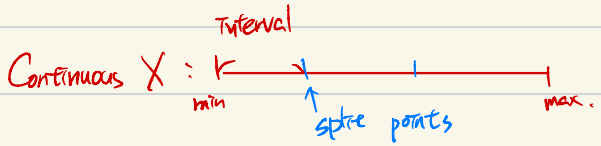
✳ Discretization : Transform continuous attributes into discrete ones.

→ Discrete 因素: 1. The number of possible value → 決定 interval 數量

2. The split points of intervals → interval 的範圍



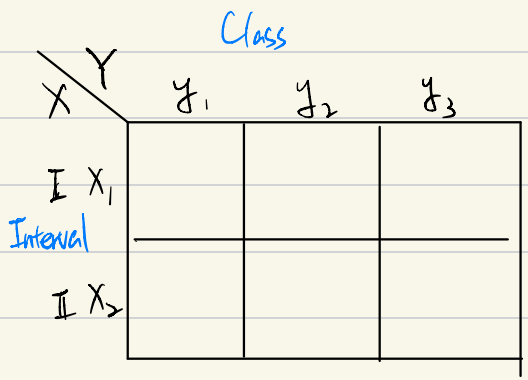
◎ Two categories of discretization methods

1. Unsupervised (非監督式): Equal-width, equal-frequency.
2. Supervised (監督式): Entropy-based, ✓chi-merge.
   - ↳ 依資訊量多寡區分
   - 大部份占優勢
   - Implementation 較複雜

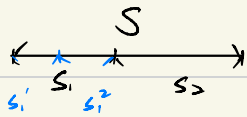
— Chimerge : 用卡方值的概念求獨立性(檢定)

The formula for computing the χ^2 value is:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}$$



✳ Instance 數 = 初始 interval 數

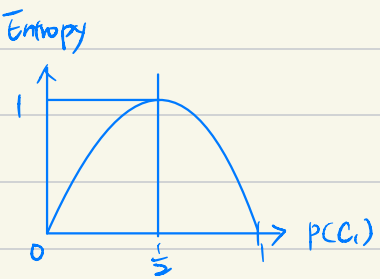
— Entropy-based : X 

, m instances

$$\text{Entropy}(S) = \sum_j -P(c_j)\log_2 P(c_j)$$

→ Evaluate the class distribution of the instances in S.

	$P(C_1)$	$P(C_2)$	Entropy
$S^{(1)}$	1	0	0
$S^{(2)}$	0	1	0
$S^{(3)}$	$\frac{1}{2}$	$\frac{1}{2}$	1



$$\text{Gain}(S,X) = \text{Entropy}(S) - \left[\frac{|S_1|}{|S|}\text{Entropy}(S_1) + \frac{|S_2|}{|S|}\text{Entropy}(S_2)\right]$$

Entropy before binary splitting

The expected entropy after binary splitting on X

分割後 class 內的 instances #

→ $\text{Gain}(S,X)$ 越大越好.

→ 演算法重複往下分割, until 停止條件.

* Unsupervised

— Equal width: $w = \frac{x_{max} - x_{min}}{n}$, n intervals (自定，一般 n = 10)

Splitting point: $x_{min} + kw$, $k = 1, 2, 3, \ldots, n-1$

— Equal frequency: m instance

The frequency in each interval $\frac{m}{n} = k$ (整数)

Sorting $x_{(1)} < x_{(2)} < x_{(3)} < \ldots < x_{(m)}$ (min … max)

Splitting point: $\frac{x_{(k)} + x_{(k+1)}}{2}$, $k = 1, 2, 3, \ldots, n-1$