Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 12: Pattern and Rule Assessment

Rule Assessment Measures: Support and Confidence

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

$$sup(X \longrightarrow Y) = sup(XY) = |\boldsymbol{t}(XY)|$$

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

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

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

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

Example Dataset: Support and Confidence

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

Frequent itemsets: $minsup = 3$

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

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

Confidence should be evaluated considering the support of the rule components. For instance, since $P(BC) = 0.67$, the rule $E \longrightarrow BC$, with a 60% confidence, has a deleterious effect on BC.

Rule Assessment Measures: Lift, Leverage and Jaccard

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,

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

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

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

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:

$$jaccard(X \longrightarrow Y) = \frac{|\boldsymbol{t}(X) \cap \boldsymbol{t}(Y)|}{|\boldsymbol{t}(X) \cup \boldsymbol{t}(Y)|} = \frac{P(XY)}{P(X) + P(Y) - P(XY)}$$

Lift and Leverage

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

lift < 1 indicates the rule support is smaller than expected, while *lift* > 1 means the reverse.

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

Lift and leverage must be evaluated together, since the same lift may refer to significantly different leverages.

Lift, Jaccard, and Confidence

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

Lift and confidence must be evaluated together, to avoid either weak rules or rules where the antecedent and consequent are independent ($lift = 1$).

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

Jaccard and Lift provide similar information, but Jaccard is bounded to the interval $[0, 1]$.

Contingency Table for X and Y

We may also define the contingency table for X and Y, and exploit their absence, represented by $\neg X$ and $\neg Y$.

	Y	$\neg Y$	
X	$sup(XY)$	$sup(X\neg Y)$	$sup(X)$
$\neg X$	$sup(\neg XY)$	$sup(\neg X\neg Y)$	$sup(\neg X)$
	$sup(Y)$	$sup(\neg Y)$	$\|\boldsymbol{D}\|$

Rule Assessment Measures: Conviction

Define $\neg X$ to be the event that X is not contained in a transaction, 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.

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

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

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.

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

$A \longrightarrow DE$ is a strong rule, confirmed by high values of both lift and conviction.
$DE \longrightarrow A$ has 100% confidence, being a trivial rule.
$E \longrightarrow C$ and $C \longrightarrow E$ are weak, but, despite the same support and lift, conviction indicates that the $E \longrightarrow C$ is stronger than $C \longrightarrow E$, while confidence indicates the reverse.

Rule Assessment Measures: Odds Ratio

The odds ratio utilizes all four entries from the contingency table. 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:

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

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

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

$$oddsratio(X \longrightarrow Y) = \frac{odds(Y|X)}{odds(Y|\neg X)} = \frac{P(XY)\cdot P(\neg X\neg Y)}{P(X\neg Y)\cdot P(\neg XY)}$$

$$= \frac{sup(XY)\cdot sup(\neg X\neg Y)}{sup(X\neg Y)\cdot sup(\neg XY)}$$

If X and Y are independent, then odds ratio has value 1.

Odds Ratio

Let us compare the odds ratio for two rules, $C \longrightarrow A$ and $D \longrightarrow A$. 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 stronger than $C \longrightarrow A$, which is also confirmed by lift and confidence.

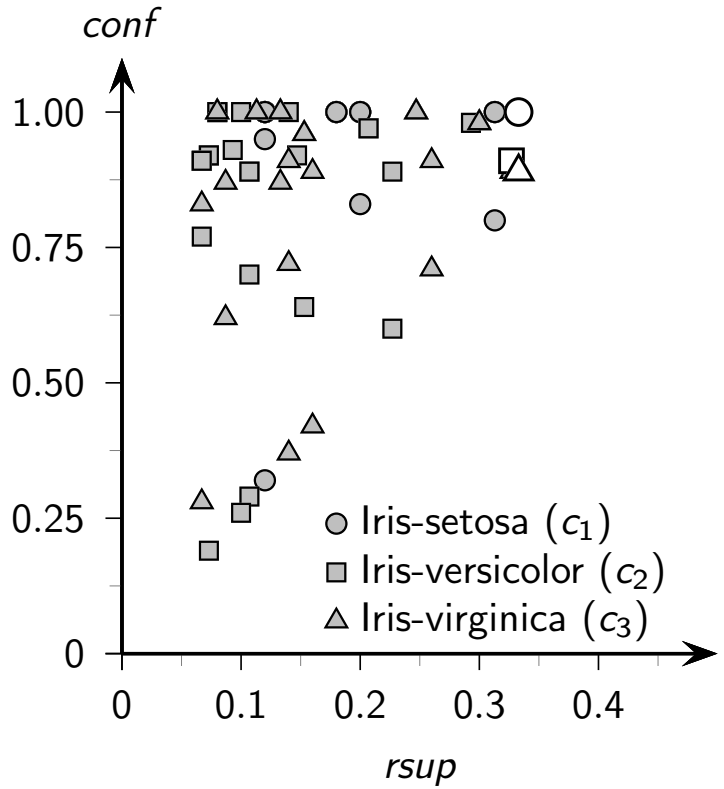
Example: Association Rules from Iris Data

