Universal Modeling: Introduction to 'Modern' MDL

Peter Grünwald
CWI Amsterdam
www.cwi.nl/~pdg

Further Reading:
P. Grünwald. *The Minimum Description Length Principle*, MIT Press, 2007.

Overview

- Introduction
- Probability and Code Length
- Universal Models
- MDL Model Selection
- Interpretation
- New Developments

Overview

- Introduction
- Probability and Code Length
- Universal Models
- MDL Model Selection
- Interpretation
- New Developments

Minimum Description Length Principle

Rissanen 1978, 1987, 1996,
Barron, Rissanen and Yu 1998

- ‘MDL’ is a method for inductive inference,
- in particular developed and suited for model selection problems
- but can do prediction/estimation as well

Minimum Description Length Principle

- MDL is based on the correspondence between ‘regularity’ and ‘compression’:
  - The more you are able to compress a sequence of data, the more regularity you have detected in the data
  - Example:

$$00100100100100100100100100 \ldots 001$$

$$01011011100100111010001010 1\ldots 010$$

Minimum Description Length Principle

- MDL is based on the correspondence between ‘regularity’ and ‘compression’:
  - The more you are able to compress a sequence of data, the more regularity you have detected in the data…
  - …and thus the more you have learned from the data:
    - ‘inductive inference’ as trying to find regularities in data (and using those to make predictions of future data)

Model Selection

Given data $x^n = x_1, \ldots, x_n$ and 'models' $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \ldots$,
which model *best explains* the data ?

- Need to take into account
  - Error (minus Goodness-of-fit)
  - Complexity of models
- Examples
  - Variable (order) selection in regression
  - Selection of order in (hidden) Markov Models

‘Modern’ MDL?

Kinds of MDL

- Algorithmic, ‘ideal’ MDL (Li and Vitányi ’97)
- MML (Wallace ‘68 (!), ‘87)
- 2-part code MDL (Rissanen ’78, ’83)
- Universal model based MDL (Rissanen ’96, Barron, Rissanen, Yu ‘98, Grünwald ‘07)

Modern MDL!

Kinds of MDL

- Algorithmic, 'ideal' MDL (Li and Vitányi '97)
- MML (Wallace '68 (!), '87)
- 2-part code MDL (Rissanen '78, '83)
- Universal model based MDL (Rissanen '96, Barron, Rissanen, Yu '98, Grünwald '07)

Overview

- Introduction
- Probability and Code Length
- Universal Models
- MDL Model Selection
- Interpretation
- Predictive MDL Estimation

Codes

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

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

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

Code Length & Probability

- Let P be a probability distribution. Since $\sum_x P(x) \leq 1$ only 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.
- This suggests an **analogy!**

Code Lengths 'are' probabilities…

- Let C be a (uniquely decodable) code over countable set $\mathcal{X}$. Then there exists a (possibly defective) probability distribution P_C such that

$$\text{for all } x : L_C(x) = -\log P_C(x)$$

- P_C is a 'proper' probability distribution iff the code C is 'complete'.

(follows from Kraft-McMillan inequality)

…and probabilities 'are' code lengths!

- Let P be a probability distribution over countable set $\mathcal{X}$. Then there exists a code C_P for $\mathcal{X}$ such that

$$\text{for all } x : L_{C_P}(x) = \lceil - \log P(x) \rceil$$

The Most Important Slide!

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

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

The Most Important Slide!

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

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

Example: P is 1st 0rder Markov Chain – if P fits data well (regularities in data are well-captured by P) , the code based on P compresses much.

Remarks

- In this correspondence, we **do not** assume that data are sampled from a probability distribution!
- Extend correspondence to continuous sample space through discretization; P may stand for *density*
- Distributions and codes over sequences of outcomes: still max. 1 bit round-off error
- Neglect difference and *identify* code length functions and probability mass functions

Overview

- Introduction
- Probability and Code Length
- Universal Models
- MDL Model Selection
- Interpretation
- Predictive MDL Estimation

Universal Codes

- $\mathcal{L}$: set of code (length function)s available to encode data $x^n = x_1, \ldots, x_n$
- Suppose we think that one of the code(length function)s in $\mathcal{L}$ allows for substantial compression of x^n
- GOAL: encode x^n using minimum number of bits!

