Aalto University
School of Science
and Technology

Lunch is never free

How Information Theory, MDL, and Statistics are connected

Nikolaj Tatti

Department of Information and Computer Science
Aalto University, School of Science and Technology

nikolaj.tatti@aalto.fi

http://users.ics.aalto.fi/ntatti/nofreelunch2014/

September 21, 2014

Why we need model selection

Many core data mining problems can be viewed as an optimization problem:
Given data D, a set of structures $\mathcal{O}$, and a score L, find a structure O minimizing $L(D, O)$.

- how good is this particular Bayesian network?
- how good is this set of patterns?
- how good is this clustering?

Why we need model selection

We need statistics to design a good score.

Two key features

- L should be small, if data contains correlation that we want to detect
- L should be large, if O suggests some correlation that does not occur in data (or very limited)

Occam's razor

Simplest model that explains data is the best:

Bayes:
Measure the goodness of a model with a posterior $p(M \mid D)$.

Compression:
Model is good if it compresses data well.

Things to do in this tutorial

- Under certain setups, compression is very close to modelling:
  - compressing data is equal to computing maximum log-likelihood
  - MDL is equal to computing maximum posterior
- You should not actually compress data!
- When designing a score function, information-theoretic concepts have little to do with compression
  - compression is related to statistics
  - statistics are related to information theory through maximum entropy models

http://users.ics.aalto.fi/ntatti/nofreelunch2014/

Why this is important

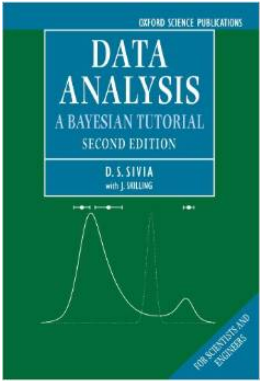
Why understanding these connections is important?

- entropy and kullback-leibler divergence are less blackbox
- tools from statistics can be used
  - p-value
  - BIC
- no more awkward justifications when using MDL
  - real values are not a problem
- no free lunch
  - every approach contains assumptions about the data
  - every approach is biased

Introduction to Bayesism

Data Analysis: A Bayesian Tutorial
by Devinderjit Sivia, John Skilling,

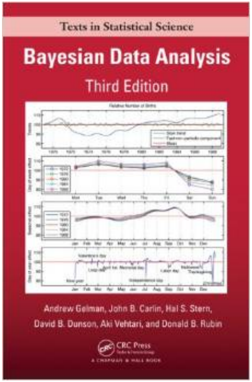
Oxford University Press
ISBN: 0198568320



Textbook on Bayesian Data Analysis

Bayesian Data Analysis
by Andrew Gelman, John B. Carlin,
Hal S. Stern and David B. Dunson

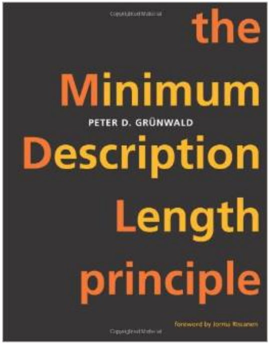
Chapman & Hall
ISBN: 1439840954



Textbook on MDL

The Minimum Description Length Principle
by Peter D. Grunwald

The MIT Press
ISBN: 0262072815

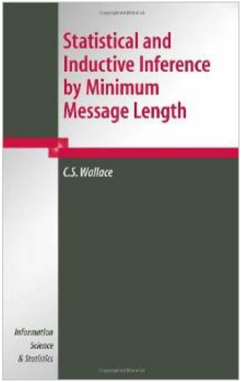


or a tutorial
http://homepages.cwi.nl/~pdg/ftp/mdlintro.ps

Textbook on MML

Statistical and Inductive Inference
by Minimum Message Length
by Chris Wallace

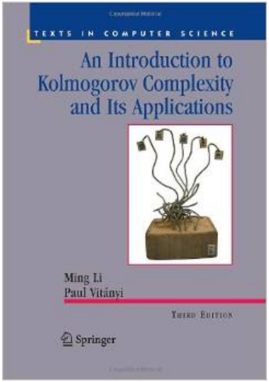
Springer
ISBN: 0071457453



Textbook on Kolmogorov Complexity

An Introduction to Kolmogorov Complexity
and Its Applications
by Ming Li and Paul Vitányi,

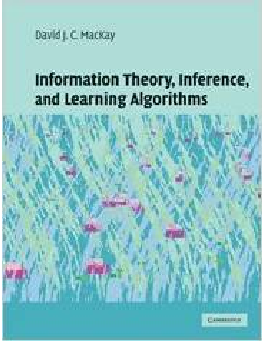
Springer
ISBN: 0387339981



Textbook on Information Theory

Information Theory, Inference and Learning Algorithms
by David J. C. MacKay

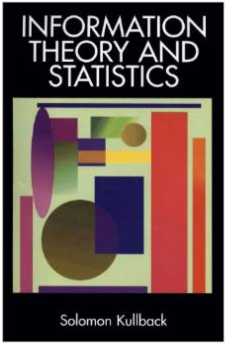
Cambridge University Press
ISBN: 0521642981



Oldie but a goodie

Information Theory and Statistics
by Solomon Kullback

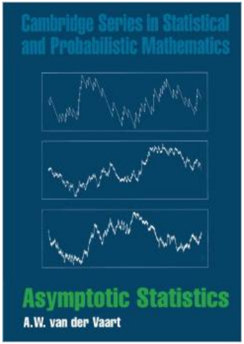
Dover Publications
ISBN: 0486696847



Everything you want to know about asymptotics

Asymptotic Statistics
by A. W. van der Vaart

Cambridge
ISBN: 0521784506



Codes and probabilities

Definitions and assumptions

$$\Omega = \text{a sample space} \quad .$$

For notational simplicity we will assume that $|\Omega|$ is finite.
For a distribution p over Ω

$$H(p) = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega) \quad .$$

Logarithms are mostly 2-base and $0 = 0 \log 0$.

Definitions and assumptions

$$\Omega = \text{a sample space} \quad .$$

For notational simplicity we will assume that $|\Omega|$ is finite.
For a distribution p over Ω

$$H(p) = -\sum_{\omega \in \Omega} p(\omega) \log p(\omega) \quad .$$

Logarithms are mostly 2-base and $0 = 0 \log 0$.
For two distributions p and q over Ω

$$KL(p \,\|\, q) = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)} \quad .$$

If there is ω s.t. $q(\omega) = 0$ and $p(\omega) > 0$, then $KL(p \,\|\, q) = \infty$.

Definitions and assumptions

For two distributions p and q over Ω

$$C(p, q) = \sum_{\omega \in \Omega} p(\omega) \log q(\omega)$$

It follows that

$$KL(p \parallel q) = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)} = -H(p) - C(p, q) \quad .$$

Kraft's Inequality

Assume we have a data D, a set of symbols $\omega \in \Omega$.
We want to trasmit D using prefix codes one-by-one.

