データサイエンスエキスパート演習
補足資料

概 要

この資料では，ページ数の関係でデータサイエンスエキスパート演習に収録できなかった補足的な内容を示す。

目 次

1 統計基礎 **1**
1.1 十分統計量 1
1.2 統計的推測の例: ポアソン分布 2

2 数学基礎 **4**
2.1 最長経路問題 4

3 情報基礎 **5**
3.1 データベースの更新時異状 5

1 統計基礎

1.1 十分統計量

統計モデルを $\{f(x;\theta), \theta \in \Theta\}$ とする。実際には，x はモデルからのサイズ n の独立標本であることが多く，その場合 $x = (x_1, \ldots, x_n)$ であり，$f(x;\theta)$ は $f_n(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ であるが，ここでは単に x および $f(x;\theta)$ と表す。統

計量 $t(x)$ が θ に関する十分統計量であるとは，$f(x;\theta)$ が以下の形に標記できることである。

$$f(x;\theta) = g(t(x);\theta)h(x)$$

すなわち，パラメータ θ と $t(x)$ のみの関数 $g(t(x);\theta)$ と，θ を含まない関数 $h(x)$ の積として分解できることである。θ も $t(x)$ も多次元であることが多い。

$t(x)$ が十分統計量ならば，$t(x) = t$ を与えた条件の下で，x の条件付き分布は θ に依存しない。逆に，この条件を十分統計量の定義とすることができる。

θ の推測においては，十分統計量のみに基づいておこなえばよい。例えば，最尤推定量は $t(x)$ のみの関数となる。以下のポアソン分布の推測の例において，十分統計量を用いる。

1.2 統計的推測の例: ポアソン分布

ここでは，統計的推測の例として，ポアソン分布に関する推測を説明する。

$X_1, \ldots, X_n$ を互いに独立にパラメータ λ のポアソン分布に従う確率変数としたとき，その同時確率関数は，$t = x_1 + \cdots + x_n$ として，

$$\begin{aligned} f(x_1, \ldots, x_n; \lambda) &= \prod_{i=1}^{n} \Big(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \Big) \\ &= \frac{\lambda^t}{x_1! \cdots x_n!} e^{-n\lambda} \\ &= \frac{(n\lambda)^t}{t!} e^{-n\lambda} \times \frac{t!}{x_1! \cdots x_n!} \Big(\frac{1}{n}\Big)^{x_1} \cdots \Big(\frac{1}{n}\Big)^{x_n} \end{aligned} \tag{1}$$

となる。式 (1) より

$$g(t;\lambda) = \frac{(n\lambda)^t}{t!} e^{-n\lambda}, \quad h(x) = \frac{t!}{x_1! \cdots x_n!} \Big(\frac{1}{n}\Big)^{x_1} \cdots \Big(\frac{1}{n}\Big)^{x_n}$$

とおけば，(a) $T = X_1 + \cdots + X_n$ は λ に対する十分統計量で，パラメータ $n\lambda$ のポアソン分布に従い，(b) $T = t$ が与えられたときの $X_1, \ldots, X_n$ の条件付き分布はパラメータ $\{t, 1/n, \ldots, 1/n\}$ の多項分布であることがわかる。

パラメータ λ の対数尤度関数は，式 (1) の t の関数 $f(t;n\lambda)$ より，

$$l(\lambda) = \log f(t;n\lambda) = t\log(n\lambda) - \log t! - n\lambda$$

であるので，これを λ で微分して 0 と置き，

$$\frac{t}{\lambda} - n = 0 \;\Rightarrow\; \hat{\lambda} = \frac{t}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

より，標本平均が最尤推定値となる。T はパラメータ $n\lambda$ のポアソン分布に従うので，最尤推定量 T/n の平均と分散は $E[T/n] = \lambda$, $V[T/n] = \lambda/n$ である。よって，標本平均 $\overline{X} = T/n$ は λ の不偏推定量であり，標準誤差 (standard error) は $SE[T/n] = \sqrt{\hat{\lambda}/n}$ となる。なお，T は確率変数の和であるので，n が大きいとき，$\hat{\lambda} = T/n$ は中心極限定理により正規分布 $N(\lambda, \lambda/n)$ で近似される。

