Sampling for Frequent Itemset Mining
The Power of PAC learning

prof. dr Arno Siebes

Algorithmic Data Analysis Group
Department of Information and Computing Sciences
Universiteit Utrecht

The Papers

Today we discuss:

- Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees
  - Proc ECML PKDD 2012, LNCS 7523
  - ACM Transactions on Knowledge Discovery from Data, Vol 8, No 4, 2014
- Mining Frequent Itemsets through Progressive Sampling with Rademacher Averages
  - Proc. KDD 2015

Note, for the first paper(s) we restrict ourselves to frequent itemsets only

The Problem

Let D be a transaction database over $\mathcal{I}$ (which we assume fixed throughout the lecture). Denote by $F(D, \theta)$ the set of frequent itemsets, i.e.,

- $I \in F(D, \theta) \Leftrightarrow supp_D(I) \geq \theta$.

We want to compute a (small) sample $S \subset D$ such that

- $F(S, \theta') \approx F(D, \theta)$

More formally, we want that our sample to yield an (ϵ, δ) approximation:

For $(\epsilon, \delta) \in (0, 1)^2$, an (ϵ, δ) approximation of $F(D, \theta)$ is a set $C = \{(I, s(I)) \mid I \in \mathcal{I}, s(I) \in (0, 1]\}$ such that with probability at least $1 - \delta$:

1. $(I, supp_D(I)) \in F(D, \theta) \rightarrow (I, s(I)) \in C$
2. $(I, s(I)) \in C \rightarrow supp_D(I) \geq \theta - \epsilon$
3. $(I, s(I)) \in C \rightarrow |supp_D(I) - s(I)| \leq \epsilon/2$

Now in Natural Language

C is an (ϵ, δ) approximation of $F(D, \theta)$ if with probability at least $1 - \delta$:

- C contains all frequent itemsets
- the non-frequent itemsets in C are almost frequent
- our "guess "of the frequency of the itemsets in C is not too far off.

This means that (with high probability)

- C may contain false positives
- but *no* false negatives
- and there is some loss of accuracy

Which means that we can compute $F(D, \theta)$

- with *one* scan over D using C.
- an advantage over Toivonen's scheme

In Terms of Samples

In the terminology of samples we can rephrase our goal now as

Find a sample S such that

$$\mathbb{P}\left(\exists I \in \mathcal{I} : |supp_D(I) - supp_S(I)| > \epsilon/2\right) < \delta$$

Because if this holds for S, then

- $F(S, \theta - \epsilon/2)$ is an (ϵ, δ) approximations of $F(D, \theta)$

For the simple reasons that

1. with probability at least $1 - \delta$: $F(D, \theta) \subset F(S, \theta - \epsilon/2)$
2. $I \in F(S, \theta - \epsilon/2) \rightarrow$
   [with probability at least $1 - \delta$: $supp_D(I) \geq supp_S(I) - \epsilon/2 \geq \theta - \epsilon$]
3. $I \in F(S, \theta - \epsilon/2) \rightarrow$
   [with probability at least $1 - \delta$: $|supp_D(I) - supp_S(I)| \leq \epsilon/2$]

How Big Should $|S|$ be?

For each itemset I, the Chernoff bound is

$$\mathbb{P}\left(|supp_D(I) - supp_S(I)| \geq \epsilon/2\right) \leq 2e^{-|S|\epsilon^2/12}$$

Since there are a priori $2^{|\mathcal{I}|}$ frequent itemsets, the union bound gives us

$$\mathbb{P}\left(\forall I : |supp_D(I) - supp_S(I)| \geq \epsilon/2\right) \leq 2^{|\mathcal{I}|}2e^{-|S|\epsilon^2/12}$$

So, to have this probability less then δ we need

$$|S| \geq \frac{12}{\epsilon^2}\left(|\mathcal{I}| + \ln(2) + \ln\left(\frac{1}{\delta}\right)\right)$$

Since $\mathcal{I}$ tends to be very large

- think of all the goods Amazon sells

this is not a very good bound

- we should be able to do better using PAC learning!

