IST 687: Association Rule Mining and Support Vector Machines

Corey Jackson

2021-03-10 19:50:59

Today's Agenda

- Announcements
- Lab 9
  - Using exploratory analysis and arules
- Tips for Homework 9
- Project meetings (Project update III next week)

Announcements

- Office Hours: Wed. 6-7pm EDT and by appointment
- HW grades/feedback on the LMS
- Exam digest
  - Grades/feedback on the LMS over the weekend
  - Comments/questions about exam

Week 9: Association Rule Mining (ARM)

Week 9: Association Rule Mining (ARM)

- An unsupervised machine learning algorithm
- Finding interesting associations by determining which items in data co-occur at some threshold and quantify these associations in rules e.g., **{milk,diapers}** $\Rightarrow$ **{beer}**
- A wide range of applications in consumer behavior e.g., cross selling, product placement, customer behavior.
- Good synthesis of ARM Explanation of the Market Basket Model.

Week 9: An example of ARM

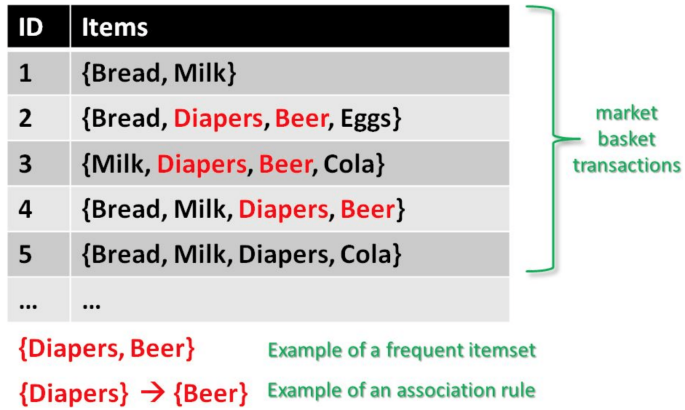
ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

Week 9: An example of ARM



ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

To evalute association rules:

- **support** indicates how frequently items appear together.
  - "How often do we observe $\{milk, diapers\}$ and $\{beer\}$ in the data" (2/5; 20%)
- **confidence** indicate how often the rule has found to be true.
  - "When we observe $\{milk, diapers\}$ how often do we also observe $\{beer\}$" (2/3; 66%)
- **lift** a measure of interestingness of the rule using confidence

Lab 9 Overview: Association Rule Mining

- Goal: Using association rules determine which words are likely to co-occur and evalute the rules using **support**, **confidence**, and **lift**

Lab 9 Overview: Rule Mining Starter Kit

Lab 9 packages packages needed:

- arules: use the `apriori()` function to find these relations based on the frequency of items bought together.
- arulesViz: helps visualize rules/word frequencies in the dataset.

Lab 9 data:

- R Dataset: a term document matrix of tweets about data mining

Lab 9 Overview: Data Munging

Data structure for ARM

- Term-document matrix: matrix that describes the frequency of terms that occur in a collection of documents.

1. Intelligent applications creates intelligent business processes
2. Bots are intelligent applications
3. I do business intelligence