Discretization of Iris Data

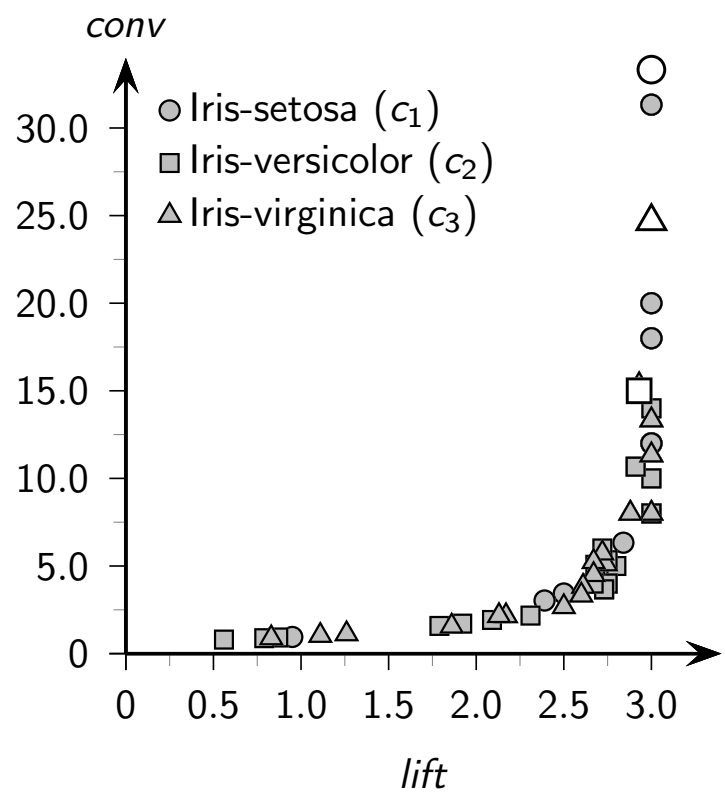
Attribute	Range or value	Label
Sepal length	4.30–5.55	sl_1
	5.55–6.15	sl_2
	6.15–7.90	sl_3
Sepal width	2.00–2.95	sw_1
	2.95–3.35	sw_2
	3.35–4.40	sw_3
Petal length	1.00–2.45	pl_1
	2.45–4.75	pl_2
	4.75–6.90	pl_3
Petal width	0.10–0.80	pw_1
	0.80–1.75	pw_2
	1.75–2.50	pw_3
Class	Iris-setosa	c_1
	Iris-versicolor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

Iris: Support vs. Confidence, and Conviction vs. Lift

minsup = 10 and *minlift* = 0.1 results in 79 rules



(a) Support vs. confidence



(b) Lift vs. conviction

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.

Iris Data: Best Class-specific Rules

Best Rules by Support and Confidence

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

Best Rules by Lift and Conviction

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

Comparing the rules for each criterion, we verify that the best rule for c_1 is the same, but the comparison between rules for c_2 and c_3 suggests a trade-off between support and novelty, represented by lift and conviction.

Pattern Assessment Measures: Support and Lift

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

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

Lift: The *lift* of a k-itemset $X = \{x_1, x_2, \ldots, x_k\}$ is defined as

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

Generalized Lift: Assume that $\{X_1, X_2, \ldots, X_q\}$ is a q-partition of X, i.e., a partitioning of X into q nonempty and disjoint itemsets X_i. 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.

Pattern Assessment Measures: Rule-based Measures

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.

For example, if Θ is rule lift, then we can define the average, maximum, and minimum lift values for X as follows:

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

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

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

Iris Data: Support Values for $\{pl_2, pw_2, c_2\}$ and its Subsets

Consider the support and relative support of itemset $X = \{pl_2, pw_2, c_2\}$ 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

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

Rules Generated from $X = \{pl_2, pw_2, c_2\}$

Consider all rules that may be generated from X:

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

We may then calculate $AvgLift(X)$:

