Iterative Data Mining

Jilles Vreeken

26 June 2014 (TADA)
Service Announcement #1

Evaluation Forms

1. Hand forms out (me)
2. Fill forms out (you)
3. Collect forms (you)
4. Put forms in envelop (you)
5. Bring envelop back to Evelyn (one ‘volunteer’ and me)
The Exam

- **type:** oral
- **when:** September 11\textsuperscript{th}
- **time:** individual
- **where:** E1.3 room 0.16
- **what:** all material discussed in the lectures, plus one assignment (your choice) per topic

The Re-Exam

- **type:** oral
- **when:** October 1\textsuperscript{st}
- **time:** individual
- **where:** E1.3 room 001
Master thesis projects

- in principle: yes!
- in practice: depending on background, motivation, interests, and grades --- plus, on whether I have time
- interested? mail me and/or Pauli

Student Research Assistant (HiWi) positions

- in principle: maybe!
- in practice: depends on background, grades, and in particular your motivation and interests
- interested? mail me and/or Pauli, include CV and grades
Service Announcement #4

**Introduction**
- Is DM science?
- DM in action

**Tensors**
- Introduction to tensors
  - Tensors in DM
  - Special topics in tensors

**Information Theory**
- MDL + patterns
- Entropy + correlation
- MaxEnt + iterative DM

**Mixed Grill**
- Influence Propagation
- Redescription Mining
  - *special request*
Service Announcement #4

Introduction
- Is DM science?
- DM in action

Tensors
- Introduction to tensors
- Tensors in DM
- Special topics in tensors

Information Theory
- MDL + patterns
- Entropy + correlation
- MaxEnt + iterative DM

Mixed Grill
- Influence Propagation
- Redescription Mining

Let us know (asap, mail) what topic you would like to see discussed
Service Announcement #5

- Introduction
- Tensors
- Information Theory
- Mixed Grill
- Wrap-up + <ask-us-anything>
<ask-us Anything>?

Yes!
Prepare questions on anything* you’ve always wanted to ask Pauli and/or me. We’ll answer on the spot

* preferably related to TADA, data mining, machine learning, science, the world, etc.
Good Reads

Data Analysis: a Bayesian Tutorial
D.S. Sivia & J. Skilling
(very good, but skip the MaxEnt stuff)

Elements of Information Theory
Thomas Cover & Joy Thomas
(very good textbook)

The Information
James Gleick
(great light reading)
Iterative Data Mining

Jilles Vreeken

26 June 2014 (TADA)
Question of the day

How can we find things that are interesting with regard to what we already know?

How can we measure subjective interestingness?
What is interesting?

something that **increases** our knowledge about the data
What is a good result?

something that reduces our uncertainty about the data (ie. increases the likelihood of the data)
What is really good?

something that, in simple terms, strongly reduces our uncertainty about the data

(maximise likelihood, but avoid overfitting)
Let’s make this visual

universe $\mathcal{D}$ of possible datasets

our dataset $D$
Given what we know

all possible datasets, given current knowledge

our dataset $D$
More knowledge...

dimensions, margins, pattern $P_1$

all possible datasets

our dataset $D$
Fewer possibilities...

dimensions, margins, patterns $P_1$ and $P_2$

all possible datasets

our dataset $D$
Less uncertainty.

dimensions, margins, the key structure

all possible datasets

our dataset $D$
Maximising certainty

dimensions, margins, patterns $P_1$ and $P_2$

all possible datasets

knowledge added by $P_2$

our dataset $D$
How can we define ‘uncertainty’ and ‘simplicity’? interpretability and informativeness are intrinsically subjective
Measuring Uncertainty

We need access to the likelihood of data $D$ given background knowledge $B$

$$P(D \mid B)$$

such that we can calculate the gain for $X$

$$P(D \mid B \cup X) - P(D \mid B)$$

...which distribution should we use?
Measuring Surprise

We need access to the likelihood of result $X$ given background knowledge $B$

$$P(X \mid B)$$

such that we can mine the data for $X$ that have a low likelihood, that are surprising

...which distribution should we use?
Measuring Surprise

We need access to the likelihood of result $X$ given background knowledge $B$

This is called the *p-value* of result $X$

such that we can mine the data for $X$ that have a low likelihood, that are *surprising*

...which distribution should we use?
Measuring Surprise

We need access to the **likelihood** of result $X$ given background knowledge $B$ such that we can mine the data for $X$ that have a low likelihood, that are **surprising**…which distribution should we use?

