left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right]$$

$$= 0.443$$

$$= Gini_{income \in \{high\}}(D).$$

Example 3: Gini index (CART)

• Similarly, Gini index values for splits on the remaining subsets are **0.458** (for the subsets {*low*, *high*} and {*medium*}) and **0.450** (for the subsets {*medium*, *high*} and {*low*}) [Exercise]. Therefore, the best binary split for attribute *income* is on {*low*, *medium*} (or {*high*}) because it **minimizes the Gini index**.

Example 3: Gini index (CART)

- Evaluating *age*, we obtain {*youth, senior*} (or {*middle_aged*}) as the best split for *age* with a Gini index of **0.357** [Exercise].
- The attributes *student* and *credit_rating* are both binary, with Gini index values of **0.367** and **0.429**, respectively [Exercise].

Example 3: Gini index (CART)

• Attribute *age* and splitting subset {*youth*, *senior*} therefore give **minimum Gini index** overall (i.e., 0.357), with a **reduction in impurity** of 0.459 - 0.357 = 0.102. Binary split "*age* ∈ {*youth*, *senior*}?" results in **maximum reduction in impurity** of the tuples in *D* and is selected as splitting criterion.

Example 3: Gini index (CART)

TID	*age*	*income*	*student*	*credit_rating*	*Class: buy_computer*
9	youth	low	yes	fair	yes
11	youth	medium	yes	excellent	yes
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
8	youth	medium	no	fair	no
7	middle_aged	low	yes	excellent	yes
13	middle_aged	high	yes	fair	yes
3	middle_aged	high	no	fair	yes
12	middle_aged	medium	no	excellent	yes
5	senior	low	yes	fair	yes
10	senior	medium	yes	fair	yes
4	senior	medium	no	fair	yes
6	senior	low	yes	excellent	no
14	senior	medium	no	excellent	no