Changing Concepts

The authors use range spaces rather than classifiers;

A range space (X, R) consists of

- a finite or infinite set of points X
- a finite of infinite family R of subsets of X, called ranges

For any $A \subset X$, the projection of R on A is given by

$$\Pi_R(A) = \{r \cap A \mid r \in R\}$$

The relation between the two approaches is simple

- each $r \in R$ has an indicator function $\mathbb{1}_r$
  - a function that assigns 1 to $x \in r$
  - and 0 to $x \notin r$
- which is a classifier on X

Thus a subset $A \subset X$ is shattered by R if $\Pi_R(A) = P(A)$

ϵ-Approximations

Our ϵ representative samples turn up as ϵ- approximations:

Let (X, R) be a range space and let $A \subset X$ be a finite subset. For $\epsilon \in (0, 1)$, a $B \subset A$ is an ϵ-approximation for A if

$$\forall r \in R : \left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \epsilon$$

without proof (it should look and feel familiar):

There is a constant c (≤ 0.5 experimentally) such that if (X, R) is a range space of VC dimension at most υ, $A \subset X$ a finite subset, $(\epsilon, \delta) \in (0, 1)^2$, then a random $B \subset A$ of cardinality

$$m \geq \min \left\{ |A|, \frac{c}{\epsilon^2} \left(\upsilon + \log \frac{1}{\delta} \right) \right\}$$

is an ϵ-approximation for A with probability at least $1 - \delta$

From Data to Range Space

To use this bound, we should first turn our data into a range space. Which is done as follows:

Let D be a transaction database over $\mathcal{I}$. The associated range space $S = (X, R)$ is defined by

1. $X = D$, the set of transactions
2. $R = \{T_D(I) \mid I \subseteq \mathcal{I}, I \neq \emptyset\}$, where $T_D(I) = \{t \in D \mid I \subseteq t\}$
   - the range of I, $T_D(I)$, is the set of transactions that support I
   - i.e., its support set.

Note that R contains exactly the support sets of the closed itemsets

- the support set of a non-closed itemset is the support set of a closed itemset as well.

Now all we have to do is to determine the VC dimension of this range set.

- as more often, we'll settle for a (tight) upper bound.

A First Result

Let D be a dataset with associated range space $S = (X, R)$. Then $VC(S) \geq \upsilon \in \mathbb{N}$ if there exists an $A \subset X$ with $|A| = \upsilon$ such that

$$\forall B \subset A : \exists I_B \subset \mathcal{I} : T_A(I_B) = B$$

$\Leftarrow$: For all $B \subset A$, $T_D(I_B) \cap A = B$, that means that $B \in \Pi_R(A)$ for all B. That means that $\Pi_R(A) = P(A)$. A is shattered and, thus, $VC(S) \geq \upsilon$.

$\Rightarrow$: If $VC(S) \geq \upsilon$ then there is an $A \subseteq D$ with $|A| = \upsilon$ such that $\Pi_R(A) = P(A)$. Hence, for every $B \subseteq A$, there is an itemset I_B such that $T_D(I_B) \cap A = B$. Since $T_A(I_B) \subseteq T_D(I_B)$, this means that $T_A(I_B) \cap A \subseteq B$. Now note that

- $T_A(I_B) \subseteq A$ and
- by construction $B \subseteq T_A(I_B)$

and thus $T_A(I_B) = B$

An Immediate Consequence

Let D be a dataset with associated range space $S = (X, R)$. Then $VC(S)$ is the largest $\upsilon \in \mathbb{N}$ such that there is an $A \subseteq D$ with $|A| = \upsilon$ such that

$$\forall B \subset A : \exists I_B \subset \mathcal{I} : T_A(I_B) = B$$

Example

$D = \{\{a, b, c, d\}, \{a, b\}, \{a, c\}, \{d\}\}$ and $A = \{\{a, b\}, \{a, c\}\}$

- $I_{\{\{a,b\}\}} = \{\{a, b\}\}$
- $I_{\{\{a,c\}\}} = \{\{a, c\}\}$
- $I_{\{\{a\}\}} = A$
- $I_{\emptyset} = \{d\}$

Larger subsets of D cannot be shattered and hence its VD dimension is 2.

