§5 クラスタリング (1)　演習問題

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．

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：

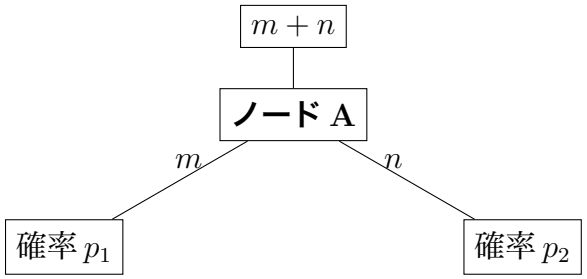
	商品購入	商品未購入	計
クーポン配布済男性	12	38	50
クーポン配布済女性	8	42	50
クーポン未配布男性	20	100	120
クーポン未配布女性	10	70	80
計	50	250	300

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．