2017年　準一級　奇数番号

問1

加重平均、幾何平均、調和平均の意味を知っているかを問う問題。計算自体は楽。

[加重平均]

$x_1, \dots, x_n$ に対して、重み $w_1, \dots, w_n$ とする。加重平均は、

$$\frac{1}{n}\sum_{i=1}^{n} w_i x_i$$

(2) は価格を x、売り上げ比率を w としたときの加重平均に相当する。

$$\frac{1}{450+700+850}(450\cdot 550 + 700\cdot 500 + 850\cdot 450) = 440$$

[幾何平均]

$x_1 \dots x_n$ に対して、$(x_1 x_2 \cdots x_n)^{\frac{1}{n}}$

(3) は4年間の伸び率は、$1.044\times 0.982\times 1.025\times 0.991 = 1.04138$ である。平均伸び率は、

$$\sqrt[4]{1.04138} \fallingdotseq 1.01$$

[調和平均]

$x_1, \dots, x_n$ に対して、$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1/x_i}$ の逆数

(1) かかった時間は $100/10 + 100/15$ より、時速は

$$\frac{2\cdot 100}{100/10 + 100/15} = \left(\frac{1}{\frac{1}{2}\left(\frac{1}{10}+\frac{1}{15}\right)}\right) = 12$$

調和平均の形

問3

交差検証法という名前を覚えているかと、L1/L2正則化の性質を言えるかを問う問題。

（1）学習用データと検証用データに分けて、検証用データに対する精度を見てハイパーパラメータを選定する方法を交差検証法と呼ぶ。

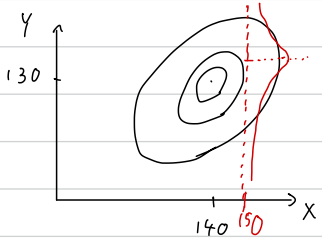
（2）L1正則化はパラメータ値が0になりやすく、スパースになりやすい。
L2正則化は正則化係数が大きくてもパラメータが厳密に0とはなりづらい。

問5

多変量正規分布における周辺分布や条件付き分布が、正規分布になることを使う問題。

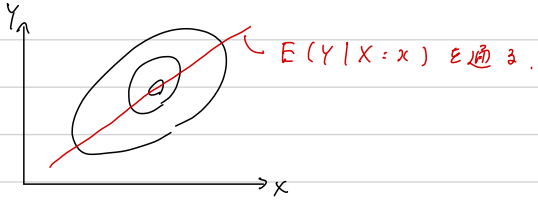
V (X - Y) = V(X) + V(Y) - 2Cov(X, Y) も覚えておくと時短になる。

(1)　X = 150 の条件の元での Y の期待値を求める。



解答では条件付き分布が $N(\mu_y + \rho\sigma_y(x-\mu_x)/\sigma_x,\ (1-\rho^2)\sigma_y^2)$ であること用いているが、これは覚えられん…

代わりに、単回帰を使って求める。X を説明変数、Y を目的変数とした回帰線を引くと、X の値における Y の期待値を通る直線が引ける。



$y = \beta_0 + \beta_1(x-\mu_x)$ とすると、回帰係数は

$$\begin{cases} \beta_0 = \mu_y = 130 \\ \beta_1 = \dfrac{\sigma_{xy}}{\sigma_x^2} = \dfrac{\sigma_y}{\sigma_x}\dfrac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}} = \dfrac{15}{15}\cdot 0.6 = 0.6 \end{cases}$$

