Statistical learning theory meets knowledge discovery

Randomized algorithms for Big Data analytics

Matteo Riondato

Thesis proposal, April 8th 2013

Outline

- Introduction
  - Thesis statement
- Mining Frequent Itemsets through sampling
- A statistical test for True Frequent Itemsets
- Proposed work
  - Efficient progressive sampling for Frequent Itemsets mining
  - Graph mining: problems and challenges

Data, data, data

- “In God we trust, all others bring data”

Prof. W.E. Deming, Statistician, (attributed)

Data, data, data

- “In God we trust, all others bring data”

Prof. W.E. Deming, Statistician, (attributed)

- “Every two days now we create as much information as we did from the dawn of civilization up until 2003”

Eric Schmidt, Google Exec. Chairman, 2009

- “Data explosion is bigger than Moore's Law”

Marissa Mayer, Yahoo! CEO, 2009

The value of data

- We (as humans)
  - create more and more data
  - store more and more data

The value of data

- We (as humans)
  - create more and more data
  - store more and more data
  - need to analyze more and more data

The value of data

- We (as humans)
  - create more and more data
  - store more and more data
  - need to analyze more and more data
- Data has an implicit value
- Explicit through data analysis (analytics)
  - We have been doing this for ages: statistics, machine learning, data mining, ...

What is data?

- Two different points of view:
  1. “One-shot”: data is the whole reality
     - Goal is extracting information from data
  2. “Scientific”: data is a collection of samples from an unknown generating process
     - Goal is understanding the process through data
- In both cases we want to find interesting patterns
  - Concept of “interesting” is different

Statistical validation of results

- Scientific point of view:
  - data = collection of samples from unknown generating process
- Want to use data to understand generating process
- Interesting in data   interesting in generating process
- Opens a whole new can of worms

Statistical validation of results

- Scientific point of view:
  - data = collection of samples from unknown generating process
- Want to use data to understand generating process
- Interesting in data interesting in generating process
- Opens a whole new can of worms
- Need for statistical validation of results
  - “Is this pattern really interesting?”
- Need to develop statistical tests to assess results

Big Data – The 3 V's

- Data is not what it used to be: now it is Big Data

Big Data – The 3 V's

- Data is not what it used to be: now it is Big Data
- Key characteristics (analytics point of view)
  - Volume: datasets are huge and growing
  - Velocity: analysis must be fast
  - Variety: data is structured (XML, graphs, ...), multidimensional, rich
- The 3V's are challenges that need to be addressed

Challenges of Big Data

- “Traditional” analytics techniques are limited
  - do not scale well (velocity) with volume
  - may not address all the variety
    - Example: few methods for structured data, graphs, uncertain/noisy data, …

Challenges of Big Data

- “Traditional” analytics techniques are limited
  - do not scale well (velocity) with volume
  - may not address all the variety
    - Example: few methods for structured data, graphs, uncertain/noisy data, …
- Need new methods/ideas to handle Big Data

How to address the scalability challenge?

- Velocity VS Volume

How to address the scalability challenge?

- Velocity VS Volume
- Idea: trade off accuracy of results for execution speed

How to address the scalability challenge?

- Velocity VS Volume
- Idea: trade off accuracy of results for execution speed

Can approximation algorithms for data analysis be fast and still guarantee high-quality results?

Thesis statement

I develop efficient and scalable approximation algorithms and statistical tests for a variety of problems in data analysis, addressing the challenges of Big Data using modern statistics and probability

Why is it difficult?

Traditional statistics / probability techniques are not powerful enough to address the challenges of Big Data

Outline

- Introduction ✓
  - Thesis statement ✓
- Mining Frequent Itemsets through sampling
- A statistical test for True Frequent Itemsets
- Proposed work
  - Efficient progressive sampling for Frequent Itemset mining
  - Graph mining: problems and challenges

What I am going to show you

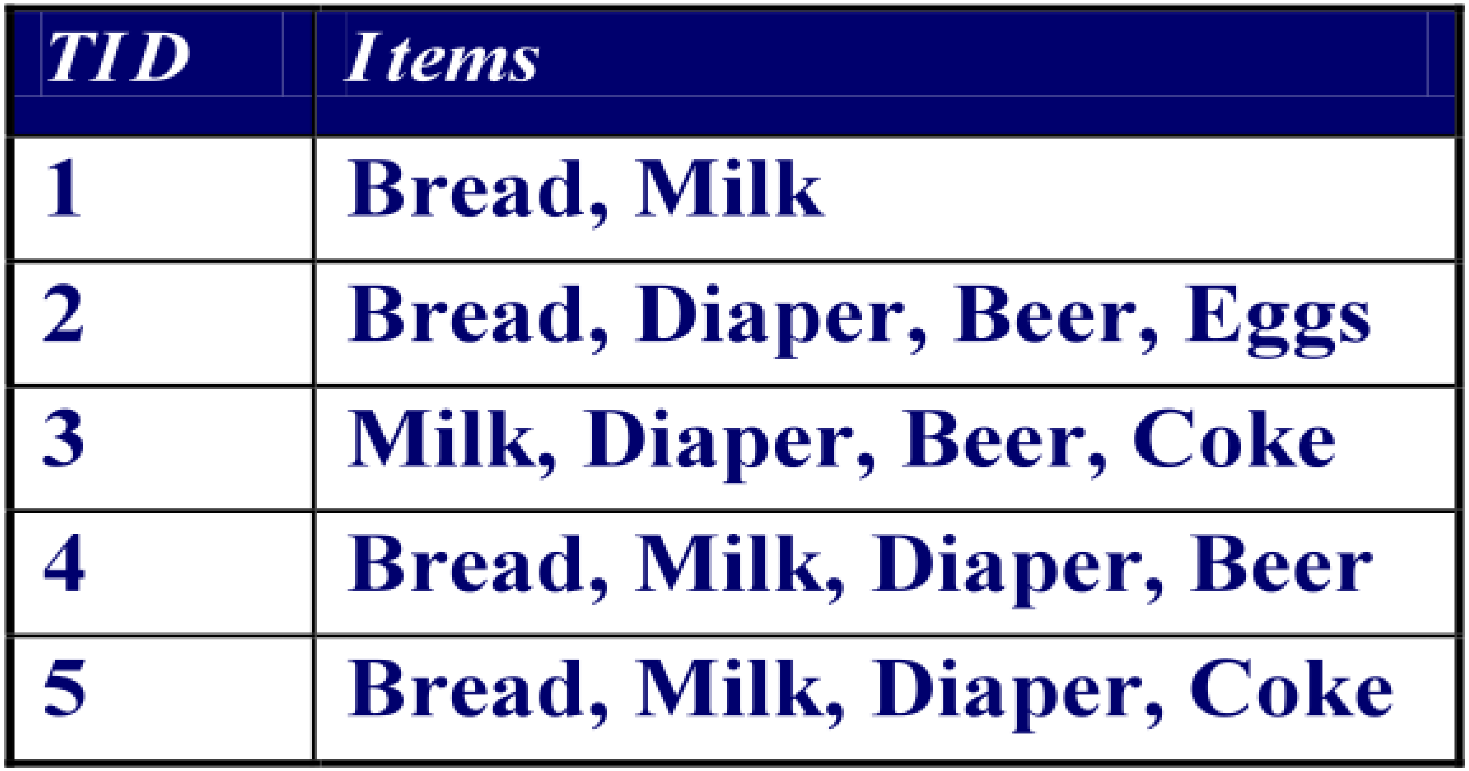
- Two algorithms for problems in data analysis
- Problems are similar but different
  - Same settings (Frequent Itemsets)
  - Different points of view on data
    - "one-shot" vs "scientific"
- Show how to use modern statistics to address Big Data challenges

Motivation

- Market Basket Analysis
- You own a grocery store
- Have copy of receipts from sales
  - For each customer, you have the list of products she bought
- Interested in what groups of products are bought together the most
  - Useful to take business decisions

Settings

- Transactional dataset D



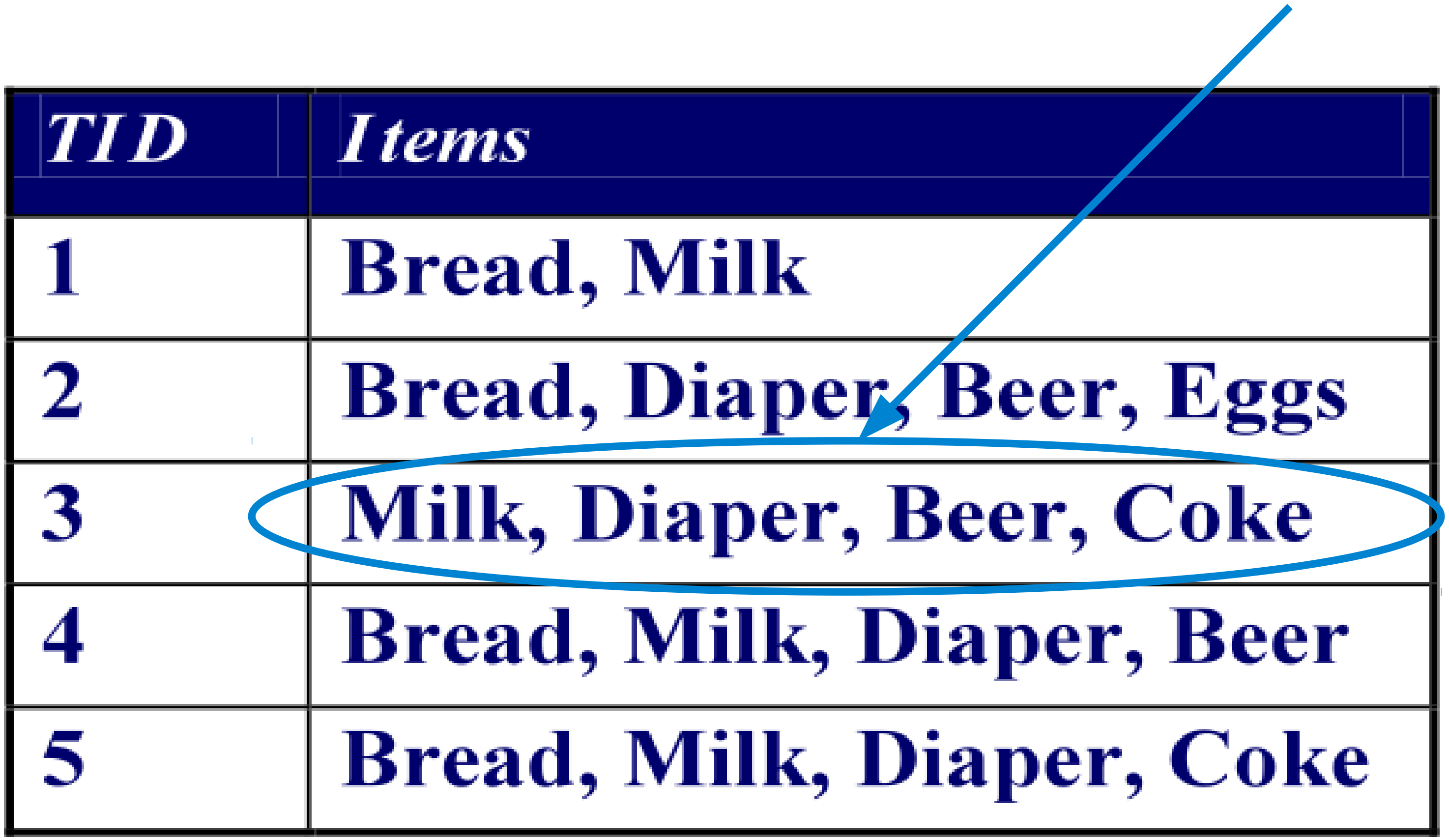
TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Figure from Tan et al. - Introduction to Data Mining

