CHAPTER 12 Pattern and Rule Assessment

In this chapter we discuss how to assess the significance of the mined frequent patterns, as well as the association rules derived from them. Ideally, the mined patterns and rules should satisfy desirable properties such as conciseness, novelty, utility, and so on. We outline several rule and pattern assessment measures that aim to quantify different properties of the mined results. Typically, the question of whether a pattern or rule is interesting is to a large extent a subjective one. However, we can certainly try to eliminate rules and patterns that are not statistically significant. Methods to test for the statistical significance and to obtain confidence bounds on the test statistic value are also considered in this chapter.

12.1 RULE AND PATTERN ASSESSMENT MEASURES

Let $\mathcal{I}$ be a set of items and $\mathcal{T}$ a set of tids, and let $\mathbf{D} \subseteq \mathcal{T} \times \mathcal{I}$ be a binary database. Recall that an *association rule* is an expression $X \longrightarrow Y$, where X and Y are itemsets, i.e., $X, Y \subseteq \mathcal{I}$, and $X \cap Y = \emptyset$. We call X the antecedent of the rule and Y the consequent.

The tidset for an itemset X is the set of all tids that contain X, given as

$$\mathbf{t}(X) = \Big\{ t \in \mathcal{T} \mid X \text{ is contained in } t \Big\}$$

The support of X is thus $sup(X) = |\mathbf{t}(X)|$. In the discussion that follows we use the short form XY to denote the union, $X \cup Y$, of the itemsets X and Y.

Given a frequent itemset $Z \in \mathcal{F}$, where $\mathcal{F}$ is the set of all frequent itemsets, we can derive different association rules by considering each proper subset of Z as the antecedent and the remaining items as the consequent, that is, for each $Z \in \mathcal{F}$, we can derive a set of rules of the form $X \longrightarrow Y$, where $X \subset Z$ and $Y = Z \setminus X$.

12.1.1 Rule Assessment Measures

Different rule interestingness measures try to quantify the dependence between the consequent and antecedent. Below we review some of the common rule assessment measures, starting with support and confidence.

Table 12.1. Example Dataset

Tid	Items
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

Table 12.2. Frequent itemsets with $minsup = 3$ (relative minimum support 50%)

sup	*rsup*	Itemsets
3	0.5	$ABD, ABDE, AD, ADE, BCE, BDE, CE, DE$
4	0.67	$A, C, D, AB, ABE, AE, BC, BD$
5	0.83	E, BE
6	1.0	B

Support

The *support* of the rule is defined as the number of transactions that contain both X and Y, that is,

$$sup(X \longrightarrow Y) = sup(XY) = |\mathbf{t}(XY)| \tag{12.1}$$

The *relative support* is the fraction of transactions that contain both X and Y, that is, the empirical joint probability of the items comprising the rule

$$rsup(X \longrightarrow Y) = P(XY) = rsup(XY) = \frac{sup(XY)}{|\mathbf{D}|}$$

Typically we are interested in frequent rules, with $sup(X \longrightarrow Y) \geq minsup$, where *minsup* is a user-specified minimum support threshold. When minimum support is specified as a fraction then relative support is implied. Notice that (relative) support is a symmetric measure because $sup(X \longrightarrow Y) = sup(Y \longrightarrow X)$.

Example 12.1. We illustrate the rule assessment measures using the example binary dataset $\mathbf{D}$ in Table 12.1, shown in transactional form. It has six transactions over a set of five items $\mathcal{I} = \{A, B, C, D, E\}$. The set of all frequent itemsets with $minsup = 3$ is listed in Table 12.2. The table shows the support and relative support for each frequent itemset. The association rule $AB \longrightarrow DE$ derived from the itemset $ABDE$ has support $sup(AB \longrightarrow DE) = sup(ABDE) = 3$, and its relative support is $rsup(AB \longrightarrow DE) = sup(ABDE)/|\mathbf{D}| = 3/6 = 0.5$.

Confidence

The *confidence* of a rule is the conditional probability that a transaction contains the consequent Y given that it contains the antecedent X:

$$conf(X \longrightarrow Y) = P(Y|X) = \frac{P(XY)}{P(X)} = \frac{rsup(XY)}{rsup(X)} = \frac{sup(XY)}{sup(X)}$$

Table 12.3. Rule confidence

Rule	*conf*
$A \longrightarrow E$	1.00
$E \longrightarrow A$	0.80
$B \longrightarrow E$	0.83
$E \longrightarrow B$	1.00
$E \longrightarrow BC$	0.60
$BC \longrightarrow E$	0.75

Typically we are interested in high confidence rules, with $conf(X \longrightarrow Y) \geq minconf$, where *minconf* is a user-specified minimum confidence value. Confidence is not a symmetric measure because by definition it is conditional on the antecedent.

Example 12.2. Table 12.3 shows some example association rules along with their confidence generated from the example dataset in Table 12.1. For instance, the rule $A \longrightarrow E$ has confidence $sup(AE)/sup(A) = 4/4 = 1.0$. To see the asymmetry of confidence, observe that the rule $E \longrightarrow A$ has confidence $sup(AE)/sup(E) = 4/5 = 0.8$.

Care must be exercised in interpreting the goodness of a rule. For instance, the rule $E \longrightarrow BC$ has confidence $P(BC|E) = 0.60$, that is, given E we have a probability of 60% of finding BC. However, the unconditional probability of BC is $P(BC) = 4/6 = 0.67$, which means that E, in fact, has a deleterious effect on BC.

Lift

Lift is defined as the ratio of the observed joint probability of X and Y to the expected joint probability if they were statistically independent, that is,

$$lift(X \longrightarrow Y) = \frac{P(XY)}{P(X) \cdot P(Y)} = \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} = \frac{conf(X \longrightarrow Y)}{rsup(Y)}$$

One common use of lift is to measure the surprise of a rule. A lift value close to 1 means that the support of a rule is expected considering the supports of its components. We usually look for values that are much larger (i.e., above expectation) or smaller than 1 (i.e., below expectation).

Notice that lift is a symmetric measure, and it is always larger than or equal to the confidence because it is the confidence divided by the consequent's probability. Lift is also not downward closed, that is, assuming that $X' \subset X$ and $Y' \subset Y$, it can happen that $lift(X' \longrightarrow Y')$ may be higher than $lift(X \longrightarrow Y)$. Lift can be susceptible to noise in small datasets, as rare or infrequent itemsets that occur only a few times can have very high lift values.

Example 12.3. Table 12.4 shows three rules and their lift values, derived from the itemset $ABCE$, which has support $sup(ABCE) = 2$ in our example database in Table 12.1.

Table 12.4. Rule lift

Rule			*lift*
AE	$\longrightarrow$	BC	0.75
CE	$\longrightarrow$	AB	1.00
BE	$\longrightarrow$	AC	1.20

The lift for the rule $AE \longrightarrow BC$ is given as

$$\mathit{lift}(AE \longrightarrow BC) = \frac{rsup(ABCE)}{rsup(AE)\cdot rsup(BC)} = \frac{2/6}{4/6 \times 4/6} = 6/8 = 0.75$$

Since the lift value is less than 1, the observed rule support is less than the expected support. On the other hand, the rule $BE \longrightarrow AC$ has lift

$$\mathit{lift}(BE \longrightarrow AC) = \frac{2/6}{2/6 \times 5/6} = 6/5 = 1.2$$

indicating that it occurs more than expected. Finally, the rule $CE \longrightarrow AB$ has lift equal to 1.0, which means that the observed support and the expected support match.

Example 12.4. It is interesting to compare confidence and lift. Consider the three rules shown in Table 12.5 as well as their relative support, confidence, and lift values. Comparing the first two rules, we can see that despite having lift greater than 1, they provide different information. Whereas $E \longrightarrow AC$ is a weak rule ($conf = 0.4$), $E \longrightarrow AB$ is not only stronger in terms of confidence, but it also has more support. Comparing the second and third rules, we can see that although $B \longrightarrow E$ has lift equal to 1.0, meaning that B and E are independent events, its confidence is higher and so is its support. This example underscores the point that whenever we analyze association rules, we should evaluate them using multiple interestingness measures.

Leverage

Leverage measures the difference between the observed and expected joint probability of XY assuming that X and Y are independent

$$leverage(X \longrightarrow Y) = P(XY) - P(X)\cdot P(Y) = rsup(XY) - rsup(X)\cdot rsup(Y)$$

Leverage gives an “absolute” measure of how surprising a rule is and it should be used together with lift. Like lift it is symmetric.

Example 12.5. Consider the rules shown in Table 12.6, which are based on the example dataset in Table 12.1. The leverage of the rule $ACD \longrightarrow E$ is

$$leverage(ACD \longrightarrow E) = P(ACDE) - P(ACD)\cdot P(E) = 1/6 - 1/6 \times 5/6 = 0.03$$

Similarly, we can calculate the leverage for other rules. The first two rules have the same lift; however, the leverage of the first rule is half that of the second rule, mainly due to the higher support of ACE. Thus, considering lift in isolation may be

Table 12.5. Comparing support, confidence, and lift

Rule	*rsup*	*conf*	*lift*
$E \longrightarrow AC$	0.33	0.40	1.20
$E \longrightarrow AB$	0.67	0.80	1.20
$B \longrightarrow E$	0.83	0.83	1.00

Table 12.6. Rule leverage

Rule	*rsup*	*lift*	*leverage*
$ACD \longrightarrow E$	0.17	1.20	0.03
$AC \longrightarrow E$	0.33	1.20	0.06
$AB \longrightarrow D$	0.50	1.12	0.06
$A \longrightarrow E$	0.67	1.20	0.11

misleading because rules with different support may have the same lift. On the other hand, the second and third rules have different lift but the same leverage. Finally, we emphasize the need to consider leverage together with other metrics by comparing the first, second, and fourth rules, which, despite having the same lift, have different leverage values. In fact, the fourth rule $A \longrightarrow E$ may be preferable over the first two because it is simpler and has higher leverage.