Let $l(\omega)$ be the code length in bits.
Then

$$\sum_{\omega \in \Omega} 2^{-l(\omega)} \leq 1 \quad .$$

Define

$$p(\omega) = c2^{-l(\omega)}, \quad \text{where } c \geq 1 \text{ is such that } \sum_{\omega} p(\omega) = 1 \quad .$$

Kraft's Inequality

Kraft implies that

$$\sum_{\omega \in D} l(\omega) = \log c - \sum_{\omega \in} \log p(\omega) \geq - \sum_{\omega \in D} \log p(\omega) \quad .$$

Kraft's Inequality

Kraft implies that

$$\sum_{\omega \in D} l(\omega) = \log c - \sum_{\omega \in} \log p(\omega) \geq - \sum_{\omega \in D} \log p(\omega) \quad .$$

Let p^* be the distribution maximizing the likelihood,

$$p^* = \arg\max_q \sum_{\omega \in D} \log q(\omega) \quad .$$

Then

$$\sum_{\omega \in D} l(\omega) \geq - \sum_{\omega \in D} \log p^*(\omega) \quad .$$

Total cost is bounded by the negative maximum log-likelihood.

Maximum log-likelihood

Empirical distribution maximizes the likelihood,

$$p^*(\omega) = \frac{1}{|D|} \left|\{t \in D \mid t = \omega\}\right| \quad .$$

Moreover,

$$-\sum_{\omega \in D} \log p^*(\omega) = -|D| \sum_{\omega \in \Omega} p^*(\omega) \log p^*(\omega) = |D|\, H(p^*) \quad .$$

Maximum log-likelihood and entropy

In this case,

$$\text{negative maximum log-likelihood} = \text{entropy} \quad .$$

This is because, we considered all possible distributions, so the optimal distribution was the empirical distribution.

If we restict ourselves to some model, then this equality does not necessarily hold...
...but it holds for log-linear models.

Two ways to get close to the bound

Round up:

Given a distribution p, there is a prefix encoding s.t.

$$l(\omega) = \lceil -\log p(\omega) \rceil \quad .$$

Huffman coding is optimal if you encode one-by-one.

Arithemtic encoding

Arithmetic encoding:
Transmit data as a whole. This will bring down the transmission cost to

$$\lceil - \log p^*(D) \rceil \quad .$$

The error is at most 1 bit.
As $|D|$ increases, the relative rounding error goes down.

Why you shouldn't round up bits?

Assume $D = 100$ coin tosses, t tails and h heads.
The maximum log-likelihood

$$-\log p^*(D) = -t \log \frac{t}{100} - h \log \frac{h}{100} \quad .$$

Why you shouldn't round up bits?

Assume $D = 100$ coin tosses, t tails and h heads.
The maximum log-likelihood

$$-\log p^*(D) = -t \log \frac{t}{100} - h \log \frac{h}{100} \quad .$$

If we need to actually transmit the data one-by-one...
...we need at least 100 bits.

100 bits will be optimal transmission cost.
Doesn't depend on t!

Why you shouldn't round up bits?

You should not round bits because:

1. the score may behave badly, if done for individual symbols
2. doesn't matter, if done for the whole dataset
3. there is no need
   - the goal is to build a score not to transmit data
   - theoretical foundation and motivation comes from statistics...
   - ...compression is just a nice interpretation

Maximum Entropy models

Maximum entropy model is a way of obtaining a distribution from statistics

- provides justification for log-linear models
- connects Information Theory with Statistics
- has nice properties
  - G-test (log-likelihood ratio test)
  - BIC
- is not unbiased

Things to discuss

- definition of the model
- log-linear form
- connection between entropy and log-likelihood
- I-projections
- G-test
- solving the model
- examples

Maximum entropy models, formally

Assume n functions $S_1, \ldots, S_n$,

$$S_i : \Omega \to \mathbb{R}$$

and n numbers, $\theta_1, \ldots, \theta_n$.

Maximum entropy models, formally

Assume n functions $S_1, \ldots, S_n$,

$$S_i : \Omega \to \mathbb{R}$$

and n numbers, $\theta_1, \ldots, \theta_n$.

Let $\mathcal{P}$ be the set of all distributions such that $p \in \mathcal{P}$ if

$$\mathrm{E}_p\left[S_i\right] = \theta_i \quad \text{for every} \quad i = 1, \ldots, n \quad .$$

Maximum entropy models, formally

Assume n functions $S_1, \ldots, S_n$,

$$S_i : \Omega \to \mathbb{R}$$

and n numbers, $\theta_1, \ldots, \theta_n$.

Let $\mathcal{P}$ be the set of all distributions such that $p \in \mathcal{P}$ if

$$\mathrm{E}_p\left[S_i\right] = \theta_i \quad \text{for every} \quad i = 1, \ldots, n \quad .$$

$\mathcal{P}$ contains infinite number of distributions.
Select the one that has the highest entropy,

$$p^* = \arg\max_{p \in \mathcal{P}} H(p) \quad .$$

Maximum entropy models

1. How do we get constrains $S_1, \ldots, S_n$?

2. How do we get targets $\theta_1, \ldots, \theta_n$?

3. Is $H(p)$ the only possible (reasonable) objective?

Maximum entropy models

1. How do we get constrains $S_1, \ldots, S_n$?
   - these are god-given
   - a (weak) form of no free lunch
2. How do we get targets $\theta_1, \ldots, \theta_n$?

3. Is $H(p)$ the only possible (reasonable) objective?

Maximum entropy models

1. How do we get constrains $S_1, \ldots, S_n$?
   - these are god-given
   - a (weak) form of no free lunch
2. How do we get targets $\theta_1, \ldots, \theta_n$?
   - from the data
   - guarantees consistency
   - entropy = log-likelihood
3. Is $H(p)$ the only possible (reasonable) objective?

Maximum entropy models

1. How do we get constrains $S_1, \ldots, S_n$?
   - these are god-given
   - a (weak) form of no free lunch
2. How do we get targets $\theta_1, \ldots, \theta_n$?
   - from the data
   - guarantees consistency
   - entropy = log-likelihood
3. Is $H(p)$ the only possible (reasonable) objective?
   - No

From data to model

Assume data D, iid. samples from unknown distribution.

1. Choose n functions $S_1, \ldots, S_n$.
2. Compute
$$\theta_i = \frac{1}{|D|} \sum_{t \in D} S_i(t) \quad .$$
3. Compute maxent using $\{S_i\}$ and $\{\theta_i\}$.

Guarantees that θ_i are consistent: $\mathcal{P} \neq \emptyset$.

