

「第4章クラスタリングとニューラルネットワーク」以降は、以下例のようにゆっくり考えながら理解していかなければならない世界なので、読み飛ばした。

【クラスタ条件】

クラスタリング対象のデータ集合を S ，クラスタの個数を K とする。

$$C_i \neq \emptyset$$

$$C_i \cap C_j = \emptyset, \quad i \neq j \text{ のとき}$$

$$\bigcup_{i=1}^K C_i = S$$

この条件を満たすようなクラスタを作成することを特にハードクラスタリングという。この条件を満たさないクラスタリングについては後述する。

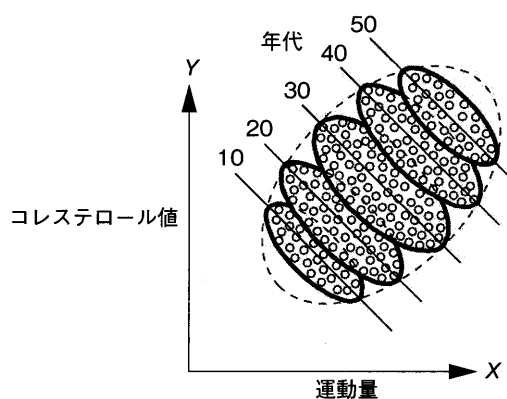


図 1.3 シンプソンのパラドクス
[Pearl 他, 2018] を改変.

ただし元のデータが、異なる特徴（年齢や性別など）を持ったグループからなるときは、各グループの傾向（相関など）がグループ間や全体とは異なる場合があるので注意が必要である。このような現象をシンプソンのパラドクス

[Pearl 他, 2018] という。例えば図 1. 3 では年代別のグループ内で 2 変数（運動量とコレステロール値）の間に負の相関（運動すればコレステロールが減る）があるが、全体としては正の相関（運動すればコレステロールが増える）が生まれている。

オッカムの剃刀(かみそり)

剃刀は余分な説明の削ぎ取り。複数の競争的仮説がある場合に、前提が最も少ない仮説を選ぶ。

KISS 格言 keep it simple, stupid!

Keep it short and simple.

- ・累積度数分布多角形（ogive）は階級の上方限界値を横軸に、それまでの階級の度数の累積を縦軸にして可視化する（図 2・4（a））。
- ・ローレンツ曲線（Lorentz curve）は小さい方から各階級（例えば資産額区分）に属するデータ件数（世帯数）の累積相対度数（全体を 1 になるように累積度数を正規化）を横軸に、各階級の値の総和（資産総額）に基づく累積相対度数を縦軸にして描いた曲線（通常は直線で近似化したもの）である。正方形（1×1）の面積からローレンツ曲線の下側の面積を 2 倍にしたものを引いた数は **GiNi 係数**とよばれ、分布の偏りを表す指標として使われる（図 2・4 川))）。

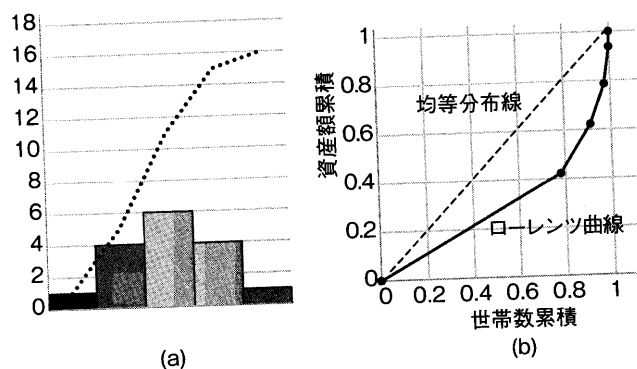


図 2.4 (a) ogive と (b) GiNi

ogive S 字曲線、累積（度数）分布曲線

演繹的推論： 一般的な原則（公理、定理を含む）を適用して具体的な結論を得る。

三段論法 syllogism

帰納的推論： 個別のデータから一般化を行って、新たな仮説を作り出す。

統計的三段論法（statistical syllogism）

集団 P の比率 Q が特徴 F を持つ。

個人 I は集団 P のメンバーである。

...

個人 I は Q というもっともらしさ（plausibility）で特徴 F を持つ（新たな仮説）。

ミンコフスキー距離の特殊な場合は下記。

ユークリッド距離

マンハッタン距離

チェビシェフ距離

距離とは一般に以下のような公理を満たすものをいう。

【距離の公理】

$d(x, y)$ はオブジェクト x と y の間の距離である。

$$d(x, x) = 0 \Leftrightarrow x = y$$

$$d(x, y) = d(y, x)$$

$$d(x, y) + d(y, z) \geq d(x, z)$$

アブダクション

A が真なら B は真である。

B は真である。

A はよりもっともらしい。

(例)