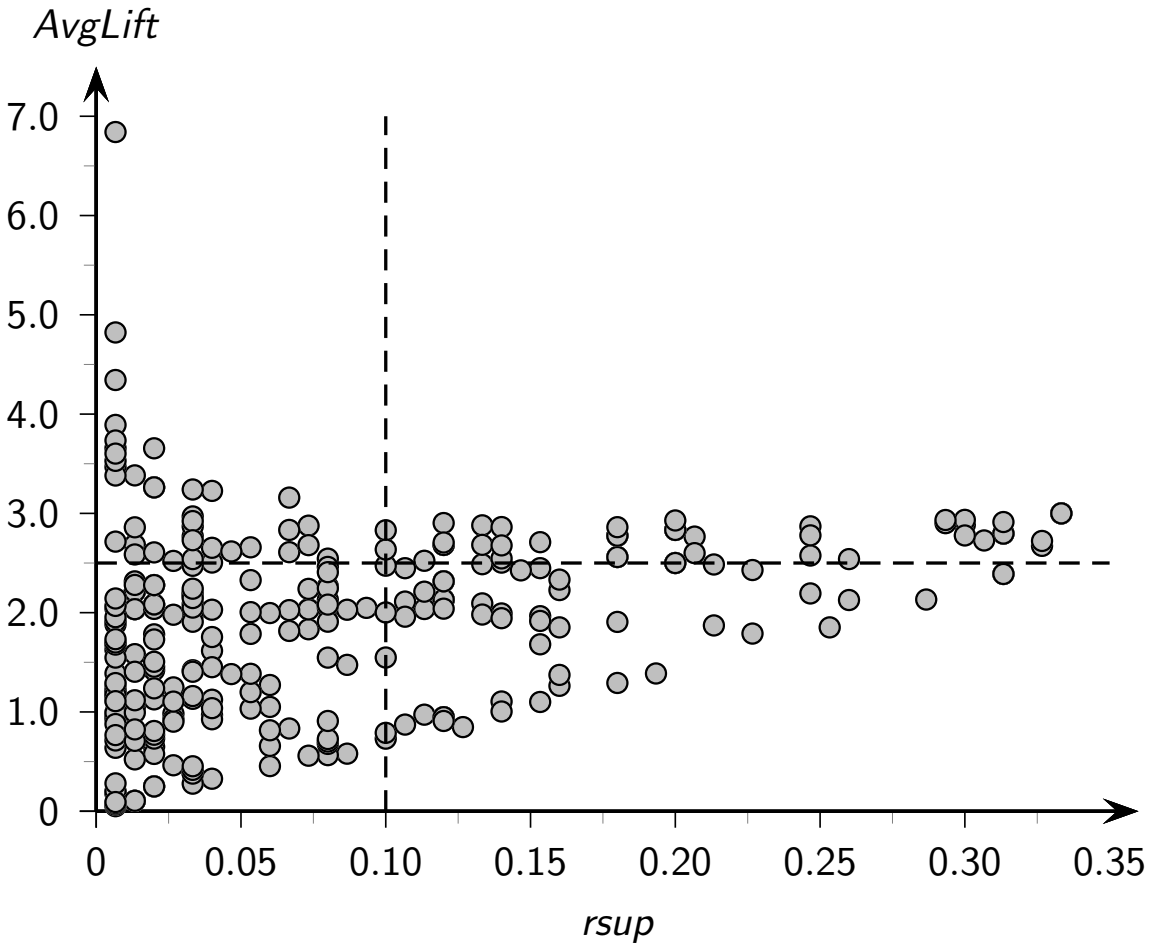
$$AvgLift(X) = avg\{2.993, 2.778, 2.933\} = 2.901$$

And also $AvgConf(X)$:

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

Iris: Relative Support and Average Lift of Patterns

306 frequent itemsets with $minsup = 1$ and $k \geq 2$



For sake of analysis, we focus on patterns with high *rsup* and then high *AvgLift*, such as $X = \{pl_1, pw_1, c_1\}$.

Comparing Itemsets: Maximal Itemsets

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

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

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.

Iris: Maximal Patterns for Average Lift

We focus on the 37 class-specific itemsets that present $rsup > 0.1$ and $AvgLift > 2.5$ and select the maximal ones:

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

For instance, for c_1, the essential items are sl_1, pl_1, pw_1 and either sw_2 or sw_3.

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.

Closed Itemsets and Minimal Generators

The support of an itemset X is

- the maximum support among all closed itemsets that contain X.
- the minimum support among all minimal generators that are subsets of X.

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*

Consider itemset AE:

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

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

Comparing Itemsets: 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

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

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

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

Comparing Itemsets: Productive Itemsets

$ABDE$ is not productive because there is at least a bipartition with $lift = 1$. For instance, the bipartition $\{B, ADE\}$:

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

ADE, on the other hand, is productive:

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

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

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

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 $sup(R) = sup(R')$.

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

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

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 $conf(R) > conf(R')$.

Comparing Rules

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, $sup(BCE) = sup(BC)$.

$BE \longrightarrow C$ 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.

Fisher Exact Test for Productive Rules

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 $\boldsymbol{D}$, conditional on the fact that W occurs, we can create a 2×2 contingency table between Z and the consequent Y

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

where

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

Fisher Exact Test for Productive Rules

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,

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

Under the null hypothesis H_0 that Z and Y are independent given W the odds ratio is 1. 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.

$$P\Big(a \Big| (a+c), (a+b), n\Big) = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$$

Fisher Exact Test: P-value

Our aim is to contrast the null hypothesis H_0 that $oddsratio = 1$ with the alternative hypothesis H_a that $oddsratio > 1$.
The *p-value* for a is given as

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

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

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 the table below:

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

Fisher Exact Test: Example

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 $\emptyset \longrightarrow c_2$. We have

$$a = sup(pw_2, c_2) = 49 \qquad b = sup(pw_2, \neg c_2) = 5$$
$$c = sup(\neg pw_2, c_2) = 1 \qquad d = sup(\neg pw_2, \neg c_2) = 95$$

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
$p\text{-}value = \sum_{i=0}^{\min(b,c)} P(a+i \mid (a+c), (a+b), n) = 1.51 \times 10^{-32}$

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.