Universal Codes

- Simply encoding x^n using the $\hat{L} \in \mathcal{L}$ that minimizes code length $\hat{L}(x^n) = \inf_{L \in \mathcal{L}} L(x^n)$ does not work (encoding cannot be decoded)
- But there exist codes $L_{\mathcal{L}}$ which, for any sequence x^n are 'almost' as good as $\inf_{L \in \mathcal{L}} L(x^n)$
- These are called 'universal codes' for $\mathcal{L}$

Universal Codes

- Example: $\mathcal{L}$ finite
- There exists a code $L_{\mathcal{L}}$ such that for some constant K, for all n, x^n, all $L \in \mathcal{L}$:

$$L_{\mathcal{L}}(x^n) \leq L(x^n) + K$$

- In particular,

$$L_{\mathcal{L}}(x^n) \leq \inf_{L \in \mathcal{L}} L(x^n) + K$$

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

Universal Models

- Let $\mathcal{M}$ be a probabilistic model, i.e. a family (set) of probability distributions
- Assume $\mathcal{M}$ finite: $\mathcal{M} = \{P(\cdot|\theta_1), \ldots, P(\cdot|\theta_M)\}$
- There exists a code $L_{\mathcal{M}}$ such that for all n, x^n, θ :
$$L_{\mathcal{M}}(x^n) \leq -\log P(x^n|\theta) + K$$
- Hence, exists distribution $P_{\mathcal{M}}$ such that
$$-\log P_{\mathcal{M}}(x^n) \leq -\log P(x^n|\theta) + K$$
- i.e. $P_{\mathcal{M}}(x^n) \geq K' \cdot P(x^n|\theta)$
- $P_{\mathcal{M}}$ is a 'universal model' (distribution) for $\mathcal{M}$

Terminology

- Statistics:
  - Model = family of distributions
- Information theory:
  - Model = single distribution
  - Model class = family of distributions
- Universal model is a single distribution acting as a representative of/defined relative to a set of distributions

Bayesian Mixtures are universal models

- Let W be a prior over $\mathcal{M}$. The Bayesian marginal likelihood P_{Bayes} is defined as:

$$P_{\text{Bayes}}(x^n|\mathcal{M}) = \sum_{j=1}^{M} P(x^n|\theta_j)W(\theta_j)$$

Bayesian Mixtures are universal models

- Let W be a prior over $\mathcal{M}$. The Bayesian marginal likelihood is defined as:

$$P_{\text{Bayes}}(x^n|\mathcal{M}) = \sum_{j=1}^{M} P(x^n|\theta_j)W(\theta_j)$$

- This is a universal model, since

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

2-part MDL code is a universal model (code)

- The ML (*maximum likelihood)* distribution $\hat{\theta}(x^n)$ is the θ achieving $\inf_{P(\cdot|\theta)\in\mathcal{M}}\{-\log P(x^n|\theta)\}$

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

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

2-part vs. Bayes universal models

- Bayes' mixture strictly 'better' universal model in that it assigns larger probability (shorter code length) to outcomes.
- What does 'better' really mean?
- What *prior* leads to short code lengths?

Optimal Universal Model

Look for P^* such that regret

$$\cdot - \log P^*(x^n) - [-\log P(x^n | \hat{\theta}(x^n))]$$

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

$$\inf_{P^*} \sup_{x^n \in \mathcal{X}^n} \{-\log P^*(x^n) - [-\log P(x^n | \hat{\theta}(x^n))]\}$$

Optimal Universal Model - II

$$\inf_{P^*} \sup_{x^n \in \mathcal{X}^n} \{-\log P^*(x^n) - [-\log P(x^n|\hat{\theta}(x^n))]\}$$

is achieved for Normalized Maximum Likelihood (NML) distribution (Shtarkov 1987):

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

MDL Model Selection

- Suppose we are given data $x^n = x_1, \ldots, x_n$
- We want to select between models $\mathcal{M}_1$ and $\mathcal{M}_2$ as explanations for the data. MDL tells us to pick the $\mathcal{M}_i$ for which the associated optimal universal model $P_{\mathrm{NML}}(\cdot|\mathcal{M}_i)$ assigns the largest probability to the data:

$$\mathcal{M}_{mdl} = \arg\sup_{\mathcal{M}_i} P_{\mathrm{NML}}(x^n|\mathcal{M}_i) = \arg\inf_{\mathcal{M}_i} -\log P_{\mathrm{NML}}(x^n|\mathcal{M}_i)$$

MDL Model Selection

Select $\mathcal{M}_i$ minimizing $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}_i)$, i.e. minimizing

