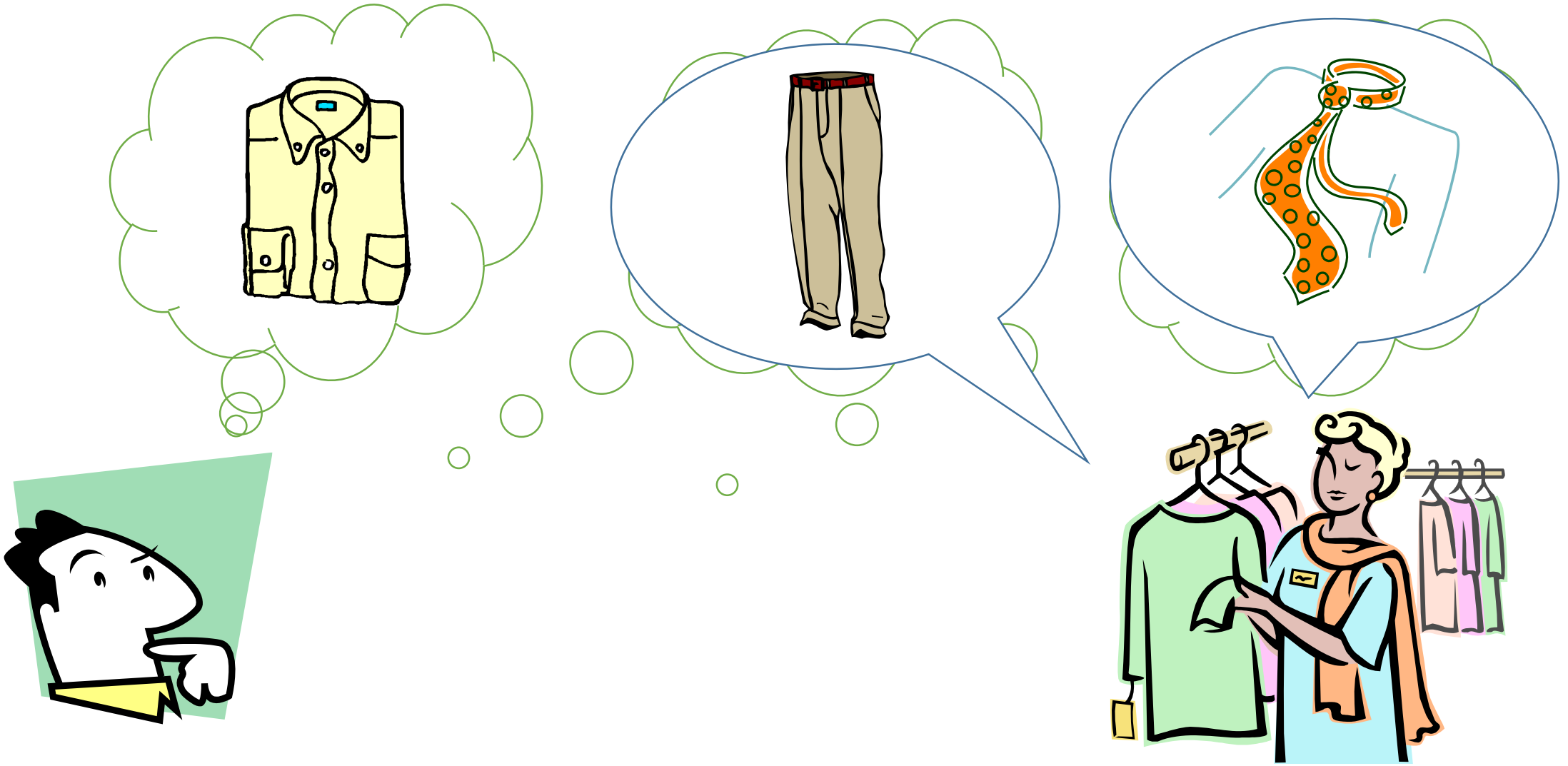
連関規則

第12回ゼミ担当　菊川翔平



連関規則とは

- 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。
  先ほどの例でいうなら、「上着を買った客はズボンを買いやすい」
- データマイニングにおいてこのような知識を「連関規則」「連想規則」
  あるいは「アソシエーションルール」と呼んでいる。
- A→Bと記す。

*Amaz〇nなどのサイトで紹介されるおすすめ商品は
  顧客ごとの購買履歴etcを元にして、連関規則を適用している。

連関規則とは(2)