Fisher Exact Test: Example

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

Fisher Exact Test: Example

$R_3' : \emptyset \longrightarrow c_2$

$Z = \{sw_1, pw_2\}$

$W = X \setminus Z = \emptyset$

$p\text{-}value = 3.55 \times 10^{-17}$

$W = \emptyset$	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 $p\text{-}value = 0.84$, 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.

Multiple Hypothesis Testing

There can be an exponentially large number of rules that need to be tested to check whether they are productive or not.

Multiple hypothesis testing problem: The sheer number of hypothesis tests leads to some unproductive rules passing the *p-value* $\leq \alpha$ threshold by random chance.

Bonferroni correction: takes into account the number of experiments performed during the hypothesis testing process.

$$\alpha' = \frac{\alpha}{\#r}$$

where $\#r$ is the number of rules to be tested or its estimate.

The rule false discovery rate becomes bounded by α, where a false discovery is to claim that a rule is productive when it is not.

Multiple Hypothesis Testing

Given the class-specific rules of discretized Iris dataset, the maximum number of class-specific rules 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, b is the maximum number of bins for any other attribute, i is the antecedent size, and there are b^i possible combinations for the chosen set of i attributes.

Since $c = 3$ and $b = 3$, the number of possible rules is:

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

Given $\alpha = 0.01$, $\alpha' = \alpha / \#r = 0.01/765 = 1.31 \times 10^{-5}$.

The rule $pw_2 \longrightarrow c_2$ has $p\text{-}value = 1.51 \times 10^{-32}$, and thus it remains productive even when we use α'.

Permutation Test for Significance: Swap Randomization

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.

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 $\boldsymbol{D}$, we randomly create k datasets that have the same row and column margins. We then mine frequent patterns in $\boldsymbol{D}$ 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.

Swap Randomization

Given a binary matrix $\boldsymbol{D} \subseteq \mathcal{T} \times \mathcal{I}$, the swap randomization method exchanges two nonzero cells of the matrix via a *swap* that leaves the row and column margins unchanged.

Consider any two transactions $t_a, t_b \in \mathcal{T}$ and any two items $i_a, i_b \in \mathcal{I}$ such that $(t_a, i_a), (t_b, i_b) \in \boldsymbol{D}$ and $(t_a, i_b), (t_b, i_a) \notin \boldsymbol{D}$, which corresponds to the 2×2 submatrix in $\boldsymbol{D}$, given as

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

After a swap operation we obtain the new submatrix

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

where we exchange the elements in $\boldsymbol{D}$ so that $(t_a, i_b), (t_b, i_a) \in \boldsymbol{D}$, and $(t_a, i_a), (t_b, i_b) \notin \boldsymbol{D}$. We denote this operation as $Swap(t_a, i_a; t_b, i_b)$.

Algorithm SwapRandomization

SwapRandomization(t, $\boldsymbol{D} \subseteq \mathcal{T} \times \mathcal{I}$):

1 **while** $t > 0$ **do**
2 Select pairs $(t_a, i_a), (t_b, i_b) \in \boldsymbol{D}$ randomly
3 **if** $(t_a, i_b) \notin \boldsymbol{D}$ *and* $(t_b, i_a) \notin \boldsymbol{D}$ **then**
4 $\boldsymbol{D} \leftarrow \boldsymbol{D} \setminus \{(t_a, i_a), (t_b, i_b)\} \cup \{(t_a, i_b), (t_b, i_a)\}$
5 $t = t - 1$
6 **return** $\boldsymbol{D}$

Swap Randomization Example

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 $\boldsymbol{D}$

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

Swap Randomization Example

We generated $k = 100$ swap randomized datasets (150 swaps).

Let the test statistic be the total number of frequent itemsets using $minsup = 3$. For $\boldsymbol{D}$, we have $|\mathcal{F}| = 19$, and for the $k = 100$ permuted datasets we find:

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

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.