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

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

Four Interpretations

- Compression interpretation
  - Select model that compresses data most, *treating all distributions within model on equal footing*; detects most (non-spurious) regularity in data
- Counting/Geometric interpretation
- Bayesian interpretation
- Predictive interpretation

Counting Interpretation of MDL

Select $\mathcal{M}_i$ minimizing $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}_i)$, i.e. minimizing

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

error (= minus fit) term

complexity term ($\leq \log M$)

Something like 'total fit' model gives to data

Log 'effective' number of distributions

Counting Interpretation of MDL

$$\sum_{y^n \in \mathcal{X}^n} P(y^n | \hat{\theta}(y^n)) = \sum_{\theta : P(\cdot|\theta) \in \mathcal{M}} \; \sum_{y_n \in \mathcal{X}^n : \hat{\theta}(y^n) = \theta} P(y^n | \theta) =$$

$$\sum_{\theta} P(\{y^n \in \mathcal{X}^n : \hat{\theta}(y^n) = \theta\} | \theta) = \sum_{\theta} [1 - P(\{x^n \in \mathcal{X}^n : \hat{\theta}(x^n) \neq \theta\} | \theta)] =$$

$$M - \sum_{\theta} P(\hat{\theta} \neq \theta | \theta)$$

number of distributions

total amount of confusion

Counting Interpretation of MDL

Select $\mathcal{M}_i$ minimizing $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}_i)$, i.e. minimizing

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

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

Something like 'total fit' model gives to data

Log number of 'distinguishable' distributions

Parametric Model Classes

- Under regularity conditions: $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}) =$

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

- Here:

k	Number of free parameters in $\mathcal{M}$
$I(\theta)$	Fisher information matrix at θ
$o(1)$	Goes to 0 as $n \to \infty$

Geometric Interpretation of MDL

- Under regularity conditions: $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}) =$

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

- Compare **BIC** (Schwartz '78), old **'MDL Criterion'** (Rissanen '78): select $\mathcal{M}$ minimizing:

$$-\log P(x^n|\hat{\theta}(x^n)) + \frac{k}{2}\log\frac{n}{2\pi}$$

Geometric Interpretation of MDL

- Under regularity conditions: $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}) =$

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

nr of parameters

Curvature term

Error Term

Complexity Terms

Bayesian Model Selection vs. MDL

- Under regularity conditions: $-\log P_{\mathrm{NML}}(x^n|\mathcal{M}) =$

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

- Under regularity conditions: $-\log P_{\mathrm{Bayes}}(x^n|\mathcal{M}) \approx$

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

- Always within O(1) ; hence, for large enough n, Bayes and MDL select the same model

Bayesian Model Selection vs. MDL

- Under regularity conditions: $- \log P_{\mathrm{NML}}(x^n | \mathcal{M}) =$

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

- Under regularity conditions: $- \log P_{\mathrm{Bayes}}(x^n | \mathcal{M}) \approx$

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

- If we take *Jeffreys-Bernardo prior*,

$$w(\theta) = \sqrt{\det I(\theta)} / \int_\theta \sqrt{\det I(\theta)} d\theta$$

  within o(1): Bayes and NML become *indistinguishable*

Bayes and MDL, remarks

- Jeffreys' prior was proposed as a 'non-informative Bayesian prior' by Jeffreys in 1939
- Jeffreys' prior is uniform prior *not* on parameter space but on the space of distributions with the 'natural metric' that measures distances between distributions by how distinguishable they are
- (but MDL is not Bayes!
  - e.g., MDL is immune to Diaconis-Freedman nonparametric inconsistency results)

Further topics

- Predictive interpretation (MDL as an automatic cross-validation like procedure)
- Comparing infinitely many models
- Predictive MDL Estimation
- Frequentist justification

Predictive Interpretation

- Interpret $-\log P(x)$ as 'loss' incurred when predicting using P while actual outcome was x

$$\text{Loss}(x, P) \equiv -\log P(x)$$

- Bayesian codelength can be rewritten as accumulated log-loss prediction error

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

$$\sum_{i=1}^{n} -\log P_{\text{Bayes}}(x_i | x_1, \ldots, x_{i-1}) = \sum_{i=1}^{n} \text{Loss}(x_i, P_{\text{Bayes}}(\cdot | x^{i-1}))$$

- Here $P_{\text{Bayes}}(\cdot | x_1, \ldots, x_{i-1})$ is the Bayesian predictive distribution (posterior mixture)