Jaccard

The Jaccard coefficient measures the similarity between two sets. When applied as a rule assessment measure it computes the similarity between the tidsets of X and Y:

$$\begin{aligned} jaccard(X \longrightarrow Y) &= \frac{|\mathbf{t}(X) \cap \mathbf{t}(Y)|}{|\mathbf{t}(X) \cup \mathbf{t}(Y)|} \\ &= \frac{sup(XY)}{sup(X) + sup(Y) - sup(XY)} \\ &= \frac{P(XY)}{P(X) + P(Y) - P(XY)} \end{aligned}$$

Jaccard is a symmetric measure.

Example 12.6. Consider the three rules and their Jaccard values shown in Table 12.7. For example, we have

$$jaccard(A \longrightarrow C) = \frac{sup(AC)}{sup(A) + sup(C) - sup(AC)} = \frac{2}{4+4-2} = 2/6 = 0.33$$

Conviction

All of the rule assessment measures we considered above use only the joint probability of X and Y. Define $\neg X$ to be the event that X is not contained in a transaction,

Table 12.7. Jaccard coefficient

Rule			*rsup*	*lift*	*jaccard*
A	$\longrightarrow$	C	0.33	0.75	0.33
A	$\longrightarrow$	E	0.67	1.20	0.80
A	$\longrightarrow$	B	0.67	1.00	0.67

that is, $X \not\subseteq t \in \mathcal{T}$, and likewise for $\neg Y$. There are, in general, four possible events depending on the occurrence or non-occurrence of the itemsets X and Y as depicted in the contingency table shown in Table 12.8.

Conviction measures the expected error of the rule, that is, how often X occurs in a transaction where Y does not. It is thus a measure of the strength of a rule with respect to the complement of the consequent, defined as

$$conv(X \longrightarrow Y) = \frac{P(X)\cdot P(\neg Y)}{P(X\neg Y)} = \frac{1}{lift(X \longrightarrow \neg Y)}$$

If the joint probability of $X\neg Y$ is less than that expected under independence of X and $\neg Y$, then conviction is high, and vice versa. It is an asymmetric measure.

From Table 12.8 we observe that $P(X) = P(XY) + P(X\neg Y)$, which implies that $P(X\neg Y) = P(X) - P(XY)$. Further, $P(\neg Y) = 1 - P(Y)$. We thus have

$$conv(X \longrightarrow Y) = \frac{P(X)\cdot P(\neg Y)}{P(X) - P(XY)} = \frac{P(\neg Y)}{1 - P(XY)/P(X)} = \frac{1 - rsup(Y)}{1 - conf(X \longrightarrow Y)}$$

We conclude that conviction is infinite if confidence is one. If X and Y are independent, then conviction is 1.

Example 12.7. For the rule $A \longrightarrow DE$, we have

$$conv(A \longrightarrow DE) = \frac{1 - rsup(DE)}{1 - conf(A)} = 2.0$$

Table 12.9 shows this and some other rules, along with their conviction, support, confidence, and lift values.

Odds Ratio

The odds ratio utilizes all four entries from the contingency table shown in Table 12.8. Let us divide the dataset into two groups of transactions – those that contain X and those that do not contain X. Define the odds of Y in these two groups as follows:

$$odds(Y|X) = \frac{P(XY)/P(X)}{P(X\neg Y)/P(X)} = \frac{P(XY)}{P(X\neg Y)}$$

$$odds(Y|\neg X) = \frac{P(\neg XY)/P(\neg X)}{P(\neg X\neg Y)/P(\neg X)} = \frac{P(\neg XY)}{P(\neg X\neg Y)}$$

Table 12.8. Contingency table for X and Y

	Y	$\neg Y$	
X	$sup(XY)$	$sup(X\neg Y)$	$sup(X)$
$\neg X$	$sup(\neg XY)$	$sup(\neg X\neg Y)$	$sup(\neg X)$
	$sup(Y)$	$sup(\neg Y)$	$\lvert\mathbf{D}\rvert$

Table 12.9. Rule conviction

Rule			*rsup*	*conf*	*lift*	*conv*
A	$\longrightarrow$	DE	0.50	0.75	1.50	2.00
DE	$\longrightarrow$	A	0.50	1.00	1.50	∞
E	$\longrightarrow$	C	0.50	0.60	0.90	0.83
C	$\longrightarrow$	E	0.50	0.75	0.90	0.68

The odds ratio is then defined as the ratio of these two odds:

$$\begin{aligned} oddsratio(X \longrightarrow Y) &= \frac{odds(Y|X)}{odds(Y|\neg X)} = \frac{P(XY)\cdot P(\neg X\neg Y)}{P(X\neg Y)\cdot P(\neg XY)} \\ &= \frac{sup(XY)\cdot sup(\neg X\neg Y)}{sup(X\neg Y)\cdot sup(\neg XY)} \end{aligned}$$

The odds ratio is a symmetric measure, and if X and Y are independent, then it has value 1. Thus, values close to 1 may indicate that there is little dependence between X and Y. Odds ratios greater than 1 imply higher odds of Y occurring in the presence of X as opposed to its complement $\neg X$, whereas odds smaller than one imply higher odds of Y occurring with $\neg X$.

Example 12.8. Let us compare the odds ratio for two rules, $C \longrightarrow A$ and $D \longrightarrow A$, using the example data in Table 12.1. The contingency tables for A and C, and for A and D, are given below:

	C	$\neg C$
A	2	2
$\neg A$	2	0

	D	$\neg D$
A	3	1
$\neg A$	1	1

The odds ratio values for the two rules are given as

$$oddsratio(C \longrightarrow A) = \frac{sup(AC)\cdot sup(\neg A\neg C)}{sup(A\neg C)\cdot sup(\neg AC)} = \frac{2\times 0}{2\times 2} = 0$$

$$oddsratio(D \longrightarrow A) = \frac{sup(AD)\cdot sup(\neg A\neg D)}{sup(A\neg D)\cdot sup(\neg AD)} = \frac{3\times 1}{1\times 1} = 3$$

Thus, $D \longrightarrow A$ is a stronger rule than $C \longrightarrow A$, which is also indicated by looking at other measures like lift and confidence:

$$conf(C \longrightarrow A) = 2/4 = 0.5 \qquad conf(D \longrightarrow A) = 3/4 = 0.75$$

$$lift(C \longrightarrow A) = \frac{2/6}{4/6 \times 4/6} = 0.75 \qquad lift(D \longrightarrow A) = \frac{3/6}{4/6 \times 4/6} = 1.125$$

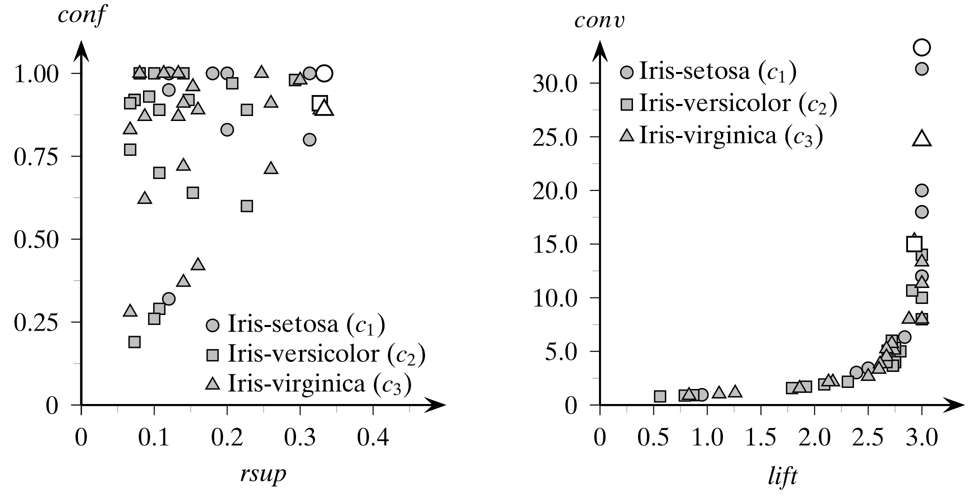
$C \longrightarrow A$ has less confidence and lift than $D \longrightarrow A$.

Example 12.9. We apply the different rule assessment measures on the Iris dataset, which has $n = 150$ examples, over one categorical attribute (`class`), and four numeric attributes (`sepal length`, `sepal width`, `petal length`, and `petal width`). To generate association rules we first discretize the numeric attributes as shown in Table 12.10. In particular, we want to determine representative class-specific rules that characterize each of the three Iris classes: `iris setosa`, `iris virginica` and `iris versicolor`, that is, we generate rules of the form $X \longrightarrow y$, where X is an itemset over the discretized numeric attributes, and y is a single item representing one of the Iris classes.