カラスがいるなら、ゴミ置き場のごみが散らかされる。

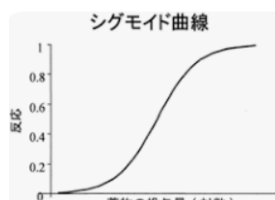
ゴミ置き場のゴミが散らかされている。

多分カラスの仕業だろう。

~~~~~

#### シグモイド曲線

字状曲線のこと。正確には、S 字の両端を左右に引き延ばしたような形状で、化学物質の毒性評価に用いられる用量-反応曲線はシグモイド曲線となる。統計論的には、出現頻度が正規分布するものを累積頻度で表すとシグモイド曲線となる。



~~~~~

1au (astronomical unit, 天文単位=1.495978707×10¹¹m, 地球と太陽の間の平均距離)

なお誤差の代わりに残差という言葉が使われることがある・残差は概念としては誤差とは異なる・誤差とは真の値と観測値との差を意味する・しかしながら真の値はわからない以上・代わりに予測値（推定値）と観測値との差を考慮ことにして、それに残差という名前がつけられた・しかしながら本書では両者を厳密に区別しない・

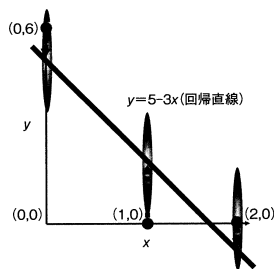


図 3.6 正規分布と回帰直線

- ある直線の上に以下のような関係を満たす値の組 (x_i, β_i) が存在する.

$$\beta_i = c + dx_i$$

すると y_i が Δy の精度で観測される確率は以下ようになる.

$$P(0 \leq \beta_i - y_i \leq \Delta y) = \Delta y \frac{\exp(-(\beta_i - y_i)^2 / 2\sigma^2)}{(2\pi)^{1/2} \sigma}$$

3 個の点 y_i すべてが Δy の精度で観測される確率は、観測値同士の独立性から、上記の確率の積になり、以下のように表せる.

$$(\Delta y)^3 \frac{\exp\left(-\sum_i (\beta_i - y_i)^2 / 2\sigma^2\right)}{((2\pi)^{1/2} \sigma)^3}$$

この式を尤度という. 今求めようとする直線はこの尤度を最大化するものである. 尤度を最大化するようにパラメータを決定することを最尤推定という. そのためには指数関数のべき指数に現れる以下の式の値を最小化すればよい.

$$E = \frac{1}{2} \sum (\beta_i - y_i)^2 = \frac{1}{2} \|A\mathbf{t} - \mathbf{y}\|^2$$

右辺に $1/2$ がかけてある理由は, E を微分すると消えて, それ以後の扱いがしやすいというだけのことである. これを残差二乗和という. E は誤差関数ともいわれる. ただし A と $\mathbf{t}$ は以下のように定義する.

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix}$$

$$\mathbf{t} = (c, d)^T$$

当時は最小二乗法ではなぜ二乗が基本であって, 四乗や六乗ではないのかと

尤度
最尤推定

では上式 E を最小にするような $\hat{\mathbf{t}} = (c, d)^T$ を求めることを考えよう。式 E は c と d を変数とする関数として表せる。

$$E = \frac{1}{2} \sum (c + dx_i - y_i)^2$$

最小値を求める常套手段として以下のように式 E を変数 c と d で偏微分して 0 となる値を求める。すなわち以下の方程式を解く。

$$\frac{\partial E}{\partial c} = 0$$

$$\frac{\partial E}{\partial d} = 0$$

ここで以下を勾配という。

$$\nabla E = \left(\frac{\partial E}{\partial c}, \frac{\partial E}{\partial d} \right)$$

∇ はナブラ、 ∂ はデルと読む。また勾配が 0 となる点を停留点という。一般に停留点で関数は極値（極小値または極大値）をとる。これを整理すると以下のようになる。

$$\begin{pmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

この方程式を正規方程式という。

行列 $A^T A$ が特異でないとするとその逆行列が存在し、以下のように解ける。

$$\hat{\mathbf{t}} = (A^T A)^{-1} A^T \mathbf{y}$$

こうして求められた直線を回帰直線という（図 3.6）。つまり回帰直線はデータをもっともらしく説明する仮説である。

ナブラ

解ける

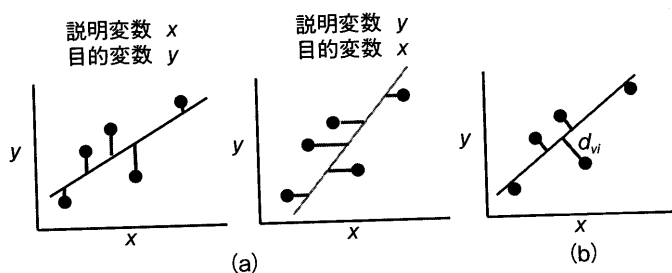


図 3.7 (a) 最小二乗法と (b) 直交回帰

3.1.4 回帰から直交回帰、そして主成分分析へ

直線を求める回帰では独立変数（説明変数） x と従属変数（目的変数） y の区別ができることを仮定している。しかしながら独立変数と従属変数の区別が明確にできない（あるいはわからない）場合もある。そうした場合に従来の回帰における各点から従属変数の軸に平行に回帰直線に下した線分の長さ（すなわち残差）ではなく、回帰直線に垂直に下した線分の長さ（それを d_{vi} とする）について最小化の対象にすることが考えられる。これはいわば変数の役割の対称性を反映した考えである（図 3.7）。

