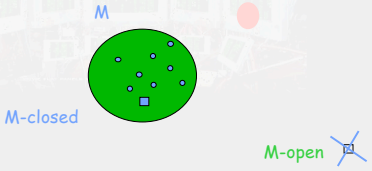
On Minimum Description Length Modeling

On Modeling



Do you believe that the data generating mechanism really is in your model class M?

"non-M-closed" predictive inference

- Explicitly include prediction (and intervention) in modeling

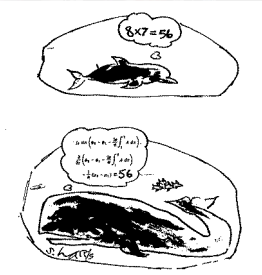
Models are a means (a language) to describe interesting properties of the phenomenon to be studied, but they are not intrinsic to the phenomenon itself.

"All models are false, but some are useful."

Minimum Description Length Principle

- "MDL" is a method related to modeling, inductive inference, machine learning…
- Rissanen 1978-; Barron, Rissanen and Yu 1998
- tasks
  - Model selection
  - Parameter estimation
  - Prediction
- "From arithmetic coding to modeling"

"Model selection"



Descriptive complexity

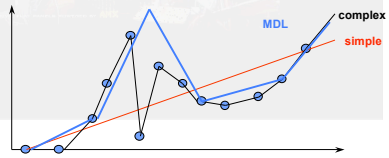
- 1000 bit strings
  - 00010001000100010001000 …….. 00010001
  - 01110100110100010101010 …….. 10101110
  - 11100111111010011111111 …….. 01111011
- Solomonoff-Kolmogorov-Chaitin complexity
  - "shortest possible encoding with the help of L"
  - code based on a universal computer language L
  - too strong a description language - uncomputability

The idea

- a good model M captures regular features (constraints) of the observed data
- any set of regularities we find reduces our uncertainty of the data D, and we can use it to encode the data in a shorter and less redundant way
- The more we are able to compress a sequence of data, the more regularity you have detected in the data and the more you have learned from the data (to make predictions of future data)

For example regression

- There is a trade-off between the model complexity and fit to the data



"Types" of MDL

- algorithmic, "ideal" MDL (Li and Vitányi '97)
- MML (Wallace '68, '87)
- two-part code MDL (Rissanen '78, '83)
- universal model based MDL (Rissanen '96, Barron, Rissanen, Yu '98, Grünwald '02)

Probability

$\mathcal{X}$ "sample" space

$\mathcal{X}^n$ set of all sequences of n outcomes

$\mathcal{X}^+$ set of arbitrary length sequences

$\mathcal{X}^\infty$ set of infinite sequences

P probability distribution over $\mathcal{X}^\infty$

Code

$\mathcal{X}$ (countable) data alphabet

A (uniquely decodable) code C is a one-to-one map from $\mathcal{X}$ to $\{0,1\}^+$

$L_C(x)$ denotes the length in bits needed to describe x

Analogy

- let P be a probability distribution. Since $\sum_x P(x) \le 1$
- only very few x can have large probability
- let C be a code for $\{0,1\}^m$. Since the fraction of sequences that can be compressed by more than k bits is less than $2^{m-k} / 2^m = 2^{-k}$

only very few symbols can have small code length

Correspondence

- there is a 1-1 correspondence between probability distributions and code length functions such that large code lengths correspond to small probabilities and vice versa

$$\text{for all } x^n \in \chi^+ : L(x^n) = -\log P(x^n)$$

Universal codes

- L is a set of code(length function)s available to encode data x^n
- assume that one of the code(length function)s in L allows for substantial compression of x^n
- TASK: encode x^n using minimum number of bits!

Universal code (more)

- For example: L finite
- There exists a code L_L such that for some constant K, for all n, x^n, all $L \in L$:

$$L_L(x^n) \le L(x^n) + K$$

- Specifically

$$L_L(x^n) \le \inf_{L \in L} L(x^n) + K$$

- K does not depend on n, while typically $L(x^n)$ grows linearly in n

Universal Models

- Let M be aprobabilistic model, i.e., a family (set) of probability distributions
- Assume M finite: $M = \{P(\cdot \mid \theta_1), \ldots, P(\cdot \mid \theta_M)\}$
- There exists a code L_M s.t. for all n, x^n, θ:

$$L_M(x^n) \le -\log P(x^n \mid \theta) + K$$

- hence, exists distribution P_M s.t.

$$-\log P_M(x^n) \le -\log P(x^n \mid \theta) + K$$

- i.e. $P_M(x^n) \ge K' \cdot (x^n \mid \theta)$

P_M is a "universal model" (distribution) for M

Bayesian mixture as a univeral model

- let W be a prior over M. The Bayesian marginal likelihood is defined as:

$$P_{Bayes}(x^n \mid M) = \sum_{j=1}^{m} P(x^n \mid \theta_j) W(\theta_j)$$

- This is a universal model, since

$$\text{For all } n, x^n, \theta : -\log P_{Bayes}(x^n \mid M) = -\log \sum_{j=1}^{m} P(x^n \mid \theta_j) W(\theta_j) \le -\log \sum_{j=1}^{m} P(x^n \mid \theta) - \log W(\theta)$$