Linear transformation

Write

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_n \end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \quad .$$

Linear transformation

Let A be $m \times n$ matrix

- $p_1^* =$ the maxent distribution using S and θ
- $p_2^* =$ the maxent distribution using AS and $A\theta$

Then,

$$\mathrm{E}_p\left[S\right] = \theta \quad \text{implies} \quad \mathrm{E}_p\left[AS\right] = A\,\mathrm{E}_p\left[S\right] = A\theta \quad .$$

So,

$$\mathcal{P}_1 \subseteq \mathcal{P}_2 \quad \text{and so} \quad H(p_2^*) \geq H(p_1^*) \quad .$$

Linear transformation

Let A be $m \times n$ matrix

- $p_1^* =$ the maxent distribution using S and θ
- $p_2^* =$ the maxent distribution using AS and $A\theta$

If A has inverse, then

$$\mathcal{P}_1 = \mathcal{P}_2 \quad \text{and so} \quad p_1^* = p_2^* \quad .$$

Redundant constraints

If we can write

$$S_n = \alpha_1 S_1 + \alpha_2 S_2 + \cdots + \alpha_{n-1} S_{n-1},$$

and

- $p_1^* =$ the maxent distribution using $S_1, \ldots, S_n$
- $p_2^* =$ the maxent distribution using $S_1, \ldots, S_{n-1}$.

Redundant constraints

If we can write

$$S_n = \alpha_1 S_1 + \alpha_2 S_2 + \cdots + \alpha_{n-1} S_{n-1},$$

and

- $p_1^* =$ the maxent distribution using $S_1, \ldots, S_n$
- $p_2^* =$ the maxent distribution using $S_1, \ldots, S_{n-1}$.

Then

$$\mathcal{P}_1 = \mathcal{P}_2 \quad \text{and so} \quad p_1^* = p_2^* \quad .$$

Redundant constraints

If we can write

$$S_n = \alpha_1 S_1 + \alpha_2 S_2 + \cdots + \alpha_{n-1} S_{n-1},$$

and

- $p_1^* =$ the maxent distribution using $S_1, \ldots, S_n$
- $p_2^* =$ the maxent distribution using $S_1, \ldots, S_{n-1}$.

Then

$$\mathcal{P}_1 = \mathcal{P}_2 \quad \text{and so} \quad p_1^* = p_2^* \quad .$$

Number of independent constraints makes sense.

Maximum entropy and log-linear models

There are parameters $r_1, \ldots, r_n$ (under some conditions) s.t.

$$p^*(\omega) = \exp\left(r_0 + \sum_{i=1}^{n} r_i S_i(\omega)\right),$$

where r_0 is a normalization constant.

Proof (handwaving)

Fix ω. Use Lagrange multipliers: optimal solution must be

$$\frac{\partial H(p)}{\partial p(\omega)} = \frac{\partial}{\partial p(\omega)} - p(\omega) \log p(\omega) = -1 - \log p(\omega)$$
$$= \lambda_0 + \sum_{i=1}^{n} \lambda_i S_i \quad .$$

Proof (handwaving)

Fix ω. Use Lagrange multipliers: optimal solution must be

$$\frac{\partial H(p)}{\partial p(\omega)} = \frac{\partial}{\partial p(\omega)} - p(\omega)\log p(\omega) = -1 - \log p(\omega)$$
$$= \lambda_0 + \sum_{i=1}^{n} \lambda_i S_i \quad .$$

Write

$$r_0 = -\lambda_0 - 1 \quad \text{and} \quad r_i = -\lambda_i \quad .$$

Log-linear form, revisited

Constraints may force $p(\omega) = 0$
but log-linear model always has $p(\omega) > 0$.

Log-linear form, revisited

Constraints may force $p(\omega) = 0$
but log-linear model always has $p(\omega) > 0$.

Revised log-linear model

$$p^*(\omega) = \begin{cases} \exp\left(r_0 + \sum_{i=1}^{n} r_i S_i(\omega)\right), & \text{if } \omega \notin Z \\ 0, & \text{if } \omega \in Z \end{cases},$$

where $Z \subseteq \Omega$ is defined as $\omega \in Z$ iff

$$p(\omega) = 0 \quad \text{for every} \quad p \in \mathcal{P} \quad .$$

From data to model

Theorem

Let D *be a dataset of* m *samples.*

Let p *be any log-linear model.*

Then

$$\log p(D) = m(r_0 + \sum_{i=1}^{n} r_i \theta_i).$$

Maxent model p^* *obtained using* D *optimizes the likelihood and*

$$-mH(p^*) = \log p^*(D) \quad .$$

Proof

Let $q \in \mathcal{P}$, let p be any log-linear model

$$
\begin{aligned}
C(q, p) &= \sum_{\omega \in \Omega} q(\omega) \log p(\omega) \\
&= \sum_{\omega \in \Omega} q(\omega) \log \exp \left(r_0 + \sum_{i=1}^{n} r_i S_i(\omega) \right) \\
&= \sum_{\omega \in \Omega} q(\omega) \left(r_0 + \sum_{i=1}^{n} r_i S_i(\omega) \right) \\
&= r_0 + \sum_{i=1}^{n} r_i \, \mathrm{E}_q \left[S_i \right] \quad . \\
&= r_0 + \sum_{i=1}^{n} r_i \theta_i
\end{aligned}
$$

$C(q, p)$ depends on p but does not depend on $q \in \mathcal{P}$

Proof

Since $p^* \in \mathcal{P}$ and p^* is a log-linear model,

$$-H(p^*) = \sum_{\omega \in \Omega} p^*(\omega) \log p^*(\omega) = C(p^*, p^*) \quad .$$

Proof

Since $p^* \in \mathcal{P}$ and p^* is a log-linear model,

$$-H(p^*) = \sum_{\omega \in \Omega} p^*(\omega) \log p^*(\omega) = C(p^*, p^*) \quad .$$

Let q_D be the empirical distribution,

$$q_D(\omega) = \frac{1}{m} \left| \{ t \in D \mid t = \omega \} \right| \quad .$$

Proof

Since $p^* \in \mathcal{P}$ and p^* is a log-linear model,

$$-H(p^*) = \sum_{\omega \in \Omega} p^*(\omega) \log p^*(\omega) = C(p^*, p^*) \quad .$$

Let q_D be the empirical distribution,

$$q_D(\omega) = \frac{1}{m} \left|\{t \in D \mid t = \omega\}\right| \quad .$$

Then $q_D \in \mathcal{P}$ and for any log-linear model p

$$\log p(D) = \log \prod_{\omega \in D} p(\omega) = \sum_{\omega \in D} \log p(\omega)$$

$$= m \sum_{\omega \in \Omega} q_D(\omega) \log p(\omega) = mC(q_D, p) = m(r_0 + \sum_{i=1}^{n} r_i \theta_i) \quad .$$