Nice, But

It is good to have a simple characterisation of the VC dimension. Nut since it puts a requirement on:

- $\forall B \subset A$

it is potentially very costly to compute

- in fact, it is known to be $O(|R||X|^{\log |R|})$

Fortunately, our corollary (the immediate consequence) suggests an alternative

- we need a set of υ transactions
- if all of them are at least υ long

we have enough freedom to make the condition hold

- for technical reasons we first assume that the transactions are independent, i.e., they form an antichain.

The d-index

Let D be a data set. The *d-index* of D is the largest integer d such that

- D contains at least d transactions of length at least d
- that form an antichain

Theorem
Let D be a dataset with d-index d. Then the range space $S = (X, R)$ associated with D has VC dimension of at most d

This upper bound is tight

- there are datasets for which the VC dimension equals their d-index

Proof Sketch

Let $l > d$ and assume that $T \subset D$ with $|T| = l$ can be shattered

- note that this means that T is an antichain: if $t_1 \subseteq t_2$ All ranges containg t_2 als contain t_1: we cannot shatter

For any $t \in T$, there are 2^{l-1} subsets of T that contain t. So, t occurs in 2^{l-1} ranges T_A.
t only occurs in T_A if $A \subset t$. Which means that T occurs in $2^{|t|} - 1$ ranges.
From the definition of d we know that T must contain a t^* such that $|t^*| < l$

- otherwise the d-index would be l

This means that $2^{|t^*|} - 1 < 2^{l-1}$, so t^* cannot appear in 2^{l-1} ranges.
This is a contradiction. So, the assumed T cannot exist. Hence, the largest set that is shattered has at most size d.

From d-Index to d-Bound

The d-index of D is still a bit hard to compute

- because of the antichain requirement

So, lets us forget about that requirement.

Let D be a dataset, its *d-bound* of D is the largest integer d such that

- D contains at least d transactions of length at least d

Theorem
Let D be a dataset with d-bound d. Then the range space $S = (X, R)$ associated with D has VC dimension of at most d

This is obvious as *d-bound* $\geq$ *d-index*

- a subset witnessing the d-index satisfies the conditions for the d-bound (but not vice versa)

Computing The d-Bound

Computing the d-bound is easy

- do a scan of the dataset
- maintaining
  - the ℓ longest (different) transactions
  - that are at least ℓ long
  - breaking ties arbitrarily

See the journal version for the full details

- and the proof!

The Sample Size (finally)

Combining all the results we have seen so far, we have:

Let D be a dataset, let d be the d-bound of D, and let $(\epsilon, \delta) \in (0,1)^2$. Let S be a random sample of D with size

$$|S| \geq \min \left\{ |D|, \frac{4c}{\epsilon^2}(d + \log \frac{1}{\delta} \right\}$$

Then $F(S, \theta - \epsilon/2)$ is an ϵ close approximation of $F(D, \theta)$ with probability at least $1 - \delta$.

Such a sample we can easily compute from D is a single scan. Hence we need two scans of D to compute a ϵ close approximation of the set of all frequent itemsets.

Good, But

Our knowledge of PAC learning has given us a bound on the sample size

- you are going to determine whether or not it is better than Toivonen's!

Yet, it isn't perfect:

- computing the d-bound requires a full scan of the database
- the d-bound is influenced by outliers
  - we could have 100 transactions of length 100
  - while the remaining 10^9 transactions are shorter than 50
  - than our sample will be (far) to big

Couldn't we sample progressively

- and detect (cheaply) that the sample is big enough?

Rademacher complexity to the rescue!

