Predictive Analytics Part 1: Clustering, Classification, Association & Frequent Itemset Mining

0.1 Cluster Analysis in R

For the following exercise, load the *BigMac2003*[1] data set from the *alr4*-package. This data set contains for 69 cities world-wide the average working hours, price level and income level for the year 1991. Your task is to perform a cluster analysis of the cities based on selected variables. This data set contains the following variables (to list them, you might use, e.g., attributes(BigMac2003)):

BigMac Minutes of labor to purchase a Big Mac

Bread Minutes of labor to purchase 1 kg of bread

Rice Minutes of labor to purchase 1 kg of rice

FoodIndex Food price index (Zurich=100)

Bus Cost in US dollars for a one-way 10 km ticket

Apt Normal rent (US dollars) of a 3 room apartment

TeachGI Primary teacher's gross income, 1000s of US dollars

TeachNI Primary teacher's net income, 1000s of US dollars

TaxRate Tax rate paid by a primary teacher

TeachHours Primary teacher's hours of work per week

For the class label, generate vector with the cities' regions (aggregated to 4 classes) as follows:

```
region=as.factor(c("EU","EU","AUNZ","AS","EU","EU","EU","SA","EU","EU","EU","EU","SA","SA","NA",
  "EU","AF","EU","EU","EU","EU","AS","AF","AS","AF","AS","EU","AS","AF","SA","EU","EU","EU","NA",
  "EU","EU","EU","AF","AS","NA","NA","EU","NA","EU","AS","AF","NA","EU","EU","EU","EU","SA","EU",
  "SA","SA","AS","AS","AS","EU","EU","AUNZ","AS","EU","AF","AS","NA","EU","EU","EU"))
region4 = region;
levels(region4) = c("AF","AS","N","N","N","SA")
```

Now perform the following steps:

1. Select the variables *FoodIndex*, *Bus*, *Apt*, *TeachNI* and *TeachHours* and scale them (using the function scale() to an arithmetic mean of 0 and a standard deviation of 1.
2. Compute the distance matrices for distances based on $(1 - |correlation|)$, L_1 (manhattan) and L_2 (euclidean) norm:
3. Compute a divisive hierarchical clustering using the R function diana() from the package *cluster* and analyse the results by studying the resulting dendrograms with the plot() function.
4. Compute an agglomerative hierarchical clustering using the R function agnes() from the package *cluster* and analyse the results by studying the resulting dendrograms with the plot() function.
5. Using only the attributes *FoodIndex* and *Apt*, compute a k-means based clustering using the R function kmeans() for $k = 3$, $k = 4$, and $k = 5$, and analyse the results in a 2-dimensional plot of the clusters and their points. Using the functons intCriteria() and extCriteria from the library *clusterCrit*, compute the Silhouette Index as internal measure for all clustering. For $k = 4$, compute also the Rand Index as external measure (against the class label vector *region4*).
6. As above, use the attributes *FoodIndex* and *Apt*, but now compute an EM-based clustering using the R functions init.EM() and emcluster() for $nclass = 3$ and $nclass = 5$ from the library *EMCluster*, and analyse the results in a 2-dimensional plot of the clusters and their points. As above, compute the Silhouette and Rand Indices.
7. Generate a knee-plot for determining the optimal number of clusters in your kmeans or EM-clusterings based on the Silhouette Index.

[1] Provided by the Union Bank of Switzerland.

0.2 How would you distinguish between flat and hierarchical partitioning?

0.3 What is probabilistic clustering?

0.4 What is the difference between soft and hard partitioning?

0.5 What are the two main approaches of flat partitioning?

0.6 What are the two main approaches of hierarchical partitioning?

0.7 Discuss the Minkowski-Distances for $r = 1$ and $r = 2$, their differences and their alternative names.

0.8 On which of the following datasets would you use (1) K-Means Clustering (or EM), (2) single-linkage agglomerative hierarchical clustering, (3) density-based clustering? Provide arguments for your recommendation.