Proof

Plug in p^* for p

$$\begin{aligned}\log p^*(D) &= mC(q_D, p^*) = mC(p^*, p^*) \\ &= m \sum_{\omega \in \Omega} p^*(\omega) \log p^*(\omega) = -mH(p^*) \quad .\end{aligned}$$

Proof

Let p be any log-linear model. Then

$$\log p^*(D) - \log p(D) = mC(q_D, p^*) - mC(q_D, p)$$

Proof

Let p be any log-linear model. Then

$$\begin{aligned}\log p^*(D) - \log p(D) &= mC(q_D, p^*) - mC(q_D, p) \\ &= mC(p^*, p^*) - mC(p^*, p)\end{aligned}$$

Proof

Let p be any log-linear model. Then

$$
\begin{aligned}
\log p^*(D) - \log p(D) &= mC(q_D, p^*) - mC(q_D, p) \\
&= mC(p^*, p^*) - mC(p^*, p) \\
&= \sum_{\omega \in \Omega} p^*(\omega) \log p^*(\omega) - \sum_{\omega \in \Omega} p^*(\omega) \log p(\omega) \\
&= KL(p^* \parallel p) \geq 0 \quad .
\end{aligned}
$$

Maximum entropy bias

Maximum entropy model is not unbiased!

It is biased towards uniform distribution.

But uniform distribution depends on Ω!

Example

Consider a dice with 6 sides, you want to have 1 or 2 to win.

You can either model the dice throw or you can model win/loss.

1. Dice throw: $\Omega = \{1, \ldots, 6\}$. Maxent model without constraints is uniform distribution.

   $$p^* \omega = 1/6 \quad .$$

   This implies that $p^*(\text{win}) = 2/6$ and $p^*(\text{loss}) = 4/6$.
2. Win/loss: $\Omega = \{\text{win}, \text{loss}\}$.
   In this case, $p^*(\text{win}) = 1/2$ and $p^*(\text{loss}) = 1/2$.

Maximum entropy bias

Assume that the raw data is a set of iid samples residing in Ω_{raw}.

Map the data from Ω_{raw} to Ω_{clean}

- data collection
- preprocessing

Maxent models in Ω_{raw} and Ω_{clean} can be different.

I-projections

Assume n functions $S_1, \ldots, S_n$,

$$S_i : \Omega \to \mathbb{R}$$

n numbers, $\theta_1, \ldots, \theta_n$,
and a target distribution q.

I-projections

Assume n functions $S_1, \ldots, S_n$,

$$S_i : \Omega \to \mathbb{R}$$

n numbers, $\theta_1, \ldots, \theta_n$,
and a target distribution q.

Let $\mathcal{P}$ be the set of all distributions such that $p \in \mathcal{P}$ if

$$E_p[S_i] = \theta_i \quad \text{for every} \quad i = 1, \ldots, n \quad .$$

I-projections

Assume n functions $S_1, \ldots, S_n$,

$$S_i : \Omega \to \mathbb{R}$$

n numbers, $\theta_1, \ldots, \theta_n$,
and a target distribution q.

Let $\mathcal{P}$ be the set of all distributions such that $p \in \mathcal{P}$ if

$$\mathrm{E}_p\left[S_i\right] = \theta_i \quad \text{for every} \quad i = 1, \ldots, n \quad .$$

Select the one that is closest to q,

$$p^* = \arg\min_{p \in \mathcal{P}} KL(p \,\|\, q) \quad .$$

Log-linear form

Log-linear model

$$p^*(\omega) = \begin{cases} q(\omega)\exp\left(r_0 + \sum_{i=1}^n r_i S_i(\omega)\right), & \text{if } \omega \notin Z \\ 0, & \text{if } \omega \in Z \end{cases},$$

where $Z \subseteq \Omega$ is defined as $\omega \in Z$ iff

$$p(\omega) = 0 \quad \text{for every} \quad p \in \mathcal{P} \quad .$$

I-projections in different universes

Assume

- two universes Ω_1 and Ω_2,
- a transformation $T : \Omega_1 \to \Omega_2$,
- set of functions $S_1, \ldots, S_n$, $S_i : \Omega_2 \to \mathbb{R}$,
- a target distribution q_1 in Ω_1.

I-projections in different universes

Assume

- two universes Ω_1 and Ω_2,
- a transformation $T : \Omega_1 \to \Omega_2$,
- set of functions $S_1, \ldots, S_n$, $S_i : \Omega_2 \to \mathbb{R}$,
- a target distribution q_1 in Ω_1.

Let

- $p_1^* =$ I-projection on Ω_1 with $S_1 \circ T, \ldots, S_n \circ T$ and q_1.
- $p_2^* =$ I-projection on Ω_2 with $S_1, \ldots, S_n$ and q_2, where

$$q_2(\omega_2) = q_1(T(\omega_1) = \omega_2) \quad .$$

Then $p_2^*(\omega_2) = p_1^*(T(\omega_1) = \omega_2)$.

I-projection and Maximum entropy

If q_1 is uniform, then I-projection $=$ maxent.

If transformation T is not 1-1, then

- maxent distribution in Ω_1 is not maxent in Ω_2
- maxent distribution in Ω_1 is I-projection to q_2 in Ω_2

Difference between two models

Let $S_1, \ldots, S_n$ be n functions.
Two distributions

- $p_1^* =$ the maxent distribution using $S_1, \ldots, S_n$
- $p_2^* =$ the maxent distribution using $S_1, \ldots, S_k$, where $k \leq n$.

We know already that

$$H(p_2^*) \geq H(p_1^*)$$

but now

$$H(p_2^*) - H(p_1^*) = KL(p_1^* \parallel p_2^*) \quad .$$

Proof

Since $p_1^* \in \mathcal{P}_1 \subseteq \mathcal{P}_2$,

$$H(p_2^*) = -C(p_2^*, p_2^*) = -C(p_1^*, p_2^*) \quad .$$

Then

$$\begin{aligned} H(p_2^*) - H(p_1^*) &= -C(p_1^*, p_2^*) - H(p_1^*) \\ &= -\sum_{\omega \in \Omega} p_1^*(\omega) \log p_2^*(\omega) + \sum_{\omega \in \Omega} p_1^*(\omega) \log p_1^*(\omega) \\ &= KL(p_1^* \parallel p_2^*) \quad . \end{aligned}$$

Difference between two models

Assume data D with m samples.
Let p_1^* and p_2^* be two maxent distributions.
Then

$$m(H(p_2^*) - H(p_1^*)) = mKL(p_1^* \parallel p_2^*) = \log \frac{p_1^*(D)}{p_1^*(D)} \quad .$$

Kullback-leibler = log-likehood ratio!