Progressive Sampling

The idea of progressive sampling is simple:

1. at iteration i create a random sample S_i
   - of some predetermined size $|S_i|$
   - sampled i.i.d. from D
2. check whether or not the current sample will yield an (ϵ, δ) approximation of $F(D, \theta)$
   - by some cheap to compute test
   - *not* by computing the full list of frequent item sets
   - or something else silly
3. If the sample size is big enough, return its frequent item sets for some appropriate threshold γ. Else perform the next sampling round in step 1.

- $\gamma = \theta - \epsilon/2$ as before
- good sample sizes will be discussed later
- we first focus on the stopping criterion

What We Want to Achieve

Like in our VC dimension approach, we want to have a sample S such that

$$\mathbb{P}\left(\sup_{I \in \mathcal{I}} |supp_D(I) - supp_S(I)| \leq \eta\right) \geq 1 - \delta$$

Because we know for such an S that if $\eta \leq \epsilon/2$ then

- $F(S, \theta - \epsilon/2)$ is an (ϵ, δ) approximation of $F(D, \theta)$

Previously, we used the VC dimension to compute a bound

- now we use the Rademacher averages

of the indicator functions $\mathbb{1}_I$

$$\mathbb{1}_I(J) = \begin{cases} 1 & \text{if } J \subseteq I \\ 0 & \text{otherwise} \end{cases}$$

for $I, J \in P(\mathcal{I}$

Why and How Rademacher

For a database D we have:

$$supp_D(I) = \frac{1}{|D|} \sum_{t \in D} \mathbb{1}_I(t)$$

For a subset $S \subset D$ of size n we have for the Rademacher variable σ

$$\mathcal{R}_S = \mathbb{E}_D \left[\sup_{I \in \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}_I(t_i) \right]$$

and then, using the error bounds we know for Rademacher complexity, we get the error bound for the support on S wrt D as:

$$\mathbb{P}\left(\sup_{I \in \mathcal{I}} |supp_D(I) - supp_S(I)| \leq 2\mathcal{R}_S + \sqrt{\frac{2\ln(2/\delta)}{n}} \right) \geq 1 - \delta$$

So, setting $\eta = 2\mathcal{R}_S + \sqrt{\frac{2\ln(2/\delta)}{n}}$ we are done

Bounding R_S

In our discussion of Rademacher complexity, we already saw that computing it exactly is rather costly

- if you don't mind an understatement

We then discussed Massart's Lemma as a way to bound it. Since this lemma "lives" in vector spaces, we have to resort to the embedding we have seen before.

For a given $S = \{t_1, \ldots, t_n\}$, for any $I \subseteq \mathcal{I}$ define the vector

$$v_S(I) = (\mathbb{1}_I(t_1), \ldots, \mathbb{1}_I(t_n))$$

And define

- $V_S = \{v_S(I) \mid I \subseteq \mathcal{I}\}$

Massart's Lemma

With this embedding we have by Massart:

$$\mathcal{R}_S \leq \max_{I \subseteq \mathcal{I}} ||v_S(I)|| \frac{\sqrt{2 \ln |V_S|}}{n}$$

To compute this bound

- we have to compute V_S

Which is rather expensive, since

$$v_S(I) = (\mathbb{1}_I(t_1), \ldots, \mathbb{1}_I(t_n))$$

That is we have to compute (very) many, (pretty) large, vectors.

So, can we compute a bound on $|V_S|$?

- and for $||v_S(I)||$ also, of course

The only direct bound is: $|V_S| \leq 2^{|\mathcal{I}|}$ meaning we get $|\mathcal{I}|$ in our bound for the sample size, which we already rejected.

A Variant of Massart's Lemma

Fortunately, from our proof of Massart's lemma it is easy to see that the following inequality also holds

Define $w : \mathbb{R}^+ \to \mathbb{R}^+$ by

$$w(s) = \frac{1}{s} \ln \left(\sum_{v \in V_S} exp(s^2 ||v||^2 / (2n^2)) \right)$$

then

$$\mathcal{R}_S \leq \min_{s \in \mathbb{R}^+} w(s)$$

This may not seem much of an improvement

- you don't need to compute $|V(s)|$
- but you still need to
  - sum over all of V_S, i.e., compute V_S
  - and compute all $||v||$ like before

A Relation with Closed Itemsets

Clearly, we are not going to compute V_S

- rather we are going to derive bounds

And the bounds we will derive depend on a relationship with closed itemsets:

If we denote by $Cl(S)$ the set of all closed frequent itemsets on S, then:

1. $V_S = \{v_S(I), I \in Cl(S)\}$
2. $|V_S| = |Cl(S)|$

Clearly, we are also not going to compute $Cl(S)$ either

- that would not yield an easy to compute stopping criterion.