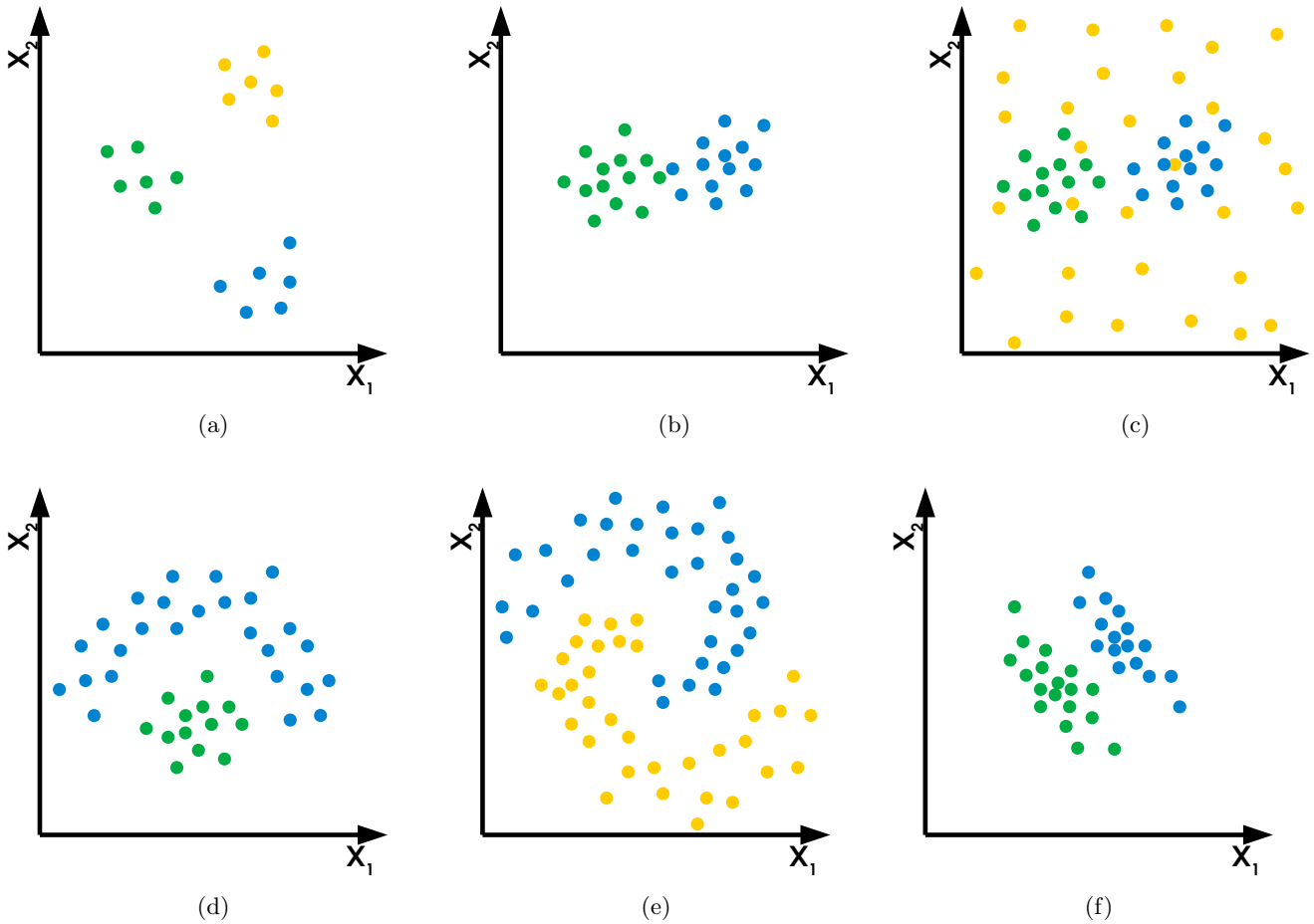


Figure 1: Scatter plots of different data sets.

0.9 Which clustering(s) tend(s) to form clusters that are chains of points?

( ) Single Linkage Agglomerative Hierarchical Clustering

( ) Complete Linkage Agglomerative Hierarchical Clustering

( ) K-Means Clustering

( ) Expectation Maximisation with Mixture of Gaussians

0.10 Which clustering(s) tend(s) to form spherical clusters?

( ) Single Linkage Agglomerative Hierarchical Clustering

( ) Complete Linkage Agglomerative Hierarchical Clustering

( ) K-Means Clustering

( ) Expectation Maximisation with Mixture of Gaussians

0.11 Which of the following similarity/distance measures is appropriate for comparing instances based on their relative attribute values?

( ) Euclidean Distance

( ) L1 norm or City-Block Metric

( ) Mahanalobis distance

( ) Q-Correlation Coefficient

0.12 Which of the following similarity/distance measures is appropriate for comparing instances based on the absolute distance on their attribute values?