Asymptotic behaviour

Assume $S_1, \ldots, S_n$ and let $k < n$.
Assume data D with m samples coming from a log-linear model based on $S_1, \ldots, S_k$.

Two distributions

- $p_1^* =$ the maxent distribution using $S_1, \ldots, S_n$
- $p_2^* =$ the maxent distribution using $S_1, \ldots, S_k$, where $k \leq n$.

Then, under conditions,

$$2mKL(p_1^* \parallel p_2^*) \to \chi^2(n-k) \quad \text{as} \quad m \to \infty \quad .$$

(here we use natural logarithm)

Conditions

1. conditions that make covariance to be finite
2. conditions that make $S_1, \ldots, S_n$ 'independent'

   Otherwise, we can duplicate the last constraint and have $p_1^* =$ the maxent distribution using $S_1, \ldots, S_n, S_n$

Asymptotic behaviour

Let p^* be the true distribution.
We can write

$$\begin{aligned}
KL(p_1^* \parallel p_2^*) &= H(p_2^*) - H(p_1^*) \\
&= H(p_2^*) + C(p_2^*, p^*) - H(p_1^*) - C(p_2^*, p^*) \\
&= H(p_2^*) + C(p_2^*, p^*) - H(p_1^*) - C(p_1^*, p^*) \\
&= KL(p_1^* \parallel p^*) - KL(p_2^* \parallel p^*) \quad .
\end{aligned}$$

Both terms converge to 0 but at different rates.

Asymptotic behaviour

Write

$$S = [S_1, \ldots, S_n], \quad \theta_1 = \frac{1}{m} \sum_{t \in D} S(t), \quad \theta_1^* = \mathrm{E}_{p^*}[S] \quad .$$

Asymptotic behaviour

Write

$$S = [S_1, \ldots, S_n], \quad \theta_1 = \frac{1}{m} \sum_{t \in D} S(t), \quad \theta_1^* = \mathrm{E}_{p^*}[S] \quad .$$

Estimate with a 2nd-degree Taylor,

$$2mKL(p_1^* \parallel p^*) \approx m(\theta_1 - \theta_1^*)^T H(\theta_1 - \theta_1^*) \to X^T C_1^{-1} X \quad .$$

where X is distributed as $N(0, C_1)$ and

$$C_1 = \text{covariance matrix for } S_1, \ldots, S_n = \mathrm{Cov}_{p^*}[S] \quad .$$

Asymptotic behaviour

Similarly, write

$$T = [S_1, \ldots, S_k], \quad \theta_2 = \frac{1}{m} \sum_{t \in D} T(t), \quad \theta_2^* = \mathrm{E}_{p^*}\left[T\right] \quad .$$

Asymptotic behaviour

Similarly, write

$$T = [S_1, \ldots, S_k], \quad \theta_2 = \frac{1}{m} \sum_{t \in D} T(t), \quad \theta_2^* = \mathrm{E}_{p^*}\left[T\right] \quad .$$

Then,

$$2mKL(p_2^* \parallel p^*) \approx m(\theta_2 - \theta_2^*)^T H(\theta_2 - \theta_2^*) \to Y^T C_2^{-1} Y \quad .$$

where Y is distributed as $N(0, C_1)$ and

$$C_2 = \text{covariance matrix for } S_1, \ldots, S_k = \mathrm{Cov}_{p^*}\left[T\right] \quad .$$

Asymptotic behaviour

After a little algebra

$$X^T C_1^{-1} X - Y^T C_2^{-1} Y = Z^T Z,$$

where Z is a vector of length $n - k$ and distributed as $N(0, I)$.
This is known to be distributed as $\chi^2(n - k)$.

Solving MaxEnt

To infer the model, we need to find $r_1, \ldots, r_n$, maximizing,

$$f(r_0, \ldots, r_n) = r_0 + \sum_i^n r_i \theta_i \quad .$$

Solving MaxEnt

To infer the model, we need to find $r_1, \ldots, r_n$, maximizing,

$$f(r_0, \ldots, r_n) = r_0 + \sum_i^n r_i \theta_i \quad .$$

Luckily,

$$\text{Hessian} = \left(\frac{\partial \log f}{\partial r_i \partial r_j} \right)_{ij} = \text{Cov}_p \left[S \right]$$

is positive semidefinite.

Solving MaxEnt

To infer the model, we need to find $r_1, \ldots, r_n$, maximizing,

$$f(r_0, \ldots, r_n) = r_0 + \sum_i^n r_i \theta_i \quad .$$

Luckily,

$$\text{Hessian} = \left(\frac{\partial \log f}{\partial r_i \partial r_j} \right)_{ij} = \text{Cov}_p[S]$$

is positive semidefinite.

Any local maximum is also a global maximum. Gradient is

$$\frac{\partial f}{\partial r_i} = \theta_i - \text{E}_p[S_i] \quad .$$

Computing $\text{E}_p[S_i]$ may be difficult if Ω is large.

Difficult to compute

The following problem is **NP**-hard:
Let $\Omega = \{0, 1\}^k$.
Assume

- a set of conjunctive constraints $S_1, \ldots, S_n$,
- targets $\theta_1, \ldots, \theta_n$,
- and a conjunctive query Q.

Compute $\mathrm{E}_{p^*}[Q]$.

Examples

No constraints: $p^* =$ uniform distribution.

Examples

No constraints: $p^* =$ uniform distribution.

Constraints define distribution completely:
If we use $|\Omega| - 1$ constraints with

$$S_i(\omega) = [i = \omega],$$

then $\mathcal{P} = \{p^*\}$ and $p^*(i) = \theta_i$.

Examples

Gaussian model: If

$$\Omega = \mathbb{R}, \quad S_1(\omega) = \omega, \quad \text{and} \quad S_2(\omega) = \omega^2,$$

then $p^* =$ Gaussian distribution.

Examples

Gaussian model: If

$$\Omega = \mathbb{R}, \quad S_1(\omega) = \omega, \quad \text{and} \quad S_2(\omega) = \omega^2,$$

then $p^* =$ Gaussian distribution.

Independence model: If

$$\Omega = \{0, 1\}^k, \quad S_i(\omega) = \omega_i, \quad \text{for} \quad i = 1, \ldots, k,$$

then $p^* =$ independence model with θ_i as margins.

Examples

Gaussian model: If

$$\Omega = \mathbb{R}, \quad S_1(\omega) = \omega, \quad \text{and} \quad S_2(\omega) = \omega^2,$$

then $p^* =$ Gaussian distribution.

Independence model: If

$$\Omega = \{0, 1\}^k, \quad S_i(\omega) = \omega_i, \quad \text{for} \quad i = 1, \ldots, k,$$

then $p^* =$ independence model with θ_i as margins.