But for the moment it is good to think that we do

- but first the proof

Proof Sketch

Let $I \in Cl(S)$, define S_I by

$$S_I = \{J \subset I \mid supp_S(J) = supp_S(I)\}$$

Since for $J \in S_I : I \subseteq t \Rightarrow J \subseteq t$, (this holds for all $J \subseteq I$ after all), we have that I and J have the same support set in S, i.e.,

$$v_S(J) = v_S(I)$$

which means that $|V_S| \leq |Cl(S)|$.

Now let $A, B \in Cl(S)$, with $A \neq B$, such that $v_S(A) = v_S(B)$, then $v_S(A \cup B) = v_S(A) = v_S(B)$

- that is, $supp_S(A) = supp_S(A \cup B)$

That is, A is not closed. Since this is a contradiction we have that all $I \in Cl(S)$ have a different v_S that is: $|V_S| \geq |Cl(S)|$

Using $Cl(S)$ to Bound w

We want to have an upperbound for w that is simple to compute. That is we can compute it with

- in one scan over the data
- while we take the sample

Realistically speaking, the type of things we can compute during that scan is limited, e.g.,

- The support of the singleton itemsets, as far as they occur, on S – denoted by $\mathcal{I}_S$
  - counting the support of larger itemsets does not make too much sense
  - which ones would you count
- the length of the transactions t that occur in t.

It turns out that with this simple measures we can

- partition $Cl(S)$
- use that partitioning to put an upperbound on $|Cl(S)|$
- and put an upperbound on $||v||$

Orders

To define the partitioning of $Cl(S)$, we are assuming two orders

- one on the items
- and one on the transactions

These two orders allow us to sort the closed frequent itemsets in groups

- that is, too partition them.

More precisely, we

- assume an order $<_{\mathcal{I}}$ on $\mathcal{I}$
  - e.g., increasing on their frequency
- for all $a \in \mathcal{I}$ assume an order $<_a$ on all transactions in S that contain a; i.e, $<_a$ is an order on $T_S(\{a\})$.
  - e.g., increasing by the number of items larger than a they contain.

Partitioning

Since we already now the support for the singletons, denote the closed singletons by $\mathcal{C}_1$.
We now focus on the closed itemsets of size 2 or more. This set is partitioned as follows:

Assign a closed itemset I to partition $\mathcal{C}_{a,t}$ such that

- $a \in I$ is its first item according to $<_{\mathcal{I}}$.
- t is the first transaction containing A according to $<_a$.

Clearly,

$$Cl(S) = \mathcal{C}_1 \cup \bigcup_{a \in \mathcal{I}} \bigcup_{t \in T_S(\{a\})} \mathcal{C}_{a,t}$$

since every closed itemset gets assigned to exactly one $\mathcal{C}_{a,t}$.

Bounding $|\mathcal{C}_{a,t}|$

Since $\forall A \in \mathcal{C}_{a,t} : A \subseteq t$ we know how many closed itemsets t can be involved in:

$$|\mathcal{C}_{a,t}| \leq 2^{|t|}$$

But notice that

- t (potentially) defines multiple partitions
- especially low support items will probably not have transactions for which they are the first element

This means that we

- seriously overestimate $|Cl(S)|$
- by using this bound for $|\mathcal{C}_{a,t}|$

By taking both the

- two orders into account, e.g., only look at the items in t that are larger than a.
- and the assignment of closed itemsets to partitions

it is easy to devise a tighter bound.

A Tighter Bound

Let $t \in T_S(\{a\})$.