( ) Euclidean Distance

( ) L1 norm or City-Block Metric

( ) Mahanalobis distance

( ) Q-Correlation Coefficient

0.13 Which of the following statements are correct?

( ) The Silhouette Index is an internal cluster quality measure.

( ) Internal cluster quality measures compare the cluster quality based on ground-truth class labels.

( ) External cluster quality measures compare the cluster quality based on ground-truth class labels.

( ) Clustering aims to maximise similarity between instances within clusters and dissimilarity between instances in different clusters.

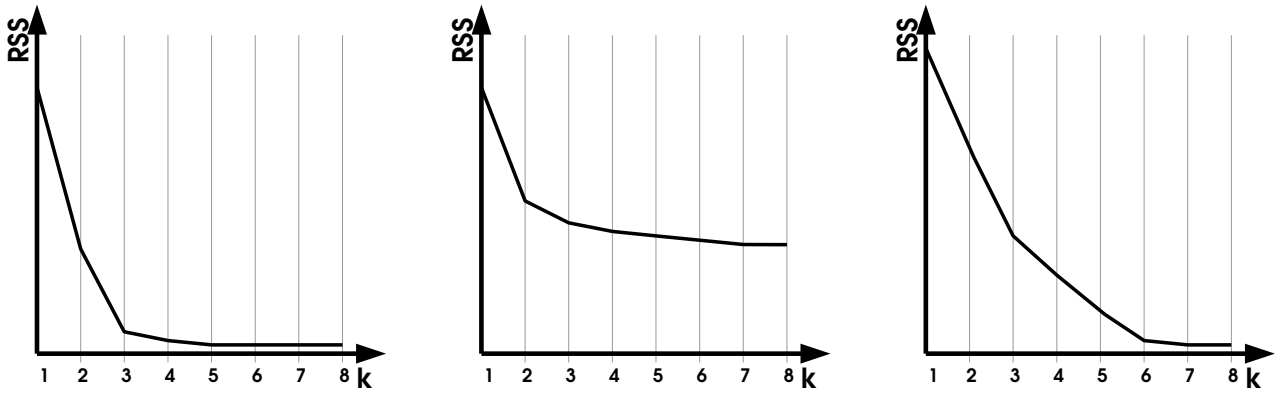
( ) The Silhouette Index considers the intra-cluster and inter-cluster distances.

0.14 You want to compare two clusterings, one with three and one with five clusters. Which external performance measures can you use?

( ) Purity

( ) Normalised Mutual Information

0.15 Based on the elbow method, determine in the plots in Fig. 2 below which number of clusters k to chose.



(a) (b) (c)

Figure 2: Elbow plots for three different data sets.

0.16 Which of the following statements are correct?

( ) In classification, the task is to predict a dependent variable (class label) based on a set of explanatory (feature) variables.

( ) A probabilistic classifier returns also estimates of the posterior class probability $\mathcal{P}r(y|x)$.

( ) A lazy learner does not construct an (abstract) model from the data.

( ) A Bayes classifier requires an estimate of either the joint probability $\mathcal{P}r(x, y)$, or estimates of the class prior probability $\mathcal{P}r(y)$ together with the class-conditional feature probability $\mathcal{P}r(x|y)$.

( ) The (unconditional) feature probability $\mathcal{P}r()$ can be computed by summing (marginalising) $\mathcal{P}r(x, y)$ over the different values of x.

( ) The (unconditional) feature probability $\mathcal{P}r()$ can be computed by summing (marginalising) $\mathcal{P}r(x, y)$ over the different values of y.

( ) In contrast to a multivariate Bayes classifier, a Naive Bayes classifier assumes conditional independence.

( ) A predictive classifier also provides a model of the underlying data distribution, which can be used to generate data.

( ) For k-fold cross validation, the data set is partitioned into k subsets, and each of them is used once for testing and the other times for training.

