Groceries に関する R 言語を使用した頻出パターン抽出及び相関ルール分析

文理学部情報科学科
5419045 高林 秀

2021 年 7 月 23 日

概要

本稿では、今年度データ科学 2 で学習した「頻出パターン抽出」及び「相関ルール分析」の手法を使用して、R 言語のライブラリである arules に付属しているデータ Groceries を対象とした頻出パターン抽出、相関ルール分析を行うものである。

1 目的

本稿では実際に、R 言語を使用しライブラリ arules 付属のデータである Groceries の頻出パターン抽出、相関ルール分析を行うことで、本年度データ科学 2 で学習した頻出パターン、相関ルール分析の手法への理解を深め、その定着を図ることを目的とする。また、1 年次に学習した latex を用いた PDF 作成の復習も兼ねるものである。

2 理論説明

今回の実験で用いた、計算理論をそれぞれ説明する。

2.1 バスケット分析

初めに、頻出パターン抽出、相関ルール分析を説明する前に「バスケット分析」について説明する。

バスケット分析とは一言でいうと、「顧客の購買記録をデータ化し分析を行うことで、顧客に共通するルールや傾向を導く」データ分析のことである。顧客の買い物データを分析しその結果を企業の販促活動などのマーケティングに関わる施策に適用するのが目的である。なお、バスケット分析はアソシエーション分析[*1]の一つとされ、マーケットバスケット分析とも呼ばれる。

[*1] データマイニングにおけるデータ間の関連性を見つける手法のこと。「もし A ならば B である」といった法則を見つけ出し、主に購買記録などから顧客の購買行動の関連性を見つけ出すのに利用される。

今回の実験は仮想マシン上で R 言語を起動し行った。下記に実験時の環境を示す。

- ホスト OS：Window10 Home Ver.20H2
- 仮想 OS：Ubuntu 20.04.2 LTS
- CPU：Intel(R)Core(TM)i7-9700K @ 3.6GHz
- GPU：Nvidia Geforce RTX2070 OC @ 8GB
- ホスト RAM：16GB
- 仮想 RAM：4GB