**Standard** practice: if the **p-value** of result $X < 0.5$ its **significant**

...which distribution should we use?
Approach 1: Randomization

1. Mine original data
2. Mine random data
3. Determine probability

Original data → Random data #1 → Random data #2 → ... → Random data #N

empirical p-value

score(\(X \mid D\))
Approach 1: Randomization

1. Mine original data
2. Mine random data

The fraction of better ‘randoms’ is the empirical p-value of result $X$

(score($X$ | $D$))
Approach 1: Randomization

1. Mine original data
2. Mine random data
3. Determine probability

- Randomization
- Original data
- Random data #1
- Random data #2
- Random data #N

The fraction of better 'randoms' is the empirical p-value of result $X$.
Random Data

So, we need data that
- maintains our background knowledge, and
- is otherwise completely random

How can we get our hands on that?
Swap Randomization

Let there be data

\[
\begin{array}{ccccccc}
1 & 1 & 1 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{array}
\]
Swap Randomization

Say we only know overall density.
How to sample random data?

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

27

(swap randomization, Gionis et al. 2005)
Didactically, let us instead consider a Monte-Carlo Markov Chain

Very simple scheme

1. select two cells at random,
2. swap values,
3. repeat until convergence.

Swap Randomization

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(swap randomization, Gionis et al. 2005)
Swap Randomization

Margins are easy understandable for binary data, how can we sample data with same margins?

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>27</td>
<td></td>
</tr>
</tbody>
</table>

(swap randomization, Gionis et al. 2005)
Swap Randomization

By MCMC!

1. randomly find submatrix

(swap randomization, Gionis et al. 2005)
Swap Randomization

By MCMC!

1. randomly find submatrix

2. swap values

(swap randomization, Gionis et al. 2005)
Swap Randomization

By MCMC!

1. randomly find submatrix
2. swap values
3. repeat until convergence

(swap randomization, Gionis et al. 2005)
Many ways to test **static** null hypothesis assuming distribution, swap-randomization, MaxEnt

What can we use this for?

**ranking** based on **static significance**

mining the **top-k** most significant patterns,

but **not** suited for iterative mining
For iterative data mining, we need models that can maintain the type of information (eg. patterns) that we mine.

Randomization is powerful:
- variations exists for many data types (Ojala ‘09, Henelius et al ‘13)
- can be pushed beyond margins (see Hanhijärvi et al 2009)
- but... has key disadvantages
Approach 2: Maximum Entropy

‘the best distribution $p^*$ satisfies the background knowledge, but makes no further assumptions’

very useful for data mining: unbiased measurement of subjective interestingness

(Jaynes 1957; De Bie 2009)
Constraints and Distributions

Let $B$ be our set of constraints

$$B = \{f_1, \ldots, f_n\}$$

Let $C$ be the set of admissible distributions

$$C = \{ p \in \mathbb{P} \mid p(f_i) = \tilde{p}(f_i) \text{ for } f_i \in B \}$$

We need the most **uniformly** distributed $p \in \mathbb{P}$
Uniformity and Entropy

Uniformity $\leftrightarrow$ Entropy

$$H(p) = -\sum_{x \in X} p(X = x) \log p(X = x)$$

tells us the **entropy** of a (discrete) distribution $p$
Maximum Entropy

We want access to the distribution $p^*$ with **maximum entropy**

$$p_B^* = \arg\max_{p \in C} H(p)$$

better known as the **maximum entropy model**

It can be shown that $p^*$ is well defined there *always* exist a unique $p^*$ with maximum entropy for any constrained set $C$

* that’s not completely true, some esoteric exceptions exist
Some examples

Mean and

- interval? uniform
- variance? Gaussian
- positive? exponential
- discrete? geometric
- ...