We start by generating all class-specific association rules using $minsup = 10$ and a minimum lift value of 0.1, which results in a total of 79 rules. Figure 12.1a plots the relative support and confidence of these 79 rules, with the three classes represented by different symbols. To look for the most surprising rules, we also plot in Figure 12.1b the lift and conviction value for the same 79 rules. For each class we select the most specific (i.e., with maximal antecedent) rule with the highest relative support and then confidence, and also those with the highest conviction and then lift. The selected rules are listed in Table 12.11 and Table 12.12, respectively. They are also highlighted in Figure 12.1 (as larger white symbols). Compared to the top rules for support and confidence, we observe that the best rule for c_1 is the same, but the rules for c_2 and c_3 are not the same, suggesting a trade-off between support and novelty among these rules.

Table 12.10. Iris dataset discretization and labels employed

Attribute	Range or value	Label
Sepal length	4.30–5.55	sl_1
	5.55–6.15	sl_2
	6.15–7.90	sl_3
Sepal width	2.00–2.95	sw_1
	2.95–3.35	sw_2
	3.35–4.40	sw_3
Petal length	1.00–2.45	pl_1
	2.45–4.75	pl_2
	4.75–6.90	pl_3
Petal width	0.10–0.80	pw_1
	0.80–1.75	pw_2
	1.75–2.50	pw_3
Class	Iris-setosa	c_1
	Iris-versicolor	c_2
	Iris-virginica	c_3



(a) Support vs. confidence

(b) Lift vs. conviction

Figure 12.1. Iris: support vs. confidence, and conviction vs. lift for class-specific rules. The best rule for each class is shown in white.

Table 12.11. Iris: best class-specific rules according to support and confidence

Rule	*rsup*	*conf*	*lift*	*conv*
$\{pl_1, pw_1\} \longrightarrow c_1$	0.333	1.00	3.00	33.33
$pw_2 \longrightarrow c_2$	0.327	0.91	2.72	6.00
$pl_3 \longrightarrow c_3$	0.327	0.89	2.67	5.24

Table 12.12. Iris: best class-specific rules according to lift and conviction

Rule	*rsup*	*conf*	*lift*	*conv*
$\{pl_1, pw_1\} \longrightarrow c_1$	0.33	1.00	3.00	33.33
$\{pl_2, pw_2\} \longrightarrow c_2$	0.29	0.98	2.93	15.00
$\{sl_3, pl_3, pw_3\} \longrightarrow c_3$	0.25	1.00	3.00	24.67

12.1.2 Pattern Assessment Measures

We now turn our focus on measures for pattern assessment.

Support

The most basic measures are support and relative support, giving the number and fraction of transactions in $\mathbf{D}$ that contain the itemset X:

$$sup(X) = |\mathbf{t}(X)| \qquad\qquad rsup(X) = \frac{sup(X)}{|\mathbf{D}|}$$

Lift

The *lift* of a k-itemset $X=\{x_1,x_2,\ldots,x_k\}$ in dataset $\mathbf{D}$ is defined as

$$lift(X,\mathbf{D})=\frac{P(X)}{\prod_{i=1}^{k}P(x_i)}=\frac{rsup(X)}{\prod_{i=1}^{k}rsup(x_i)} \tag{12.2}$$

that is, the ratio of the observed joint probability of items in X to the expected joint probability if all the items $x_i\in X$ were independent.

We may further generalize the notion of lift of an itemset X by considering all the different ways of partitioning it into nonempty and disjoint subsets. For instance, assume that the set $\{X_1,X_2,\ldots,X_q\}$ is a q-partition of X, i.e., a partitioning of X into q nonempty and disjoint itemsets X_i, such that $X_i\cap X_j=\emptyset$ and $\cup_i X_i=X$. Define the generalized lift of X over partitions of size q as follows:

$$lift_q(X)=\min_{X_1,\ldots,X_q}\left\{\frac{P(X)}{\prod_{i=1}^{q}P(X_i)}\right\}$$

This is, the least value of lift over all q-partitions X. Viewed in this light, $lift(X)=lift_k(X)$, that is, lift is the value obtained from the unique k-partition of X.

Rule-based Measures

Given an itemset X, we can evaluate it using rule assessment measures by considering all possible rules that can be generated from X. Let Θ be some rule assessment measure. We generate all possible rules from X of the form $X_1\longrightarrow X_2$ and $X_2\longrightarrow X_1$, where the set $\{X_1,X_2\}$ is a 2-partition, or a bipartition, of X. We then compute the measure Θ for each such rule, and use summary statistics such as the mean, maximum, and minimum to characterize X. If Θ is a symmetric measure, then $\Theta(X_1\longrightarrow X_2)=\Theta(X_2\longrightarrow X_1)$, and we have to consider only half of the rules. For example, if Θ is rule lift, then we can define the average, maximum, and minimum lift values for X as follows:

$$AvgLift(X)=\operatorname*{avg}_{X_1,X_2}\left\{lift(X_1\longrightarrow X_2)\right\}$$

$$MaxLift(X)=\max_{X_1,X_2}\left\{lift(X_1\longrightarrow X_2)\right\}$$

$$MinLift(X)=\min_{X_1,X_2}\left\{lift(X_1\longrightarrow X_2)\right\}$$

We can also do the same for other rule measures such as leverage, confidence, and so on. In particular, when we use rule lift, then $MinLift(X)$ is identical to the generalized lift $lift_2(X)$ over all 2-partitions of X.

Example 12.10. Consider the itemset $X=\{pl_2,pw_2,c_2\}$, whose support in the discretized Iris dataset is shown in Table 12.13, along with the supports for all of its subsets. Note that the size of the database is $|\mathbf{D}|=n=150$.

Using Eq. (12.2), the lift of X is given as

$$lift(X)=\frac{rsup(X)}{rsup(pl_2)\cdot rsup(pw_2)\cdot rsup(c_2)}=\frac{0.293}{0.3\cdot 0.36\cdot 0.333}=8.16$$

Table 12.13. Support values for $\{pl_2, pw_2, c_2\}$ and its subsets

Itemset	*sup*	*rsup*
$\{pl_2, pw_2, c_2\}$	44	0.293
$\{pl_2, pw_2\}$	45	0.300
$\{pl_2, c_2\}$	44	0.293
$\{pw_2, c_2\}$	49	0.327
$\{pl_2\}$	45	0.300
$\{pw_2\}$	54	0.360
$\{c_2\}$	50	0.333

Table 12.14. Rules generated from itemset $\{pl_2, pw_2, c_2\}$

Bipartition	Rule	*lift*	*leverage*	*conf*
$\big\{\{pl_2\}, \{pw_2, c_2\}\big\}$	$pl_2 \longrightarrow \{pw_2, c_2\}$	2.993	0.195	0.978
	$\{pw_2, c_2\} \longrightarrow pl_2$	2.993	0.195	0.898
$\big\{\{pw_2\}, \{pl_2, c_2\}\big\}$	$pw_2 \longrightarrow \{pl_2, c_2\}$	2.778	0.188	0.815
	$\{pl_2, c_2\} \longrightarrow pw_2$	2.778	0.188	1.000
$\big\{\{c_2\}, \{pl_2, pw_2\}\big\}$	$c_2 \longrightarrow \{pl_2, pw_2\}$	2.933	0.193	0.880
	$\{pl_2, pw_2\} \longrightarrow c_2$	2.933	0.193	0.978

Table 12.14 shows all the possible rules that can be generated from X, along with the rule lift and leverage values. Note that because both of these measures are symmetric, we need to consider only the distinct bipartitions of which there are three, as shown in the table. The maximum, minimum, and average lift values are as follows:

$$MaxLift(X) = \max\{2.993, 2.778, 2.933\} = 2.998$$

$$MinLift(X) = \min\{2.993, 2.778, 2.933\} = 2.778$$

$$AvgLift(X) = \text{avg}\{2.993, 2.778, 2.933\} = 2.901$$

We may use other measures too. For example, the average leverage of X is given as

$$AvgLeverage(X) = \text{avg}\{0.195, 0.188, 0.193\} = 0.192$$

However, because confidence is not a symmetric measure, we have to consider all the six rules and their confidence values, as shown in Table 12.14. The average confidence for X is

$$AvgConf(X) = \text{avg}\{0.978, 0.898, 0.815, 1.0, 0.88, 0.978\} = 5.549/6 = 0.925$$

Example 12.11. Consider all frequent itemsets in the discretized Iris dataset from Example 12.9, using *minsup* $= 1$. We analyze the set of all possible rules that can be generated from these frequent itemsets. Figure 12.2 plots the relative support and average lift values for all the 306 frequent patterns with size at least 2 (since nontrivial

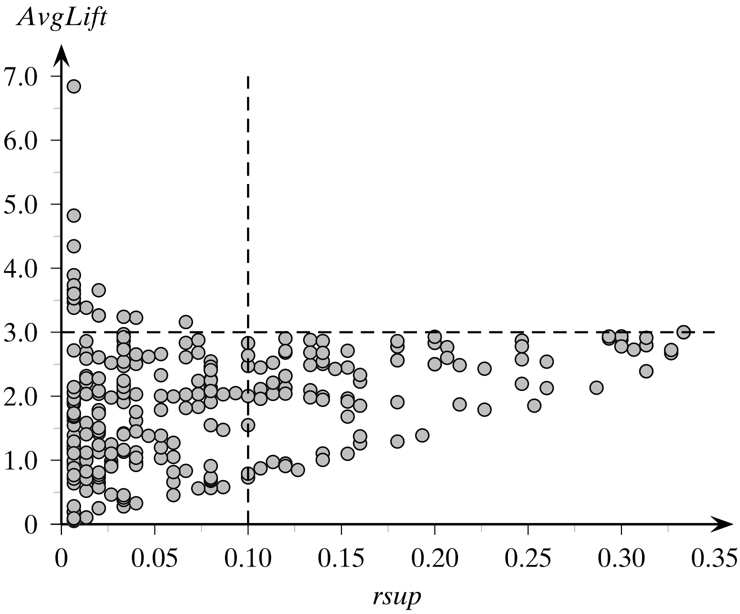


Figure 12.2. Iris: support and average lift of patterns assessed.

rules can only be generated from itemsets of size 2 or more). We can see that with the exception of low support itemsets, the average lift value is bounded above by 3.0. From among these we may select those patterns with the highest support for further analysis. For instance, the itemset $X = \{pl_1, pw_1, c_1\}$ is a maximal itemset with support $rsup(X) = 0.33$, all of whose subsets also have support $rsup = 0.33$. Thus, all of the rules that can be derived from it have a lift of 3.0, and the minimum lift of X is 3.0.