Transactions

- Transactional dataset D

transaction

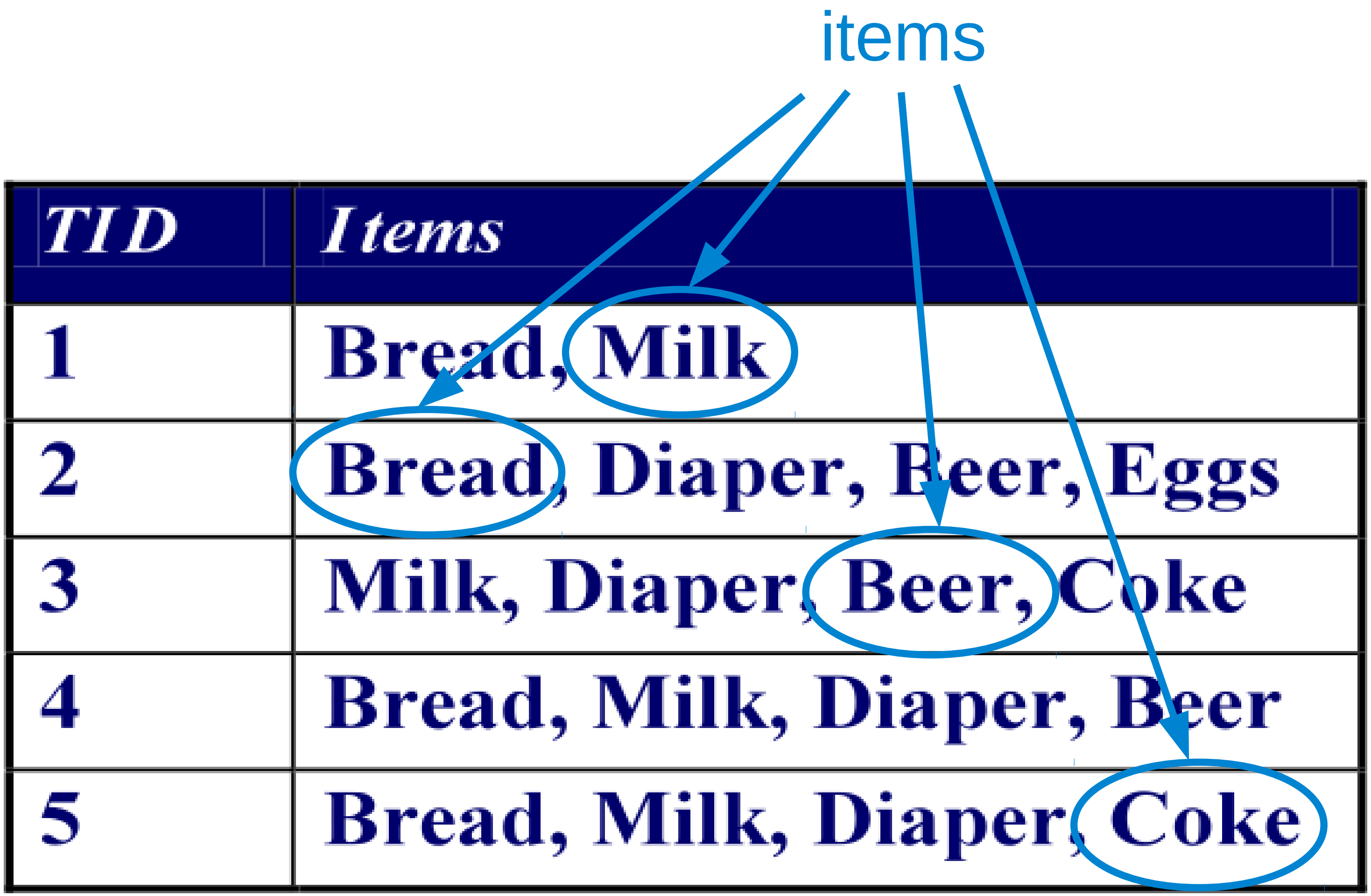


TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Figure from Tan et al. - Introduction to Data Mining

Items

- Transactions are built on items from a domain



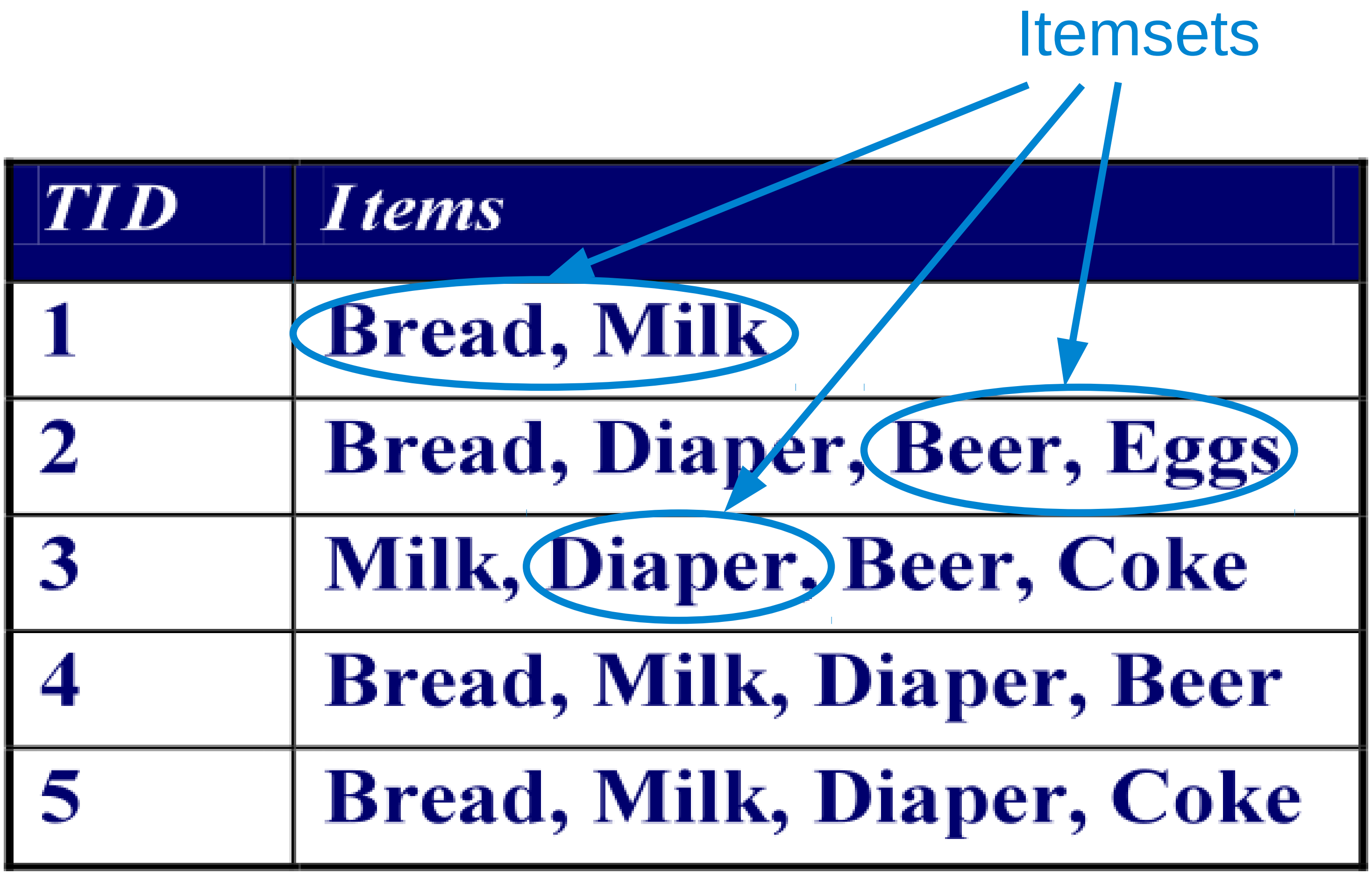
items

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Figure from Tan et al. - Introduction to Data Mining

Itemsets

- Sets of items are called itemsets



Itemsets

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

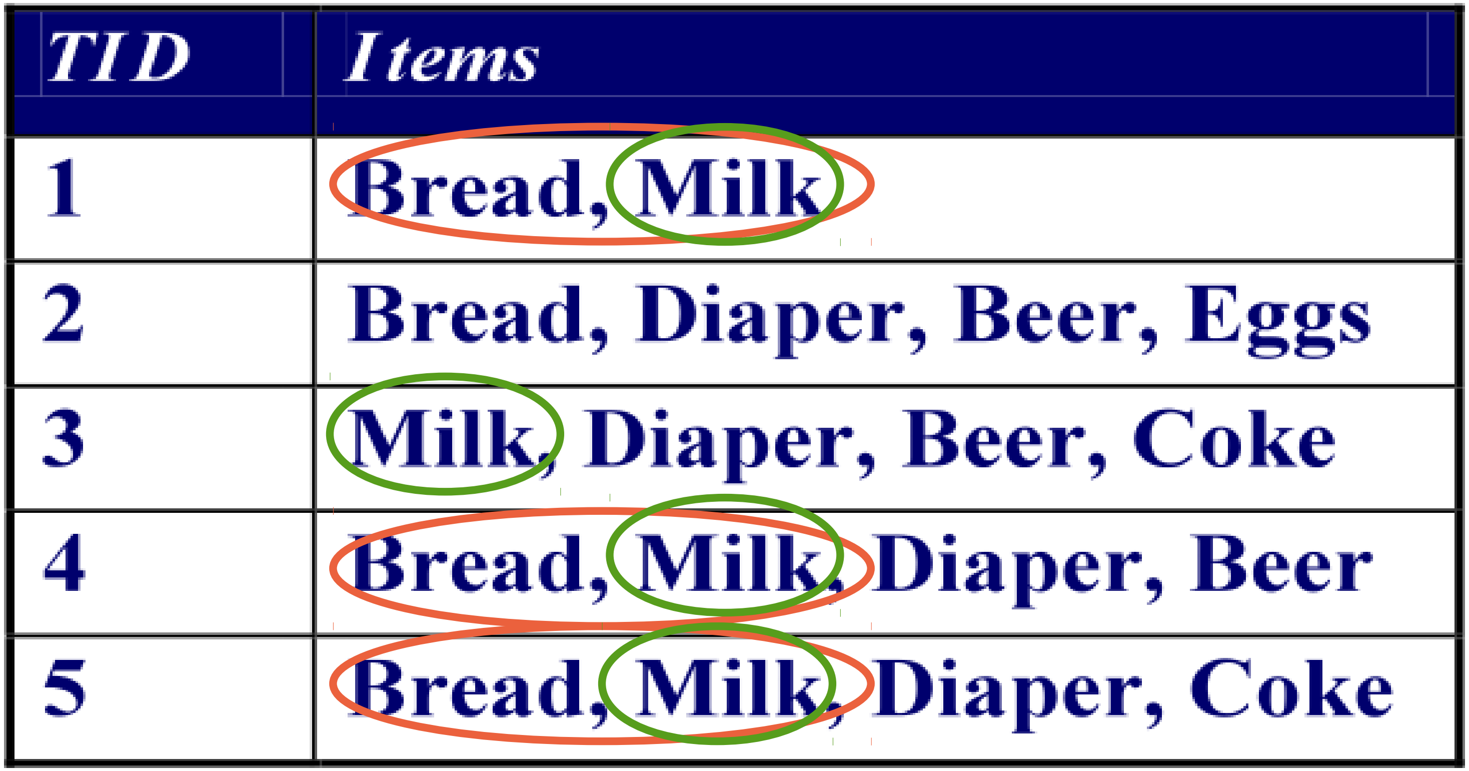
Figure from Tan et al. - Introduction to Data Mining

Frequency

- Frequency of an itemset X in dataset D:
  - $f_D(X)$: fraction of transactions of D containing X

Frequency

- Frequency of an itemset X in dataset D:
  - $f_D(X)$: fraction of transactions of D containing X
  - Example: Milk: 4/5, {Bread, Milk}: 3/5



TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Figure from Tan et al. - Introduction to Data Mining

Frequent Itemsets

- Frequent Itemsets mining with respect to $\theta \in [0, 1]$
  - Find all itemsets X with frequency $f_D(X) \geq \theta$

Frequent Itemsets

- Frequent Itemsets mining with respect to $\theta \in [0, 1]$
  - Find all itemsets X with frequency $f_D(X) \geq \theta$
- “One-shot” point of view: data is whole reality
- Interesting patterns = frequent itemsets

Frequent Itemsets

- Frequent Itemsets mining with respect to $\theta \in [0, 1]$
  - Find all itemsets X with frequency $f_D(X) \geq \theta$
- "One-shot" point of view: data is whole reality
- Interesting patterns = frequent itemsets
- Variants:
  - Top-K Frequent Itemsets
    - Find all itemsets at least as frequent as the kth most frequent
  - Association Rules
    - Inference rules involving itemsets

Frequent Itemsets mining

- Well studied classical problem in data analysis
- There are algorithms to extract exact collection of FI's
  - APriori [AS94], FPGrowth [HPY00], Eclat [Zaki00]

Frequent Itemsets mining

- Well studied classical problem in data analysis
- There are algorithms to extract exact collection of FI's
  - APriori [AS94], FPGrowth [HPY00], Eclat [Zaki00]
- Running time (velocity) depends on number of transactions in the dataset and on the number of frequent itemsets (volume)
  - 10^8 transactions considered “normal” size
  - 10^4 items   2^(10^4) itemsets total
  - A transaction with d items contains 2^d itemsets

How to speed up mining?