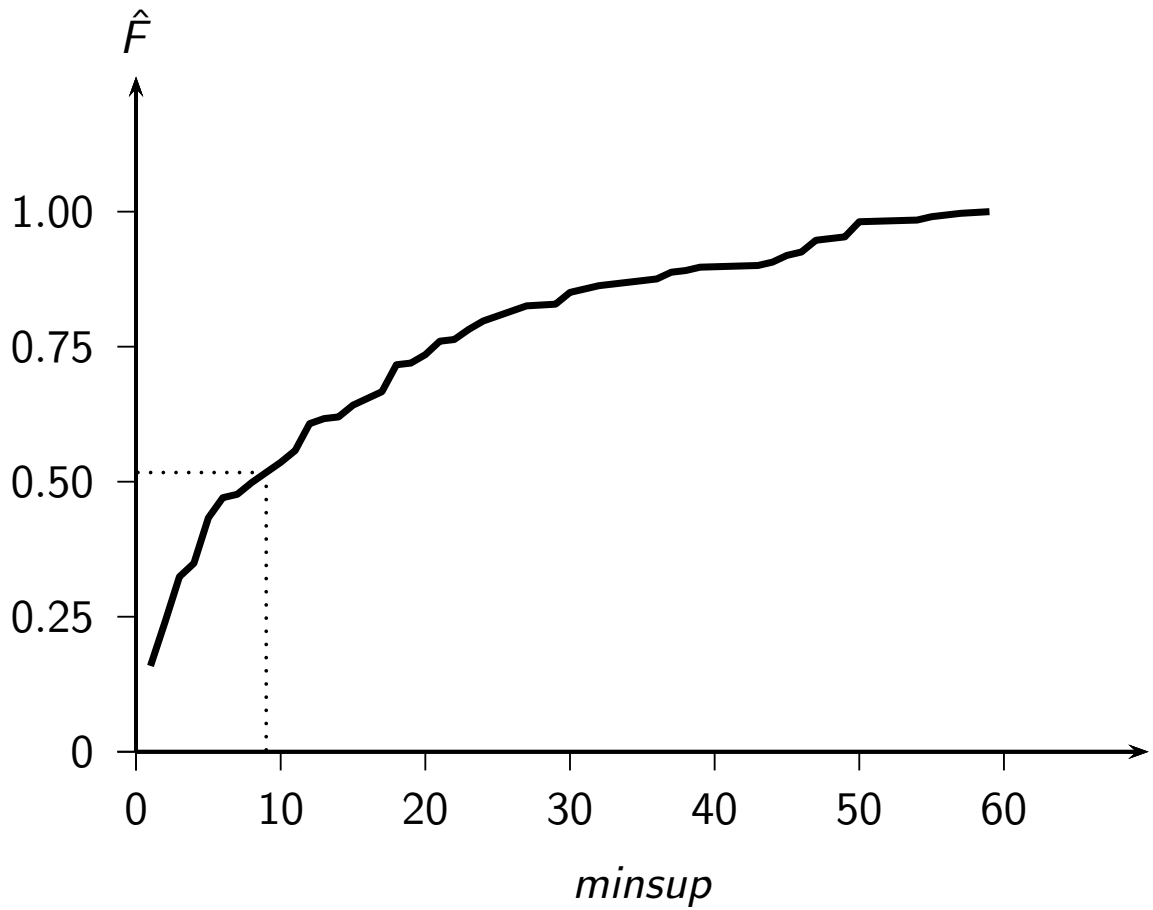
Consider $ABDE$, where $sup(ABDE) = 3$ and the probability that $ABDE$ is frequent is $17/100 = 0.17$. As this probability is not very low, $ABDE$ is not a statistically significant pattern.

Consider BCD, where $sup(BCD) = 2$. The empirical PMF is given as

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

Since 54% indicates BCD is infrequent, we may assume it.

CDF for Number of Frequent Itemsets: Iris



We choose $minsup = 10$, for which we have $\hat{F}(10) = P(sup < 10) = 0.517$, that is, 48.3% of the itemsets that occur at least once are frequent.