Two-part MDL code as a universal model

- The ML (maximum likelihood) distribution is $\hat{\theta}(x^n)$

$$\inf_{P(\cdot \mid \theta) \in M} \{-\log P(x^n \mid \theta)\}$$

- code x^n by first coding $\hat{\theta}(x^n)$, then coding x^n with the help of $\hat{\theta}(x^n)$:

$$L_{2p}(x^n \mid M) = -\log W(\hat{\theta}(x^n)) - \log P(x^n \mid \hat{\theta}(x^n))$$

- Bayes' mixture assigns larger probability (shorter code length) to outcomes....
- what prior leads to short code lengths?

Optimal Universal Model

Look for P* such that regret

$$-\log P^*(x^n) - \left(-\log P(x^n \mid \hat{\theta}(x^n))\right)$$

is small no matter what x^n are; i.e. look for

$$\inf_{P^*} \sup_{x^n \in \mathcal{X}^n} \left\{-\log P^*(x^n) - \left(-\log P(x^n \mid \hat{\theta}(x^n))\right)\right\}$$

Optimal Universal Model

$$\inf_{P^*} \sup_{x^n \in \mathcal{X}^n} \left\{-\log P^*(x^n) - \left(-\log P(x^n \mid \hat{\theta}(x^n))\right)\right\}$$

is achieved by Normalized Maximum Likelihood (NML) distribution

$$P_{NML}(x^n \mid M) = \frac{P(x^n \mid \hat{\theta}(x^n))}{\sum_{y^n \in \mathcal{X}^n} P(y^n \mid \hat{\theta}(y^n))}$$

MDL Model Selection

Select M_i minimizing $-\log P_{NML}(x^n \mid M_i)$, i.e.

$$-\log P(x^n \mid \hat{\theta}_i(x^n)) + \log \sum_{y^n \in \mathcal{X}^n} P(y^n \mid \hat{\theta}_i(y^n))$$

error(=minus fit) term — complexity term $(\leq \log M)$

Geometric Interpretation of MDL

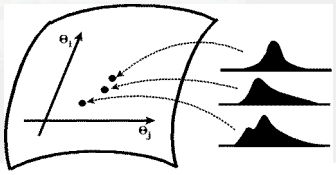
Under regularity conditions $-\log P_{NML}(x^n \mid M) =$

$$-\log P(x^n \mid \hat{\theta}_i(x^n)) + \frac{k}{2}\log\frac{n}{2\pi} + \log \int \sqrt{\det I(\theta)}\, d\theta + o(1)$$

error term — #parameters — "volume" of model viewed as manifold in space of all distributions

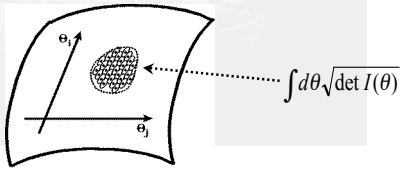
Space of probability distributions

The family of probability distributions forms a Riemannian manifold (information geometry; Rao, 1945; Efron, 1975; Amari, 1980).



Count only "distinguishable" distributions

The Riemannian volume measure is related to the number of all possible '*distinguishable*' probability distributions that are indexed by the model family (Balasubramanian, 1997):



Bayes vs. MDL

Under regularity conditions $-\log P_{NML}(x^n \mid M) =$

$$-\log P(x^n \mid \hat{\theta}_i(x^n)) + \frac{k}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det I(\theta)}d\theta + o(1)$$

Under regularity conditions $-\log P_{Bayes}(x^n \mid M) \approx$

$$-\log P(x^n \mid \hat{\theta}_i(x^n)) + \frac{k}{2}\log\frac{n}{2\pi} - \log w(\hat{\theta}) + \log\sqrt{\det I(\theta)}d\theta + o(1)$$

If we take Jeffrey's prior

$$w(\theta) = \sqrt{\det I(\theta)} \Big/ \int_\theta \sqrt{\det I(\theta)}d\theta \quad \text{..... ☺}$$

Predictive Interpretation

- interpret $-\log P(x)$ as loss incurred when predicting using P while actual outcome was x
- Bayesian marginal likelihood can be written as cumulative log-loss prediction error

$$-\log P_{Bayes}(x^n) = -\log\prod_1^n \frac{P_{Bayes}(x^i)}{P_{Bayes}(x^{i-1})} =$$

$$\sum_1^n -\log P_{Bayes}(x_i \mid x_1,\ldots,x_{i-1}) = \sum_{i=1}^n Loss(x_i, P_{Bayes}(\bullet \mid x^{i-1}))$$

Philosophy

- what do we do when the data generating mechanism is not in the family of models M we consider? (what is prior?)
- MDL priors are technical in nature
- Jeffreys' prior is uniform prior on the space of distributions with the "natural metric" that measures distances between distributions by how distinguishable they are