母集団分布がポアソン分布であると想定されるとき，λ_0 を λ のある値として，帰無仮説 $H_0 : \lambda = \lambda_0$，対立仮説 $H_1 : \lambda > \lambda_0$ の検定は，T の実現値が t^* のとき，P 値が $P(T \geq t^* | \lambda = \lambda_0)$ で与えられることを用いて実行できる。

ポアソン性の検定は，ポアソン分布では平均と分散が等しいので，$X_1, \ldots, X_n$ の標本平均と不偏標本分散を

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

として，それらの比

$$D = \sum_{i=1}^{n}(X_i - \overline{X})^2/\overline{X} = (n-1)S^2/\overline{X} \tag{2}$$

を用いて検定できる。式 (2) の D は分散指標と言われる。また，ポアソン性がなりたつとき，式 (1) より，$h(x_1, \ldots, x_n | t)$ はパラメータ $\{t, 1/n, \ldots, 1/n\}$ の多項分布に従うので，$x_1, \ldots, x_n$ がこの多項分布から得られたものかどうかを調べればよい。このときの適合度のカイ二乗統計量は

$$Y = \sum_{i=1}^{n} \frac{(X_i - t/n)^2}{t/n} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\overline{x}}$$

と分散指標 D と同じ形となる。したがって，式 (2) の D は，H_0 の下で近似的に自由度 $n-1$ のカイ 2 乗分布に従う。

2　数学基礎

2.1　最長経路問題

ここでは 3.3.2 項の図 3.6 に関する【問題 2】の解が長さ 8 であることを示す。図 3.6 の各ノードに図 1 のように番号をつける。

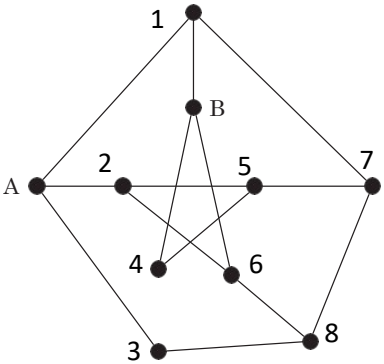

図 1: 番号づけ

まず目視で以下のようなノード 3 以外のすべてのノードを通る A から B への長さ 8 のルート (A → 1 → 7 → 8 → 6 → 2 → 5 → 4 → B) が見つかる。

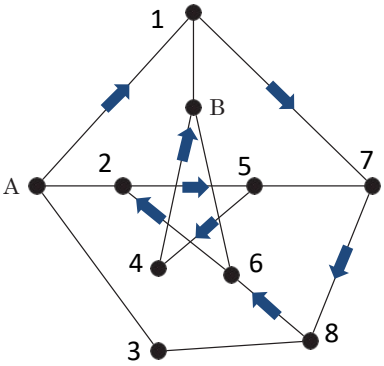

図 2: 長さ 8 のルート

A,B 以外の 8 頂点すべてを通るルートが存在しないことは，次のように場合分けで確認できる。まず A から (ノード)1 に出て行く場合を考える。このときは 1 → 7 と進む必要がある。ここで 3 について考える。3 に行くには 8 から行かなければならないが，そこで行き詰まりになってしまい，B に到達することができない。したがって，上で示した長さ 8 のルートのように，3 を通ることができない。

次に A から 2 に出て行く場合を考える。ここでも 3 を考慮すると，やはり 8 から行かなければならず，上と同様に行き詰まりになる。

最後に A から 3 に出て行く場合を考える。このときは 3 → 8 と進む必要がある。ここで (a) 8 → 7 と進む場合，と (b) 8 → 6 と進む場合，を考える。まず (a) については，7 の次に 1 に進むと B に到着せざるを得なくなり，2,4,5,6 を通ることができない。7 の次に 5 に進むと，もう 7 に戻ることはできないから，1 を通ることができない。次に (b) については，6 → 2 → 5 と進む必要がある。5 の後 7 に進むと次は 1 に進まなければならず 4 を通ることができない。逆に 5 の後 4 に進むと B で終わらざるを得ず，7 を通ることができない。

以上で，場合を尽くしたので，A,B 以外の 8 頂点すべてを通るルートは存在しない。

3 情報基礎

3.1 データベースの更新時異状

リレーションは第 1 正規形であるだけでは不十分でありタプルの追加，削除，修正を行う際にリレーションの**更新時異状**とよばれる問題が発生する可能性がある。例として，表 1 のリレーション「ソフト注文」の更新について考える。このリレーションは第 1 正規形であり，属性集合 { 学校名，ソフトウェア } が主キーであることに注意する。