ここでは求める回帰直線に下した線分について、以下の値を最小化する。

$$\sum d_{vi}^2$$

通常の最小二乗法（ordinary least squares または classical least squares）と区別して、こうした回帰を直交回帰（orthogonal regression）あるいは垂直最小二乗法（perpendicular least squares）とよぶ。

直交回帰

よく使われる仮説検証（例えば後述する t 検定）は以下のように行う。

- ① 検証したい仮説を対立仮説として、その逆を帰無仮説として構築する。
- ② 有意水準 α を決定する。
- ③ サンプルデータに基づいて、検定統計量（この例では t 値）および p 値を計算する。
- ④ p 値が有意水準より小さければ、帰無仮説を棄却して、有意差があるとしてその逆（対立仮説）を立証する。
- ⑤ さもなければ、仮説の間には有意差がないと判断する。

ここで p 値は、帰無仮説が正しいとして、検定統計量より極端な値が観察される確率である。 p 値は、検定統計量から計算される。一方帰無仮説が正しい場合に、それを誤って棄却してしまう確率が有意水準であり、その値として 0.1, 0.05, 0.01 などがとられる。有意水準は分析者自身があらかじめ選択できる判断基準である。

p 値については注意も必要である。 p 値はその定義により、サンプル数の増大に伴って減少する傾向がある。すなわちサンプル数が増えればそれだけ有意水準を下げる事が可能になる。言い換えれば、純粋に統計学的には、それだけ対立仮説が確からしくなる。このことは相関係数についてもいえて、相関係数の値が同じならデータ数が大きいほど有意水準を減少できる。

つまりビッグデータ時代になって、データ数が増えたという理由だけで、仮説の有意水準を低く設定できることになる。

ネイマンの尤度と最尤推定

ネイマンはすでに導入した尤度を定義したことでも知られる。すなわちモデルのパラメータ θ に関する尤度は、連続型ランダム変数 x_i に関するパラメータ θ を含む確率密度関数 $f(x_i, \theta)$ の積として一般には以下のように表される。

$$L(\theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$

この尤度を最大にするようなパラメータを求めることを最尤推定という。尤度の対数をとったものを対数尤度という。対数尤度を最大化するパラメータは元の尤度を最大化する。こうした考え方はガウスが最小二乗法を考案したときにもすでに使われている。また対数尤度の値はオッカムの剃刀の関連で紹介した AIC の中で使われている。

尤度
最尤推定

5.3.3 医療画像の差分解析

特に医療分野では時間差のある同一部位の画像を経時差分画像とよぶ。

経時差分画像では特にそれらの間の位置合わせ（マッチング）を行うことが重要である。なぜなら、撮影機器の違いや被写体の姿勢の違いから、被写体の全体的位置や骨、臓器の位置のずれが生じるのだが、差分画像の位置のずれはアーチファクト（偽像）といわれるノイズを生み出し、それが目指す病変の発見を邪魔するという問題を引き起こすからである。

通常経時差分画像のマッチングでは全体的な位置合わせと局所的な位置合

差分解析のわかり
やすい例…経時差
分画像

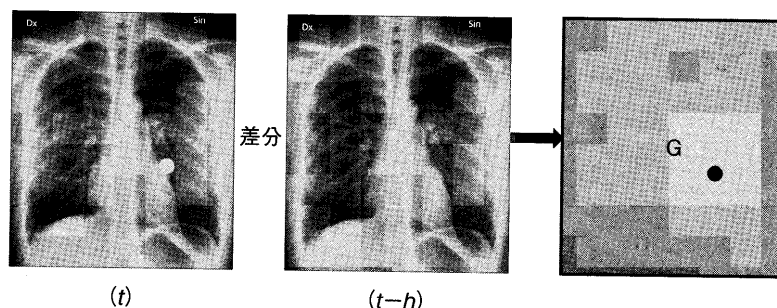


図 5.13 経時差分画像（イメージ）

出典：https://en.wikipedia.org/wiki/Projectional_radiography

せ（変形）とを組み合わせるアプローチが提案されている [板井 他, 2007].

病変の進度に応じて h は数カ月から数年である。

【応用】

同一患者の同一部位画像の差分をとること（経時差分処理）により、新たな病巣・変異を発見することが可能となる（図 5.13）。