- let t contain exactly $k_{a,t}$ items larger than a
- let t be the $l_{a,t}$'th such transaction, i.e., t is the $l_{a,t}$'th transaction in $T_S(\{a\})$ containing exactly $k_{a,t}$ items larger than a
- let $g_{a,r}$ be the number of transactions in $T_S(\{a\})$ containing exactly r items larger than a
- let χ_a be the largest r such that at least one transaction in $T_S(\{a\})$ contains exactly r items larger than a; i.e., $\chi_a = \max\{r \mid g_{a,r} > 0\}$
- let $h_{a,r}$ be the number of transactions in $T_S(\{a\})$ that contain at least r items larger than a, i.e., $h_{a,r} = \sum_{j \geq r}^{\chi_a} g_{a,j}$

Then

$$|\mathcal{C}_{a,t}| \leq 2^{\min\{k_{a,t}, h_{a,k_{a,t}} - l_{a,t}\}}$$

How About $||v_S||$?

So, we have now have a bound for

- $|CL(S)| = |V_S|$

But how about a bound for $||v||$?

That one is easier. First we have:

- $\forall A \subset \mathcal{I} : ||v_S(A)|| = \sqrt{n \times supp_S(A)}$
- because $v_S(A)$ contains a 1 for every $t \in T_S(A)$

Next note that

- $\forall a \in A : supp_S(\{a\}) \geq supp_S(A)$

Hence

$$\forall A \in \mathcal{C}_{a,t} : ||v_S(A)|| \leq \sqrt{n \times supp_S(a)}$$

Bounding the Rademacher Complexity

With these two bounds we get:

Let $\tilde{w} : \mathbb{R}^+ \to \mathbb{R}^+$ be the function

$$\tilde{w}(s) = \frac{1}{s} \ln \left(\sum_{a \in \mathcal{I}_S} \left(\left(\sum_{r=1}^{\chi_a} \sum_{j=1}^{g_{a,r}} 2^{r, h_{a,r} - j} \right) e^{\frac{s^2 supp_S(a)}{2n}} \right) \right)$$

then

$$\mathcal{R}_S \leq \min_{s \in \mathbb{R}^+} \tilde{w}(s)$$

Note that we can compute all constants

- with one scan over the sample S
- we do not need to compute $Cl(S)$ at all!

The Stopping Condition

With this bound on the Rademacher Complexity we can finally formulate our stopping condition

Let i be the minimal index for which

$$2\tilde{w}(s^*) + \sqrt{\frac{2\ln(2/\delta)}{|S_i|}} \leq \epsilon/2$$

where s^* is the $s \in \mathbb{R}^+$ achieving $\min_{s \in \mathbb{R}^+} \tilde{w}(s)$.
Then $F(S, \theta - \epsilon/2)$ is an (ϵ, δ) approximation of $F(D, \theta)$

The last thing we have to do is to determine how big each sample should be and we have our algorithm

The Sampling Scheme

Note that

$$\epsilon/2 \geq 2\tilde{w}(s^*) + \sqrt{\frac{2\ln(2/\delta)}{|S_i|}} \geq \sqrt{\frac{2\ln(2/\delta)}{|S_i|}}$$

Hence, we choose S_0 such that

$$|S_0| = \frac{8\ln(2/\delta)}{\epsilon^2}$$

For the next round, notice that

$$\eta_i = 2\tilde{w}(s^*) + \sqrt{\frac{2\ln(2/\delta)}{|S_i|}}$$

is an upperbound to our current error. Hence we take S_{i+1}

$$|S_{i+1}| = \left(\frac{2\eta_i}{\epsilon}\right)^2 |S_i|$$

We are Done?

So, we have our progressive sampling approach to probably approximately correct frequent itemset mining

- do we?

Well, we do if you know how to compute

- $\min_{s \in \mathbb{R}^+} \tilde{w}(s)$

Since $\tilde{w}$ is a (twice) differentiable convex function, this is an easy exercise in optimization – many packages can do this for you

- for those who actually know about this stuff

It has been brought to my attention that

- not all of you actually do!