Partition model: If $\Omega = \Omega_1 \times \cdots \times \Omega_k$
and S_i depends only on Ω_i, then

$$p^*(t) = p_1^*(t_{\Omega_1}) \cdots p_k^*(t_{\Omega_k}) \quad .$$

Mutual information

Assume

- two random categorical variables X and Y,
- $n =$ number of possible states for X,
- $m =$ number of possible states for Y.

Let

$$\begin{aligned} p &= \text{ joint empirical distribution of } X \text{ and } Y, \\ p_1 &= \text{marginal distribution of } X, \\ p_2 &= \text{marginal distribution of } Y \quad . \end{aligned}$$

Mutual information

$$H(p_1) + H(p_2) - H(p) \quad .$$

Mutual information

Joint distribution

$$p = \text{ maxent distribution with } nm - 1 \text{ constraints.}$$

Independent distribution

$$\begin{aligned} p_{\mathrm{ind}}(x, y) &= p_1(x) p_2(y) \\ &= \text{ maxent distribution with } n + m - 2 \text{ constraints.} \end{aligned}$$

Mutual information

Joint distribution

$$p = \text{ maxent distribution with } nm - 1 \text{ constraints.}$$

Independent distribution

$$\begin{aligned} p_{\text{ind}}(x, y) &= p_1(x)p_2(y) \\ &= \text{ maxent distribution with } n + m - 2 \text{ constraints.} \end{aligned}$$

Mutual information

$$H(p_1) + H(p_2) - H(p) = M(p) = H(p_{\text{ind}}) - H(p) = KL(p \parallel p_{\text{ind}}) \quad .$$

Mutual information

Joint distribution

$$p = \text{ maxent distribution with } nm - 1 \text{ constraints.}$$

Independent distribution

$$\begin{aligned} p_{\text{ind}}(x, y) &= p_1(x) p_2(y) \\ &= \text{ maxent distribution with } n + m - 2 \text{ constraints.} \end{aligned}$$

Mutual information

$$H(p_1) + H(p_2) - H(p) = M(p) = H(p_{\text{ind}}) - H(p) = KL(p \parallel p_{\text{ind}}) \quad .$$

Also,

$$2\left|D\right| KL(p \parallel p_{\text{ind}}) \to \chi^2(nm - n - m + 1)$$

can be used as statistical test.

Rasch models

Let $\Omega = \{0, 1\}^{k \times m}$ be the universe of $k \times m$ matrices.

Rasch models

Let $\Omega = \{0, 1\}^{k \times m}$ be the universe of $k \times m$ matrices. Define $k + m$ constraints,

$$S_i(\omega) = \sum_{j=1}^{m} \omega_{ij}, \quad \text{for} \quad i = 1, \ldots, k \quad .$$

and

$$T_j(\omega) = \sum_{i=1}^{k} \omega_{ij}, \quad \text{for} \quad j = 1, \ldots, m \quad .$$

Given a single binary dataset D, compute target row and column sums.

Rasch models

Let $\Omega = \{0, 1\}^{k \times m}$ be the universe of $k \times m$ matrices. Define $k + m$ constraints,

$$S_i(\omega) = \sum_{j=1}^{m} \omega_{ij}, \quad \text{for} \quad i = 1, \ldots, k \quad .$$

and

$$T_j(\omega) = \sum_{i=1}^{k} \omega_{ij}, \quad \text{for} \quad j = 1, \ldots, m \quad .$$

Given a single binary dataset D, compute target row and column sums.

$$p^*(\omega) = \prod_{i,j} p^*_{ij}(\omega_{ij}) \quad .$$

Chow-Liu tree model

Assume $\Omega = \{0, 1\}^k$. Let $2k - 1$ constraints,

- margins of individual variables,

$$S_i(\omega) = \omega_i,$$

- selected co-occurences,

$$C_(i, j)(\omega) = \omega_i \omega_j,$$

- co-occurences must form a tree T.

Chow-Liu tree model

Assign a root r to a tree, p^* is a bayesian network, where a node can have only one parent,

$$p^*(\omega) = p^*(\omega_r) \prod_{i \neq r} p^*(\omega_i \mid \omega_{par(i)}) \quad .$$

Rewrite

$$\begin{aligned} p^*(\omega) &= p^*(\omega_r) \prod_{i \neq r} \frac{p^*(\omega_i, \omega_{par(i)})}{p^*(\omega_{par}(i))} \\ &= \Big[\prod_i p^*(\omega_i)\Big]\Big[\prod_{i \neq r} \frac{p^*(\omega_i, \omega_{par(i)})}{p^*(\omega_i) p^*(\omega_{par}(i))}\Big] \\ &= \Big[\prod_i p^*(\omega_i)\Big]\Big[\prod_{(i,j) \in T} \frac{p^*(\omega_i, \omega_j)}{p^*(\omega_i) p^*(\omega_j)}\Big] \quad . \end{aligned}$$

Chow-Liu tree model

The entropy is equal to

$$H(p^*) = \sum_{i=1}^{k} H(\Omega_i) + \sum_{(i,j) \in T} H(\Omega_i, \Omega_j) - H(\Omega_i) - H(\Omega_j) \quad .$$

Define

$$w(i, j) = H(\Omega_i) + H(\Omega_j) - H(\Omega_i, \Omega_j) \quad .$$

Optimal T = smallest spanning tree.
Finding optimal structure for general Bayesian network is **NP**-hard.

Model Selection

3 'variants' of model selection

1. Bayes (1812)
2. Kolmogorov (1960–1965)
3. MML (1968)
4. MDL (1978)

Bayesian model selection

A statistical model = set of distributions parameterized by parameters

meaning that we have the likelihood

$$p(D \mid M, \theta) \quad \text{and the prior} \quad p(M, \theta) \quad .$$

Assume two models M_1 and M_2

- M_1 is parameterized by θ_1
- M_2 is parameterized by θ_2

Compute

$$p(M_i \mid D) \propto p(D \mid M_i)p(M_i) = p(M_i) \int p(D \mid M_i, \theta_i)p(\theta_i \mid M_i)d\theta_i \quad .$$

Bayesian model selection

$$\int p(D \mid M_i, \theta_i) p(\theta_i \mid M_i) d\theta_i$$

punishes complex models.

A flexible model will contain many different distributions.

For a fixed dataset D, there will be some distributions that have high likelihood...

...but most of them will have bad likelihood.

Bayesian model selection example

Model a coin toss.

First model: Bernoulli variable with 0.5 probability,
no parameters.
Second model: Bernoulli variable with a unknown probability,
one parameter, uniform prior.

Assume $p(M_1) = p(M_2) = 1/2$ (i.e., ignore them).

Dataset with t tails and h heads.