相関係数作り出す

よって X = 150 における Y の期待値は、

$y = 130 + 0.6(150-140) = 136$ 。

(2) ランダムに選ばれた兄弟のうち弟の身長の確率分布は、(X, Y) の確率分布における Yの周辺分布に相当する。Yの分布は、 $Y \sim N(130, 15^2)$ であることを使う。

$$Z = \frac{Y-130}{15} \sim N(0,1) \quad \text{とすると},$$

$$P(Y > 115 \mid Y \sim N(130, 15^2))$$

$$= P\left(Z \geq \frac{115-130}{15} \mid Z \sim N(0,1)\right)$$

$$= P(Z \geq -1)$$

$$= 0.8413$$

(3) XとYが正規分布に従うので, X − Y も正規分布に従うことを使う。求める確率は、

$P(X-Y \geq 20)$ である。

X − Y の期待値と分散を求める.

$$E(X-Y) = E(X) - E(Y) = 140 - 130 = 10$$

$$V(X-Y) = V(X) + V(Y) - 2\,Cov(X,Y)$$

$$= 15^2 + 15^2 - 2 \cdot \sqrt{V(X)V(Y)} \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

$$= 15^2 + 15^2 - 2 \cdot 15^2 \cdot 0.6$$

$$= 180$$

したがって $X - Y \sim N(10, 180)$

これより, 求める確率は,

$$P(X-Y \geq 20) = P\left(Z > \frac{20-10}{\sqrt{180}} \mid Z \sim N(0,1)\right)$$

$$= P(Z > 0.745 \mid Z \sim N(0,1))$$

$$= 0.23$$

問7

乱塊法は、ブロック因子を導入することで、本来検出したい因子の検出力を高める方法である。

なぜ検出しやすくなるかを説明できることが問われる。

実験条件Aは4水準 A1, A2, A3, A4 あり、ブロック因子Rは3日間 R1, R2, R3 ある。

	R_1	R_2	R_3
A_1	9	0	9
A_2	0	0	9
A_3	0	9	0
A_4	0	9	0

まず、ブロック因子を考慮せず、Aの一元配置分散分析をする場合の分散分析表を考える。

要因	平方和	自由度	平均平方	F値
A	S_A	3 (=4-1)	$V_A = S_A/3$	$V_A/V_E \sim F(3,8)$
誤差	S_E	8 (=11-3)	$V_E = S_E/8$	
合計	S_T	11 (=12-1)		

これを、ブロック因子Rを追加した場合の分散分析表と比較する。

要因	平方和	自由度	平均平方	F値
A	S_A	3 (=4-1)	$V_A = S_A/3$	$V_A/V_E^* \sim F(3,6)$
R	S_R^*	2 (=3-1)	$V_R = S_R^*/2$	
誤差	S_E^*	6 (=11-5)	$V_E^* = S_E^*/6$	
合計	S_T	11 (=12-1)		

注目すべきは次の2点である。

1元配置の平方和 S_E が、2元配置における $S_E = S_R^* + S_E^*$ に分解される。

2元配置では誤差分散の自由度が減る。

が分解されることにより、F値が大きくなり、検出力が向上する。その一方で、誤差分散の自由度が減ることによって、棄却域は正の方向にのび、検出力は逆に下がる。ブロック因子の影響が大きいほど、一つ目の効果が大きく現れ、全体的には検出力が向上する。

これに基づいて、①〜⑤の真偽を判断する。

①残差分散は小さくなるはずなので×

②F値が変化するため×

③Aの平方和は変化しないので×

④残差分散の自由度は小さくなるので×

⑤○

問9

有限母集団修正による効果がわかっているかを問う問題。分散の式自体は設問に書いてくれているので、問題自体はよく読めば解答できるが、修正の係数とその効果は知っておいて損はなさそう。

$$V[\bar{X}] = \frac{N-n}{N-1}\frac{1}{n}\sigma^2$$

$\frac{N-n}{N-1}$：有限修正の項　　$\frac{1}{n}\sigma^2$：復元抽出標本 or 無限標本での分散

[1] (ア)　$n = \frac{N}{2}$ とすると、

$$V(\bar{x}) = \frac{N-\frac{N}{2}}{N-1}\frac{1}{\frac{N}{2}}\sigma^2 = \frac{1}{N-1}\sigma^2$$