文献 [1] の p.134 のコラム「更新時異状と情報無損失分解」において，タプル挿入時異状について説明したが，以下ではタプル削除時異状とタプル修正時異状について例をあげて説明する。

たとえば，C 高専からデータ処理ソフトの注文がキャンセルされ，リレーションからタプル（C 高専，データ処理ソフト，4， 80,000，320,000）を削除した

表 1: リレーション「ソフト注文」

学校名	ソフトウェア	ライセンス数	単価	総額
A 大学	統計処理ソフト	5	100,000	500,000
B 大学	統計処理ソフト	2	100,000	200,000
B 大学	数式処理ソフト	3	150,000	450,000
C 高専	データ処理ソフト	4	80,000	320,000

とする。この削除の処理自体は問題ないが，データ処理ソフトの注文が他の学校からなかったとすると，データ処理ソフトの単価のデータまで失われてしまう。単価のデータを残すためにタプル（–，データ処理ソフト，– ，80,000，–）を保存しようとするとキー制約に抵触してしまい，単価のデータを残すことができない。このような異状をタプル削除時異状という。

タプルを修正する場合には 2 種類の異状が発生する可能性がある。まず，統計処理ソフトの単価が 100,000 から 90,000 に変更になったとすると 1 番目と 2 番目のタプルの両方の単価（および総額）を修正する必要がある。つまり，実世界では統計処理ソフトの単価の変更という 1 つの事象が起こっているだけであるにもかかわらず，リレーションでは複数のタプルを修正する必要がある（ただし，単一のタプル内で単価と総額の両方を修正する必要があることはやむを得ない）。また，C 高専の注文がデータ処理ソフトから数式処理ソフトに変更されたとすると，タプル削除時異状の場合と同じ理由でデータ処理ソフトの単価のデータが失われてしまうという問題がある。これらをタプル修正時異状という。

上の例のリレーション「ソフト注文」で生じた更新時異状の原因について考えてみると，1 つのリレーションにソフトウェアの単価とソフトウェアの注文という本来独立な 2 つの事象に関するデータが格納されていることがその原因であることが分かる。実際，表 1 のリレーションを，表 2 および表 3 に示すリレーション「ソフト注文 [学校名，ソフトウェア，ライセンス数，総額]」とリレーション「ソフト注文 [ソフトウェア，単価]」という 2 つの射影に分解すると，上で挙げたいずれの更新時異状も解消されていることが確認できる。

表 2 と表 3 のリレーションに対して，それぞれの属性 ソフトウェア の等しいものについて等結合を行い，さらに重複する属性 ソフトウェア を 1 つ削除

表 2: リレーション「ソフト注文 [学校名，ソフトウェア，ライセンス数，総額]」

学校名	ソフトウェア	ライセンス数	総額
A 大学	統計処理ソフト	5	500,000
B 大学	統計処理ソフト	2	200,000
B 大学	数式処理ソフト	3	450,000
C 高専	データ処理ソフト	4	320,000

表 3: リレーション「ソフト注文 [ソフトウェア，単価]」

ソフトウェア	単価
統計処理ソフト	100,000
数式処理ソフト	150,000
データ処理ソフト	80,000

する射影を行うことで，元のリレーション「ソフト注文」を得ることができる。このような結合演算（自然結合という）によって元のリレーションに戻せるようなリレーションの分解は情報無損失分解とよばれる。つまり，上記の表 1 のリレーションを表 2 と表 3 の 2 つのリレーションに分ける分解は情報無損失分解である。このように，情報無損失分解によって，より正規化度の高いリレーションを得ることで更新時異状を解消することも正規化とよばれる。

リレーションの正規形には第 1 正規形，第 2 正規形，第 3 正規形，ボイス－コッド正規形，第 4 正規形，第 5 正規形の 6 つがある。これらは，タプルの更新時異状を解消する目的で，情報無損失分解によって第 1 正規形から順に，より正規化度を高めることで導出されたものであるが，これ以上正規化度を上げることができない第 5 正規形に至っても，タプルの更新時異状を完全には解消できないことが明らかとなっている。このため，現在では，バーンシュタインの合成的手法とよばれる方法を利用することで，リレーションの正規化は第 3 正規形にとどめられることが多い。

参考文献

[1] 日本統計学会編，データサイエンス発展演習，東京図書 (2024)