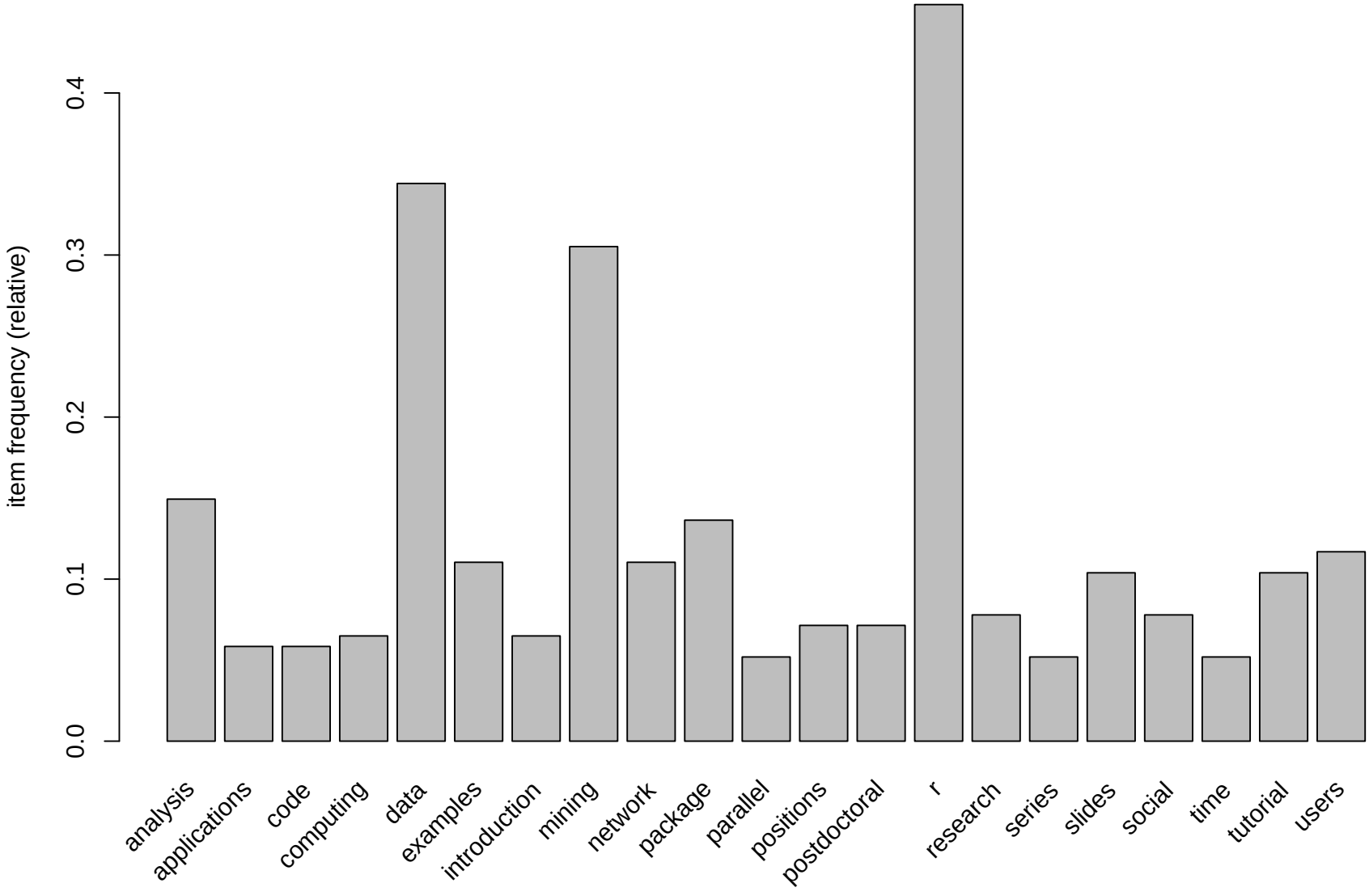
The Document Term Matrix for the three documents is:

	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0	1	0	0	0	1	1	1

- I've created the transposed document for you to begin constructing arules here -> Lab 9 Starter Code

Lab 9 Overview: Finding association rules

```
itemFrequencyPlot(d)
```



Lab 9 Overview: Finding association rules

```
rules <- apriori(df,parameter=list(support=0.01, +
  confidence=0.5))
```

To check rules: `inspect(rules)`

```
##       lhs                  rhs                support confi
## [1]   {applications} => {mining}        0.03896104  0.666
## [2]   {applications} => {data}          0.04545455  0.777
## [3]   {parallel}     => {computing}     0.04545455  0.875
## [4]   {computing}    => {parallel}      0.04545455  0.700
## [5]   {parallel}     => {r}             0.04545455  0.875
## [6]   {research}     => {data}          0.03896104  0.500
## [7]   {code}         => {examples}      0.03896104  0.666
## [8]   {code}         => {r}             0.05194805  0.888
## [9]   {computing}    => {r}             0.05844156  0.900
## [10]  {series}       => {time}          0.05194805  1.000
## [11]  {time}         => {series}        0.05194805  1.000
## [12]  {series}       => {analysis}      0.02597403  0.500
```