- Decrease dependency on number of transactions:

How to speed up mining?

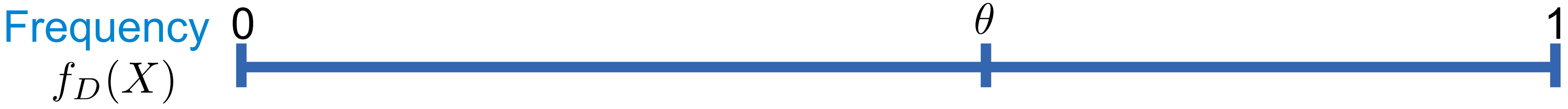
- Decrease dependency on number of transactions:
    1. create random sample of dataset
    2. extract frequent itemsets from sample with lower $\theta' < \theta$

How to speed up mining?

- Decrease dependency on number of transactions:
  1. create random sample of dataset
  2. extract frequent itemsets from sample with lower $\theta' < \theta$
- Sampling force us to accept approximate results
- Need to quantify what is acceptable

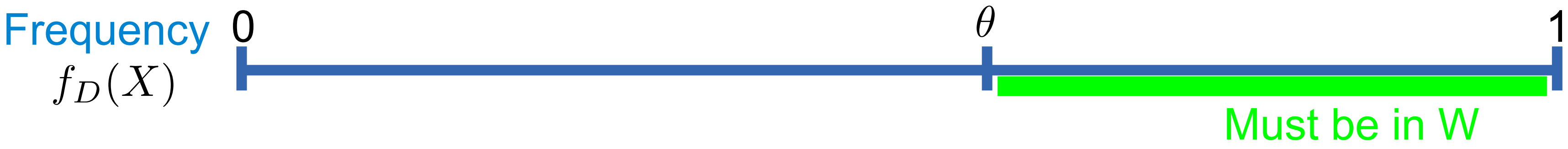
How to speed up mining?

- Decrease dependency on number of transactions:
  1. create random sample of dataset
  2. extract frequent itemsets from sample with lower $\theta' < \theta$
- Sampling force us to accept approximate results
- Need to quantify what is acceptable
- Given $\varepsilon \in [0, 1]$, extract collection W of FI's such that:



How to speed up mining?

- Decrease dependency on number of transactions:
  1. create random sample of dataset
  2. extract frequent itemsets from sample with lower $\theta' < \theta$
- Sampling force us to accept approximate results
- Need to quantify what is acceptable
- Given $\varepsilon \in [0, 1]$, extract collection W of FI's such that:



Frequency $f_D(X)$

0

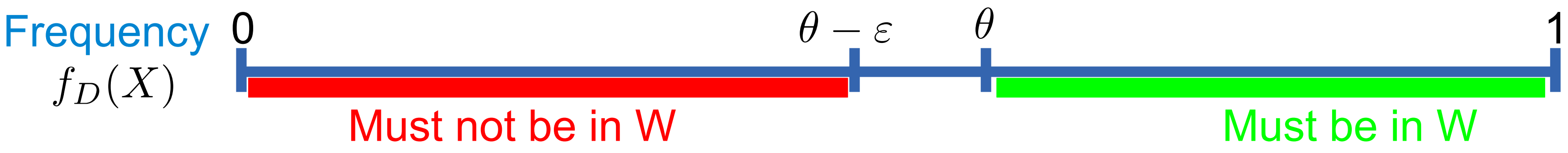
θ

1

Must be in W

How to speed up mining?

- Decrease dependency on number of transactions:
  1. create random sample of dataset
  2. extract frequent itemsets from sample with lower $\theta' < \theta$
- Sampling force us to accept approximate results
- Need to quantify what is acceptable
- Given $\varepsilon \in [0, 1]$, extract collection W of FI's such that:



Frequency $f_D(X)$

0

$\theta - \varepsilon$

θ

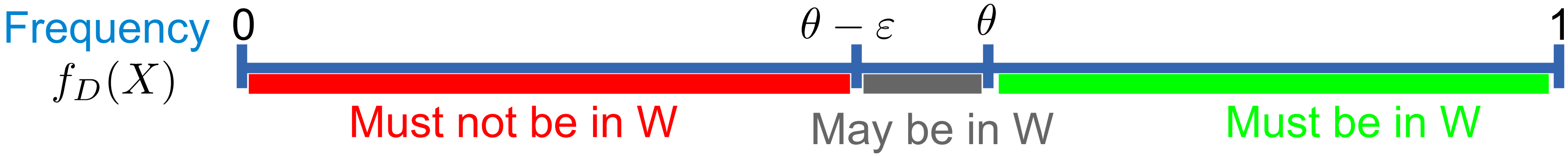
1

Must not be in W

Must be in W

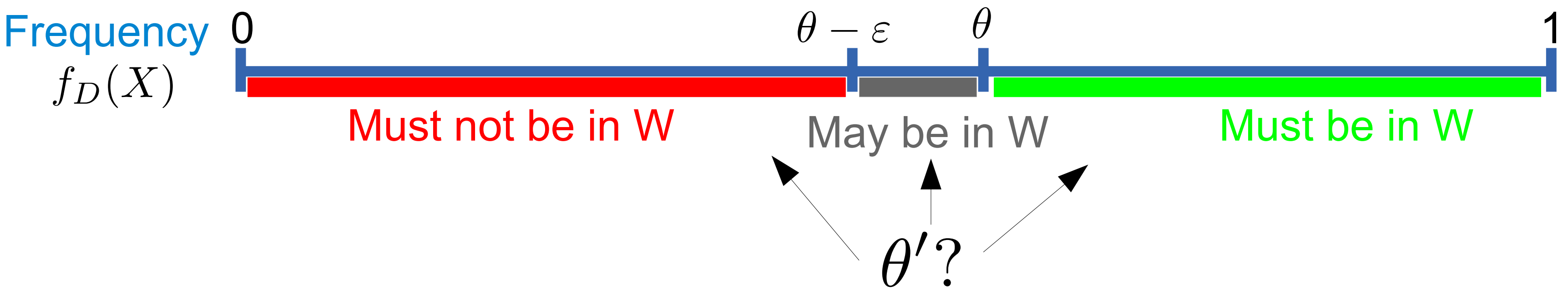
How to speed up mining?

- Decrease dependency on number of transactions:
  1. create random sample of dataset
  2. extract frequent itemsets from sample with lower $\theta' < \theta$
- Sampling force us to accept approximate results
- Need to quantify what is acceptable
- Given $\varepsilon \in [0, 1]$, extract collection W of FI's such that:



How to speed up mining?

- Decrease dependency on number of transactions:
  1. create random sample of dataset
  2. extract frequent itemsets from sample with lower $\theta' < \theta$
- Sampling force us to accept approximate results
- Need to quantify what is acceptable
- Given $\varepsilon \in [0, 1]$, extract collection W of itemset from sample, such that:



Frequency $f_D(X)$

0

$\theta - \varepsilon$

θ

1

Must not be in W

May be in W

Must be in W

θ'?

- Problem: choose right sample size and right $\theta' < \theta$

Choosing θ'

- If we have, for all itemsets X simultaneously

$$|f_S(X) - f_D(X)| \leq \varepsilon/2$$

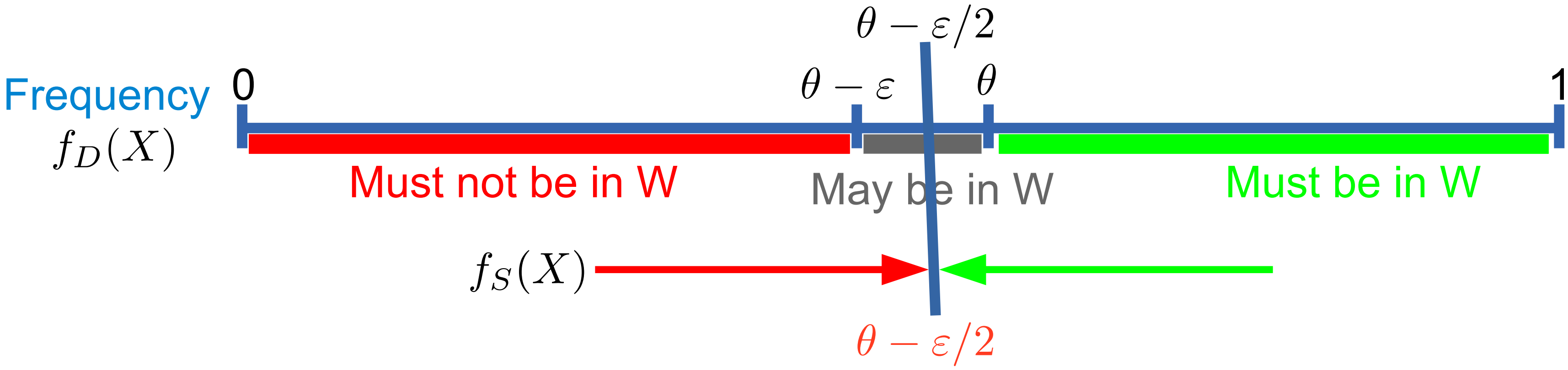
then $\theta' = \theta - \varepsilon/2$ does the trick

Choosing θ'

- If we have, for all itemsets X simultaneously

$$|f_S(X) - f_D(X)| \leq \varepsilon/2$$

then $\theta' = \theta - \varepsilon/2$ does the trick

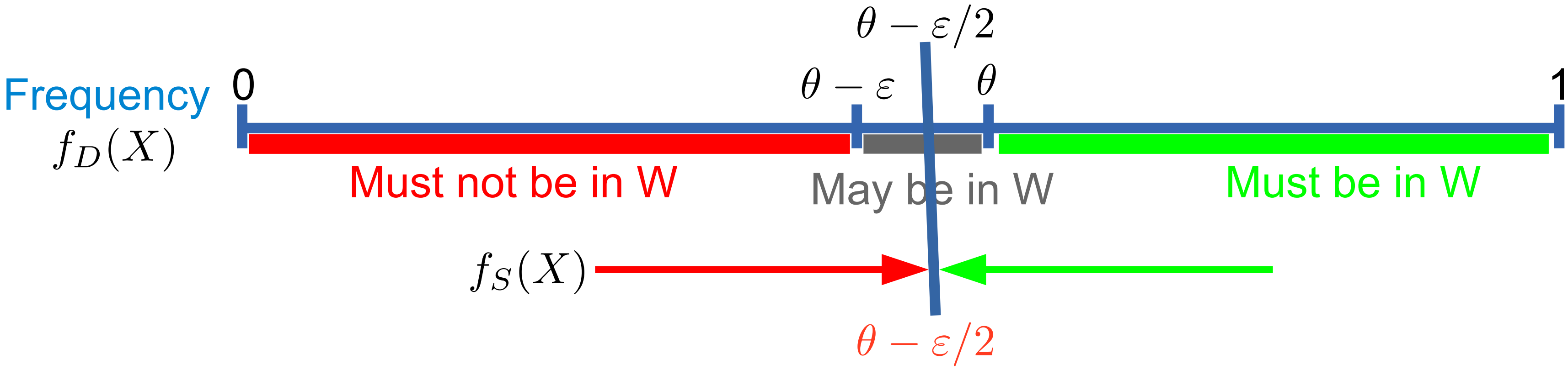


Choosing θ'