12.1.3 Comparing Multiple Rules and Patterns

We now turn our attention to comparing different rules and patterns. In general, the number of frequent itemsets and association rules can be very large and many of them may not be very relevant. We highlight cases when certain patterns and rules can be pruned, as the information contained in them may be subsumed by other more relevant ones.

Comparing Itemsets

When comparing multiple itemsets we may choose to focus on the maximal itemsets that satisfy some property, or we may consider closed itemsets that capture all of the support information. We consider these and other measures in the following paragraphs.

Maximal Itemsets An frequent itemset X is *maximal* if all of its supersets are not frequent, that is, X is maximal iff

$$sup(X) \geq minsup, \text{ and for all } Y \supset X, sup(Y) < minsup$$

Table 12.15. Iris: maximal patterns according to average lift

Pattern	Avg. lift
$\{sl_1, sw_2, pl_1, pw_1, c_1\}$	2.90
$\{sl_1, sw_3, pl_1, pw_1, c_1\}$	2.86
$\{sl_2, sw_1, pl_2, pw_2, c_2\}$	2.83
$\{sl_3, sw_2, pl_3, pw_3, c_3\}$	2.88
$\{sw_1, pl_3, pw_3, c_3\}$	2.52

Given a collection of frequent itemsets, we may choose to retain only the maximal ones, especially among those that already satisfy some other constraints on pattern assessment measures like lift or leverage.

Example 12.12. Consider the discretized Iris dataset from Example 12.9. To gain insights into the maximal itemsets that pertain to each of the Iris classes, we focus our attention on the class-specific itemsets, that is, those itemsets X that contain a class as one of the items. From the itemsets plotted in Figure 12.2, using $minsup(X) \geq 15$ (which corresponds to a relative support of 10%) and retaining only those itemsets with an average lift value of at least 2.5, we retain 37 class-specific itemsets. Among these, the maximal class-specific itemsets are shown in Table 12.15, which highlight the features that characterize each of the three classes. For instance, for class c_1 (`Iris-setosa`), the essential items are sl_1, pl_1, pw_1 and either sw_2 or sw_3. Looking at the range values in Table 12.10, we conclude that `Iris-setosa` class is characterized by `sepal-length` in the range $sl_1 = [4.30, 5.55]$, `petal-length` in the range $pl_1 = [1, 2.45]$, and so on. A similar interpretation can be carried out for the other two Iris classes.

Closed Itemsets and Minimal Generators An itemset X is *closed* if all of its supersets have strictly less support, that is,

$$sup(X) > sup(Y), \text{ for all } Y \supset X$$

An itemset X is a *minimal generator* if all its subsets have strictly higher support, that is,

$$sup(X) < sup(Y), \text{ for all } Y \subset X$$

If an itemset X is not a minimal generator, then it implies that it has some redundant items, that is, we can find some subset $Y \subset X$, which can be replaced with an even smaller subset $W \subset Y$ without changing the support of X, that is, there exists a $W \subset Y$, such that

$$sup(X) = sup(Y \cup (X \setminus Y)) = sup(W \cup (X \setminus Y))$$

One can show that all subsets of a minimal generator must themselves be minimal generators.

Table 12.16. Closed itemsets and minimal generators

sup	Closed Itemset	Minimal Generators
3	$ABDE$	AD, DE
3	BCE	CE
4	ABE	A
4	BC	C
4	BD	D
5	BE	E
6	B	B

Example 12.13. Consider the dataset in Table 12.1 and the set of frequent itemsets with $minsup = 3$ as shown in Table 12.2. There are only two maximal frequent itemsets, namely $ABDE$ and BCE, which capture essential information about whether another itemset is frequent or not: an itemset is frequent only if it is a subset of one of these two.

Table 12.16 shows the seven closed itemsets and the corresponding minimal generators. Both of these sets allow one to infer the exact support of any other frequent itemset. The support of an itemset X is the maximum support among all closed itemsets that contain it. Alternatively, the support of X is the minimum support among all minimal generators that are subsets of X. For example, the itemset AE is a subset of the closed sets ABE and $ABDE$, and it is a superset of the minimal generators A, and E; we can observe that

$$sup(AE) = \max\{sup(ABE), sup(ABDE)\} = 4$$

$$sup(AE) = \min\{sup(A), sup(E)\} = 4$$

Productive Itemsets An itemset X is *productive* if its relative support is higher than the expected relative support over all of its bipartitions, assuming they are independent. More formally, let $|X| \geq 2$, and let $\{X_1, X_2\}$ be a bipartition of X. We say that X is productive provided

$$rsup(X) > rsup(X_1) \times rsup(X_2), \text{ for all bipartitions } \{X_1, X_2\} \text{ of } X \tag{12.3}$$

This immediately implies that X is productive if its minimum lift is greater than one, as

$$MinLift(X) = \min_{X_1, X_2} \left\{ \frac{rsup(X)}{rsup(X_1) \cdot rsup(X_2)} \right\} > 1$$

In terms of leverage, X is productive if its minimum leverage is above zero because

$$MinLeverage(X) = \min_{X_1, X_2} \left\{ rsup(X) - rsup(X_1) \times rsup(X_2) \right\} > 0$$

Example 12.14. Considering the frequent itemsets in Table 12.2, the set $ABDE$ is not productive because there exists a bipartition with lift value of 1. For instance, for its bipartition $\{B, ADE\}$ we have

$$lift(B \longrightarrow ADE) = \frac{rsup(ABDE)}{rsup(B) \cdot rsup(ADE)} = \frac{3/6}{6/6 \cdot 3/6} = 1$$

On the other hand, ADE is productive because it has three distinct bipartitions and all of them have lift above 1:

$$lift(A \longrightarrow DE) = \frac{rsup(ADE)}{rsup(A) \cdot rsup(DE)} = \frac{3/6}{4/6 \cdot 3/6} = 1.5$$

$$lift(D \longrightarrow AE) = \frac{rsup(ADE)}{rsup(D) \cdot rsup(AE)} = \frac{3/6}{4/6 \cdot 4/6} = 1.125$$

$$lift(E \longrightarrow AD) = \frac{rsup(ADE)}{rsup(E) \cdot rsup(AD)} = \frac{3/6}{5/6 \cdot 3/6} = 1.2$$

Comparing Rules

Given two rules $R : X \longrightarrow Y$ and $R' : W \longrightarrow Y$ that have the same consequent, we say that R is *more specific* than R', or equivalently, that R' is *more general* than R provided $W \subset X$.

Nonredundant Rules We say that a rule $R : X \longrightarrow Y$ is *redundant* provided there exists a more general rule $R' : W \longrightarrow Y$ that has the same support, that is, $W \subset X$ and $sup(R) = sup(R')$. On the other hand, if $sup(R) < sup(R')$ over all its generalizations R', then R is *nonredundant*.

Improvement and Productive Rules Define the *improvement* of a rule $X \longrightarrow Y$ as follows:

$$imp(X \longrightarrow Y) = conf(X \longrightarrow Y) - \max_{W \subset X}\Big\{conf(W \longrightarrow Y)\Big\}$$

Improvement quantifies the minimum difference between the confidence of a rule and any of its generalizations. A rule $R : X \longrightarrow Y$ is *productive* if its improvement is greater than zero, which implies that for all more general rules $R' : W \longrightarrow Y$ we have $conf(R) > conf(R')$. On the other hand, if there exists a more general rule R' with $conf(R') \geq conf(R)$, then R is *unproductive*. If a rule is redundant, it is also unproductive because its improvement is zero.

The smaller the improvement of a rule $R : X \longrightarrow Y$, the more likely it is to be unproductive. We can generalize this notion to consider rules that have at least some minimum level of improvement, that is, we may require that $imp(X \longrightarrow Y) \geq t$, where t is a user-specified minimum improvement threshold.

Example 12.15. Consider the example dataset in Table 12.1, and the set of frequent itemsets in Table 12.2. Consider rule $R : BE \longrightarrow C$, which has support 3, and confidence $3/5 = 0.60$. It has two generalizations, namely

$$R'_1 : E \longrightarrow C, \quad sup = 3, conf = 3/5 = 0.6$$
$$R'_2 : B \longrightarrow C, \quad sup = 4, conf = 4/6 = 0.67$$

Thus, $BE \longrightarrow C$ is redundant w.r.t. $E \longrightarrow C$ because they have the same support, that is, $sup(BCE) = sup(BC)$. Further, $BE \longrightarrow C$ is also unproductive, since $imp(BE \longrightarrow C) = 0.6 - \max\{0.6, 0.67\} = -0.07$; it has a more general rule, namely R'_2, with higher confidence.