Lab 9 Overview: Finding association rules

- Set your parameters so you generate at least 20 rules
- Check rules using `summary(rules)` and visualize using `plot(rules)`

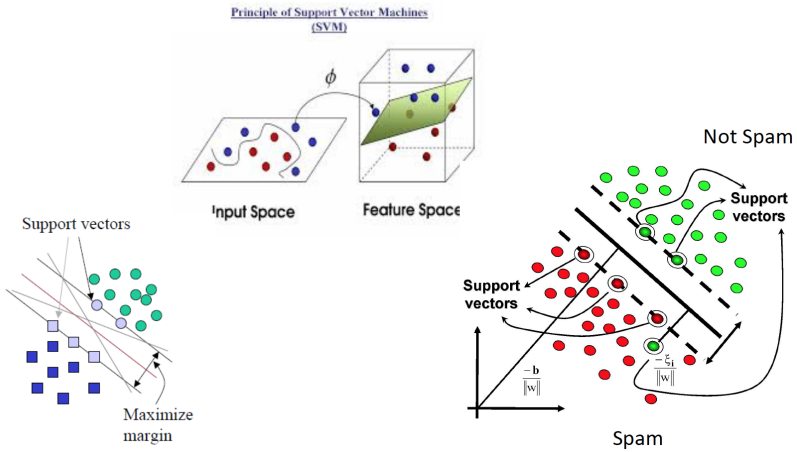
Code to build arules chapter 17 in course textbook and more code examples for AR rules and visualization here: r-statistics.co

Questions???

Week 9: Support Vector Machines (SVMs)

Week 9: Support Vector Machines (SVMs)

- Support Vector Machines
  - Classification of cases based on input data. Use cases may be in facial detection, handwriting, categorizing articles on the web.



- Introduce you to several data mining algorithms: naive bayes `naiveBayes()`, support vector machines `ksvm()`/`svm()`, linear modeling `lm()`

Packages needed: `e1071`, `ggplot2`, `kernlab`, and `gridExtra`

Homework 9 Tips: Comparing machine learning models

- There are many machine learning models to choose from and each has pros/cons.
- Using the same dataset and organization of independent and dependent variables you should compare several models.

Method	Sensitivity	Specificity	F-Measure	Accuracy	AUC
CART	0.813	0.807	0.811	0.810	0.855
Naive Bayesian	0.589	0.893	0.710	0.743	0.811
BayesianNet	0.604	0.914	0.727	0.760	0.839
J48	0.858	0.843	0.850	0.850	0.892
Logistic	0.628	0.893	0.738	0.762	0.845
Neural Network	0.659	0.849	0.742	0.754	0.833
SVM	0.565	0.902	0.695	0.735	0.734

Homework 9 Tips: Comparing model output

Compare the performance of three regression (lm, svm, ksvm) and three classification (naiveBayes, svm, ksvm) algorithms. For each model you should:

- Create the model using the training data,
- Evaluate the model on the test dataset (Step 5)
  - Regression: compute model error using rmse
  - Classification compute accuracy via confusion matrix using `predict()`
- plot the model error using `ggplot()`

Remember: In Step 3, your dependent variable is continuious. In Step 4, the dependent variable is discrete.

Homework 9 Tips: Converting a continuious variable to discrete (Step 4)

We may be interested in predicting a discrete class (e.g., good vs. bad).

Scenario: Given continuious data about the number of fawns birthed (HW 8), is the number of fawns in that year good or bad?

Approach: If the fawns is less than the mean. . . it was a bad year ($= 0$), otherwise it was a good year ($= 1$).

Homework 9 Tips: Converting a continuious variable to discrete (Step 4)

```
##   fawn adult precipitation severity
## 1  2.9   9.2          13.2        2
## 2  2.4   8.7          11.5        3
```

```
df$goodyear <- as.factor(ifelse(df$fawn <
mean(df$fawn), 0, 1))
```

```
##   fawn adult precipitation severity goodyear
## 1  2.9   9.2          13.2        2        1
## 2  2.4   8.7          11.5        3        0
```

- In Step 4, the dependent variable is goodozone

Next week

Asynchronous Materials

- Week 10: Text Mining
- Submit HW 9

Live Session

- Lab 10 on Text Mining
- Project Updates