Predictive Interpretation, II

- Idea (Dawid/Rissanen): for large n, Bayesian predictive distribution resembles ML distribution more and more; therefore, may try to approximate $P_{\text{Bayes}}(\cdot|x_1, \ldots, x_{i-1})$ by

$$P(\cdot|\hat{\theta}(x_1, \ldots, x_{i-1}))$$

or more generally by

$$P(\cdot|\tilde{\theta}(x_1, \ldots, x_{i-1}))$$

for any 'likelihood-based estimator' $\tilde{\theta}$

Predictive Interpretation, III

- It turns out that (under regularity conditions)

$$-\sum_{i=1}^{n} \log P(x_i|\hat{\theta}(x^{i-1})) = -\log P(x^n|\hat{\theta}(x^n)) + \frac{k}{2}\log n + O(1)$$

- Hence, ‘predictive code’ is a universal code
- MDL model selection picks the model $\mathcal{M}$ such that sequential prediction of the future given the past within the observed data leads to lowest accumulated sequential prediction error.

Predictive Interpretation, IV

- MDL can be cast in terms of prequential validation (Dawid '84)
- similar to leave-one-out cross-validation
- essential difference: in MDL/prequential validation, if value of x_i is used in prediction of x_j, then value of x_j not used in prediction of x_i
- If number of models under consideration is finite and constant, but $n \longrightarrow \infty$, then
  - Prequential validation/MDL like BIC
  - Leave-One-Out CV like AIC

Comparing infinitely many models

Select $\mathcal{M}_i$ minimizing $-\log P_{\mathrm{nml}}(x^n|\mathcal{M}_i) + L(i)$
i.e. minimizing

$$-\log P(x^n|\hat{\theta}_i(x^n)) + \log \sum_{y^n \in \mathcal{X}^n} P(y^n|\hat{\theta}_i(y^n)) + \log i + 2 \log \log i$$

Reason: **whole** procedure should be interpretable as minimizing codelength for data.

- We implicitly used uniform code to encode $\mathcal{M}_i$ before.

Comparing infinitely many models

- Better not use two-part code for parameters
  - NML, Bayes give smaller regret (relative code-lengths)
- We are *forced* to use two-part code for encoding model index
  - Because we want to select a model, we explicity have to encode it
  - Note: complexity of models *not* due to model index!

Overview

- Introduction
- Probability and Code Length
- Universal Models
- MDL Model Selection
- Interpretation
- Overview of New Developments
- Predictive MDL Estimation/Justification

New Developments

- Efficient Calculation of NML for some model classes (Myllymaki, 11.00 today)
- **What if NML distribution is undefined?**
  - luckiness principle
    (Luckiness NML, Conditional NML)
  - Sequential NML (Silander, 14.30 today)
- Nonparametrics (Seeger, 16.30 today)
- Inherent improvement by adopting different notion of universality; solving AIC-BIC dilemma
  (De Rooij, 11.30, Grunwald, 9.30 tomorrow)

Luckiness Principle

Luckiness Principle

- Define "luckines" or "slack" function $a(\theta)$ and define

$$\bar{P}_{\mathsf{LNML}} =$$

$$\arg\min_{\bar{P}} \max_{x^n \in \mathcal{X}^n} \left\{ - \log \bar{P}(x^n) - [\min_{\theta \in \Theta} - \log P(x^n \mid \theta) + a(\theta)] \right\}$$

- $a(\theta)$ uniform: Luckiness NML = NML
- $a(\theta)$ nonuniform, $\mathcal{M}$ parametric:

$$\bar{P}_{\mathsf{LNML}}(x^n) = \frac{\max_\theta P_\theta(x^n) 2^{-a(\theta)}}{\sum_{x^n} \max_\theta P_\theta(x^n) 2^{-a(\theta)}}$$

corresponds to Bayes with tilted Jeffreys'prior

Luckiness Principle

- Define "luckines" or "slack" function $a(\theta)$ and define

$$\bar{P}_{\text{LNML}} = \arg\min_{\bar{P}} \max_{x^n \in \mathcal{X}^n} \left\{ -\log \bar{P}(x^n) - \left[\min_{\theta \in \Theta} -\log P(x^n \mid \theta) + a(\theta)\right] \right\}$$

- $a(\theta)$ uniform: Luckiness NML = NML
- $a(\theta)$ nonuniform, $\mathcal{M}$ parametric:
- But can do this also for large, nonparametric $\mathcal{M}$

