Exercise 2: Apriori, Confidence, Itemsets and Association Rules

Exercise 2-1: Combinatoric explosion

(a) A database contains transactions over the following items: “apples”, “bananas”, and “cherries”. How many different combinations of these items can exist (i.e., how many different transactions could possibly occur in the database)?

(We do not distinguish whether a transaction contains a fruit once or several times, e.g., if someone bought one apple or several apples would just result in the transaction containing “apples”.)

(b) The database now also contains the items “dates”, “eggplants”, “figs”, and “guavas”. How many possible transactions do we have now?

(c) How many combinations (possible different transactions) do we have with n items?

(d) How many transactions with exactly two items (i.e., 2-itemsets) can we have when the database contains 3 items? When it contains 5 items? How many k-itemsets do we have when the database contains n items?

Exercise 2-2: Itemsets and Association Rules

Given a set of transactions T according to the following table:

Set of transactions T

Transaction ID	items in basket
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies }
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

(a) What are the support and the confidence of $\{\text{Milk}\} \Rightarrow \{\text{Diapers}\}$?

(b) What are the support and the confidence of $\{\text{Diapers}\} \Rightarrow \{\text{Milk}\}$?

(c) What is the maximum number of size-3 itemsets that can be derived from this data set?

(d) What is the maximum number of association rules that can be extracted from this dataset (including rules, that have zero support)?

(e) What is the maximum size of frequent itemsets that can be extracted (assuming $\sigma > 0$)?

(f) Find an itemset (of size 2 or larger) that has the largest support.

(g) Find a pair of items, a and b, such that the rules $\{a\} \Rightarrow \{b\}$ and $\{b\} \Rightarrow \{a\}$ have the same confidence.

Exercise 2-3: Apriori candidate generation

Given the frequent 3-itemsets:

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,5\},\{2,3,4\},\{2,3,5\},\{2,3,6\},\{2,5,6\},\{3,4,5\},\ \{3,5,6\}$$

List all candidate 4-itemsets following the Apriori joining and pruning procedure.

Exercise 2-4: The monotonicity of confidence

Theorem 2.1 in the Lecture states:

Given:

— itemset X

— $Y \subset X, Y \neq \emptyset$

If $\mathrm{conf}(Y \Rightarrow (X \setminus Y)) < c$, then $\forall Y' \subset Y$:

$$\mathrm{conf}(Y' \Rightarrow (X \setminus Y')) < c.$$

(a) Prove the theorem.

(b) Sketch an algorithm (pseudo code) that generates all association rules with support σ or above and a minimum confidence of c, provided the set F of all frequent itemsets (w.r.t. σ) with their support, efficiently using the pruning power of the given theorem.

Exercise 2-5: Tools

(a) Install python packages: `scikit-learn, numpy, matplotlib, metrics`, and `linear-model` from `scikit-learn`, then load `diabetes` dataset from `sklearn`.

(b) Reserve a randomly chosen 80% of the data for training and the remaining for test using `sklearn.model_selection.train_test_split`, then assign data as x and target as y and investigate the shapes of the data.

(c) Normalize data using `StandardScaler` from `sklearn.preprocessing`.

(d) Fit a linear regression model to the training set and make prediction.

(e) Evaluate mean squared error (MSE) of the fitted model on the test set.

(f) Plot the fitted model as a line and print its intercept and slope.

(g) Comment on the outcome. Could the model fit to data accurately enough?