12.2 SIGNIFICANCE TESTING AND CONFIDENCE INTERVALS

We now consider how to assess the statistical significance of patterns and rules, and how to derive confidence intervals for a given assessment measure.

12.2.1 Fisher Exact Test for Productive Rules

We begin by discussing the Fisher exact test for rule improvement. That is, we directly test whether the rule $R : X \longrightarrow Y$ is productive by comparing its confidence with that of each of its generalizations $R' : W \longrightarrow Y$, including the default or trivial rule $\emptyset \longrightarrow Y$.

Let $R : X \longrightarrow Y$ be an association rule. Consider its generalization $R' : W \longrightarrow Y$, where $W = X \setminus Z$ is the new antecedent formed by removing from X the subset $Z \subseteq X$. Given an input dataset $\mathbf{D}$, conditional on the fact that W occurs, we can create a 2×2 contingency table between Z and the consequent Y as shown in Table 12.17. The different cell values are as follows:

$$a = sup(WZY) = sup(XY) \qquad b = sup(WZ\neg Y) = sup(X\neg Y)$$
$$c = sup(W\neg ZY) \qquad d = sup(W\neg Z\neg Y)$$

Here, a denotes the number of transactions that contain both X and Y, b denotes the number of transactions that contain X but not Y, c denotes the number of transactions that contain W and Y but not Z, and finally d denotes the number of transactions that contain W but neither Z nor Y. The marginal counts are given as

$$\text{row marginals: } a + b = sup(WZ) = sup(X), \quad c + d = sup(W\neg Z)$$
$$\text{column marginals: } a + c = sup(WY), \quad b + d = sup(W\neg Y)$$

where the row marginals give the occurrence frequency of W with and without Z, and the column marginals specify the occurrence counts of W with and without Y. Finally, we can observe that the sum of all the cells is simply $n = a + b + c + d = sup(W)$. Notice that when $Z = X$, we have $W = \emptyset$, and the contingency table defaults to the one shown in Table 12.8.

Given a contingency table conditional on W, we are interested in the odds ratio obtained by comparing the presence and absence of Z, that is,

$$oddsratio = \frac{a/(a+b)}{b/(a+b)} \bigg/ \frac{c/(c+d)}{d/(c+d)} = \frac{ad}{bc} \tag{12.4}$$

Table 12.17. Contingency table for Z and Y, conditional on $W = X \setminus Z$

W	Y	$\neg Y$	
Z	a	b	$a+b$
$\neg Z$	c	d	$c+d$
	$a+c$	$b+d$	$n = sup(W)$

Recall that the odds ratio measures the odds of X, that is, W and Z, occurring with Y versus the odds of its subset W, but not Z, occurring with Y. Under the null hypothesis H_0 that Z and Y are independent given W the odds ratio is 1. To see this, note that under the independence assumption the count in a cell of the contingency table is equal to the product of the corresponding row and column marginal counts divided by n, that is, under H_0:

$$a = (a+b)(a+c)/n \qquad b = (a+b)(b+d)/n$$

$$c = (c+d)(a+c)/n \qquad d = (c+d)(b+d)/n$$

Plugging these values in Eq. (12.4), we obtain

$$oddsratio = \frac{ad}{bc} = \frac{(a+b)(c+d)(b+d)(a+c)}{(a+c)(b+d)(a+b)(c+d)} = 1$$

The null hypothesis therefore corresponds to $H_0 : oddsratio = 1$, and the alternative hypothesis is $H_a : oddsratio > 1$. Under the null hypothesis, if we further assume that the row and column marginals are fixed, then a uniquely determines the other three values b, c, and d, and the probability mass function of observing the value a in the contingency table is given by the hypergeometric distribution. Recall that the hypergeometric distribution gives the probability of choosing s successes in t trails if we sample *without replacement* from a finite population of size T that has S successes in total, given as

$$P(s \mid t, S, T) = \binom{S}{s} \cdot \binom{T-S}{t-s} \bigg/ \binom{T}{t}$$

In our context, we take the occurrence of Z as a success. The population size is $T = sup(W) = n$ because we assume that W always occurs, and the total number of successes is the support of Z given W, that is, $S = a+b$. In $t = a+c$ trials, the hypergeometric distribution gives the probability of $s = a$ successes:

$$\begin{aligned} P\Big(a \Big| \; (a+c), (a+b), n\Big) &= \frac{\binom{a+b}{a} \cdot \binom{n-(a+b)}{(a+c)-a}}{\binom{n}{a+c}} = \frac{\binom{a+b}{a} \cdot \binom{c+d}{c}}{\binom{n}{a+c}} \\ &= \frac{(a+b)!\,(c+d)!}{a!\,b!\,c!\,d!} \bigg/ \frac{n!}{(a+c)!\,(n-(a+c))!} \\ &= \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!} \end{aligned} \tag{12.5}$$

Table 12.18. Contingency table: increase a by i

W	Y	$\neg Y$	
Z	$a+i$	$b-i$	$a+b$
$\neg Z$	$c-i$	$d+i$	$c+d$
	$a+c$	$b+d$	$n = sup(W)$

Our aim is to contrast the null hypothesis H_0 that $oddsratio = 1$ with the alternative hypothesis H_a that $oddsratio > 1$. Because a determines the rest of the cells under fixed row and column marginals, we can see from Eq. (12.4) that the larger the a the larger the odds ratio, and consequently the greater the evidence for H_a. We can obtain the *p-value* for a contingency table as extreme as that in Table 12.17 by summing Eq. (12.5) over all possible values a or larger:

$$p\text{-}value(a) = \sum_{i=0}^{\min(b,c)} P(a+i \mid (a+c), (a+b), n)$$

$$= \sum_{i=0}^{\min(b,c)} \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,(a+i)!\,(b-i)!\,(c-i)!\,(d+i)!}$$

which follows from the fact that when we increase the count of a by i, then because the row and column marginals are fixed, b and c must decrease by i, and d must increase by i, as shown in Table 12.18. The lower the *p-value* the stronger the evidence that the odds ratio is greater than one, and thus, we may reject the null hypothesis H_0 if $p\text{-}value \leq \alpha$, where α is the significance threshold (e.g., $\alpha = 0.01$). This test is known as the *Fisher Exact Test.*

In summary, to check whether a rule $R: X \longrightarrow Y$ is productive, we must compute $p\text{-}value(a) = p\text{-}value(sup(XY))$ of the contingency tables obtained from each of its generalizations $R' : W \longrightarrow Y$, where $W = X \setminus Z$, for $Z \subseteq X$. If $p\text{-}value(sup(XY)) > \alpha$ for any of these comparisons, then we can reject the rule $R : X \longrightarrow Y$ as nonproductive. On the other hand, if $p\text{-}value(sup(XY)) \leq \alpha$ for all the generalizations, then R is productive. However, note that if $|X| = k$, then there are $2^k - 1$ possible generalizations; to avoid this exponential complexity for large antecedents, we typically restrict our attention to only the immediate generalizations of the form $R' : X \setminus z \longrightarrow Y$, where $z \in X$ is one of the attribute values in the antecedent. However, we do include the trivial rule $\emptyset \longrightarrow Y$ because the conditional probability $P(Y|X) = conf(X \longrightarrow Y)$ should also be higher than the prior probability $P(Y) = conf(\emptyset \longrightarrow Y)$.

Example 12.16. Consider the rule $R : pw_2 \longrightarrow c_2$ obtained from the discretized Iris dataset. To test if it is productive, because there is only a single item in the antecedent, we compare it only with the default rule $\emptyset \longrightarrow c_2$. Using Table 12.17, the various cell values are

$$a = sup(pw_2, c_2) = 49 \qquad b = sup(pw_2, \neg c_2) = 5$$

$$c = sup(\neg pw_2, c_2) = 1 \qquad d = sup(\neg pw_2, \neg c_2) = 95$$

with the contingency table given as

	c_2	$\neg c_2$	
pw_2	49	5	54
$\neg pw_2$	1	95	96
	50	100	150

Thus the *p-value* is given as

$$\begin{aligned} p\text{-}value &= \sum_{i=0}^{\min(b,c)} P(a+i \mid (a+c),(a+b),n) \\ &= P(49 \mid 50,54,150) + P(50 \mid 50,54,150) \\ &= \binom{54}{49}\cdot\binom{96}{95}\Big/\binom{150}{50} + \binom{54}{50}\cdot\binom{96}{96}\Big/\binom{150}{50} \\ &= 1.51\times10^{-32} + 1.57\times10^{-35} = 1.51\times10^{-32} \end{aligned}$$

Since the *p-value* is extremely small, we can safely reject the null hypothesis that the odds ratio is 1. Instead, there is a strong relationship between $X = pw_2$ and $Y = c_2$, and we conclude that $R: pw_2 \longrightarrow c_2$ is a productive rule.