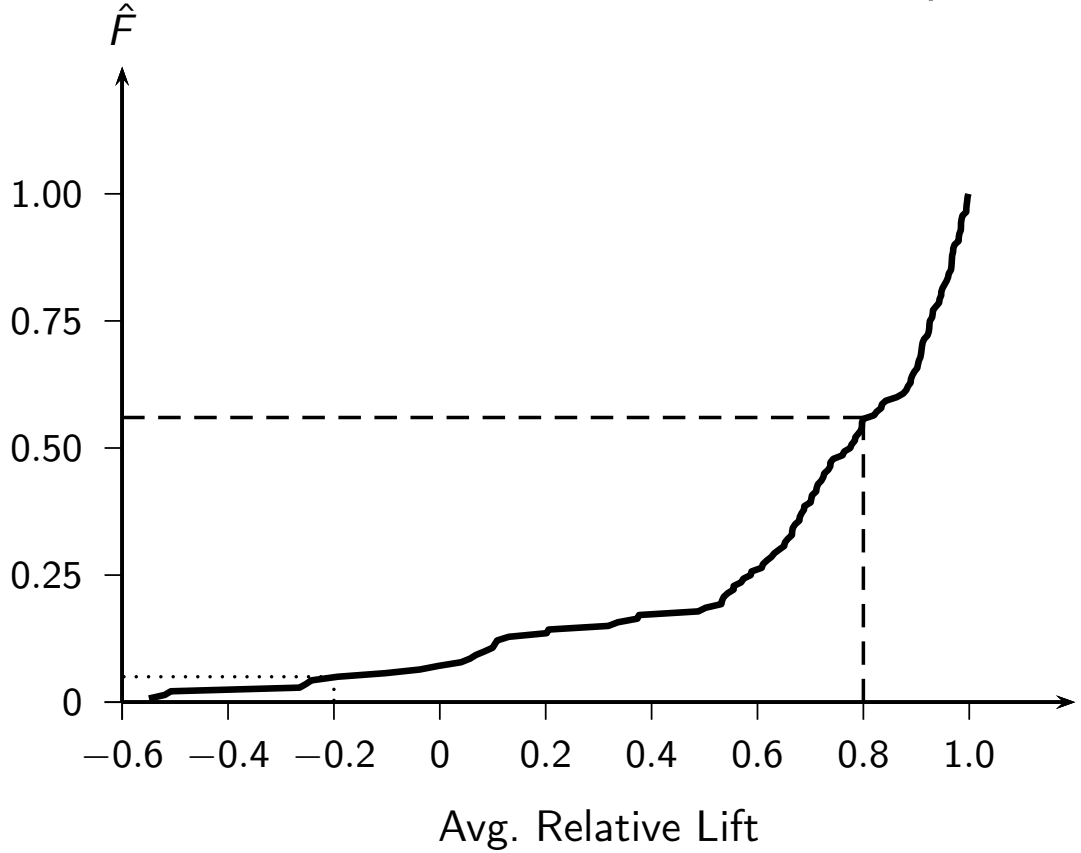
CDF for Average Relative Lift: Iris

$k = 100$ swap randomization steps, 140 frequent itemsets

The relative lift statistic is

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

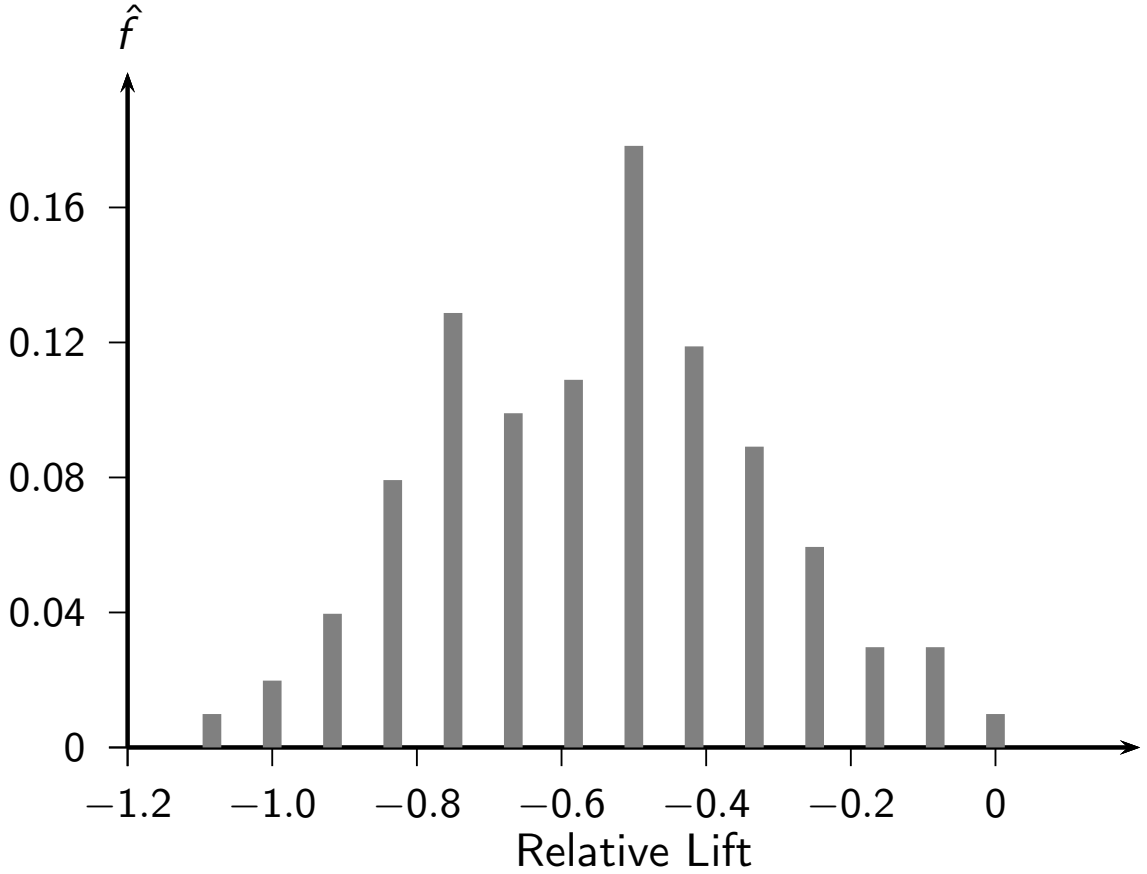
$\boldsymbol{D}_i$ is ith swap randomized dataset obtained after k steps.