Bayesian model selection example

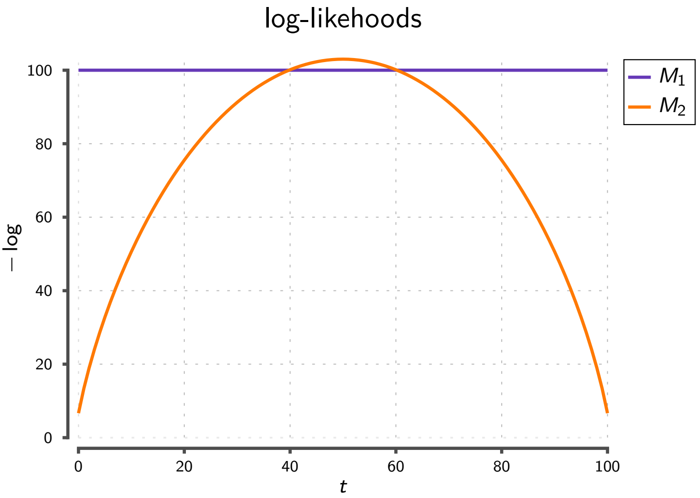
First model:

$$p(M_1 \mid D) \propto (1/2)^t (1/2)^h.$$

Second model:

$$p(M_2 \mid D) \propto \int \theta^t (1-\theta)^h d\theta = \binom{t+h+1}{t}^{-1} \quad .$$

Bayesian model selection example



Bayesian model selection

Integral is sometimes easy to compute

- especially with conjugate priors

but with complex models, it may be intractable

- so we need to estimate it, most likely with MCMC
- this is very slow and problematic...
- ...even worse we need this in an inner loop

Bayesian Information Criterion (BIC)

Assume a log-linear model M with n constraints.

Then for almost any prior,

$$\int p(D \mid \theta)p(\theta)d\theta = \log p^*(D) - \log |D| \, n/2 + o(\log |D|) \quad .$$

We can estimate the $p(D)$ using the first two terms

- No need to integrate
- No need to select the prior for θ

Kolmogorov Complexity

Defined (for convienience) for bit strings.

Assume a (Turing complete) description language M.

- Your favorite program language
- Universal Turing machine
- ...

The complexity of an object s

$$K(s) = |\text{the shortest input sequence for } M \text{ to produce } s| \quad .$$

Kolmogorov Complexity (invariance theorem)

Theorem

Assume two description languages. Then for any object s

$$|K_1(s) - K_2(s)| \leq c,$$

where c is a constant depending only on the languages.

Kolmogorov Complexity (invariance theorem)

Theorem

Assume two description languages. Then for any object s

$$|K_1(s) - K_2(s)| \leq c,$$

where c is a constant depending only on the languages.

Proof.

Make a simulator for K_1 using K_2, let its length be c. Then

$$K_2(s) \leq c + K_1(s) \quad .$$

□

Kolmogorov Complexity

Use universal Turing machine. $K(s)$ consists of

- a bit sequence and
- a Turing machine that produces s.

Kolmogorov Complexity

Use universal Turing machine. $K(s)$ consists of

- a bit sequence and
- a Turing machine that produces s.

Theorem

$K(s)$ is not computable.

Kolmogorov Complexity

Use universal Turing machine. $K(s)$ consists of

- a bit sequence and
- a Turing machine that produces s.

Theorem

$K(s)$ is not computable.

- you can still use it for theoretical analysis
- you can estimate $K(s)$ with your favorite compressor (zip)
  - simple and stupid hack...
  - ...but works quite often.

Minimum Description Length (MDL)

Assume data D that you want to transmit.
For receiver to decode D, it needs to know the distribution p.

1. receiver and transmitter have agreed about some model M.
2. transmitter sends parameters θ to pin point p from M.
3. transmitter sends data according to p.

Transmitter tries to optimize the cost for both parameters and the data.

MDL length

Let

$$L'(D \mid \theta) = \text{length of data transmission}$$

and

$$L'(\theta) = \text{length of parameter transmission} \quad .$$

These are valid prefix encodings if

$$\sum_D 2^{-L'(D|\theta)} \leq 1 \quad \text{and} \quad \sum_\theta 2^{-L'(\theta)} \leq 1$$

and $L'(D \mid \theta)$ and $L'(\theta)$ are integers.

MDL length

MDL does not care about actual encodings!

Problem (MDL)

Given D, $L(D \mid \theta)$ *and* $L(\theta)$ *such that*

$$\sum_D 2^{-L(D|\theta)} \leq 1 \quad \textit{and} \quad \sum_\theta 2^{-L(\theta)} \leq 1$$

find θ *maximizing*

$$L(D, \theta) = L(D \mid \theta) + L(\theta) \quad .$$

MDL

$L(D \mid \theta)$ and $L(\theta)$ does not need to be integers.

- $L(D, \theta)$ is a lower bound for optimal prefix encoding
- optimal prefix encoding is an upper bound for Kolmogorov complexity

Bad time


IF YOU USE MDL AND ROUND BITS
INSTRUCTO
YOU'RE GOING TO HAVE A BAD TIME
memegenerator.net

MDL and Bayesian modelling

We can express MDL using Bayesian formalism:
Write

$$p(D \mid \theta) = c_1 2^{-L(D|\theta)} \quad \text{and} \quad p(\theta) = c_2 2^{-L(\theta)},$$

where $c_1, c_2 \geq 1$ guarantee that $p(D \mid \theta)$ are $p(\theta)$ sum to 1.

- $p(D \mid \theta)$ is the probability of D given θ.
- $p(\theta)$ is the prior probability of θ.

Find θ maximizing

$$\log p(D, \theta) = \log p(D \mid \theta) + \log p(\theta) = -L(D \mid \theta) - L(\theta) + \log c_1 + \log c_2 \quad .$$

MDL and Bayesian modelling

MDL: choose an encoding
Bayes: choose a statistical model

MDL and Bayesian modelling

MDL: choose an encoding
Bayes: choose a statistical model

MDL: choose an encoding for parameters
Bayes: choose a prior

MDL and Bayesian modelling

MDL: choose an encoding
Bayes: choose a statistical model

MDL: choose an encoding for parameters
Bayes: choose a prior

MDL: encode samples in your data one-by-one
Bayes: assume that they are independent

MDL and Bayesian modelling

MDL: choose an encoding
Bayes: choose a statistical model

MDL: choose an encoding for parameters
Bayes: choose a prior

MDL: encode samples in your data one-by-one
Bayes: assume that they are independent

MDL: optimal data encoding is given by entropy
Bayes: maximum log-likelihood is given by empirical distribution

MDL and Bayesian modelling

Bayes: choose the model with the highest

$$p(M_i \mid D) \propto p(D \mid M_i)p(M_i) = p(M_i) \int p(D \mid M_i, \theta_i)p(\theta_i \mid M_i)d\theta_i \quad .$$

MDL: choose the model with the highest

$$\log p(D, \theta^*) = \log p(D \mid \theta^*) + \log p(\theta^*) \quad .$$

Punishing complex models:

- Bayes: many models will have a low likelihood
- MDL: compressing parameters will be more expensive

Normalized Maximum Likelihood (NML)

Two-part encoding:

$$\log p(D \mid \theta) + \log p(\theta) \quad .$$

Seems arbitrary (Bayes is equally arbitrary!)