これは母集団の標本数 N に依存するので ×

(イ)　有限修正により、分散は常に小さくなるので ○

(ウ)　無限母集団は $N \to \infty$ とみなせる。このとき $V[\bar{x}] = \frac{1}{n}\sigma^2$ なので ○

[2]　$N = 7270$，$\sigma^2 = 500$，$n = 800$ である

$$V_1 = \frac{N-n}{N-1}\frac{1}{n}\sigma^2 = \frac{6470}{7269}\frac{1}{800}500 = 0.55625$$

$$V_2 = \frac{1}{n}\sigma^2 = \frac{500}{800} = 0.625$$

→ ② が正解．

問11

マルコフ連鎖の問題。推移確率行列 $\mathbb{P}$ について、定常分布 π の性質 $\pi\mathbb{P} = \pi$ がわかっていれば問題なく解けるはず。

(1) P_{ij} : 行く前がi-1本、行った先でj-1本となる確率

P_{ij} を冷静に計算すれば答えが出る。 ⟶ ①

$$\mathbb{P} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1-\theta & \theta \\ 1-\theta & \theta & 0 \end{pmatrix}$$

(2) 定義通り、尤度を計算する。

$1 \to 1 \to 2 \to 0 \to 2 \to 1 \to 1 \to 1$

$P(X_1=1) \cdot P_{22} \quad P_{23} \quad P_{31} \quad P_{13} \quad P_{32} \quad P_{32} \quad P_{22}$

$P = 1 \cdot (1-\theta) \quad \theta \cdot (1-\theta) \cdot 1 \cdot \theta \cdot (1-\theta) \cdot (1-\theta)$

$= (1-\theta)^4 \theta^2$

これを最大にする θ は、P の停留点を求めれば良い。

$$\frac{d}{d\theta} P = -4(1-\theta)^3\theta^2 + (1-\theta)^4 \cdot 2\theta = 0$$