- If we have, for all itemsets X simultaneously

$$|f_S(X) - f_D(X)| \leq \varepsilon/2$$

then $\theta' = \theta - \varepsilon/2$ does the trick



- Need sample size |S| s.t., with prob. at least $1 - \delta$, for all itemsets X, we have

$$|f_D(X) - f_S(X)| \leq \varepsilon/2$$

Choosing the sample size

- Naïve approach
- Given itemset X, frequency of X in sample S is distributed like a binomial: $f_S(X) \sim \mathcal{B}(|S|, f_D(X))$
- Use Chernoff bound to bound $|f_S(X) - f_D(X)|$

$$\Pr(|f_S(X) - f_D(X)| \geq \varepsilon/2) \leq e^{-|S| f_D(X) \varepsilon^2/8}$$

- Apply Union bound over all itemsets to find |S| such that

$$\Pr(\exists X \ : \ |f_S(X) - f_D(X)| \geq \varepsilon/2) \leq \delta$$

Choosing the sample size

- Naïve approach
- Given itemset X, frequency of X in sample S is distributed like a binomial: $f_S(X) \sim \mathcal{B}(|S|, f_D(X))$
- Use Chernoff bound to bound $|f_S(X) - f_D(X)|$

$$\Pr(|f_S(X) - f_D(X)| \geq \varepsilon/2) \leq e^{-|S| f_D(X) \varepsilon^2/8}$$

- Apply Union bound over all itemsets to find |S| such that

$$\Pr(\exists X \ : \ |f_S(X) - f_D(X)| \geq \varepsilon/2) \leq \delta$$

- Problem: There's an exponential number of itemsets
  - Sample size would depend on it and be very large

How to get around this?

- Probability and statistics didn't stop 50 years ago

How to get around this?

- Probability and statistics didn't stop 50 years ago
- Statistical Learning Theory
  - studies necessary and sufficient conditions for learning (i.e. approximating) a function from “small” samples
- Main results: VC-Dimension, Rademacher averages, Structural risk minimization, ...

Vapnik-Chervonenkis Dimension

- Combinatorial property of a collection of subsets from a domain
- Measures the “richness”, “expressivity” of the subsets
- If we know the VC-dim of a collection of subsets, we can compute the sample size sufficient to approximate the sizes of the subsets using a sample

Range spaces

- VC-Dimension is defined on range spaces
- (B,R): range space
  - B: domain
  - R: collection of subsets from B (ranges)
- No restrictions:
  - B can be infinite
  - R can be infinite
  - R can contain infinitely-large subsets of B

Vapnik-Chervonenkis Dimension

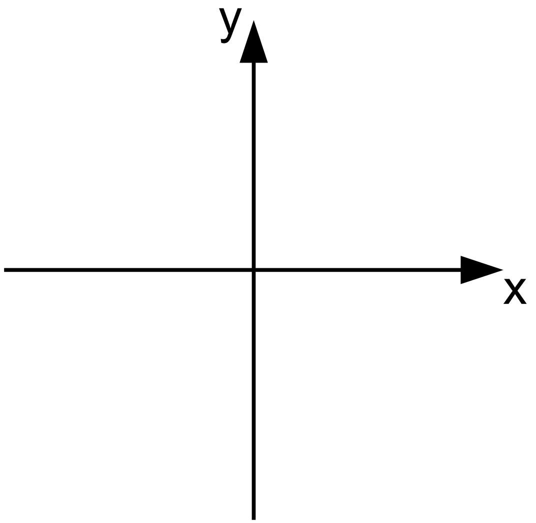
- Range space (B, R)
- For any $C \subseteq B$, define
$$P_C = \{C \cap F \; : \; F \in R\}$$
- C is shattered if $P_C = 2^C$
- The VC-Dimension of (B,R) is the size of the largest shattered subset of B

Example of VC-Dimension

- B = $\mathbb{R}^2$

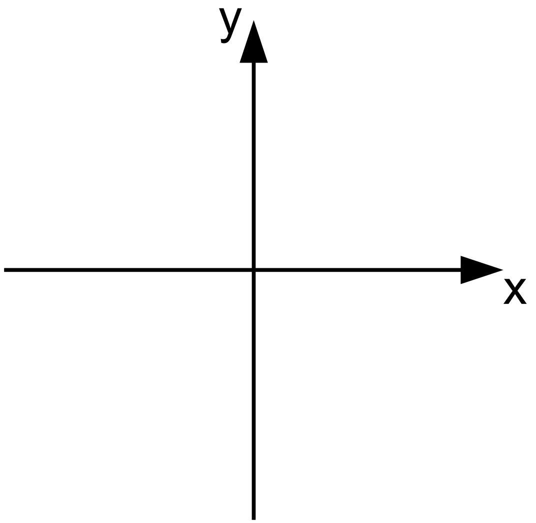
Example of VC-Dimension

- B = $\mathbb{R}^2$



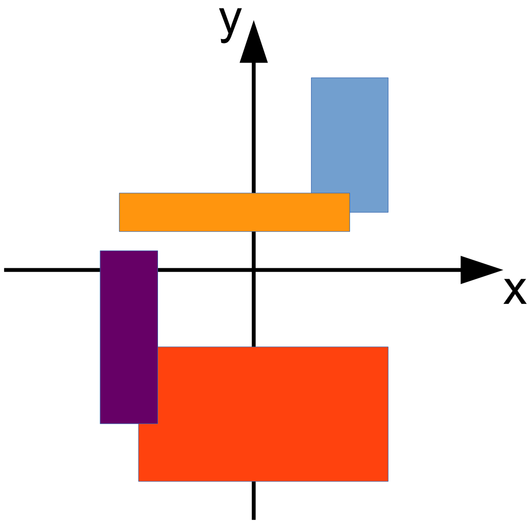
Example of VC-Dimension

- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$



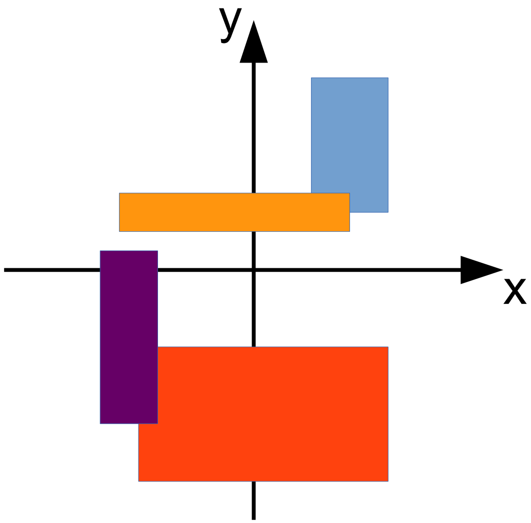
Example of VC-Dimension

- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$



Example of VC-Dimension

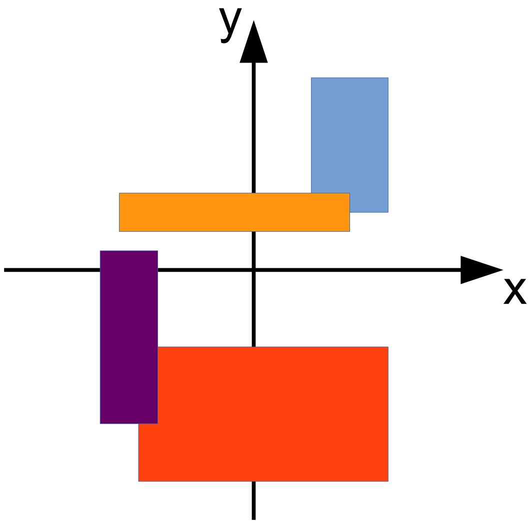
- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$



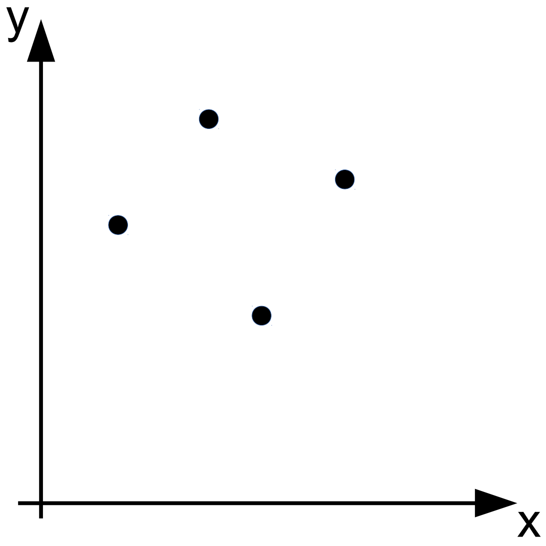
- Shattering 4 points: Easy
  - Take any 4 points s.t. no 3 of them are aligned

Example of VC-Dimension

- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$

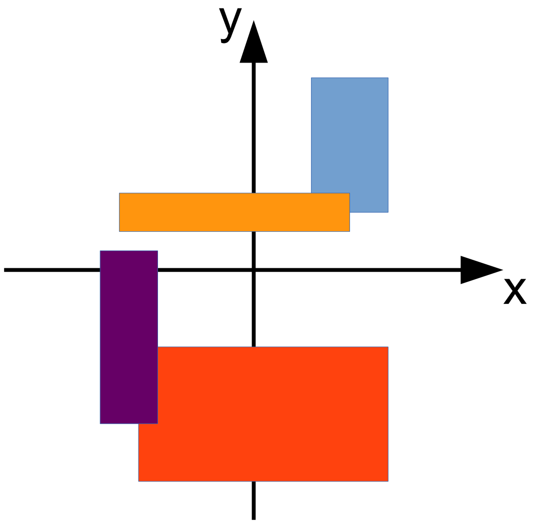


- Shattering 4 points: Easy
  - Take any 4 points s.t. no 3 of them are aligned

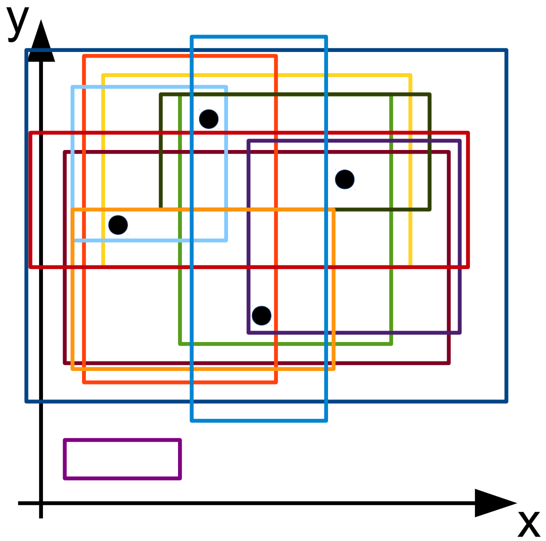


Example of VC-Dimension

- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$



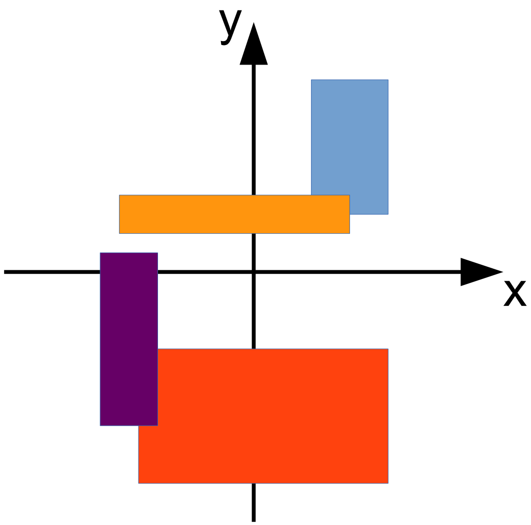
- Shattering 4 points: Easy
  - Take any 4 points s.t. no 3 of them are aligned



Need 16 rectangles

Example of VC-Dimension

- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$

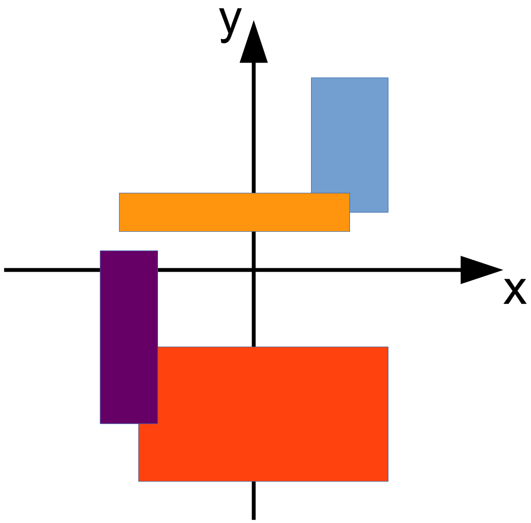


- Shattering 5 points?