$L(D \mid \theta^*)$ violates Kraft's inequality, so we cannot use it alone. Normalize it so it sums to 1:

$$p(D) = \frac{2^{-L(D|\theta^*)}}{\sum_{D'} 2^{-L(D'|\theta^*)}} \quad .$$

NML

Define

$$L_{NML}(D) = -\log p(D) = L(D \mid \theta^*) + STOC,$$

where $STOC$ is stochastic complexity,

$$STOC = \log \sum_{D'} 2^{-L(D' \mid \theta^*)} \quad .$$

Punishing complex models:
Complex model will yield small $L(D' \mid \theta^*)$ for many D'.
This will make $STOC$ large.

Example

Dataset

$$D = 100 \text{ coin tosses} \quad .$$

Log-likelihood,

$$L(D \mid \theta^*) = -t \log \frac{t}{100} - h \log \frac{h}{100} \quad .$$

Stochastic complexity,

$$STOC = \log \sum_{t=0}^{100} \binom{100}{t} \left[\frac{t}{100}\right]^t \left[\frac{100-t}{100}\right]^{100-t} \quad .$$

MDL for categorical data

Assume a random variable of k possible outcomes.
Assume data with n observations.

Then, stochastic complexity is,

$$\begin{aligned} STOC(k, n) &= \sum_{n_1 + \cdots + n_k = n} \frac{n!}{n_1! \cdots n_k!} \prod_i \left[\frac{n_i}{n}\right]^{n_i} \\ &= \sum_{n_1 + m = n} \frac{n!}{n_1! m!} \frac{m^m}{n^n} \sum_{n_2 + \cdots + n_k = m} \frac{m!}{n_1! \cdots n_k!} \prod_i \left[\frac{n_i}{m}\right]^{n_i} \\ &= \sum_{n_1 + m = n} \frac{n!}{n_1! m!} \frac{m^m}{n^n} STOC(k-1, m) \quad . \end{aligned}$$

Can be computed iteratively.

MDL for independent variables

Assume that we are modelling two variables independently:
Sample space, $\Omega = \Omega_1 \times \Omega_2$ and

$$
\begin{aligned}
p(D \mid \theta) &= p(D_1 \mid \theta_1) p(D_2 \mid \theta_2) \quad \text{or, alternatively} \\
L(D \mid \theta) &= L(D_1 \mid \theta_1) + L(D_2 \mid \theta_2) \quad .
\end{aligned}
$$

MDL for independent variables

Assume that we are modelling two variables independently:
Sample space, $\Omega = \Omega_1 \times \Omega_2$ and

$$p(D \mid \theta) = p(D_1 \mid \theta_1) p(D_2 \mid \theta_2) \quad \text{or, alternatively}$$
$$L(D \mid \theta) = L(D_1 \mid \theta_1) + L(D_2 \mid \theta_2) \quad .$$

Then,

$$\begin{aligned}
STOC(\theta) &= \log \sum_D p(D \mid \theta^*) = \log \sum_D p(D_1 \mid \theta_1) p(D_2 \mid \theta_2) \\
&= \log \sum_{D_1} p(D_1 \mid \theta_1) \sum_{D_2} p(D_2 \mid \theta_2) \\
&= \log \sum_{D_1} p(D_1 \mid \theta_1) + \log \sum_{D_2} p(D_2 \mid \theta_2) \\
&= STOC(\theta_1) + STOC(\theta_2) \quad .
\end{aligned}$$

MDL and BIC

Assume a log-linear model M with n constraints.

Then, under certain conditions

$$STOC(M) = \frac{n}{2}\log|D| - \frac{k}{2}\log 2\pi + \log \int \sqrt{\det I(\theta)}d\theta + o(1),$$

where $I(\theta)$ is the Fisher Information.

MDL and BIC

Assume a log-linear model M with n constraints.

Then, under certain conditions

$$STOC(M) = \frac{n}{2} \log |D| - \frac{k}{2} \log 2\pi + \log \int \sqrt{\det I(\theta)} d\theta + o(1),$$

where $I(\theta)$ is the Fisher Information.

- the first term is BIC penalty
- the second term is constant
- the third term is integral that can be sometimes computed

NML

More generally, NML is very hard to compute.

NML gives you way to encode (select a prior) for θ...
...but it is not unbiased.

- the stochastic complexity goes over all possible dataset
- this makes it depend on Ω
- preprocessing data may change the complexity term
- similar bias as with maximum entropy models

Complexity measures, summary

Bayes:

$$p(M_i \mid D) \propto p(D \mid M_i)p(M_i) = p(M_i) \int p(D \mid M_i, \theta_i)p(\theta_i \mid M_i)d\theta_i \quad .$$

Kolmogorov:

$$K(s) = |\text{sequence to produce } s| + |\text{Turing machine that produces } s| \quad .$$

MDL:

$$\log p(D, \theta^*) = \log p(D \mid \theta^*) + \log p(\theta^*) \quad .$$

MDL (NML):

$$\log p(D, \theta^*) = \log p(D \mid \theta^*) + STOC(\theta) \quad .$$

Wrap up

Occam's razor is the guideline to design model scores

- Bayesian/statistical approach
- Compression arguments

Maximum entropy models provide a bridge between information theory and statistics

- log-linear models
- BIC and G-test
- biased towards uniformity in the current space (I-projections)

MDL is a statistical method disguised as compression method

Afterthoughts

All methods make choices

- assumptions on a data (i.i.d, etc)
- model to use / encoding scheme
- prior assumptions on the parameters

There is no free lunch..

Afterthoughts

All methods make choices

- assumptions on a data (i.i.d, etc)
- model to use / encoding scheme
- prior assumptions on the parameters

There is no free lunch..

... but maybe it is a good thing

Afterthoughts

Statisticians abhor making choices

- a lot of research on uniformative priors

... but choices are good for us, data miners:

- we can design our score that values what we are interested in
- MDL / statistics will provide tools but not answers
- not all scores labelled as MDL are automatically good

Afterthoughts

MDL gives you tools to make choices

- if your algorithm has parameters, then you can use MDL as a guide
- but it may be a good idea to make using MDL optional: Top-k vs. MDL

If Kolmogorov Complexity was computable, would it be useful?