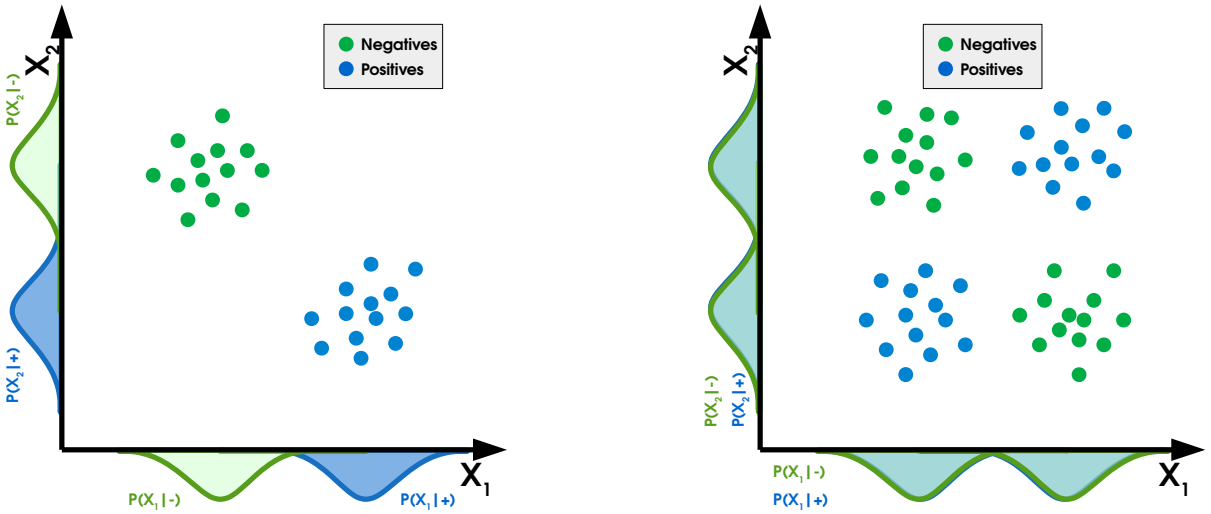
( ) For obtaining a reliable estimate of the performance on new data, we need to evaluate a classifier on its *training* data.

( ) False positives are instances that are classified as negative, but actually are positive.

( ) In a region, where we classify all instances as positive, our false negative rate is zero.

0.17 Curse of Dimensionality: If you have three feature variables X_1,X_2, and X_3, which you divide each into two intervals (low vs. high values), in how many cells does this split your feature space? Alternatively, how many training instances do you need to place exactly one in each cell? How does this change with four features X_1,X_2,X_3 and X_4?

0.18 **In Fig. 3, the scatterplots of two data sets (a) and (b) are given. In which of these two is conditional independence assumption clearly violated?**



(a) (b)

Figure 3: Scatterplots of two datasets.

0.19 Based on the AUC/ROC plotted in Fig. 4, which of the following statement(s) is/are correct?

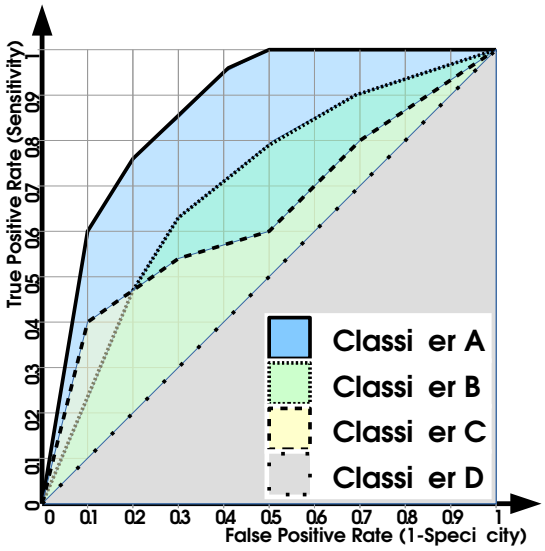


Figure 4: Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) Plots.

( ) Classifier A performs best.

( ) Classifier A is a *perfect classifier*.

( ) Classifier D is a *perfect classifier*.

( ) Classifier A performs like a *randomly guessing classifier*.

( ) Classifier D performs like a *randomly guessing classifier*.

( ) Classifier A dominates classifiers B,C,D.

( ) Classifier B dominates classifiers C,D.

