06t4027h 斎藤高行

5章 連関規則

5.1 同時確率/条件付き確率
5.2 温泉の効能の連関規則
5.3 シングルモルトの特徴
5.4「ことわざ」を創る！

連関規則

例：紳士服量販店にて

「2本目からスラックスが半額」という広告

1本目のズボンを買った客が、2本目を買ってくれる確率が高くなる

連関規則

「属性Aを持つオブザベーションは属性Bを持つ傾向にある」という知識を連関規則という。

(連想規則やアソシエーションルールとも言う)

・連関規則は　A→B　と記す

・Aをルール・ヘッド, Bをルール・ボディという

連関規則の有用性

有用な連関規則A→Bを見つけ出すための、4個の確率

- 前提確率　$p(A)$
- 条件付き確率　$p(B|A)$
- 同時確率　$p(A,B)$
- 事前確率　$p(B)$

連関規則の有用性

前提確率 $p(A)$

この確率が高いほどAは頻繁に観察され、連関規則を適用できるチャンスが増える

条件付き確率 $p(B|A)$

Aというオブザベーションの中で、Bという属性を持っているオブザベーションの確率を示す

規則のヒット率であり、信頼度とも言われる

連関規則の有用性

同時確率　$p(A, B)$

前提確率と条件付き確率の積である

同時確率はこの二つを同時に考慮した指標であり、高くなることが望ましい

事前確率　$p(B)$

この確率と条件付き確率とを比較し、この確率が低かった分だけ連関規則は効果があると言える

バスケット分析

膨大な購買記録の中から有用な連関規則を見つけることを、バスケット分析という

買った場合を1, 買わなかった場合を0とする「客の数×商品の種類」のデータ行列を分析し、連関規則を見つける

この行列の要素は、殆どが0 (疎行列)である

バスケット分析

連関規則を見つけるためには、データ行列から前提確率と条件付き確率を計算する

ヘッドとなる列データの 1 の比率が前提確率となる

ヘッドが 1 であるとき、ボディとなる別の列データが 1 となる確率が条件付き確率となる

この組合せにより連関規則が求められる

温泉の効能の連関規則

バスケット分析は、購買データ以外のデータに対しても使うことができる

例えば温泉の効能表にも適用できる

「客」を「温泉」、「商品」を「効能」と置き換え、同様にダミー変数を用いてバスケット分析を行う

p150-151 (表5.1, 表5.2)は、温泉の効能表に対してのバスケット分析の例であり、前頁の求め方を用いる

95温泉・28効能の行列データを分析している

温泉の効能の連関規則

ルール1について確認してみる

「高血圧」に効能のある温泉

17ヶ所・・・17/95＝0.1789≒0.18

「高血圧」に効能がある温泉の中で、「動脈硬化」にも効能がある温泉

15ヶ所・・・15/17＝0.8823≒0.88

結果は表5.2のものと一致する

温泉の効能の連関規則

		前提,条件付き
1.高血圧	→動脈硬化	(0.18, 0.88)
2.動脈硬化	→高血圧	(0.20, 0.79)
3.リウマチ&動脈硬化	→高血圧	(0.15, 0.86)
4.リウマチ&高血圧	→動脈硬化	(0.15, 0.86)
5.筋肉痛	→関節痛	(0.04, 1.00)
6.関節痛	→筋肉痛	(0.05, 0.80)
7.リウマチ&婦人病&創傷	→運動器障害	(0.15, 0.64)
8.婦人病&創傷	→運動器障害	(0.17, 0.56)
9.神経痛&胃潰瘍	→婦人病	(0.11, 0.90)

