< GINI INDEX > Measure of impurity

※ DATA (Before splitting)

NY : 111 SF : 139 TOTAL : 111 + 139 = 250

$Gini_{before} = 1 - \left(\frac{111}{250}\right)^2 - \left(\frac{139}{250}\right)^2 = 1 - 0.197136 - 0.309136$
$= 0.493728$

① split at ~~elev~~ elevation = 73 m

(111 NY, 139 SF)

(111 NY, 92 SF) (0 NY, 47 SF)

$Gini_{left} = 1 - \left(\frac{111}{203}\right)^2 - \left(\frac{92}{203}\right)^2$
$= 1 - 0.29899 - 0.20539$
$= 0.49562$

$Gini_{right} = 1 - \left(\frac{0}{47}\right)^2 - \left(\frac{47}{47}\right)^2$
$= 0$

$Gini_{weighted} = \left(\frac{203}{250}\right) \cdot 0.49562 + \left(\frac{47}{250}\right) \cdot 0$
$= 0.40244$

Information gain after splitting = $Gini_{before} - Gini_{weighted}$
$= 0.493728 - 0.40244$
$= \boxed{0.091288}$

② split at elevation = 21.6 m

(111 NY, 139 SF)

(105 NY, 40 SF) (6 NY, 99 SF)

$Gini_{left} = 1 - \left(\frac{105}{145}\right)^2 - \left(\frac{40}{145}\right)^2$
$= 1 - 0.52438 - 0.0761$
$= 0.39952$

$Gini_{right} = 1 - \left(\frac{6}{105}\right)^2 - \left(\frac{99}{105}\right)^2$
$= 1 - 0.00327 - 0.88898$
$= 0.10775$

$Gini_{weighted} = \left(\frac{145}{250}\right) \cdot 0.39952 + \left(\frac{105}{250}\right) \cdot 0.10775$
$= 0.23172 + 0.045255$
$= 0.276975$

Information gain after splitting = $Gini_{before} - Gini_{weighted}$
$= 0.493728 - 0.276975$
$= \boxed{0.216753}$

∴ ② is better!!

⟨CROSS ENTROPY⟩ Measure of disorder

※ DATA (Before splitting)

NY : 111 SF : 139 TOTAL : 250

$$ENTROPY_{before} = -\left(\frac{111}{250}\right)\log_2\left(\frac{111}{250}\right) - \left(\frac{139}{250}\right)\log_2\left(\frac{139}{250}\right)$$

$$= -0.444 * (-0.812) - (0.556)(-0.587)$$

$$= 0.361 + 0.3264 = 0.6874$$

① split at elevation = ~~93 m~~ 21.6 m

(111 NY, 139 SF)

(105 NY, 40 SF) (6 NY, 99 SF)

$$ENTROPY_{left} = -\left(\frac{105}{145}\right)\log_2\left(\frac{105}{145}\right) - \left(\frac{40}{145}\right)\log_2\left(\frac{40}{145}\right)$$

$$= -0.7241 * (-0.323) - 0.276 * (-1.288)$$

$$= 0.234 + 0.355$$

$$= 0.589$$

$$ENTROPY_{right} = -\left(\frac{6}{105}\right)\log_2\left(\frac{6}{105}\right) - \left(\frac{99}{105}\right)\log_2\left(\frac{99}{105}\right)$$

$$= -0.057 * (-2.862) - 0.943 * (-0.059)$$

$$= 0.163 + 0.056$$

$$= 0.219$$

$$ENTROPY_{weighted} = \left(\frac{145}{250}\right) \cdot 0.589 + \left(\frac{105}{250}\right) \cdot 0.219$$

$$= 0.342 + 0.092 = 0.434$$

Information gain = $0.6874 - 0.434 = 0.2534$