Example of VC-Dimension

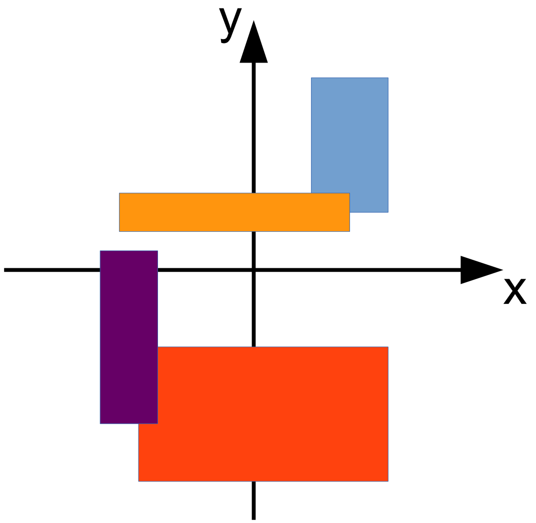
- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$

- Shattering 5 points: impossible

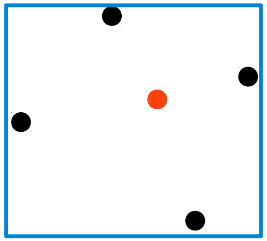


Example of VC-Dimension

- B = $\mathbb{R}^2$
- R = all axis-aligned rectangles in $\mathbb{R}^2$



- Shattering 5 points: impossible
  - Take any 5 points
  - One of them that is contained in all rectangles containing the other four
  - Impossible to find a rectangle containing only the other four
- VC(B,R)=4



Vapnik-Chervonenkis Dimension

- Combinatorial property of a collection of subsets from a domain
- Measures the “richness”, “expressivity”
- If we know the VC-dim of a collection of subsets, we can compute the minimum sample size needed to approximate the sizes of the ranges using a sample

Approximating sizes of ranges

- Sample Theorem:
  - Let (B,R) have VC(B,R) d. Given $\varepsilon, \delta \in [0,1]$, let S be a collection of points from B sampled uniformly at random. If
  $$|S| \geq \frac{1}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)$$
  then,
  $$\Pr\left(\exists F \in R \ : \ \left|\frac{|F|}{|B|} - \frac{|F \cap S|}{|S|}\right| > \varepsilon\right) \leq \delta$$
- Can approximate sizes of all $F \in R$ simultaneously
  - No need of union bound

VC-Dimension for Frequent Itemsets

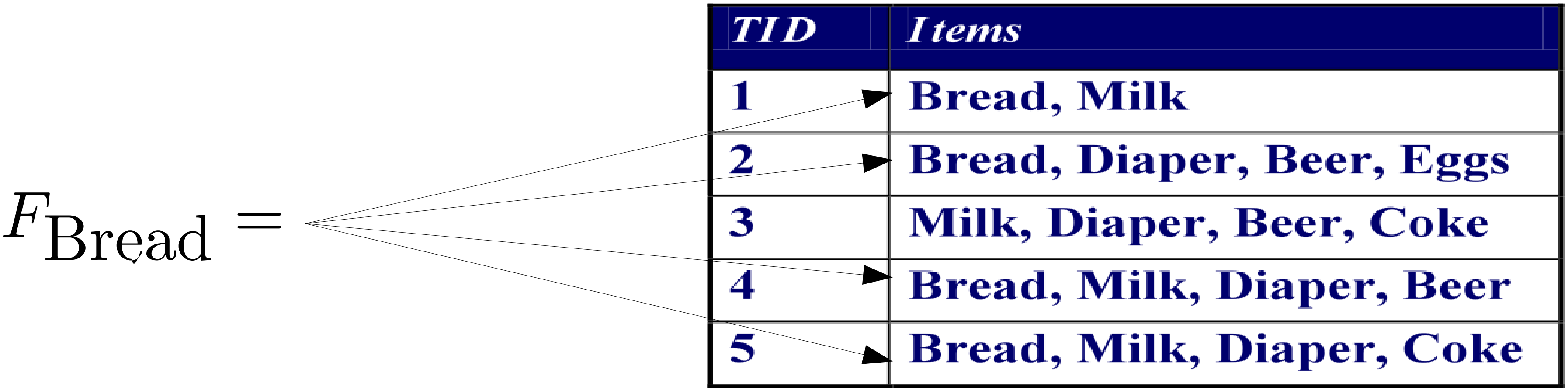
- B = dataset D (set of transactions)

VC-Dimension for Frequent Itemsets

- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X

VC-Dimension for Frequent Itemsets

- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X

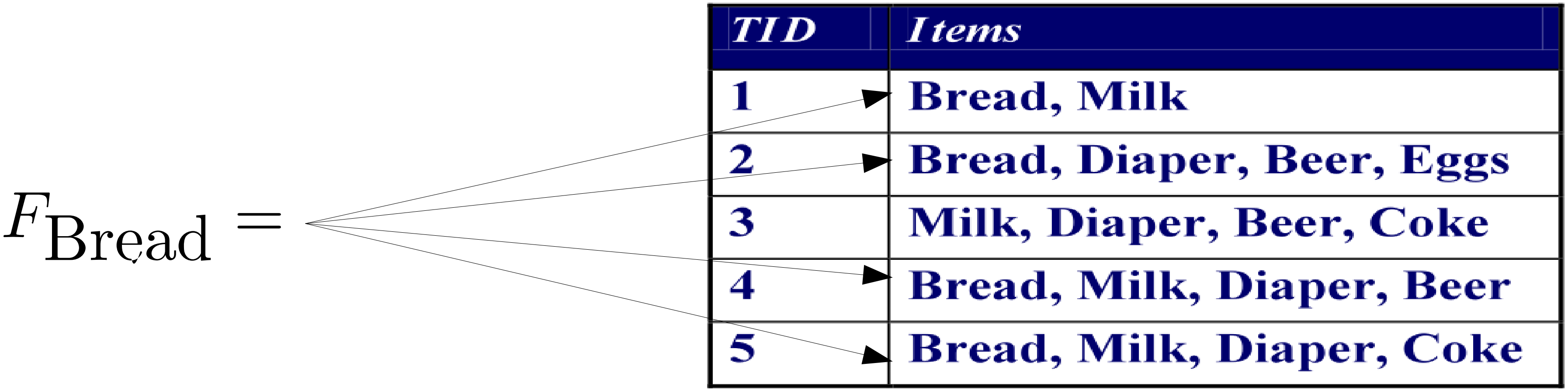


$F_{\text{Bread}} =$

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

VC-Dimension for Frequent Itemsets

- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X
- $R = \{F_X, \forall X\}$

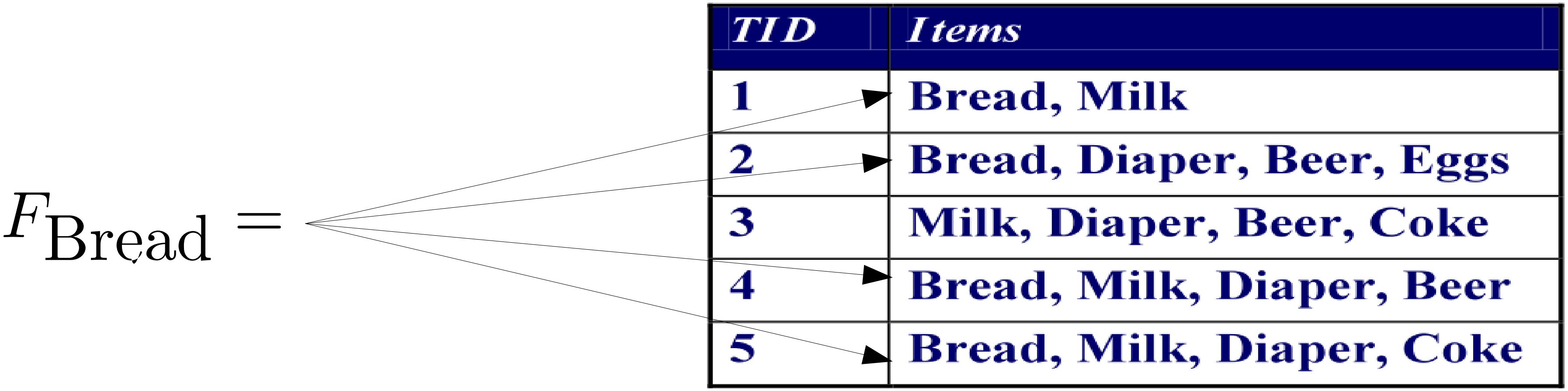


$F_{\text{Bread}} =$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

VC-Dimension for Frequent Itemsets

- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X
- $R = \{F_X, \forall X\}$



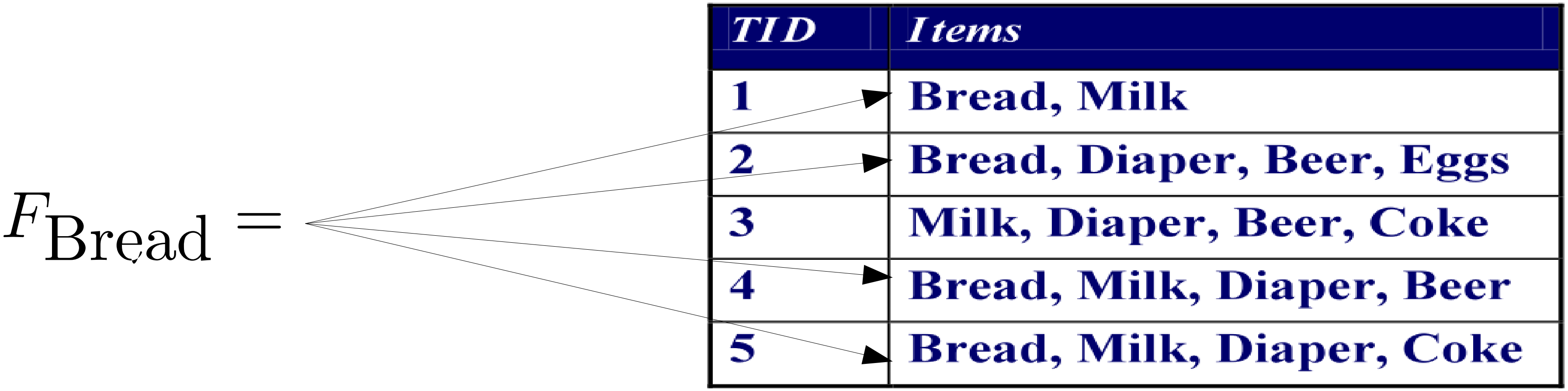
$$F_{\text{Bread}} =$$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$$\boxed{f_D(X) = \frac{|F_X|}{|B|}, f_S(X) = \frac{|F_X \cap S|}{|S|}}$$

