§5 クラスタリング (1)　演習問題　解答

問題の難易度の目安【易】★☆☆　【基礎】★★☆　【標準】★★★

1 (★☆☆)(クロス集計表)

属性 A のカテゴリーを $A_1, \ldots, A_k$，属性 B のカテゴリーを $B_1, \ldots, B_\ell$ とする．調査・実験したときのサンプルサイズ (試行回数) を n とし，$A_i \cap B_j$ の観測度数 (確率変数と考える) を X_{ij} とする．以下の表を $k \times \ell$-**クロス表**という．

	B_1	B_2	$\cdots$	B_ℓ	計
A_1	X_{11}	X_{12}	$\cdots$	$X_{1\ell}$	F_1
A_2	X_{21}	X_{22}	$\cdots$	$X_{2\ell}$	F_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_k	X_{k1}	X_{k2}	$\cdots$	$X_{k\ell}$	F_k
計	G_1	G_2	$\cdots$	G_ℓ	n

ここで，

$$\sum_{j=1}^{\ell} X_{ij} = F_i, \quad \sum_{i=1}^{k} X_{ij} = G_j \quad \text{for} \quad i = 1, \ldots, k\,;\, j = 1, \ldots, \ell$$

$$\sum_{i=1}^{k} F_i = \sum_{j=1}^{\ell} G_j = n.$$

このとき，次の問いに答えよ．

(1) クロス集計表において項目数を増やす，すなわち k, ℓ を増やすことにより生じるデメリットを答えよ．

(2) 有病率が 10% の集団に，感度 90%，特異度 75% のスクリーニング検査を実施した．このとき，陽性適中率を 2×2 クロス集計表を作成して求めよ．

解 (1)　k, ℓ を大きくすると詳細な分析が行える一方で，各項目に含まれるデータの数が少なくなってしまい，分析結果の信頼性が低くなる可能性がある点．

(2)　スクリーニング検査における 2×2-クロス集計表は以下のようになる：

	病気に罹患している人	病気に罹患していない人	計
陽性	x	z	$x + z$
陰性	y	w	$y + w$
計	$x + y$	$z + w$	$x + y + z + w$

有病率 $= \dfrac{x+y}{x+y+z+w} = \dfrac{1}{10}$，感度 $= \dfrac{x}{x+y} = \dfrac{9}{10}$，特異度 $= \dfrac{w}{z+w} = \dfrac{3}{4}$ であるから，

$$\begin{cases} 10(x+y) = x+y+z+w \\ 10x = 9(x+y) \\ 4w = 3(z+w) \end{cases} \iff \begin{cases} 9(x+y) - z - w = 0 & \cdots ① \\ x = 9y & \cdots ② \\ w = 3z & \cdots ③ \end{cases}$$

②, ③を①へ代入して，$10x - 4z = 0$，すなわち $z = \frac{5}{2}x$．したがって

$$\text{陽性適中率} = \frac{x}{x+z} = \frac{x}{x+\frac{5}{2}x} = \frac{\mathbf{2}}{\mathbf{7}} \approx \mathbf{29}\%.$$

■

2 (★★☆)(決定木分析)

以下はクーポン配布と商品購入に関するクロス集計表である：

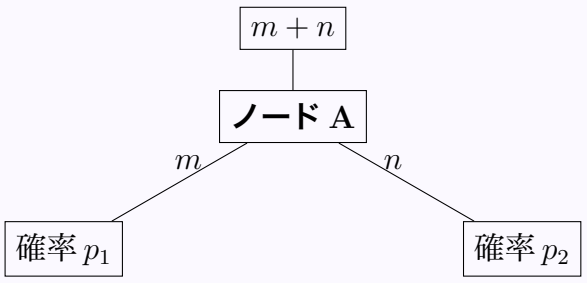
	商品購入	商品未購入	計
クーポン配布済男性	12	38	50
クーポン配布済女性	8	42	50
クーポン未配布男性	20	100	120
クーポン未配布女性	10	70	80
計	50	250	300

ノード A に関する Gini 指標 $\mathcal{G}_G(A)$ を

$$\mathcal{G}_G(A) := \frac{m}{m+n}\mathcal{G}_G(p_1) + \frac{n}{m+n}\mathcal{G}_G(p_2)$$

で定める．ただし，

$$\mathcal{G}_G(p_i) := 2p_i(1-p_i), \quad i = 1, 2.$$



このとき，次の問いに答えよ．

(1) 性別によるノードの Gini 指標を求めよ．

(2) クーポン配布によるノードの Gini 指標を求めよ．

(3) (1),(2) より，このクロス集計表に関する決定木を作成せよ．

解 (1)　男性の商品購入率を p_1，女性の商品購入率を p_2 とすると，与えられたクロス集計表により，

$$p_1 = \frac{32}{170} = \frac{16}{85}, \quad p_2 = \frac{18}{130} = \frac{9}{65}$$

であるから，性別に関するノードの Gini 指標は

$$\mathcal{G}_G(S) = \frac{170}{300}\mathcal{G}_G\left(\frac{16}{85}\right) + \frac{130}{300}\mathcal{G}_G\left(\frac{9}{65}\right) = \mathbf{\frac{1528}{5525} \approx 0.28}.$$

(2)　クーポン配布済の商品購入率を q_1，クーポン未配布の商品購入率を q_2 とすると，与えられたクロス集計表により，

$$q_1 = \frac{20}{100} = \frac{1}{5}, \quad q_2 = \frac{30}{200} = \frac{3}{20}$$

であるから，クーポン配布に関するノードの Gini 指標は

$$\mathcal{G}_G(C) = \frac{100}{300}\mathcal{G}_G\left(\frac{1}{5}\right) + \frac{200}{300}\mathcal{G}_G\left(\frac{3}{20}\right) = \mathbf{\frac{83}{300} \approx 0.28}.$$

(3)　(1),(2) より

$$\mathcal{G}_G(C) = \frac{18343}{66300} > \frac{18336}{66300} = \mathcal{G}_G(S)$$

であるから，性別に関する Gini 指標 $\mathcal{G}_G(S)$ がクーポン配布に関する Gini 指標 $\mathcal{G}_G(C)$ よりも (わずかに) 小さい．したがって，決定木の最上段には性別に関するノードを配置し，2 段目にクーポン配布に関するノードをつける．以上より，求める決定木は以下の通り：