Example 12.17. Consider another rule $\{sw_1, pw_2\} \longrightarrow c_2$, with $X = \{sw_1, pw_2\}$ and $Y = c_2$. Consider its three generalizations, and the corresponding contingency tables and *p-values*:

$R'_1 : pw_2 \longrightarrow c_2$

$Z = \{sw_1\}$
$W = X \setminus Z = \{pw_2\}$
$p\text{-}value = 0.84$

$W = pw_2$	c_2	$\neg c_2$	
sw_1	34	4	38
$\neg sw_1$	15	1	16
	49	5	54

$R'_2 : sw_1 \longrightarrow c_2$

$Z = \{pw_2\}$
$W = X \setminus Z = \{sw_1\}$
$p\text{-}value = 1.39 \times 10^{-11}$

$W = sw_1$	c_2	$\neg c_2$	
pw_2	34	4	38
$\neg pw_2$	0	19	19
	34	23	57

$R'_3 : \emptyset \longrightarrow c_2$

$Z = \{sw_1, pw_2\}$
$W = X \setminus Z = \emptyset$
$p\text{-}value = 3.55 \times 10^{-17}$

$W = \emptyset$	c_2	$\neg c_2$	
$\{sw_1, pw_2\}$	34	4	38
$\neg\{sw_1, pw_2\}$	16	96	112
	50	100	150

We can see that whereas the *p-value* with respect to R_2' and R_3' is small, for R_1' we have *p-value* $= 0.84$, which is too high and thus we cannot reject the null hypothesis. We conclude that $R: \{sw_1, pw_2\} \longrightarrow c_2$ is not productive. In fact, its generalization R_1' is the one that is productive, as shown in Example 12.16.

Multiple Hypothesis Testing

Given an input dataset $\mathbf{D}$, there can be an exponentially large number of rules that need to be tested to check whether they are productive or not. We thus run into the multiple hypothesis testing problem, that is, just by the sheer number of hypothesis tests some unproductive rules will pass the $p\text{-}value \leq \alpha$ threshold by random chance. A strategy for overcoming this problem is to use the *Bonferroni correction* of the significance level that explicitly takes into account the number of experiments performed during the hypothesis testing process. Instead of using the given α threshold, we should use an adjusted threshold $\alpha' = \frac{\alpha}{\#r}$, where $\#r$ is the number of rules to be tested or its estimate. This correction ensures that the rule false discovery rate is bounded by α, where a false discovery is to claim that a rule is productive when it is not.

Example 12.18. Consider the discretized Iris dataset, using the discretization shown in Table 12.10. Let us focus only on class-specific rules, that is, rules of the form $X \to c_i$. Since each example can take on only one value at a time for a given attribute, the maximum antecedent length is four, and the maximum number of class-specific rules that can be generated from the Iris dataset is given as

$$\#r = c \times \left(\sum_{i=1}^{4} \binom{4}{i} b^i \right)$$

where c is the number of Iris classes, and b is the maximum number of bins for any other attribute. The summation is over the antecedent size i, that is, the number of attributes to be used in the antecedent. Finally, there are b^i possible combinations for the chosen set of i attributes. Because there are three Iris classes, and because each attribute has three bins, we have $c = 3$ and $b = 3$, and the number of possible rules is

$$\#r = 3 \times \left(\sum_{i=1}^{4} \binom{4}{i} 3^i \right) = 3(12 + 54 + 108 + 81) = 3 \cdot 255 = 765$$

Thus, if the input significance level is $\alpha = 0.01$, then the adjusted significance level using the Bonferroni correction is $\alpha' = \alpha/\#r = 0.01/765 = 1.31 \times 10^{-5}$. The rule $pw_2 \longrightarrow c_2$ in Example 12.16 has *p-value* $= 1.51 \times 10^{-32}$, and thus it remains productive even when we use α'.

12.2.2 Permutation Test for Significance

A *permutation* or *randomization* test determines the distribution of a given test statistic Θ by randomly modifying the observed data several times to obtain a random sample of datasets, which can in turn be used for significance testing. In the context of pattern assessment, given an input dataset $\mathbf{D}$, we first generate k randomly permuted datasets $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_k$. We can then perform different types of significance tests. For instance, given a pattern or rule we can check whether it is statistically significant by first computing the empirical probability mass function (EPMF) for the test statistic Θ by computing its value θ_i in the ith randomized dataset $\mathbf{D}_i$ for all $i \in [1,k]$. From these values we can generate the empirical cumulative distribution function

$$\hat{F}(x) = \hat{P}\,(\Theta \le x) = \frac{1}{k}\sum_{i=1}^{k} I(\theta_i \le x)$$

where I is an indicator variable that takes on the value 1 when its argument is true, and is 0 otherwise. Let θ be the value of the test statistic in the input dataset $\mathbf{D}$, then $p\text{-}value(\theta)$, that is, the probability of obtaining a value as high as θ by random chance can be computed as

$$p\text{-}value(\theta) = 1 - F(\theta)$$

Given a significance level α, if $p\text{-}value(\theta) > \alpha$, then we accept the null hypothesis that the pattern/rule is not statistically significant. On the other hand, if $p\text{-}value(\theta) \le \alpha$, then we can reject the null hypothesis and conclude that the pattern is significant because a value as high as θ is highly improbable. The permutation test approach can also be used to assess an entire set of rules or patterns. For instance, we may test a collection of frequent itemsets by comparing the number of frequent itemsets in $\mathbf{D}$ with the distribution of the number of frequent itemsets empirically derived from the permuted datasets $\mathbf{D}_i$. We may also do this analysis as a function of *minsup*, and so on.

Swap Randomization

A key question in generating the permuted datasets $\mathbf{D}_i$ is which characteristics of the input dataset $\mathbf{D}$ we should preserve. The *swap randomization* approach maintains as invariant the column and row margins for a given dataset, that is, the permuted datasets preserve the support of each item (the column margin) as well as the number of items in each transaction (the row margin). Given a dataset $\mathbf{D}$, we randomly create k datasets that have the same row and column margins. We then mine frequent patterns in $\mathbf{D}$ and check whether the pattern statistics are different from those obtained using the randomized datasets. If the differences are not significant, we may conclude that the patterns arise solely from the row and column margins, and not from any interesting properties of the data.

Given a binary matrix $\mathbf{D} \subseteq \mathcal{T} \times \mathcal{I}$, the swap randomization method exchanges two nonzero cells of the matrix via a *swap* that leaves the row and column margins unchanged. To illustrate how swap works, consider any two transactions $t_a, t_b \in \mathcal{T}$ and any two items $i_a, i_b \in \mathcal{I}$ such that $(t_a, i_a), (t_b, i_b) \in \mathbf{D}$ and $(t_a, i_b), (t_b, i_a) \notin \mathbf{D}$, which corresponds to the 2×2 submatrix in $\mathbf{D}$, given as

$$\mathbf{D}(t_a, i_a; t_b, i_b) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

ALGORITHM 12.1. Generate Swap Randomized Dataset

```
SwapRandomization($t$, $\mathbf{D} \subseteq \mathcal{T} \times \mathcal{I}$):
while $t > 0$ do
    Select pairs $(t_a, i_a), (t_b, i_b) \in \mathbf{D}$ randomly
    if $(t_a, i_b) \notin \mathbf{D}$ and $(t_b, i_a) \notin \mathbf{D}$ then
        $\mathbf{D} \leftarrow \mathbf{D} \setminus \{(t_a, i_a), (t_b, i_b)\} \cup \{(t_a, i_b), (t_b, i_a)\}$
    $t = t - 1$
return $\mathbf{D}$
```

After a swap operation we obtain the new submatrix

$$\mathbf{D}(t_a, i_b; t_b, i_a) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

where we exchange the elements in $\mathbf{D}$ so that $(t_a, i_b), (t_b, i_a) \in \mathbf{D}$, and $(t_a, i_a), (t_b, i_b) \notin \mathbf{D}$. We denote this operation as $Swap(t_a, i_a; t_b, i_b)$. Notice that a swap does not affect the row and column margins, and we can thus generate a permuted dataset with the same row and column sums as $\mathbf{D}$ through a sequence of swaps. Algorithm 12.1 shows the pseudo-code for generating a swap randomized dataset. The algorithm performs t swap trials by selecting two pairs (t_a, i_a), $(t_b, i_b) \in \mathbf{D}$ at random; a swap is successful only if both (t_a, i_b), $(t_b, i_a) \notin \mathbf{D}$.

Example 12.19. Consider the input binary dataset $\mathbf{D}$ shown in Table 12.19a, whose row and column sums are also shown. Table 12.19b shows the resulting dataset after a single swap operation $Swap(1, D; 4, C)$, highlighted by the gray cells. When we apply another swap, namely $Swap(2, C; 4, A)$, we obtain the data in Table 12.19c. We can observe that the marginal counts remain invariant.

From the input dataset $\mathbf{D}$ in Table 12.19a we generated $k = 100$ swap randomized datasets, each of which is obtained by performing 150 swaps (the product of all possible transaction pairs and item pairs, that is, $\binom{6}{2} \cdot \binom{5}{2} = 150$). Let the test statistic be the total number of frequent itemsets using $minsup = 3$. Mining $\mathbf{D}$ results in $|\mathcal{F}| = 19$ frequent itemsets. Likewise, mining each of the $k = 100$ permuted datasets results in the following empirical PMF for $|\mathcal{F}|$:

$$P\big(|\mathcal{F}| = 19\big) = 0.67 \qquad\qquad P\big(|\mathcal{F}| = 17\big) = 0.33$$

Because $p\text{-}value(19) = 0.67$, we may conclude that the set of frequent itemsets is essentially determined by the row and column marginals.

Focusing on a specific itemset, consider $ABDE$, which is one of the maximal frequent itemsets in $\mathbf{D}$, with $sup(ABDE) = 3$. The probability that $ABDE$ is frequent is $17/100 = 0.17$ because it is frequent in 17 of the 100 swapped datasets. As this probability is not very low, we may conclude that $ABDE$ is not a statistically significant pattern; it has a relatively high chance of being frequent in random datasets. Consider another itemset BCD that is not frequent in $\mathbf{D}$ because

$sup(BCD) = 2$. The empirical PMF for the support of BCD is given as

$$P(sup = 2) = 0.54 \qquad P(sup = 3) = 0.44 \qquad P(sup = 4) = 0.02$$

In a majority of the datasets BCD is infrequent, and if $minsup = 4$, then $p\text{-}value(sup = 4) = 0.02$ implies that BCD is highly unlikely to be a frequent pattern.