VC-Dimension for Frequent Itemsets

- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X
- $R = \{F_X, \forall X\}$



$F_{\text{Bread}} =$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$$\boxed{f_D(X) = \frac{|F_X|}{|B|}, f_S(X) = \frac{|F_X \cap S|}{|S|}}$$

To approximate using the sample theorem, need upper bound to VC(B,R)

Upper Bound to VC-Dim for FI's

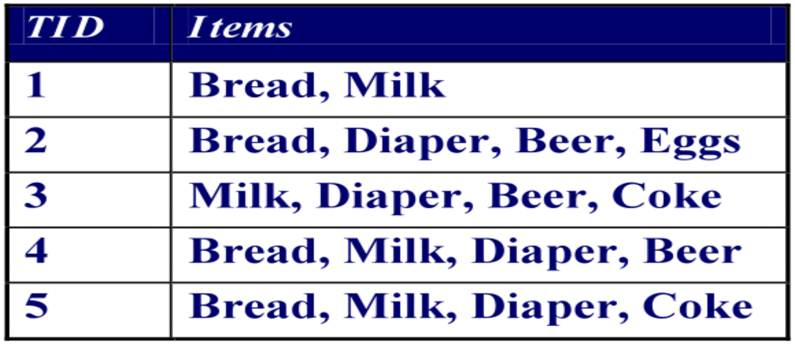
- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X
- $R = \{F_X, \forall X\}$
- Theorem:
  - Let d be the maximum integer such that D contains at least d transactions of length at least d. Then

VC(B,R)   d

Upper Bound to VC-Dim for FI's

- B = dataset D (set of transactions)
- Itemset X, F_X = transactions of D containing X
- $R = \{F_X, \forall X\}$
- Theorem:
  - Let d be the maximum integer such that D contains at least d transactions of length at least d. Then

VC(B,R) d



TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

VC(B,R) ≤ 4

d can be computed with a single linear scan of dataset (or even online)

Choosing the right sample size

- Theorem
  - Let $\varepsilon, \delta \in [0,1]$
  - Let D be a dataset and d be the max integer such that D contains at least d transactions of length at least d
  - Let S be a collection of transactions of D sampled independently and uniformly at random, with

$$|S| \geq \frac{4}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)$$

  - Then

$$\Pr(\exists X \ : \ |f_S(X) - f_D(X)| \geq \varepsilon/2) \leq \delta$$

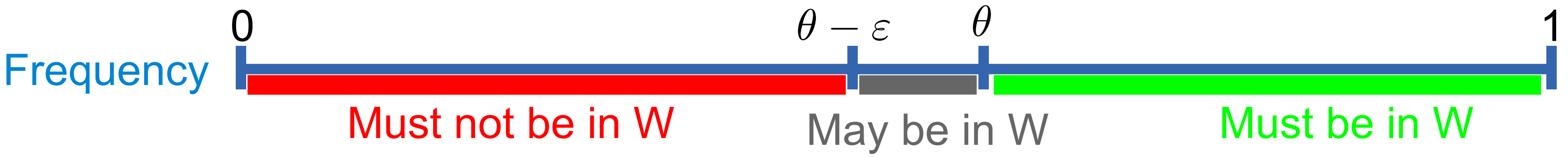
Algorithm

- To extract approximate collection of freq itemsets:
  1. Compute d for the dataset D
  2. Compute sample size given d
  3. Create random sample S from D
  4. Extract Freq Itemsets from S using $\theta' = \theta - \frac{\varepsilon}{2}$

Algorithm

- To extract approximate collection of Freq. Itemsets:
  1. Compute d for the dataset D
  2. Compute sample size given d
  3. Create random sample S from D
  4. Extract Frequent Itemsets from S using $\theta' = \theta - \frac{\varepsilon}{2}$

- Theorem: With probability at least $1-\delta$, the returned set W of itemsets satisfies the desired property



A closer look...

- Expression of sample size: $|S| \geq \frac{4}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)$

A closer look...

- Expression of sample size: $|S| \geq \frac{4}{\varepsilon^2}\left(d + \log \frac{1}{\delta}\right)$

- d: the maximum integer such that D contains at least d transactions of length at least d

A closer look...

- Expression of sample size: $|S| \geq \frac{4}{\varepsilon^2}\left(d + \log \frac{1}{\delta}\right)$
- d: the maximum integer such that D contains at least d transactions of length at least d
  - does not depend on |D|
  - does not depend on θ
  - does not depend on number of itemsets

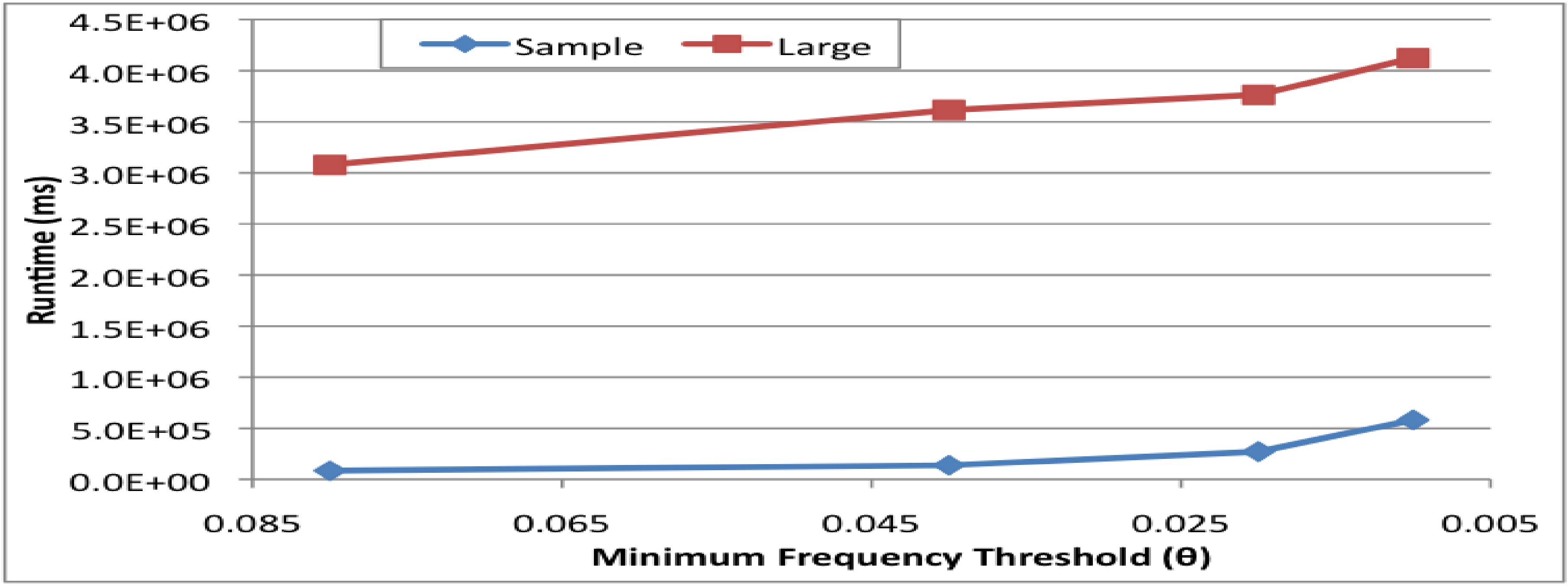
A closer look...

- Expression of sample size: $|S| \geq \frac{4}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)$

- d: the maximum integer such that D contains at least d transactions of length at least d
  - does not depend on |D|
  - does not depend on θ
  - does not depend on number of itemsets
- Then sample size does not depend on these factors
  - The time to mine S does not depend on |D| !!!
    - Velocity VS Volume challenge addressed!

Experiments

- Sample always fits into main memory
- Output always satisfies required approx. guarantees
  - Frequency accuracy even better than guaranteed
- Mining time significantly improved



Recap

We showed how VC-dimension, a concept from statistical learning theory, can help in developing an efficient algorithm to approximate the collection of frequent itemsets, addressing one of the Big Data challenges through sampling

Joint work with Eli Upfal. Published in ECML PKDD 2012

Outline

- Introduction ✓
  - Thesis statement ✓
- Mining Frequent Itemsets through sampling ✓
- A statistical test for True Frequent Itemsets
- Proposed work
  - Efficient progressive sampling for frequent itemset mining
  - Graph mining: problems and challenges

Data as a sample

- Recall the Market Basket Analysis motivation
  - You own a grocery store
  - You collect lists of products bought by each costumer
  - You want to know the groups of products that are sold together the most

Data as a sample

- Recall the Market Basket Analysis motivation
  - You own a grocery store
  - You collect lists of products bought by each costumer
  - You want to know the groups of products that are sold together the most
- Transactions collected in a specific day do not fully describe the unknown generating process
  - There are fluctuations: not everyone is your “average” customer
- You want to know the high-selling groups of products in the long term, not just on a specific day

Convergence to average

- How to solve this?
- You could compute Frequent Itemsets over the transactions collected in multiple days
- We expect aggregation over multiple days to converge to the average…
  - … at the cost of more transactions to process
  - … and how many days should we collect transactions for?
- Can we avoid this cost?

True Frequent Itemsets

- Formal settings and problem:
  - $\mathcal{I}$ = set of items
  - p(): distribution over $2^{\mathcal{I}}$ (generating process)

True Frequent Itemsets

- Formal settings and problem:
  - $\mathcal{I}$ = set of items
  - p(): distribution over $2^{\mathcal{I}}$ (generating process)
  - true frequency of itemset X:

$$r_p(X) = \sum_{\tau \in 2^{\mathcal{I}}, X \subseteq \tau} p(\tau)$$

True Frequent Itemsets

- Formal settings and problem:
  - $\mathcal{I}$ = set of items
  - p(): distribution over $2^{\mathcal{I}}$ (generating process)
  - true frequency of itemset X:

$$r_p(X) = \sum_{\tau \in 2^{\mathcal{I}}, X \subseteq \tau} p(\tau)$$

  - dataset D: collection of i.i.d. samples from p
  - $f_D(X)$ = frequency of X in D, $\mathbf{E}[f_D(X)] = r_p(X)$

True Frequent Itemsets

- Formal settings and problem:
  - $\mathcal{I}$ = set of items
  - p(): distribution over $2^{\mathcal{I}}$ (generating process)
  - true frequency of itemset X:

$$r_p(X) = \sum_{\tau \in 2^{\mathcal{I}}, X \subseteq \tau} p(\tau)$$

  - dataset D: collection of i.i.d. samples from p
  - $f_D(X)$ = frequency of X in D, $\mathbf{E}[f_D(X)] = r_p(X)$
- True Frequent Itemsets with respect to θ:

$$\{X \ : \ r_p(X) \geq \theta\}$$

Issues and Goal

- Issues:
  - $r_p()$ and p() are unknown
  - $f_D(X) \approx r_p(X)$
    - … but how well?
    - can be greater, can be smaller
- Goal:
  - Identify (almost) all and only the True Frequent Itemsets
- How:
  - Develop a statistical test to identify TrueFrequent Itemsets with few false positives and few false negatives

Statistical testing

- A statistical test is a procedure to accept or reject an hypothesis H using data:

Statistical testing

- A statistical test is a procedure to accept or reject an hypothesis H using data:
  - Define an acceptance region A
  - Compute a statistic s_H related to H from the data
  - If $s_H \in A$ accept H, otherwise reject