But... what about distributions for like data, patterns, and stuff?
MaxEnt Theory

To use MaxEnt, we need theory for modelling data given background knowledge.

Patterns
- itemset frequencies (Tatti ’06, Mampaey et al. ’11)

Binary Data
- margins (De Bie ’09)
- tiles (Tatti & Vreeken, ’12)

Real-valued Data
- margins (Kontonasios et al. ’11)
- sets of cells (Kontonasios et al. ’13)
Let \( p \) be a probability density satisfying the constraints

\[
\int_S p(x) f_i(x) dx = \alpha_i \quad \text{for } 1 \leq i \leq m .
\]

Then we can write the MaxEnt distribution as

\[
p^*(x) = p_\lambda(x) = \alpha \left\{ \begin{array}{ll}
\exp \left( \lambda_0 + \sum_{f_i \in B} \lambda_i \cdot f_i(x) \right) & D \notin \mathcal{Z} \\
0 & D \in \mathcal{Z}
\end{array} \right.,
\]

where we choose the lambdas to satisfy the constraints

(Csizar 1975)
The problem is convex – yay!

This means we can use any convex optimization strategy.

Standard approaches include
- iterative scaling,
- gradient descent,
- conjugate gradient descent,
- Newton’s method,
- etc.
Inferring the Model

Optimization requires calculating $p$

for datasets and tiles
this is easy

for itemsets and frequencies, however,
this is PP-hard
MaxEnt Theory

To use MaxEnt, we need theory for modelling data given background knowledge.

Binary Data
- margins (De Bie, ’09)
- tiles (Tatti & Vreeken, ’12)

Real-valued Data
- margins (Kontonasios et al. ’11)
- arbitrary sets of cells (now)

allow for iterative mining
MaxEnt for Real-Valued Data

Current state of the art can incorporate means, variance, and higher order moments, as well as histogram information over arbitrary sets of cells

(Kontonasios et al. 2013)
MaxEnt for Real-Valued Data

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>0.7</td>
<td>0.9</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8</td>
<td>0.7</td>
<td>0.8</td>
<td>0.4</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0.9</td>
<td>0.8</td>
<td>0.3</td>
<td>0.8</td>
<td>0.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>
MaxEnt for Real-Valued Data

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.9</td>
<td>.8</td>
<td>.7</td>
<td>.4</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.7</td>
<td>.8</td>
<td>.9</td>
<td>.3</td>
<td>.5</td>
<td>.3</td>
<td>.5</td>
</tr>
<tr>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.3</td>
<td>.4</td>
<td>.2</td>
</tr>
<tr>
<td>.7</td>
<td>.9</td>
<td>.7</td>
<td>.7</td>
<td>.3</td>
<td>.2</td>
<td>.5</td>
</tr>
<tr>
<td>.2</td>
<td>.8</td>
<td>.7</td>
<td>.8</td>
<td>.4</td>
<td>.4</td>
<td>.1</td>
</tr>
<tr>
<td>.3</td>
<td>.6</td>
<td>.9</td>
<td>.8</td>
<td>.3</td>
<td>.8</td>
<td>.3</td>
</tr>
<tr>
<td>.2</td>
<td>.1</td>
<td>.3</td>
<td>.4</td>
<td>.5</td>
<td>.3</td>
<td>.2</td>
</tr>
</tbody>
</table>