$$-4\theta \quad +2(1-\theta) = 0$$

$$6\theta = 2$$

$$\theta = \frac{1}{3} = 0.33$$

問13

クラスタリング手法の方法の違いを知っているかを設問と、K–means法の性質を問う設問である。

（1）クラスタ間の距離を定義の仕方によってデンドログラムの形が変わる。よく出る距離の定義は次のものがある。

最短距離法

2つのクラスタ内のサンプルのうち、最も近いサンプル同士の距離を使う。

最長距離法

2つのクラスタ内のサンプルのうち，最も遠いサンプル同士の距離を使う。

群平均法

2つのクラスタ内のすべてのサンプルの全組み合わせの距離の平均を使う。

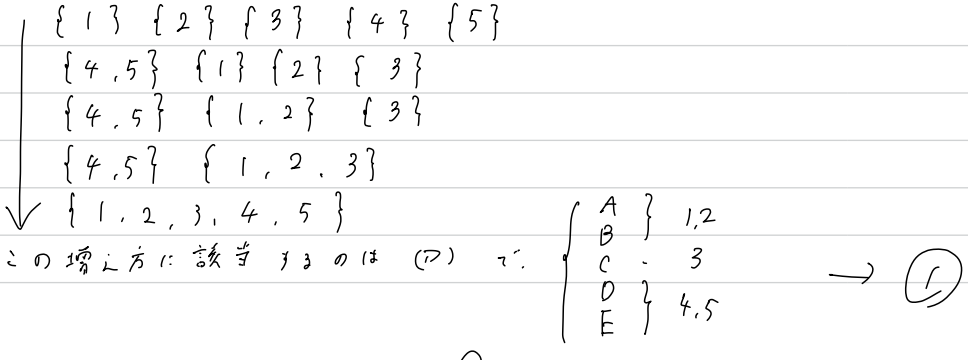
重心法

2つのクラスタの平均値の距離を使う。

ウォード法

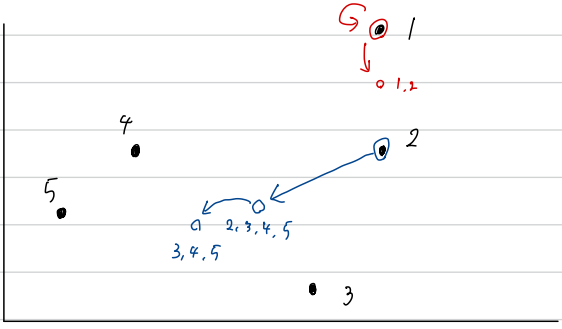
2つのクラスタを結合した際の平方和の増加分を距離とする。

最短距離法によるクラスタリングを行うと、次のようにクラスタが構築されていくことがわかる。



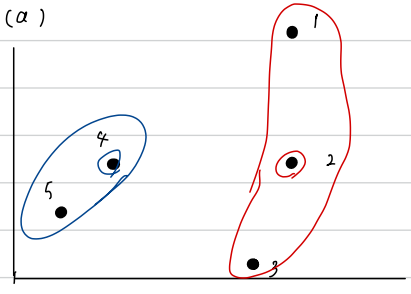
［2］K–means 法の手続きとその性質を理解しているかを問う設問である。K–means の手続きは設問を参照する。初期点の選び方によって、最終的な出力が変わりうる点に注意が必要である。

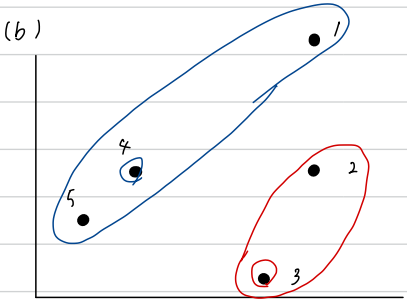
（1）図を使いながら、クラスタの中心がどう変化するかを見ていく。



→ (1)

（2）同じく図を使いながらクラスタがどうなるかを見ていく。初期点に割り当てられたクラスタがそのまま出力となるので、初期展における割り当てのみ考える。





→ (4)

論述問1

主成分分析での設問だが、必要な情報は与えられているので、特に知識なくても解けるはず。

(1) 分散共分散行列の数字から容易に計算可能

(2) 第一主成分は総合的な学力を表現し、第二主成分は理系科目における学力の高さを表現していると考えられる。各主成分の情報の大きさは、主成分における固有値を見て判断できる。正規化した固有値を寄与率とよび、第二主成分まで用いた際の累積寄与率について議論すれば良い。

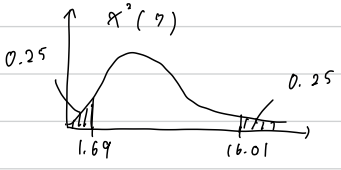
$$(\text{第二主成分までの累積寄与率}) = \frac{798+560}{798+560+160+10.5}$$

$$= 0.889$$

[なぜ固有値を情報の大きさと考えて良いか]

主成分分析は、分散が最大となるような射影の仕方を決定する手法であった。この最大化における射影ベクトルと射影後の分散の大きさが、それぞれ相関係数行列の固有ベクトルと固有値に相当する。したがって、固有値は主成分がどの程度ばらつきを説明できるかの値である。

(3) η_1 / λ_1 が近似的に自由度7の χ^2 分布に従うことをそのまま使って求めることができる。



上の分布の形状より、

$P(1.69 \leq \eta_1 / \lambda_1 \leq 16.01) = 0.95$

$$\Leftrightarrow \quad \frac{7\ell_1}{16.01} \leq \lambda_1 \leq \frac{7\ell_1}{1.69}$$

$$\Leftrightarrow \quad 348.9 \leq \lambda_1 \leq 3305$$

(4) AICの値が最も小ちいさなモデルを選べば良い。(スクリープロットは特に見る必要がなさそうに思える。)