While I like that topic as well

- and I do think it is important and all of you should know

It would be completely off topic to the rest of the course

- hence, I decided to make your life a bit easier by changing the exam slightly.

Your Essay

The essay you have to submit still consists of three parts

1. A part explaining the results we achieved today
2. A part doing experiments – this part is changed slightly
3. A part explaining a choice of course material not needed for part 1

I'll now explain all three parts again to you.

Part 1

Today we learned that based on the d-bound we need a sample size of

$$|S| \geq \min\left\{|D|, \frac{4c}{\epsilon^2}(d + \log\frac{1}{\delta}\right\}$$

to get a good approximation.And for the same result based on the Rademacher complexity that we can do progressive sampling with

- $|S_0| = \frac{8\ln(2/\delta)}{\epsilon^2}$ and
- $|S_{i+1}| = \left(\frac{2\eta_i}{\epsilon}\right)^2 |S_i|$

and stopping criterion

$$2\tilde{w}(s^*) + \sqrt{\frac{2\ln(2/\delta)}{|S_i|}} \leq \epsilon/2$$

In your first part, you should explain what these results mean and why they are true

- to a Master student who did not follow this course

Part 1, Continued

- For the first part you can/should use 5 - 7 pages.
- Use Math only where necessary, formalisms where you need them, natural language to make things intuitive.
- Not a list of definitions and theorems, but an explanation your colleague would actually be able to follow

To give some examples

- the results rely on the VC dimension and on the Rademacher complexity – so, you should introduce and explain these concepts
- it is all based on PAC learning, which we studied for classification, so you should probably start there

The grade you get for this part

- depends on how well you can convince me that you actually understand what we have achieved.

Part 2

On Wednesday we saw Toivonen's bounds

- $n \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}$
- lower the threshold by $\sqrt{\frac{1}{2n} \ln \frac{1}{\mu}}$

Today we saw the d-bound sample size

- how do these **two** approaches compare
  - how big a samples do they prescribe
  - how good are the approximate results
- You should perform experiments to find out
- Various datasets and various (ϵ, δ) settings
- you can mine frequent itemsets in any way you want.

The Rademacher paper of today gives good suggestions on the data

Part 2, What to Write

You should explain your set up

- what data did you use, what (ϵ, δ) settings
- what were the results, in terms of sample size and final quality

And, most of all:

what do you conclude from your experiments

The marks you score here depend on

- the way you experimented
- the quality of your conclusions
- the comprehensibility of what you write

You should be able to do well in 2 - 3 pages

Part 2, Bonus

If you do know (or are willing to learn) how to compute

- $\min_{s \in \mathbb{R}^+} \tilde{w}(s)$

You can add the third set of bounds to the mix

- you compare three approaches rather than two

and your mark will be raised

You may wonder:

- what if I already score 10/10
- and I did the bonus part as well and very good!

The system cannot handle numbers larger than 10

- but you will get eternal glory

Part 3

For Part 1 you need to explain large parts of the material we covered in the course, but not everything. For example, boosting or structural risk minimization are not needed.

- pick any such topic and explain it to the same student
- in 3 - 5 pages
- in the same way as for part 1
- it should be in your words: no plagiarism!
- if you tell more than I told you: you get more points
- it should be on Master level, no popular science, please

There is one thing you cannot write about

- Pattern set mining
- that is covered by another course

You can do Computational Philosophy of Induction

- but it should be a serious treatise

Are We Done?

You might think we are done, after all

- you know how to overcome the problems of Big Data
- in a concrete case: pattern mining

But, there is a bit of an embarrassing problem

- if you mine with a high minimal support, you'll get a few – often well-known – patterns
- if you mine with low support

  you may end up with more patterns than you have transactions!

The reason is simple:

- one transaction can/will support multiple patterns

Somehow this doesn't feel like a solution to Big Data

- we want a small set of characteristic patterns

That is what we are going to do next week

- based on Algorithmic Information Theory