Statistical testing

- A statistical test is a procedure to accept or reject an hypothesis H using data:
  - Define an acceptance region A
  - Compute a statistic s_H related to H from the data
  - If $s_H \in A$ accept H, otherwise reject
- In our case: for all itemsets X
  - H_X = "Itemset X has $r_p(X) < \theta$"
  - $A = [0, \theta + \varepsilon]$
  - $s_{H_X} = f_D(X)$
  - If $f_D(X) < \theta + \varepsilon$ accept H_X, else mark X as TFI

Failure and power

- A test may fail in two ways
    1. Reject a “true” hypothesis
    2. Accept a “false” hypothesis

Failure and power

- A test may fail in two ways
  1. Reject a “true” hypothesis
  2. Accept a “false” hypothesis
- Acceptance region A chosen so that

$$\text{Pr(test rejects “true” hypothesis)} \quad \delta$$

Failure and power

- A test may fail in two ways
  1. Reject a “true” hypothesis
  2. Accept a “false” hypothesis
- Acceptance region A chosen so that

  Pr(test rejects “true” hypothesis) δ

- Statistical Power of a test:

  1 – Pr(test accept “false” hypothesis)

  - Goal is maximize power
  - difficult to evaluate analytically, usually done experimentally

Multiple hypotheses testing

- We have one hypothesis for each itemset X
  - H_X: "Itemset X has $r_p(X) < \theta$ "
  - Exponential number of hypotheses

Multiple hypotheses testing

- We have one hypothesis for each itemset X
  - H_X: “Itemset X has $r_p(X) < \theta$ ”
  - Exponential number of hypotheses
- We want to make sure that

$$\Pr(\exists Y \text{ with } r_p(Y) < \theta \ : \ H_Y \text{ rejected}) \leq \delta$$

Multiple hypotheses testing

- We have one hypothesis for each itemset X
  - H_X: "Itemset X has $r_p(X) < \theta$ "
  - Exponential number of hypotheses
- We want to make sure that

$$\Pr(\exists Y \text{ with } r_p(Y) < \theta \ : \ H_Y \text{ rejected}) \leq \delta$$

- i.e. we want to control the Family-Wide Error Rate
  - FWER: probability of rejecting a true hypothesis among those to be tested

Controlling the FWER

- Traditionally done through the Bonferroni Correction (Union Bound):
  - Choose new acceptance region so that

    Pr(test rejects true hypothesis) $\delta/n \longleftarrow$ # of hypotheses to be tested

Controlling the FWER

- Traditionally done through the Bonferroni Correction (Union Bound):
  - Choose new acceptance region so that

    Pr(test rejects true hypothesis) δ/n ← # of hypotheses to be tested
- Unsuitable for Big Data problems:
  - Loose in the case of correlated hypotheses
    - Our case
  - Does not scale well with number of hypotheses
    - Number of itemsets is exponential in number of items
  - Acceptance region too large   small power

Our Goal

- Develop statistical test to identify TFI's with FWER at most δ

Our Goal

- Develop statistical test to identify TFI's with FWER at most δ
- In our test, for each itemset X, we ask:

  “Is $f_D(X) < \theta + \varepsilon$?”

  - If Yes: accept hypothesis $r_p(X) < \theta$
  - If No: reject hypothesis. Mark X as TFI

Our Goal

- Develop statistical test to identify TFI's with FWER at most δ
- In our test, for each itemset X, we ask:

  "Is $f_D(X) < \theta + \varepsilon$?"

  - If Yes: accept hypothesis $r_p(X) < \theta$
  - If No: reject hypothesis. Mark X as TFI

- To have FWER δ, we need to find ε such that

$$\Pr(\exists Y \text{ with } r_p(Y) < \theta \text{ s.t. } f_D(Y) \geq \theta + \varepsilon) < \delta$$

Finding ε

- ε should be the minimum possible to guarantee the FWER, in order to maximize the statistical power
  - Competing goals

Finding ε

- ε should be the minimum possible to guarantee the FWER, in order to maximize the statistical power
  - Competing goals
- We developed a two-phases algorithm to find ε

1. Find ε' such that all TFI's have frequency in the dataset $\geq \theta - \varepsilon'$
2. Using ε', find $\varepsilon'' < \varepsilon'$ such that all non-TFI's have frequency in the dataset $\leq \theta + \varepsilon''$

“Backwards” sample theorem

- Recall the sample theorem:
- Let (B,R) have VC(B,R) d. Given $\varepsilon, \delta \in [0,1]$, let S be a collection of points from B sampled uniformly at random. If

$$|S| \geq \frac{1}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)$$

then,

$$\Pr\left(\exists F \in R \ : \ \left|\frac{|F|}{|B|} - \frac{|F \cap S|}{|S|}\right| > \varepsilon\right) \leq \delta$$

“Backwards” sample theorem

- Recall the sample theorem:
- Let (B,R) have VC(B,R)   d. Given $\varepsilon, \delta \in [0,1]$, let S be a collection of points from B sampled uniformly at random. If

$$|S| \geq \frac{1}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)$$

then,

$$\Pr\left(\exists F \in R \;:\; \left|\frac{|F|}{|B|} - \frac{|F \cap S|}{|S|}\right| > \varepsilon\right) \leq \delta$$

- Can be used “backwards”: given $|S|$, d, and δ , compute ε for which the above holds

First phase

- The dataset D is a random sample!

First phase

- The dataset D is a random sample!
- Define a range space (B,R)

First phase

- The dataset D is a random sample!
- Define a range space (B,R)
- B = $2^{\mathcal{I}}$ (all possible transactions)
- R = $\{F_Y \ : \ Y \in 2^{\mathcal{I}}\}$
  - $F_Y = \{Z \in 2^{\mathcal{I}} \ : \ Y \subseteq Z\}$
- VC(B,R) $\leq |\mathcal{I}| - 1$ (# of items - 1)

First phase

- The dataset D is a random sample!
- Define a range space (B,R)
- B = $2^{\mathcal{I}}$ (all possible transactions)
- R =$\{F_Y \ : \ Y \in 2^{\mathcal{I}}\}$
  - $F_Y = \{Z \in 2^{\mathcal{I}} \ : \ Y \subseteq Z\}$
- VC(B,R) $\leq |\mathcal{I}| - 1$ (# of items - 1)
- Given |D|, $|\mathcal{I}|$ and δ, we can compute ε' such that

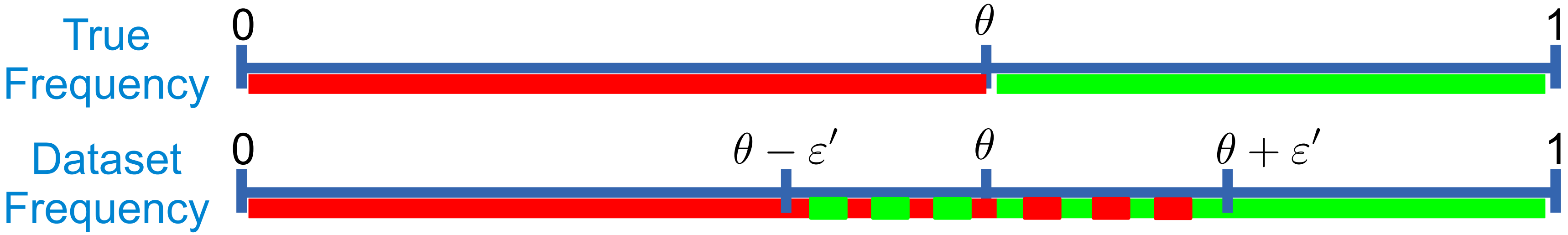
$$\Pr(\exists Y \ : \ |r_p(Y) - f_D(Y)| > \varepsilon') < \delta$$

Implications

- This means that, with probability at least $1 - \delta$
  - for all TFI's X we have $f_D(X) \geq \theta - \varepsilon'$
  - for all Y with $r_p(Y) < \theta$ we have $f_D(Y) < \theta + \varepsilon'$

Implications

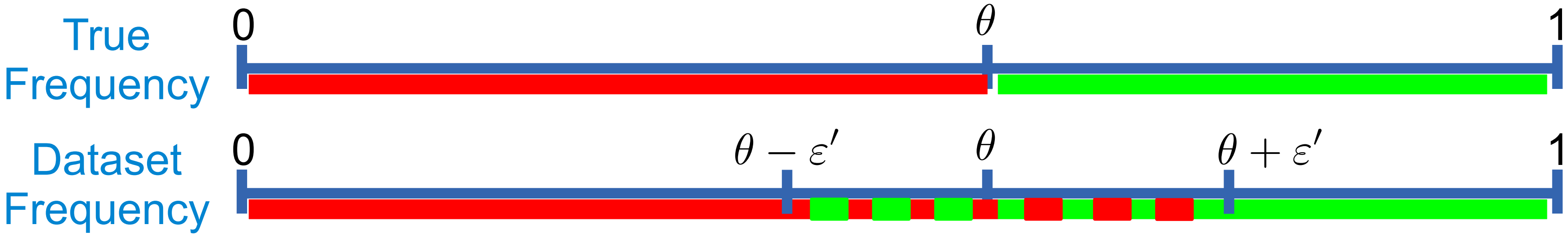
- This means that, with probability at least $1-\delta$
  - for all TFI's X we have $f_D(X) \geq \theta - \varepsilon'$
  - for all Y with $r_p(Y) < \theta$ we have $f_D(Y) < \theta + \varepsilon'$



- We could use ε' in our statistical test!

Implications

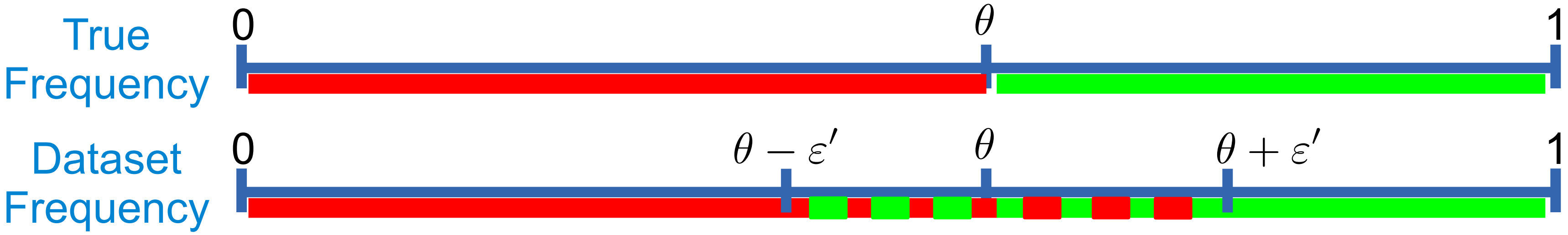
- This means that, with probability at least $1 - \delta$
  - for all TFI's X we have $f_D(X) \geq \theta - \varepsilon'$
  - for all Y with $r_p(Y) < \theta$ we have $f_D(Y) < \theta + \varepsilon'$



- We could use ε' in our statistical test!
  - Why don't we?