Pattern 1
- \{1-3\} x \{1-4\}
- mean 0.8

Pattern 2
- \{2,3\} x \{3-5\}
- mean 0.8

Pattern 3
- \{5-7\} x \{3-5\}
- mean 0.3
MaxEnt for Real-Valued Data

<table>
<thead>
<tr>
<th></th>
<th>.9</th>
<th>.8</th>
<th>.7</th>
<th>.4</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
<th>.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.7</td>
<td>.8</td>
<td>.9</td>
<td>.3</td>
<td>.5</td>
<td>.3</td>
<td>.5</td>
<td>.6</td>
</tr>
<tr>
<td></td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.3</td>
<td>.4</td>
<td>.2</td>
<td>.6</td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>.9</td>
<td>.7</td>
<td>.7</td>
<td>.3</td>
<td>.2</td>
<td>.5</td>
<td>.6</td>
</tr>
<tr>
<td></td>
<td>.2</td>
<td>.8</td>
<td>.7</td>
<td>.8</td>
<td>.4</td>
<td>.4</td>
<td>.1</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.6</td>
<td>.9</td>
<td>.8</td>
<td>.3</td>
<td>.8</td>
<td>.3</td>
<td>.6</td>
</tr>
<tr>
<td></td>
<td>.2</td>
<td>.1</td>
<td>.3</td>
<td>.4</td>
<td>.5</td>
<td>.3</td>
<td>.2</td>
<td>.3</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.7</td>
<td>.7</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.3</td>
<td>.5</td>
</tr>
</tbody>
</table>

Pattern 1
- \{1-3\} \times \{1-4\}
- mean 0.8

Pattern 2
- \{2,3\} \times \{3-5\}
- mean 0.8

Pattern 3
- \{5-7\} \times \{3-5\}
- mean 0.3
MaxEnt for Real-Valued Data

<table>
<thead>
<tr>
<th>.5</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
</tbody>
</table>

- **Pattern 1**
  - \(\{1-3\} \times \{1-4\}\)
  - Mean 0.8

- **Pattern 2**
  - \(\{2,3\} \times \{3-5\}\)
  - Mean 0.8

- **Pattern 3**
  - \(\{5-7\} \times \{3-5\}\)
  - Mean 0.3
MaxEnt for Real-Valued Data

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.6</td>
<td>.7</td>
<td>.7</td>
<td>.7</td>
<td>.5</td>
<td>.6</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>.7</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>.6</td>
<td>.7</td>
<td>.7</td>
<td>.6</td>
<td>.5</td>
<td>.5</td>
<td>.3</td>
<td>.6</td>
</tr>
<tr>
<td>.6</td>
<td>.6</td>
<td>.7</td>
<td>.6</td>
<td>.5</td>
<td>.4</td>
<td>.5</td>
<td>.6</td>
</tr>
<tr>
<td>.6</td>
<td>.6</td>
<td>.7</td>
<td>.6</td>
<td>.5</td>
<td>.4</td>
<td>.3</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.7</td>
<td>.6</td>
<td>.6</td>
<td>.5</td>
<td>.4</td>
<td>.3</td>
<td>.5</td>
</tr>
<tr>
<td>.5</td>
<td>.7</td>
<td>.7</td>
<td>.6</td>
<td>.5</td>
<td>.6</td>
<td>.3</td>
<td>.6</td>
</tr>
<tr>
<td>.3</td>
<td>.6</td>
<td>.6</td>
<td>.3</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.3</td>
</tr>
</tbody>
</table>

Pattern 1
- \(\{1-3\} \times \{1-4\}\)
- mean 0.8

Pattern 2
- \(\{2,3\} \times \{3-5\}\)
- mean 0.8

Pattern 3
- \(\{5-7\} \times \{3-5\}\)
- mean 0.3

(Kontonasios et al., 2011)
MaxEnt for Real-Valued Data

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>.2</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.4</td>
<td>.5</td>
<td>.4</td>
<td>.5</td>
</tr>
<tr>
<td>.3</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
</tr>
<tr>
<td>.1</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.4</td>
<td>.4</td>
<td>.3</td>
<td>.3</td>
</tr>
<tr>
<td>.5</td>
<td>.7</td>
<td>.7</td>
<td>.6</td>
<td>.4</td>
<td>.4</td>
<td>.4</td>
<td>.5</td>
</tr>
</tbody>
</table>

Pattern 1
- \{1-3\} x \{1-4\}
- mean 0.8

Pattern 2
- \{2,3\} x \{3-5\}
- mean 0.8

Pattern 3
- \{5-7\} x \{3-5\}
- mean 0.3

(Kontonasios et al. 2013)
Simplicity?