Table 12.19. Input data **D** and swap randomization

Tid	Items					Sum
	A	*B*	*C*	*D*	*E*	
1	1	1	0	1	1	4
2	0	1	1	0	1	3
3	1	1	0	1	1	4
4	1	1	1	0	1	4
5	1	1	1	1	1	5
6	0	1	1	1	0	3
Sum	4	6	4	4	5	

(a) Input binary data **D**

Tid	Items					Sum
	A	*B*	*C*	*D*	*E*	
1	1	1	1	0	1	4
2	0	1	1	0	1	3
3	1	1	0	1	1	4
4	1	1	0	1	1	4
5	1	1	1	1	1	5
6	0	1	1	1	0	3
Sum	4	6	4	4	5	

(b) $Swap(1, D; 4, C)$

Tid	Items					Sum
	A	*B*	*C*	*D*	*E*	
1	1	1	1	0	1	4
2	1	1	0	0	1	3
3	1	1	0	1	1	4
4	0	1	1	1	1	4
5	1	1	1	1	1	5
6	0	1	1	1	0	3
Sum	4	6	4	4	5	

(c) $Swap(2, C; 4, A)$

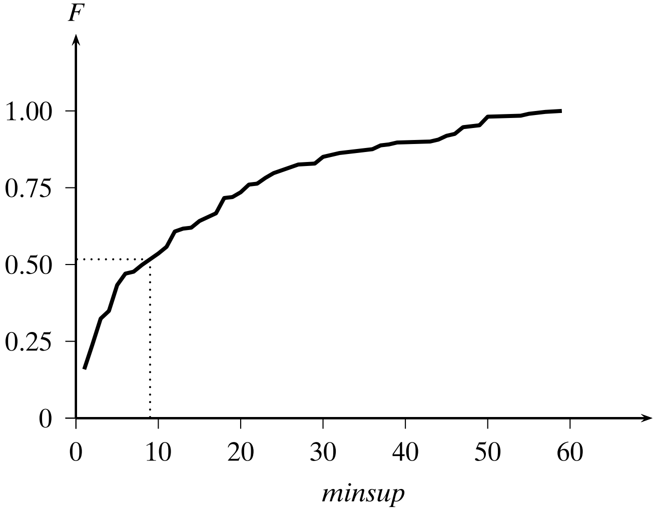


Figure 12.3. Cumulative distribution of the number of frequent itemsets as a function of minimum support.

Example 12.20. We apply the swap randomization approach to the discretized Iris dataset. Figure 12.3 shows the cumulative distribution of the number of frequent itemsets in **D** at various minimum support levels. We choose $minsup = 10$, for which we have $\hat{F}(10) = P(sup < 10) = 0.517$. Put differently, $P(sup \geq 10) = 1 - 0.517 = 0.483$, that is, 48.3% of the itemsets that occur at least once are frequent using $minsup = 10$.

Define the test statistic to be the *relative lift*, defined as the relative change in the lift value of itemset X when comparing the input dataset $\mathbf{D}$ and a randomized dataset $\mathbf{D}_i$, that is,

$$rlift(X, \mathbf{D}, \mathbf{D}_i) = \frac{lift(X, \mathbf{D}) - lift(X, \mathbf{D}_i)}{lift(X, \mathbf{D})}$$

For an m-itemset $X = \{x_1, \dots, x_m\}$, by Eq. (12.2) note that

$$lift(X, \mathbf{D}) = rsup(X, \mathbf{D}) \Big/ \prod_{j=1}^{m} rsup(x_j, \mathbf{D})$$

Because the swap randomization process leaves item supports (the column margins) intact, and does not change the number of transactions, we have $rsup(x_j, \mathbf{D}) = rsup(x_j, \mathbf{D}_i)$, and $|\mathbf{D}| = |\mathbf{D}_i|$. We can thus rewrite the relative lift statistic as

$$rlift(X, \mathbf{D}, \mathbf{D}_i) = \frac{sup(X, \mathbf{D}) - sup(X, \mathbf{D}_i)}{sup(X, \mathbf{D})} = 1 - \frac{sup(X, \mathbf{D}_i)}{sup(X, \mathbf{D})}$$

We generate $k = 100$ randomized datasets and compute the average relative lift for each of the 140 frequent itemsets of size two or more in the input dataset, as lift values are not defined for single items. Figure 12.4 shows the cumulative distribution for average relative lift, which ranges from -0.55 to 0.998. An average relative lift close to 1 means that the corresponding frequent pattern hardly ever occurs in any of the randomized datasets. On the other hand, a larger negative average relative lift value means that the support in randomized datasets is higher than in the input dataset. Finally, a value close to zero means that the support of the itemset is the same in both the original and randomized datasets; it is mainly a consequence of the marginal counts, and thus of little interest.

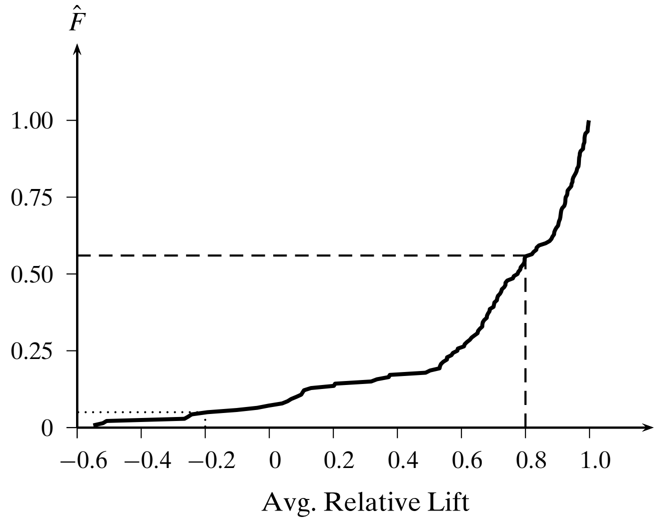


Figure 12.4. Cumulative distribution for average relative lift.

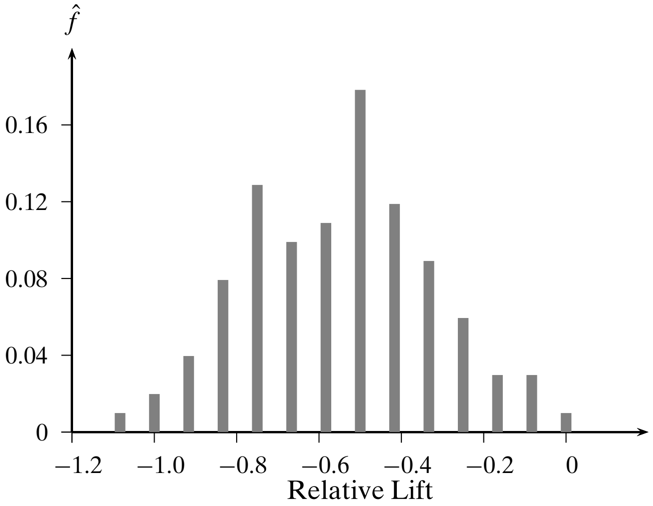


Figure 12.5. PMF for relative lift for $\{sl_1, pw_2\}$.

Figure 12.4 indicates that 44% of the frequent itemsets have average relative lift values above 0.8. These patterns are likely to be of interest. The pattern with the highest lift value of 0.998 is $\{sl_1, sw_3, pl_1, pw_1, c_1\}$. The itemset that has more or less the same support in the input and randomized datasets is $\{sl_2, c_3\}$; its average relative lift is −0.002. On the other hand, 5% of the frequent itemsets have average relative lift below −0.2. These are also of interest because they indicate more of a dis-association among the items, that is, the itemsets are more frequent by random chance. An example of such a pattern is $\{sl_1, pw_2\}$. Figure 12.5 shows the empirical probability mass function for its relative lift values across the 100 swap randomized datasets. Its average relative lift value is −0.55, and $p\text{-}value(-0.2) = 0.069$, which indicates a high probability that the itemset is disassociative.

12.2.3 Bootstrap Sampling for Confidence Interval

Typically the input transaction database $\mathbf{D}$ is just a sample from some population, and it is not enough to claim that a pattern X is frequent in $\mathbf{D}$ with support $sup(X)$. What can we say about the range of possible support values for X? Likewise, for a rule R with a given lift value in $\mathbf{D}$, what can we say about the range of lift values in different samples? In general, given a test assessment statistic Θ, bootstrap sampling allows one to infer the confidence interval for the possible values of Θ at a desired confidence level α.