PMF for Relative Lift: $\{sl_1, pw_2\}$

$k = 100$ swap randomization steps

Its average relative lift value is -0.55, and $p\text{-}value(-0.2) = 0.069$, which indicates that the itemset is likely to be disassociative.



Bootstrap Sampling for Confidence Interval

We can generate k bootstrap samples from $\boldsymbol{D}$ using sampling *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 $\boldsymbol{D}_i$.

From these values we can generate the empirical cumulative distribution function for the statistic

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

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. In other words, the interval $\left[v_{1-\alpha/2}, v_{\alpha/2}\right]$ encompasses $1-\alpha$ fraction of the probability mass, and therefore it is called the $100(1-\alpha)\%$ confidence interval for the chosen test statistic Θ.

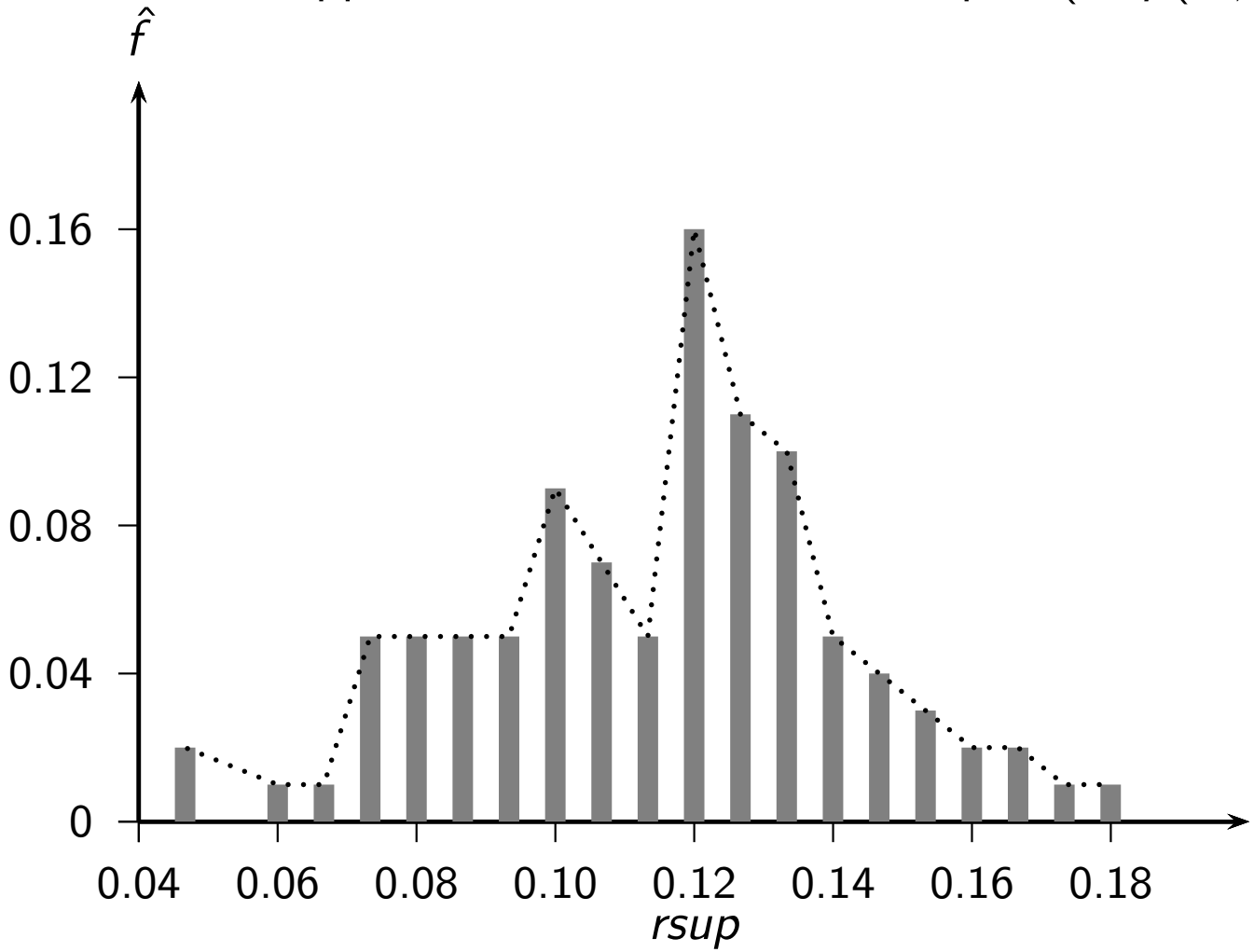
Bootstrap Confidence Interval Algorithm

Bootstrap-ConfidenceInterval(X, α, k, $\boldsymbol{D}$):