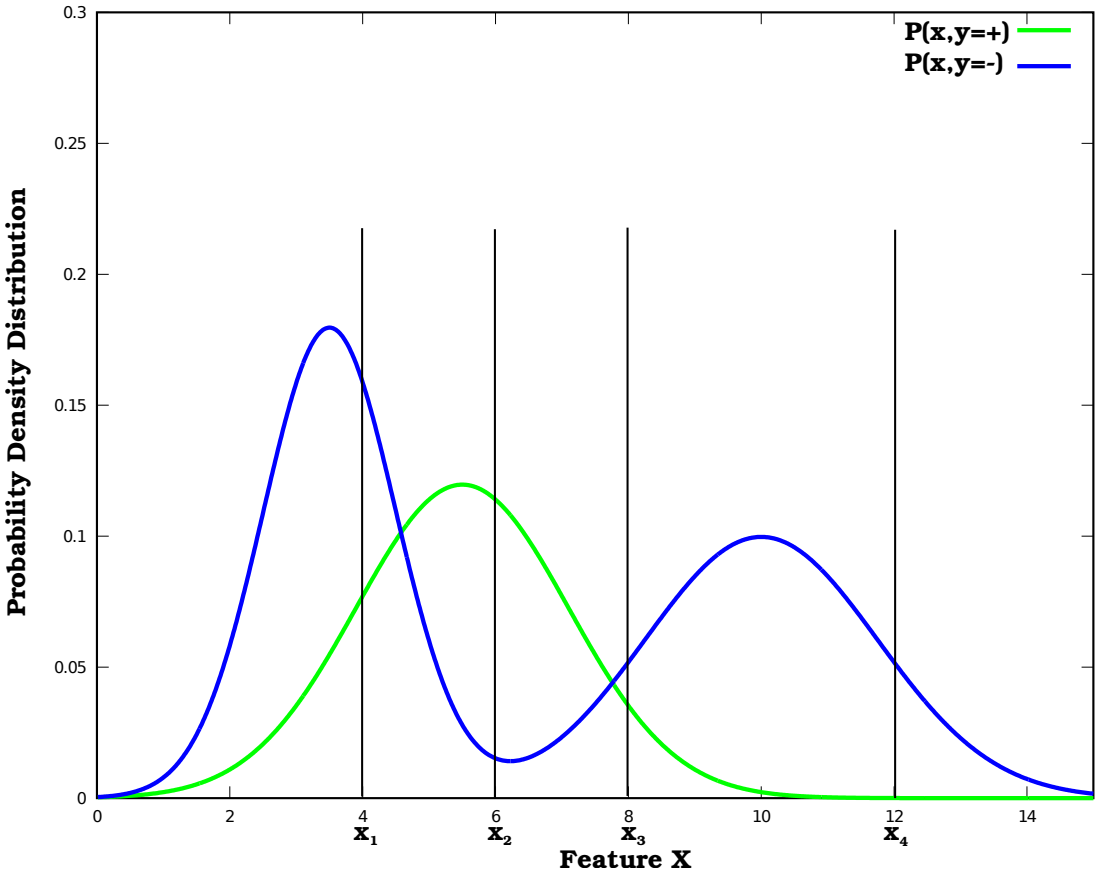
( ) Classifier D dominates classifiers A,B,C.

( ) Classifier D is dominated by classifiers A,B, and C.

( ) Neither classifier B dominates classifier C, nor classifier C dominates classifier B.

0.20 (Univariate) Bayesian Classification

Consider a binary classification problem, with class variable $Y \in \{pos, neg\}$ and feature variable $X \in (0, 15)$. The joint distributions of $P(X, Y)$ are given in the plot below ($P(X, Y = pos)$ is plotted in green, $P(X, Y = neg)$ in blue).



a) Plot the feature distribution $P(X)$

b) Can you derive the posterior class distributions $P(Y = pos|X)$ and $P(Y = neg|X)$ from the presented plot? Sketch them in the plot.

c) Mark the Bayes' optimal decision boundary/boundaries. How would you classify the following instances: $x_1 = 4$, $x_2 = 6$, $x_3 = 8$, $x_4 = 12$

d) Indicate the Bayes' error rate. Assuming a Bayes' optimal decision, annotate the number of false negatives, and the number of true negatives in the plot.