Implications

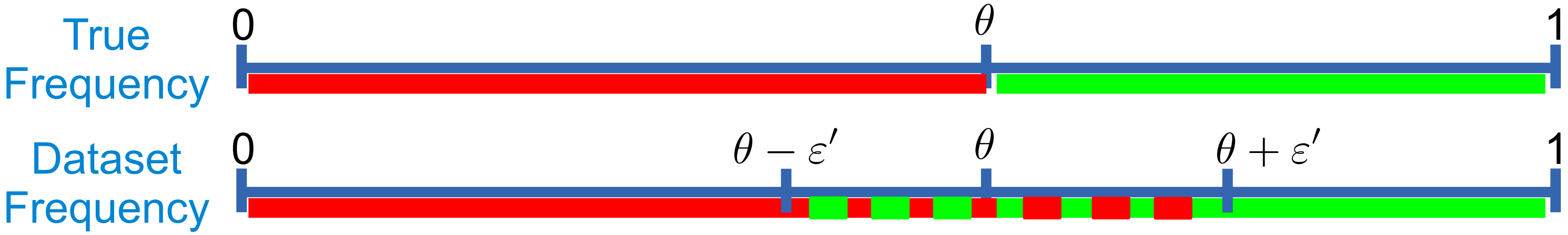
- This means that, with probability at least $1 - \delta$
  - for all TFI's X we have $f_D(X) \geq \theta - \varepsilon'$
  - for all Y with $r_p(Y) < \theta$ we have $f_D(Y) < \theta + \varepsilon'$



- We could use ε' in our statistical test!
  - Why don't we?
    - Statistical Power! We want the minimum ε

Implications

- This means that, with probability at least $1-\delta$
  - for all TFI's X we have $f_D(X) \geq \theta - \varepsilon'$
  - for all Y with $r_p(Y) < \theta$ we have $f_D(Y) < \theta + \varepsilon'$



- We could use ε' in our statistical test!
  - Why don't we?
    - Statistical Power! We want the minimum ε
    - Can we compute a $\varepsilon'' < \varepsilon'$ for which

$$\Pr\left(\exists Y \text{ with } r_p(Y) < \theta \; : \; |r_p(Y) - f_D(Y)| > \varepsilon''\right) < \delta?$$

Second phase

- Consider the set of itemsets

$$A_{\varepsilon'} = \{Z \ : \ \theta - \varepsilon' \leq f_D(Z) \leq \theta + \varepsilon'\}$$

Second phase

- Consider the set of itemsets

$$A_{\varepsilon'} = \{Z \ : \ \theta - \varepsilon' \leq f_D(Z) \leq \theta + \varepsilon'\}$$

- Take the negative border $N_{\varepsilon'}$ of $A_{\varepsilon'}$
  - $N_{\varepsilon'}$= itemsets s.t. all their subsets are in $A_{\varepsilon'}$
- Let $E_{\varepsilon'} = A_{\varepsilon'} \cup N_{\varepsilon'}$

Second phase

- Consider the set of itemsets

$$A_{\varepsilon'} = \{Z \ : \ \theta - \varepsilon' \leq f_D(Z) \leq \theta + \varepsilon'\}$$

- Take the negative border $N_{\varepsilon'}$ of $A_{\varepsilon'}$
  - $N_{\varepsilon'}$= itemsets s.t. all their subsets are in $A_{\varepsilon'}$
- Let $E_{\varepsilon'} = A_{\varepsilon'} \cup N_{\varepsilon'}$
- Lemma:

  If there is ε'' s.t. $|r_p(Z) - f_D(Z)| \leq \varepsilon'' \forall Z \in E_{\varepsilon'}$

  then for any Y s.t. $r_p(Y) < \theta$ we have

$$\boxed{f_D(Y) < \theta + \varepsilon''}$$

Second phase

- Consider the set of itemsets

$$A_{\varepsilon'} = \{Z \ : \ \theta - \varepsilon' \leq f_D(Z) \leq \theta + \varepsilon'\}$$

- Take the negative border $N_{\varepsilon'}$ of $A_{\varepsilon'}$
  - $N_{\varepsilon'}$= itemsets s.t. all their subsets are in $A_{\varepsilon'}$
- Let $E_{\varepsilon'} = A_{\varepsilon'} \cup N_{\varepsilon'}$
- Lemma:

  If there is ε'' s.t. $|r_p(Z) - f_D(Z)| \leq \varepsilon'' \forall Z \in E_{\varepsilon'}$

  then for any Y s.t. $r_p(Y) < \theta$ we have

$$\boxed{f_D(Y) < \theta + \varepsilon''}$$

- Question: find ε'' s.t. $|r_p(Z) - f_D(Z)| \leq \varepsilon'' \forall Z \in E_{\varepsilon'}$

Second phase

- Consider the range space ($2^{\mathcal{I}}$,$R_{\varepsilon'}$) where

$$R_{\varepsilon'} = \{F_Y, Y \in N_{\varepsilon'}\}$$

(F_Y = transactions from $2^{\mathcal{I}}$ containing Y)

Second phase

- Consider the range space ($2^{\mathcal{I}}$,$R_{\varepsilon'}$) where

$$R_{\varepsilon'} = \{F_Y, Y \in N_{\varepsilon'}\}$$

(F_Y = transactions from $2^{\mathcal{I}}$ containing Y)

- Bound VC($2^{\mathcal{I}}$, $R_{\varepsilon'}$) by solving an Anti-chain Set-Union Knapsack Problem
  - Probably NP but CPLEX solves it easily

Second phase

- Consider the range space ($2^{\mathcal{I}}$,$R_{\varepsilon'}$) where
$$R_{\varepsilon'} = \{F_Y, Y \in N_{\varepsilon'}\}$$
(F_Y = transactions from $2^{\mathcal{I}}$ containing Y)
- Bound VC($2^{\mathcal{I}}$, $R_{\varepsilon'}$) by solving an Anti-chain Set-Union Knapsack Problem
  - Probably NP but CPLEX solves it easily
- Apply the “backwards” sample theorem again to get ε''
- $\Pr\left(\exists Y \in N_{\varepsilon'} \ : \ f_D(Y) \geq \theta + \varepsilon''\right) < \delta$

Second phase

- Consider the range space ($2^{\mathcal{I}}$,$R_{\varepsilon'}$) where

$$R_{\varepsilon'} = \{F_Y, Y \in N_{\varepsilon'}\}$$

(F_Y = transactions from $2^{\mathcal{I}}$ containing Y)

- Bound VC($2^{\mathcal{I}}$, $R_{\varepsilon'}$) by solving an Anti-chain Set-Union Knapsack Problem
  - Probably NP but CPLEX solves it easily
- Apply the “backwards” sample theorem again to get ε''
- $\Pr\left(\exists Y \in N_{\varepsilon'} \ : \ f_D(Y) \geq \theta + \varepsilon''\right) < \delta$
- $\Pr\left(\exists Y \text{ with } r_p(Y) < \theta \ : \ f_D(Y) \geq \theta + \varepsilon''\right) < \delta$ ←

We can control the FWER

Second phase

- Consider the range space ($2^{\mathcal{I}}$,$R_{\varepsilon'}$) where

$$R_{\varepsilon'} = \{F_Y, Y \in N_{\varepsilon'}\}$$

(F_Y = transactions from $2^{\mathcal{I}}$ containing Y)

- Bound VC($2^{\mathcal{I}}$, $R_{\varepsilon'}$) by solving an Anti-chain Set-Union Knapsack Problem
  - Probably NP but CPLEX solves it easily
- Apply the “backwards” sample theorem again to get ε''
- $\Pr\left(\exists Y \in N_{\varepsilon'} \ : \ f_D(Y) \geq \theta + \varepsilon''\right) < \delta$
- $\Pr\left(\exists Y \text{ with } r_p(Y) < \theta \ : \ f_D(Y) \geq \theta + \varepsilon''\right) < \delta$ ← We can control the FWER
- In the statistical test, compare $f_D(X)$ to $\theta + \varepsilon''$

Recap

We developed a statistical test to identify True Frequent Itemsets which controls the Family-Wide Error Rate and whose acceptance region is not dependent on the number of hypotheses tested

Outline

- Introduction ✓
  - Thesis statement ✓
- Mining Frequent Itemsets through sampling ✓
- A statistical test for True Frequent Itemsets ✓
- Proposed work
  - Efficient progressive sampling for Frequent Itemset mining
  - Graph mining: problems and challenges

Progressive sampling algorithm for Frequent Itemsets mining

Progressive sampling algorithm for Frequent Itemsets mining

- First part of the talk: algorithm to mine frequent itemsets with single random sample
  - assumes worst case scenario sample size large enough to accommodate it

Progressive sampling algorithm for Frequent Itemsets mining

- First part of the talk: algorithm to mine frequent itemsets with single random sample
  - assumes worst case scenario sample size large enough to accommodate it
- More reasonable:
  - start from a small sample, check stopping condition expressing convergence/stability, enlarge sample, loop progressive sampling
  - Use info from data smaller final sample size

Progressive sampling algorithm for Frequent Itemsets mining

- Key issue: develop “good” stopping condition
  - evaluate fast
  - stop as early as possible

Progressive sampling algorithm for Frequent Itemsets mining

- Key issue: develop “good” stopping condition
  - evaluate fast
  - stop as early as possible
- Statistical Learning Theory to the rescue
  - Data-dependent sample complexity bounds
    - derived from Rademacher averages
  - Like VC-Dimension but only need info from sample

Progressive sampling algorithm for Frequent Itemsets mining

- Key issue: develop “good” stopping condition
  - evaluate fast
  - stop as early as possible
- Statistical Learning Theory to the rescue
  - Data-dependent sample complexity bounds
    - derived from Rademacher averages
  - Like VC-Dimension but only need info from sample
- Stopping rule will use info on the entire distribution of transactions lengths in the sample
- We expect very fast convergence

Graph mining

Graph mining

- Graphs are everywhere
  - Web, Internet, social networks, protein networks
  - They are huge: 10^7 nodes, 10^8 edges(sparse)
- Many problems on graphs:
  - Finding interesting subgraphs (motifs)
  - Measure properties (e.g. vertex/edge centralities)
  - Summarizing graphs (graph kernels)
  - Problems on graph sequences
  - Problems on evolving graphs

Graph mining and sampling

- Many open questions about the use of sampling and statistical validation for graphs problems
  - How to efficiently sample subgraphs from a graph?
  - How centrality measures and interestingness change as effect to sampling?
  - How much should we sample to obtain a good approximation?
  - What should we sample?
    - Nodes, vertices, induced subgraphs, …
  - What are good models to take in order to assess the statistical validity of results?

Graph mining

We are looking at these and similar questions to develop algorithms that can take graph mining up to speed and address the challenges posed by Big Data

Timeline

- Spring '13: Graph mining
- Summer '13: Internship at Yahoo! Research Barcelona, Web Mining Group
- Fall '13: Progressive sampling algorithm for frequent itemsets mining
- Spring '14: Dissertation writing

Conclusions

It is possible to use tools from Statistical Learning Theory to develop efficient and scalable approximation algorithms for data analysis problems, addressing the challenges posed by Big Data.

We propose to continue on this line of research to explore other problems using different and more recent tools from Statistical Learning Theory.

Publications

- Thesis related:
  - Riondato, Vandin. *Finding the True Frequent Itemsets.* Under submission
  - Riondato, Upfal. *Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees*. ECML PKDD 2012
  - Riondato, DeBrabant, Fonseca, Upfal. *PARMA: A Parallel Randomized Algorithm for Approximate Association Rules Mining in MapReduce*. CIKM 2012
  - Riondato, Akdere, Çetintemel, Zdonik, Upfal. *The VC-Dimension of SQL Queries and Selectivity Estimation Through Sampling*. ECML PKDD 2011
- Others:
  - Pietracaprina, Pucci, Riondato, Silvestri, Upfal. *Space-Round Tradeoffs for MapReduce Computations*. ICS 2012
  - Akdere, Çetintemel, Riondato, Upfal, Zdonik. *Learning-based Query Performance Modeling and Prediction*. ICDE 2012
  - Akdere, Çetintemel, Riondato, Upfal, Zdonik. *The Case for Predictive Database Systems: Opportunities and Challenges*. CIDR 2011
  - Pietracaprina, Riondato, Upfal, Vandin. *Mining Top-K Frequent Itemsets Through Progressive Sampling.* DMKD 21(2), 2010

The End

Please ask questions

Proof (Intuition)

- For a set of k transactions to be shattered, each transaction must appear in 2^(k-1) different F_X 's where X is an itemset
- A transaction only appears in the F_X's of the itemsets X it contains
- A transaction of length w contains 2^(w)-1 itemsets
- Need w k for the transaction to belong to a shattered set of size k
- To shatter k transactions they must all have length   k
- Max k for which it happens is upper bound to VC-Dim