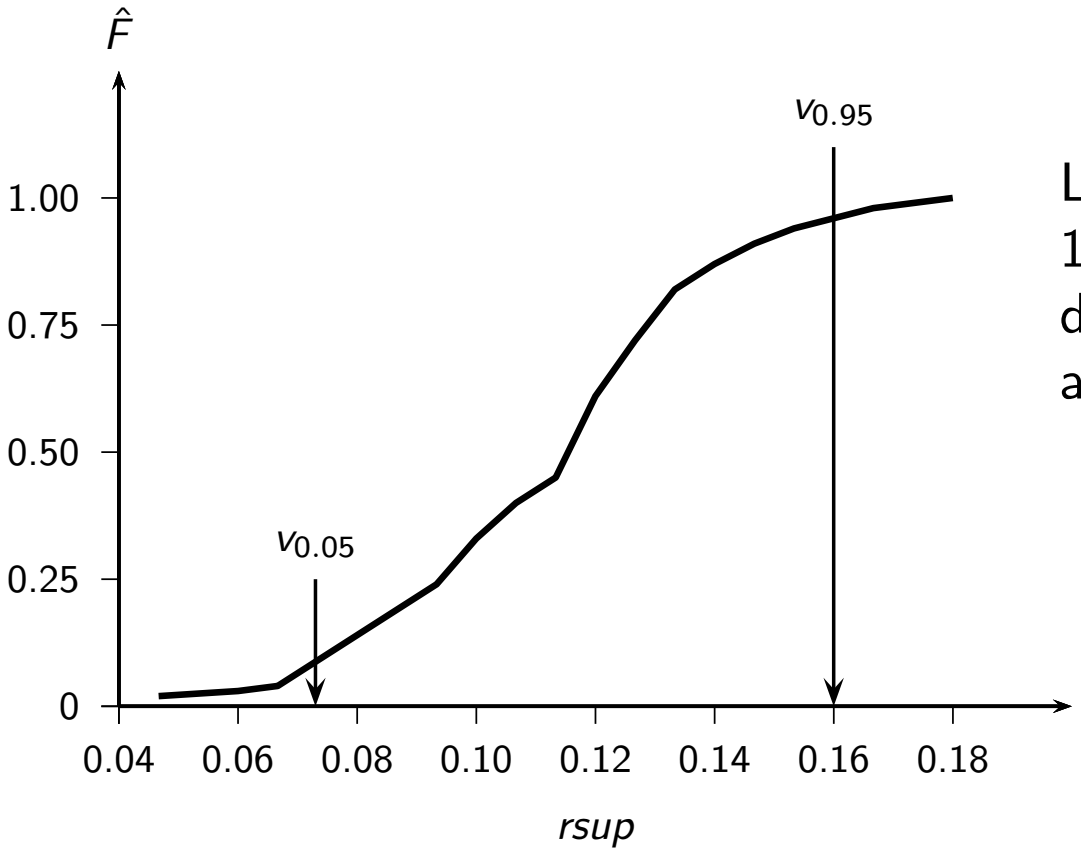
for $i \in [1, k]$ **do**
  $\boldsymbol{D}_i \leftarrow$ sample of size n with replacement from $\boldsymbol{D}$
  $\theta_i \leftarrow$ compute test statistic for X on $\boldsymbol{D}_i$
$\hat{F}(x) = P(\Theta \leq x) = \frac{1}{k} \sum_{i=1}^{k} I(\theta_i \leq x)$
$v_{(1-\alpha)/2} = \hat{F}^{-1}\big((1-\alpha)/2\big)$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

Empirical PMF for RelSupport: $X = \{sw_1, pl_3, pw_3, cl_3\}$

Given $rsup(X, \boldsymbol{D}) = 0.113$ (or $sup(X, \boldsymbol{D}) = 17$) and $k = 100$ bootstrap samples, we compute the relative support of X in each of the samples ($rsup(X, \boldsymbol{D}_i)$).



Empirical CDF for RelSupport: $X = \{sw_1, pl_3, pw_3, cl_3\}$



Let the confidence level be $1 - \alpha = 0.9$, thus $\alpha = 0.1$, discarding the values that account for $\alpha/2 = 0.05$:

$$v_{1-\alpha/2} = v_{0.95} = 0.073$$
$$v_{\alpha/2} = v_{0.05} = 0.16$$

The 90% confidence interval for $rsup(X) \in [0.073, 0.16]$, i.e., $sup \in [11, 24]$.

Note that $rsup(X, \boldsymbol{D}) = 0.113$, with $p\text{-}value(0.113) = 0.45$, and $\mu_{rsup(X)} = 0.115$.

Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 12: Pattern and Rule Assessment