- A→Bの部品の、Aを「ルール・ヘッド」、Bを「ルール・ボディ」という。
- A、Bは単一の属性でなくてよい。
  簡単に言えば、「木曜日 &ビールを買う」→「紙おむつを買う」
- A、Bを確率事象と見なして、膨大なトランザクション(購買)の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。

＊バスケット分析とは、本来は、「客が買い物かごに一緒に入れる商品は何かを分析する」という意味であるが、買い物以外の分野のデータに適用することも可能である。

連関規則とは(3)

- 有用な連関規則を見つけ出すために4つの確率を考慮する。

- 前提確率p(A)：Aが起きる確率
- 条件付き確率p(B|A)：Aが起きた時にBが起きる確率
- 同時確率p(A,B)：前提確率と条件付き確率の指標を
  同時に考慮した指標、p(A,B) = p(A) x p(B|A)
- 事前確率p(B)：Bが起きる確率

前提確率p(A)

- A（ルール・ヘッド）の確率の高い連関規則はいい規則である。
- Aが頻繁に観察され、その規則を適用するチャンスが多い。

- 例えば、Aが「ズボンを一本買った客」である場合と
  「ズボンを二本買った客」である場合を比較すると
  前者のほうが生じる確率が高く、規則を適用できるチャンスが多い。

条件付き確率p(B|A)

- 条件付き確率の高い連関規則は良い規則である。
- 条件付き確率は「信頼度(confidence)」などと呼ばれることもある。

- 例えば「パチンコ台の特定の釘が曲がってる」→「大儲け」という連関規則は、たとえ前提確率が低くても条件付確率が高ければ有用である。

同時確率p(A,B)

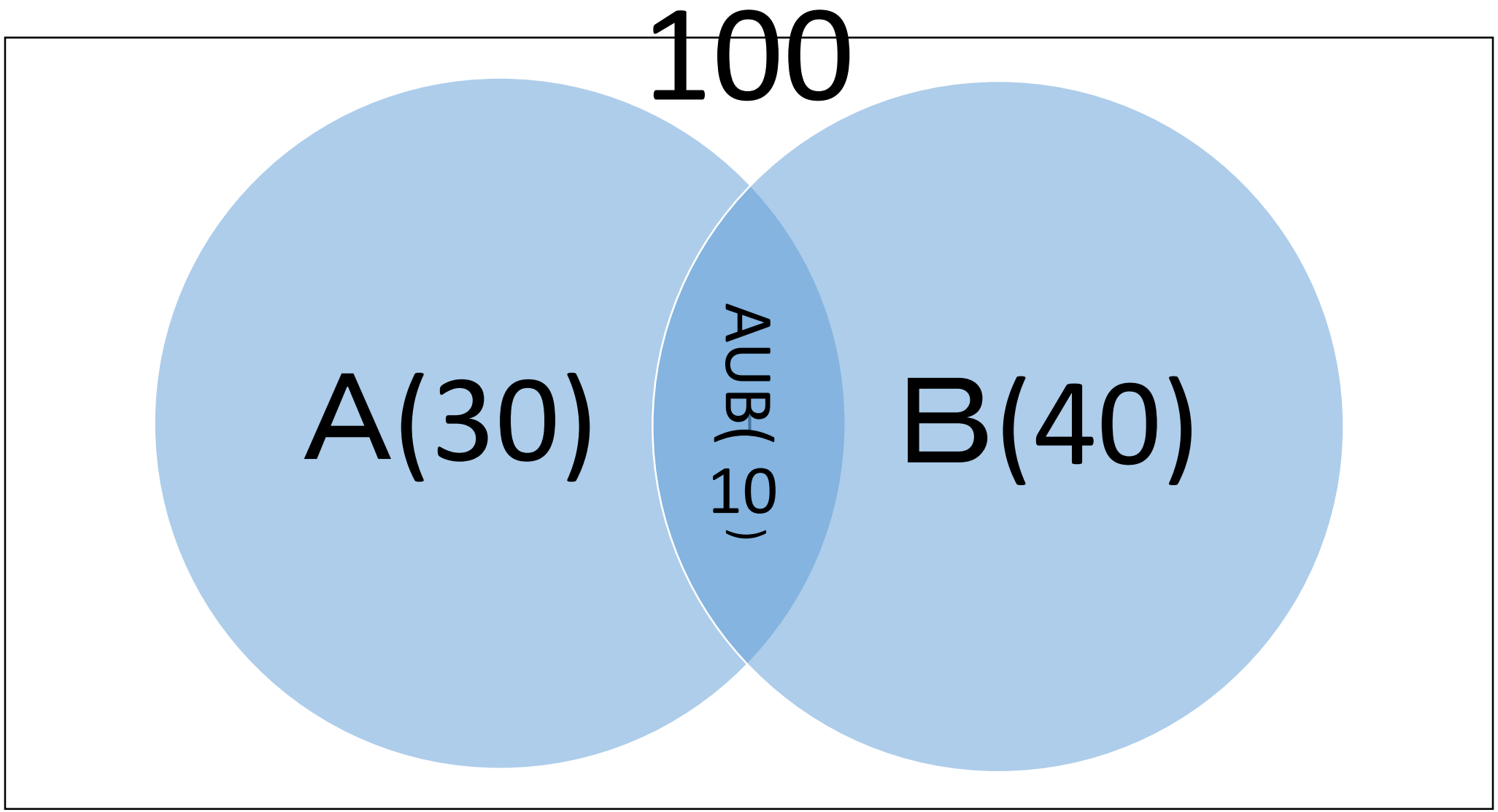
- 同時確率の高い連関規則は良い規則である。
- 同時確率は「サポート(support)」などと呼ばれることもある。
- 先ほど説明した通り、同時確率とはともに高くなってほしい
  前提確率と条件付き確率の指標を同時に考慮した指標である。

