Maximum Entropy & Subjective Interestingness

Jilles Vreeken

26 June 2015
Questions 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, dimensions, margins.
More knowledge...

dimensions, margins, pattern $P_1$

all possible datasets

our dataset $D$
Fewer possibilities...

all possible datasets

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

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?
Approach 2: Maximum Entropy

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

(Jaynes 1957; De Bie 2009)
Approach 2: Maximum Entropy

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

in other words,
\( p^* \) assigns the correct probability mass to the background knowledge instances:

\( p^* \) **is a maximum likelihood estimator**

(Jaynes 1957; De Bie 2009)
Approach 2: Maximum Entropy

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

in other words, $p^*$ spreads probability mass around as evenly as possible:

$p^*$ does not have any specific bias

(Jaynes 1957; De Bie 2009)
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
\[ C = \{p \in \mathcal{P} \mid p(f_i) = \tilde{p}(f_i) \text{ for } f_i \in B\} \]

We need the most **uniformly** distributed $p \in \mathcal{P}$
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 \textbf{maximum entropy}

\[
p^*_B = \underset{p \in C}{\arg \max} H(p)
\]

better known as the \textbf{maximum entropy model}

for constraints set \( B \)
Maximum Entropy

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

$$p^* = \arg\max_{p \in C} p$$

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

maximum entropy model

for constraints set $B$

(that’s not completely true, some esoteric exceptions exist)
Does this make sense?

Any distribution with less-than-maximal entropy must have a reason for this.

Less entropy means not-as-uniform-as-possible, that is, undue peaks of probability mass.

That is, reduced entropy = latent assumptions, exactly what we want to avoid!
Recall that through Kraft’s inequality, the probability distribution ↔ encoding

The MaxEnt distribution for $B$ gives the minimum worst-case expected encoded length over any distribution that satisfies this background knowledge.
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)
Finding the MaxEnt distribution

You can finding the MaxEnt distribution by solving the following system of linear constraints

\[
\begin{align*}
\max_{p(x)} & \quad - \sum_x p(x) \log p(x) \\
\text{s.t.} & \quad \sum_x p(x)f_i(x) = \alpha_i \quad \text{for all } i \\
& \quad \sum_x p(x) = 1
\end{align*}
\]

* for discrete data
Exponential Form

Let $p$ be a probability density satisfying the constraints

$$\int p(x)f_i(x)dx = \alpha_i \quad \text{for} \quad 1 \leq i \leq m$$

then we can write the MaxEnt distribution as

$$p^* = p_\lambda(x) \propto \begin{cases} \exp \left( \lambda_0 + \sum_{f_i \in B} \lambda_i f_i(x) \right) & D \notin Z \\ 0 & D \in Z \end{cases}$$

where we choose the lambdas, Lagrange multipliers, to satisfy the constraints, and where $Z$ is a collection of databases s.t. $p(D) = 0$ for all $p \in P$
Solving the MaxEnt

The Lagrangian is

\[ L(p(x), \mu, \lambda) = -\sum_x p(x) \log p(x) + \sum_i \lambda_i \left( \sum_i p(x)f_i(x) - \alpha_i \right) + \mu \left( \sum_x p(x) - 1 \right) \]

We set the derivative w.r.t. \( p(x) \) to 0 and get

\[ p(x) = \frac{1}{Z(\lambda)} \exp \left( \sum_i \lambda_i f_i(x) \right) \]

where \( Z(\lambda) = \sum_x \exp(\sum_i \lambda_i f_i(x)) \) is called the partition function
En Garde!

We may substitute $p(x)$ in the Lagrangian to obtain the dual objective

$$L(\lambda) = \log(Z(\lambda)) - \sum_i \lambda_i \alpha_i$$

Minimizing the dual gives the maximal solution to the original problem. Moreover, it is convex.
Inferring the Model

The problem is convex 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 $\rho$

for datasets and $tiles$
this is easy

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

Constraints: the **expected** row and column margins

\[
\sum_{D \in \{0,1\}^{n \times m}} p(D) \left( \sum_{j=1}^{m} d_{ij} \right) = r_i
\]

\[
\sum_{D \in \{0,1\}^{n \times m}} p(D) \left( \sum_{j=1}^{n} d_{ij} \right) = c_j
\]

(De Bie 2010)
Using the Lagrangian, we can solve $p(D)$

$$p(D) = \prod_{i,j} \frac{1}{Z(\lambda_i^r, \lambda_j^c)} \exp(d_{ij}(\lambda_i^r + \lambda_j^c))$$

where

$$Z(\lambda_i^r, \lambda_j^c) = \sum_{d_{ij} \in \{0,1\}} \exp(d_{ij}(\lambda_i^r + \lambda_j^c))$$
MaxEnt for Binary Databases

Using the Lagrangian, we can solve $p(D)$

$$p(D) = \prod_{i,j} \frac{1}{Z(\lambda_i^r, \lambda_j^c)} \exp(d_{ij}(\lambda_i^r + \lambda_j^c))$$

Hey! $p(D)$ is a product of independent elements! That’s handy!
We **did not** enforce this property, it’s a consequence of MaxEnt.

Following, every element is hence Bernoulli distributed, with a success probability of

$$\exp \frac{\lambda_i^r + \lambda_j^c}{1+\exp(\lambda_i^r + \lambda_j^c)}$$
Okay, say we have this $p^*$, what is it useful for?

Given $p^*$ we can

- sample data from $p^*$, and compute empirical p-values (just like with swap randomization)
- compute the likelihood of the observed data, and
- compute how surprising our findings are given $p^*$, and compute exact p-values
Expected vs. Actual

Swap randomization and MaxEnt can both maintain margins.

MaxEnt constrains the expected margins.
Swap randomization constrains the actual margins.

Does this matter?
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>.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>
MaxEnt for Real-Valued Data

<table>
<thead>
<tr>
<th>.9</th>
<th>.8</th>
<th>.7</th>
<th>.4</th>
<th>.5</th>
<th>.5</th>
<th>.5</th>
</tr>
</thead>
<tbody>
<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
- \( \{1 - 3\} \times \{1 - 4\} \)
- 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

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

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

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

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. 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

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

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

\[
\text{DescLength}(p) = L(\text{rows}(p)) + L(\text{cols}(p)) + L(\text{stat}(p))
\]
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          histograms
- dashed line      means/var
Conclusions

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

Maximum Entropy modelling
■ complex yet powerful
■ inferring can be NP-hard
■ analytical model can calculate exact probabilities
■ can be defined for anything ...if you can derive the model...

Iterative Data Mining
■ mine most informative thingy, update model, repeat.
Thank you!

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

Maximum Entropy modelling
- complex yet powerful
- inferring can be NP-hard
- analytical model can calculate exact probabilities
- can be defined for **anything** ...if you can derive the model...

Iterative Data Mining
- mine most informative thingy, update model, repeat.