Likelihood alone is insufficient
does not take size, or complexity into account

as practical example of our model:

Information Ratio
for tiles in real valued data
Information Ratio

\[
\frac{\text{Information Content}}{\text{Description Length}}
\]

\[
\text{InfContent}(p) = L(D \mid B) - L(D \mid B + p)
\]

\[
\text{DescLength}(p) = L(\text{rows}(p)) + L(\text{cols}(p)) + L(\text{stat}(p))
\]
## Results

<table>
<thead>
<tr>
<th></th>
<th>It 1</th>
<th>It 2</th>
<th>It 3</th>
<th>It 4</th>
<th>It 5</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>A2</td>
<td>B3</td>
<td>A3</td>
<td>B2</td>
<td>C3</td>
<td>A2</td>
</tr>
<tr>
<td>2.</td>
<td>A4</td>
<td>B4</td>
<td>B2</td>
<td>C3</td>
<td>C4</td>
<td>B3</td>
</tr>
<tr>
<td>3.</td>
<td>A3</td>
<td>B2</td>
<td>C3</td>
<td>C4</td>
<td>C2</td>
<td>A3</td>
</tr>
<tr>
<td>4.</td>
<td>B3</td>
<td>A3</td>
<td>C4</td>
<td>C2</td>
<td>D2</td>
<td>B2</td>
</tr>
<tr>
<td>5.</td>
<td>B4</td>
<td>C3</td>
<td>C2</td>
<td>B4</td>
<td>D4</td>
<td>C3</td>
</tr>
<tr>
<td>6.</td>
<td>B2</td>
<td>C4</td>
<td>B4</td>
<td>D2</td>
<td>D3</td>
<td>C2</td>
</tr>
<tr>
<td>7.</td>
<td>C3</td>
<td>C2</td>
<td>D2</td>
<td>D4</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>8.</td>
<td>C4</td>
<td>D2</td>
<td>D4</td>
<td>D3</td>
<td>A5</td>
<td>D3</td>
</tr>
<tr>
<td>9.</td>
<td>C2</td>
<td>D4</td>
<td>D3</td>
<td>D1</td>
<td>21</td>
<td>A5</td>
</tr>
<tr>
<td>10.</td>
<td>D2</td>
<td>D3</td>
<td>B1</td>
<td>A5</td>
<td>B5</td>
<td>B5</td>
</tr>
</tbody>
</table>

### Synthetic Data
- random Gaussian
- 4 ‘complexes’ (ABCD) of 5 overlapping tiles
- \(x2 + x3\) big with low overlap

### Patterns
- real + random tiles

### Task
- Rank on InfRatio, add best to model, iterate
Results

Real Data
- gene expression

Patterns
- Bi-clusters from external study

Legend:
- solid line
- dashed line
- histograms
- means/var

x $10^{-3}$ Information Ratio for the top–10 tiles for Alon dataset

Negative Log–Likelihood of the Alon data
Conclusions

Significance testing is important
- choosing a good model (and test) is difficult

Randomization
- simple yet powerful – difficult to extend – empirical p-values

Maximum Entropy modelling
- complex yet powerful – inferring can be NP-hard – exact p-values
- can be defined for anything ... if you can derive the model...

Iterative Data Mining
- mine most informative thingy, update model, repeat.
Significance testing is important

- choosing a good model (and test) is difficult

Randomization

- simple yet powerful – difficult to extend – *empirical* p-values

Maximum Entropy modelling

- complex yet powerful – inferring can be NP-hard – *exact* p-values
- can be defined for *anything* ... if you can derive the model...

Iterative Data Mining

- mine most informative thingy, update model, repeat.

Thank you!