温泉の効能の連関規則

この分析で確認できた連関規則の特徴は以下

A→BとB→Aを比較したときの特徴

　この二つの連関規則における前提確率と条件付き確率はそれぞれ一致しない

　同時確率は共に等しい

ルール・ヘッドの条件数を増やしたときの特徴

　信頼度が上がることもあるが、下がることもある

多値変数でのバスケット分析

バスケット分析はダミー変数以外にも、多値であるカテゴリカル変数を分析することができる

カテゴリカル変数を分析する場合は、ひとつの商品アイテムのなかでどのブランドを買ったかという点に注目して分析する

P.154-155 (表5.3)は、カテゴリカル変数を用いたウイスキーの特徴を、バスケット分析した例である

109銘柄・5変数の行列を分析している

シングルモルトの特徴

温泉の効能表と同様に、ルール1を検証する

「口当たり(軽やか)」の属性を持つ銘柄

34種・・・34/109＝0.3119≒0.31

「口当たり(軽やか)」の属性を持つ銘柄の中で、「味(甘い)」の属性を持つ銘柄

19種・・・19/34＝0.5588≒0.56

結果は表5.4のものと一致する

シングルモルトの特徴

		前提,条件付き
1.口当たり(軽やか)	→味(甘い)	(0.31,0.56)
2.色(濃金色)	→味(甘い)	(0.25, 0.56)
3.香り(芝生)	→味(甘い)	(0.15, 0.56)
4.後味(名残惜く)	→味(甘い)	(0.06, 0.67)
5.口当たり(軽やか)&色(濃金色)	→味(甘い)	(0.07, 0.88)
6.口当たり(軽やか)&香り(シェリー)	→味(甘い)	(0.10, 0.67)

連関規則と改善率

連関規則の条件付き確率は、その値だけでは高いか低いかは判断しにくい

例：

宝くじの1等当選確率の連関規則の条件付き確率=0.01%・・・異常に高い

風邪の患者の生存率の連関規則の条件付き確率=99.0%・・・異常に低い

連関規則と改善率

条件付き確率の妥当性を測る上で、事前確率 $p(B)$ を用いる

先述の例でも、確率の高さや低さを判断する基準として事前確率を用いている

特に、条件付き確率が事前確率を下回る連関規則については存在価値がない

連関規則と改善率

連関規則の価値の判断基準として、改善率がある
(改善率はリフトliftとも呼ばれる)

$$\text{改善率} = p(B \mid A) / p(B)$$
$$= p(B, A) / \left(p(A)p(B)\right)$$

改善率が1.0より大きい連関規則のみを利用し、1.0以下の連関規則を破棄する方法もある

連関規則と改善率

改善率について、シングルモルトの特徴を例に挙げる

「味(甘い)」に属する銘柄の比率は、先程の例におけるルール・ボディであり、事前確率と捉えられる

「味(甘い)」に属する銘柄・・・51種

51/109＝0.4678≒0.47

すなわち、条件付き確率が0.47を下回る連関規則が破棄される対象である

破棄される連関規則

		前提, 条件付き
7.香り(シェリー)	→味(甘い)	(0.21, 0.43)
8.香り(甘い)	→味(甘い)	(0.06, 0.43)
9.色(白ワイン)	→味(甘い)	(0.07, 0.38)
10.後味(余韻ある)	→味(甘い)	(0.22, 0.38)

「ことわざ」を創る！

天気に関することわざは多く存在する

「雨蛙が鳴くと雨が降る」

「夕焼けの次の日は晴れる」

「月が暈を被ると雨が降る」

過去の天気データから法則を抽出し、連関規則探しを試みる

「ことわざ」を創る！

分析対象のデータ

1987.1.1～1996.12.31の3650日の東京地方の天気データ

オブザベーションは「1日分の天気」

対象となる天気は晴、雨、曇、雪のカテゴリカル変数

変数は5つ・・・「前日天気」「前々日天気」「3日前天気」「1週間前天気」「月」

(最古の1週間分のデータは棄却・・・実際は3643日)

「ことわざ」を創る！

天気データの概要は以下のようである

	事前確率
晴：1965日	0.539
雨：999日	0.274
曇：650日	0.178
雪：36日	0.010

「ことわざ」を創る！

晴の連関規則

適用事例の多いルール・・・

「前日が晴→晴」	適用可能事例1962日(68.3%)
「二日晴が続く→晴」	適用可能事例1339日(71.7%)
「三日晴が続く→晴」	適用可能事例 951日(73.9%)

「12月→晴」	適用可能事例86.2%
「12月＆二日晴が続く→晴」	適用可能事例88.5%

「ことわざ」を創る！

雨の連関規則

1年中使える規則・・・

「前日が雨→雨」	適用可能事例44.1%
「曇、雨と続く→雨」	適用可能事例48.3%
「曇、雨と続く→雨」	適用可能事例48.3%
「9月&前日が雨→雨」	適用可能事例59.1%
「6月&前日が雨→雨」	適用可能事例46.3%
「7月&前日が雨→雨」	適用可能事例48.5%

「ことわざ」を創る！

雪の連関規則

東京では雪が降る日が少なく、規則は作りづらい

「1月→雪」　　適用可能事例3.3%

「2月→雪」　　適用可能事例5.5%

「雪」の事前確率は0.91%であり、非常に高いといえる

「2月&前日が雪→雪」　　適用可能事例31.3%

しかし降雪率は10年でわずか16日,希少事例に関する有益な規則を見つけるのは難しい・・・