0.21 Multivariate Bayesian Classification

Consider the following, simplified golf-player classification example[2]:

Temp	**Humidity**	**Windy**	**Play**
hot	high	false	no
hot	high	true	no
hot	high	false	*yes*
mild	high	false	*yes*
cool	normal	false	*yes*
cool	normal	true	no
cool	normal	true	*yes*
mild	high	false	no
cool	normal	false	*yes*
mild	normal	false	*yes*
mild	normal	true	*yes*
mild	high	true	*yes*
hot	normal	false	*yes*
mild	high	true	no

a) Calculate and **interpret** the following probabilities:

- $\Pr(X_{\text{temp}} = \text{hot})$
- $\Pr(y = \text{yes})$
- $\Pr(X_{\text{temp}} = \text{hot}, y = \text{yes})$
- $\Pr(X_{\text{temp}} = \text{hot} | y = \text{yes})$
- $\Pr(y = \text{yes} | X_{\text{temp}} = \text{hot})$

b) Explain the classification of a **multivariate Bayes-optimal** classifier for an instance with $X_{\text{temp}} = \text{cool}$ and $X_{\text{hum}} = \text{normal}$ and calculate the necessary probabilities.

c) Explain the classification of a **Naive Bayes** classifier for an instance with $X_{\text{temp}} = \text{cool}$ and $X_{\text{hum}} = \text{normal}$ and calculate the necessary values.

[2]Based on a fictitious dataset from [Witten et al., 2011, chapter 1]

0.22 Classification and Regression in R

In R, replicate the Golf-Player data set from the exercise above with the following command:

```
golf_train<-data.frame(
  temp=c("hot","hot","hot","mild","cool","cool","cool","mild","cool","mild","mild","mild","hot","mild"),
  hum=c("h","h","h","h","n","n","n","h","n","n","n","h","n","h"),
  windy=c("f","t","f","f","f","t","t","f","f","f","t","t","f","t"),
  play=c("n","n","y","y","y","n","y","n","y","y","y","y","y","n"))
```

Next, generate a test set with the following command:

```
golf_test<-data.frame(temp=c("cool","cool","hot"),hum=c("n","n","h"),
  windy=c("t","f","t"),play=c("n","y","n"))
```

a) In R, use the naiveBayes() and predict() functions from the package *e1071* to train a Naive Bayes classifier on the attributes *temp* and *hum*, and to predict the class label of the test instances. Compare the results with your calculation in the exercise above!

b) In R, use the rpart() and predict() functions from the package *rpart* to train a decision tree on the attributes *temp* and *hum*, and to predict the class label of the test instances. Compare the results with above!
Note: Due to the small training set size, you will have to adjust some of rpart's control paramters: $minsplit = 2$, $minbucket = 1$, and $xval = 3$ (for reasonably large data sets you should use higher settings!).
Finally, use the HMeasure() function from the package *hmeasure* to compute the error rate and AUC.

c) In R, use the build-in lm() function to fit a linear regression model on the build-in *iris* data set. This model should use *Sepal.Length* as dependent (response) variable, and *Petal.Length* as well as *Petal.Width* as explanatory variables. Print a summary of the model and plot a histogram of the residuals. Based on your model, discuss the relationship between *Petal.Length*,*Petal.Width* and *Sepal.Length*. What is the predicted *Sepal.Length* for a plant with ($Petal.Length = 4.3, Petal.Width = 1.3$)?

0.23 Frequent Itemset Mining

Apply the apriori algorithm to the basket data in Table 1, using the thresholds $\sigma \geq \frac{2}{8}$ and $c \geq \frac{1}{2}$. Based on the algorithm, which recommendations (in descending order of relevance; also state the recommendations' confidences!) should be made to the following customers:

a) Customer I, who has bought B.

b) Customer II, who has bought B and C .

IID	Item Name
A	Grey's Anatomy
B	The Big Bang Theory
C	Castle
D	Downton Abbey
G	Game of Thrones
H	How I Met Your Mother

(a) List of Items

TID	Items in Baskez
1	**A,B,C,D,G,H**
2	**A,B,C,G**
3	**A,D,H**
4	**B,C**
5	**B,C,D,G**
6	**B,C,G**
7	**B,D,G**
8	**B,G**

(b) Transactions / Baskets

Table 1: Items and Baskets for Frequent Itemset Mining

0.24 Extra: Frequent Itemset Mining in R

In R, use the function apriori() from the *arules*-package to mine the *Groceries*-data set for frequent itemsets. As parameter values, use 0.5 for support and 0.9 for confidence.
Note: This requires a version of $R \geq 3.4.0$

0.25 Extra: Text Mining:

In R, use the function sentiment() from the *sentimentr*-package to perform sentiment analysis on the text in the course_evaluation-data set.
Note: This requires a version of $R \geq 3.4.0$

Appendix

References

[Witten et al., 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3 edition.