The main idea is to generate k bootstrap samples from $\mathbf{D}$ using sampling *with replacement*, that is, assuming $|\mathbf{D}| = n$, each sample $\mathbf{D}_i$ is obtained by selecting at random n transactions from $\mathbf{D}$ with replacement. Given pattern X or rule $R : X \longrightarrow Y$, we can obtain the value of the test statistic in each of the bootstrap samples; let θ_i denote the value in sample $\mathbf{D}_i$. From these values we can generate the empirical cumulative distribution function for the statistic

$$\hat{F}(x) = \hat{P}(\Theta \le x) = \frac{1}{k}\sum_{i=1}^{k} I(\theta_i \le x)$$

where I is an indicator variable that takes on the value 1 when its argument is true, and 0 otherwise. Given a desired confidence level α (e.g., $\alpha = 0.95$) we can compute the interval for the test statistic by discarding values from the tail ends of $\hat{F}$ on both sides that encompass $(1-\alpha)/2$ of the probability mass. Formally, let v_t denote the critical value such that $\hat{F}(v_t) = t$, which can be obtained from quantile function as $v_t = \hat{F}^{-1}(t)$. We then have

$$\begin{aligned} P\Big(\Theta \in [v_{(1-\alpha)/2}, v_{(1+\alpha)/2}]\Big) &= \hat{F}\Big((1+\alpha)/2\Big) - \hat{F}\Big((1-\alpha)/2\Big) \\ &= (1+\alpha)/2 - (1-\alpha)/2 = \alpha \end{aligned}$$

Thus, the α% confidence interval for the chosen test statistic Θ is

$$[v_{(1-\alpha)/2}, v_{(1+\alpha)/2}]$$

The pseudo-code for bootstrap sampling for estimating the confidence interval is shown in Algorithm 12.2.

ALGORITHM 12.2. Bootstrap Resampling Method

```
BOOTSTRAP-CONFIDENCEINTERVAL($X$, $\alpha$, $k$, $\mathbf{D}$):
for $i \in [1,k]$ do
    $\mathbf{D}_i \leftarrow$ sample of size $n$ with replacement from $\mathbf{D}$
    $\theta_i \leftarrow$ compute test statistic for $X$ on $\mathbf{D}_i$
$\hat{F}(x) = P(\Theta \le x) = \frac{1}{k}\sum_{i=1}^{k} I(\theta_i \le x)$
$v_{(1-\alpha)/2} = \hat{F}^{-1}\big((1-\alpha)/2\big)$
$v_{(1+\alpha)/2} = \hat{F}^{-1}\big((1+\alpha)/2\big)$
return $[v_{(1-\alpha)/2}, v_{(1+\alpha)/2}]$
```

Example 12.21. Let the relative support *rsup* be the test statistic. Consider the itemset $X = \{sw_1, pl_3, pw_3, cl_3\}$, which has relative support $rsup(X,\mathbf{D}) = 0.113$ (or $sup(X,\mathbf{D}) = 17$) in the Iris dataset.

Using $k = 100$ bootstrap samples, we first compute the relative support of X in each of the samples ($rsup(X,\mathbf{D}_i)$). The empirical probability mass function for the relative support of X is shown in Figure 12.6 and the corresponding empirical cumulative distribution is shown in Figure 12.7. Let the confidence level be $\alpha = 0.9$. To obtain the confidence interval we have to discard the values that account for 0.05 of the probability mass at both ends of the relative support values. The critical values

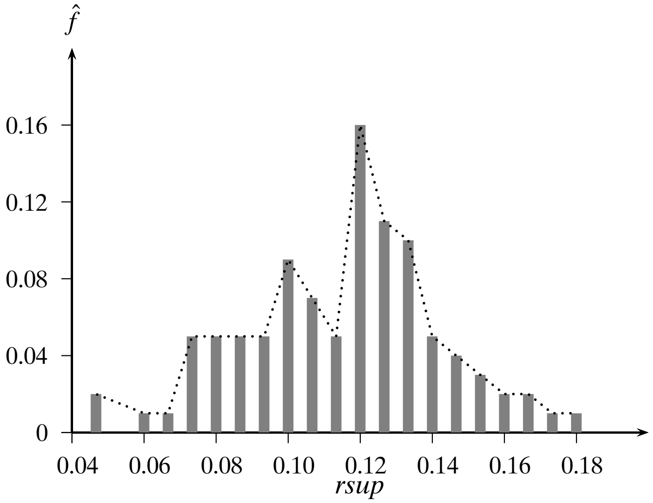


Figure 12.6. Empirical PMF for relative support.

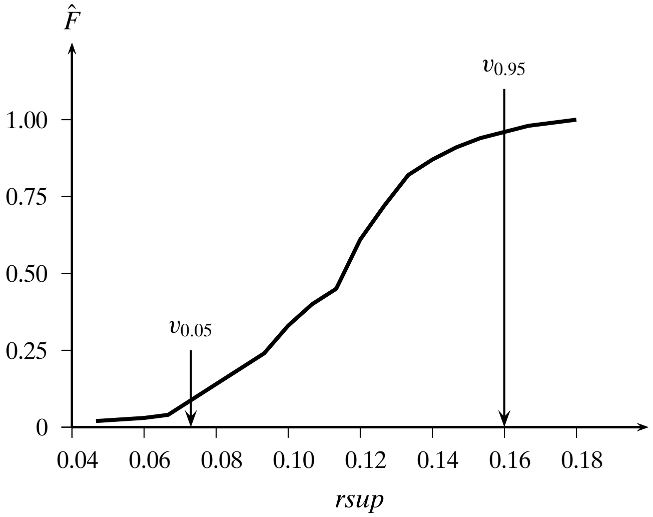


Figure 12.7. Empirical cumulative distribution for relative support.

at the left and right ends are as follows:

$$v_{(1-\alpha)/2} = v_{0.05} = 0.073$$

$$v_{(1+\alpha)/2} = v_{0.95} = 0.16$$

Thus, the 90% confidence interval for the relative support of X is $[0.073, 0.16]$, which corresponds to the interval $[11, 24]$ for its absolute support. Note that the relative support of X in the input dataset is 0.113, which has $p\text{-}value(0.113) = 0.45$, and the expected relative support value of X is $\mu_{rsup} = 0.115$.

12.3 FURTHER READING

Reviews of various measures for rule and pattern interestingness appear in Tan, Kumar, and Srivastava (2002); Geng and Hamilton (2006) and Lallich, Teytaud, and Prudhomme (2007). Randomization and resampling methods for significance testing and confidence intervals are described in Megiddo and Srikant (1998) and Gionis et al. (2007). Statistical testing and validation approaches also appear in Webb (2006) and Lallich, Teytaud, and Prudhomme (2007).

Geng, L. and Hamilton, H. J. (2006). "Interestingness measures for data mining: A survey." *ACM Computing Surveys*, 38 (3): 9.

Gionis, A., Mannila, H., Mielikäinen, T., and Tsaparas, P. (2007). "Assessing data mining results via swap randomization." *ACM Transactions on Knowledge Discovery from Data*, 1 (3): 14.

Lallich, S., Teytaud, O., and Prudhomme, E. (2007). "Association rule interestingness: measure and statistical validation." In *Quality Measures in Data Mining,* (pp. 251–275). New York: Springer Science + Business Media.

Megiddo, N. and Srikant, R. (1998). "Discovering predictive association rules." *In Proceedings of the 4th International Conference on Knowledge Discovery in Databases and Data Mining*, pp. 274–278.

Tan, P.-N., Kumar, V., and Srivastava, J. (2002). "Selecting the right interestingness measure for association patterns." *In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, pp. 32–41.

Webb, G. I. (2006). "Discovering significant rules." *In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, pp. 434–443.

12.4 EXERCISES

Q1. Show that if X and Y are independent, then $conv(X \longrightarrow Y) = 1$.

Q2. Show that if X and Y are independent then $oddsratio(X \longrightarrow Y) = 1$.

Q3. Show that for a frequent itemset X, the value of the relative lift statistic defined in Example 12.20 lies in the range

$$\left[1 - |\mathbf{D}|/minsup,\ 1\right]$$

Q4. Prove that all subsets of a minimal generator must themselves be minimal generators.

Q5. Let $\mathbf{D}$ be a binary database spanning one trillion (10^9) transactions. Because it is too time consuming to mine it directly, we use Monte Carlo sampling to find the bounds on the frequency of a given itemset X. We run 200 sampling trials $\mathbf{D}_i$ ($i = 1\ldots200$), with each sample of size $100,000$, and we obtain the support values for X in the various samples, as shown in Table 12.20. The table shows the number of samples where the support of the itemset was a given value. For instance, in 5 samples its support was 10,000. Answer the following questions:

Table 12.20. Data for Q5

Support	No. of samples
10,000	5
15,000	20
20,000	40
25,000	50
30,000	20
35,000	50
40,000	5
45,000	10

(a) Draw a histogram for the table, and calculate the mean and variance of the support across the different samples.

(b) Find the lower and upper bound on the support of X at the 95% confidence level. The support values given should be for the entire database $\mathbf{D}$.

(c) Assume that $minsup = 0.25$, and let the observed support of X in a sample be $sup(X) = 32500$. Set up a hypothesis testing framework to check if the support of X is significantly higher than the $minsup$ value. What is the p-$value$?

Q6. Let A and B be two binary attributes. While mining association rules at 30% minimum support and 60% minimum confidence, the following rule was mined: $A \longrightarrow B$, with $sup = 0.4$, and $conf = 0.66$. Assume that there are a total of 10,000 customers, and that 4000 of them buy both A and B; 2000 buy A but not B, 3500 buy B but not A, and 500 buy neither A nor B.

Compute the dependence between A and B via the χ^2-statistic from the corresponding contingency table. Do you think the discovered association is truly a strong rule, that is, does A predict B strongly? Set up a hypothesis testing framework, writing down the null and alternate hypotheses, to answer the above question, at the 95% confidence level. Here are some values of chi-squared statistic for the 95% confidence level for various degrees of freedom (df):

df	χ^2
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07
6	12.59