- 例えば、「上着を買う」→「ズボンを買う」の連関規則の場合、
  同時確率は「上着を買い、なおかつズボンも買う」の確率である。

事前確率p(B)

- 条件付き確率と比較して事前確率の低い連関規則は良い規則である。
- 条件付き確率が高くても、事前確率も同程度に高いと
  ルール・ヘッドの吟味は必要なくなってしまうからである。

- 例えば、「x」→「y」という連関規則の
  条件付き確率と事前確率が同程度だとすると、「x」という
  ルール・ヘッドはどうでもよくなってしまう。
  つまり「x」とは関係なく「y」ということは起こる。



Supp(A)=30/100

Supp(A∪B)=10/100

Conf(A→B)=Supp(A∪B)/Supp(A)=10/30

Lift(A→B)=Supp(A∪B)/（Supp(A)*Supp(B)）=10/12

バスケット分析

- バスケット分析では数十万の顧客、数万種の商品といった膨大なデータを解析する。
- 連関規則は、買った場合は1、買わない場合は0という形式で整えられた「人x物」の膨大なデータ行列から発見される。
- 解析される行列の要素は、ほとんどが0であり、これは「疎行列」と呼ばれる。
- 典型的な買い物データは非常に大きな疎行列となる。

演習

今回の演習では連関規則を発見するためのパッケージ
“arules”を使う。

```
library(arules)
OnsenData<-read.csv(“onsen.csv”,header=TRUE,row.names=1)
OnsenData<-as.matrix(OnsenData)
```

row.namesはデータの一列目を行名にする。

演習

OnsenTransaction <- as(OnsenData,”transactions”)

summary(OnsenTransaction)

inspect(OnsenTransaction)

as()という関数はデータの形式を変える関数。
これを使用してトランザクションデータへと変換している。

演習

itemFrequency(OnsenTransaction)

itemFrequencyPlot(OnsenTransaction,ylim=c(0,1))

itemFrequency()の命令を使えば疾患の出現確率が表示される。

itemFrequencyPlot()を使えば棒グラフの形で表示することができる。

演習

```
OnsenRule <- apriori(OnsenTransaction,parameter=list(maxlen=4,
        support=0.04,confidence=0.55,ext=TRUE))
```

apriori()はトランザクションデータから連関規則を抽出する関数。
maxlenは1つの連関規則に含まれる最大項目数(ルールの長さ)。
supportは規則のサポートの下限の値。
confidenceは規則の信頼度の下限の値。

ext=TRUEは前提を示すlhs.supportの項を示すようになる。
なおlhsはルール・ヘッドのことを指す。

演習

summary(OnsenRule)

inspect(OnsenRule)

以上が連関規則の見つけ出し方である。

Liftは連関規則の有用性を確かめる一つの指標。

宿題

演習で出てきた連関規則を比較してみよう。
ただし数が多いのでいくつか抽出する。

inspect(OnsenRule [c(20, 21, 70, 69),])
これは20,21,70,69番目の連関規則を抽出して表示している。

抽出した4つの連関規則を比較することで、様々なことがわかる。
→いらない連関規則を探してみよう。

感想

- データを入力するだけで連関規則を簡単に見つけだすことができるのは便利だと思った。

- 連関規則の有用性は人間が判断しなければいけないので、そこが手間のかかる部分だと思った。