Overview

- Introduction
- Probability and Code Length
- Universal Models
- MDL Model Selection
- Interpretation
- Overview of New Developments
- Predictive MDL Estimation/Justification

Universal Models as Estimators

- Let $\overline{P}$ be a distribution on $\mathcal{X}^\infty$
- Suppose $\overline{P}$ is a universal model relative to some model class $\mathcal{M}$, i.e. for all $P^* \in \mathcal{M}$,

$$\sup_{x^n} \{ -\log \overline{P}(x^n) + \log P^*(x^n) \} = o(n)$$

  – now think of $\mathcal{M}$ as a countably infinite union of parametric models, or as a "nonparametric" class
- Suppose $X_1, X_2, \ldots \sim P^*$
- We can think of $\overline{P}(X_{n+1} \mid x^n)$ as an estimator of P^*

Predictive MDL Estimation

- We can think of $\overline{P}_n = \overline{P}(X_{n+1} = \cdot \mid X^n)$ as estimator of P^*
- Example: if $\overline{P}$ is a Bayesian universal model, then this is the posterior predictive distribution:
  - a mixture of $P \in \mathcal{M}$, e.g. the Laplace estimator
  - should 'converge' to 'true' P^*
- Theorem (Barron, 1998)
  - If $\overline{P}$ is universal relative to $\mathcal{M}$, then the estimator $\overline{P}(X_{n+1} \mid x^n)$ must be consistent
  - what counts is universality, not Bayesianity...

Barron's Theorem

- Let $\overline{P}_0, \overline{P}_1, \overline{P}_2, \ldots$ denote any estimator, $\overline{P}_n : \mathcal{X}^n \rightarrow \mathcal{P}$
- The KL-risk of this estimator is

$$\text{risk}_n := E_{X_1,\ldots,X_n \sim P^*}[D(P^* \| \overline{P}_n)]$$

- The Cesaro KL-risk of this estimator is

$$\text{c-risk}_n := \frac{1}{n} \sum_{i=0}^{n-1} \text{risk}_i$$

- Barron's Theorem: suppose

$$\sup_{x^n} \{ -\log \overline{P}(x^n) + \log P^*(x^n) \} \leq f(n)$$

Then

$$\text{c-risk}_n \leq \frac{1}{n} f(n)$$

Barron's Theorem

- If $\sup_{x^n} \{ -\log \overline{P}(x^n) + \log P^*(x^n) \} \leq f(n)$

  then $\text{c-risk}_n \leq \frac{1}{n} f(n)$

- Example: $\mathcal{M}$ parametric.
  - Bayesian universal model achieves $f(n) = \frac{k}{2} \log n + O(1)$
  - Then for all $P^* \in \mathcal{M}$: $\text{c-risk}_n = O\left(\frac{\log n}{n}\right)$
  - Suggests Bayes achieves optimal rate of $O\left(\frac{1}{n}\right)$

Frequentist Justification of MDL

- MDL based on designing universal model/code $\bar{P}$ relative to model class $\mathcal{M}$
- If $\bar{P}$ universal, then consistency automatic
- Let $P^* \in \mathcal{M}$
- The better $\bar{P}$ compresses data from P^*, the faster the estimator $\overline{P}_n$ converges to P^*

Frequentist Justification of MDL

- In other words:

Good Compression implies Fast Learning!

Thank you for your attention!

Overview – part II

- Justification
- What if NML distribution undefined?
- MDL and Bayes; philosophy of MDL

Does it ‘work’ in frequentist sense?

- rule of thumb: MDL procedures are ‘consistent’ whenever Bayes’ procedures are consistent
  - rates of convergence comparable to Bayes
    - in our simple case, ‘consistency’ means that if countably infinite number of models is compared, the ‘true’ model is eventually selected.
    - Surprising exception: (Csiszár, Shields 2000)
- Barron & Cover (1991) show consistency of MDL density estimation in parametric, nested parametric and non-parametric cases
    - Rate of convergence within log of minimax optimal
    - Recently improved by Zhang (2004); shows rate of convergence is minimax optimal

Does it ‘work’ in frequentist sense?

- NOTE: the nested parametric and non-parametric cases include many cases in which maximum likelihood would be dreadfully inconsistent, severely overfitting irrespective of the amount of available data
- Example:
  - order selection/parameter estimation among all Markov chains of each order
  - Finding the best Gaussian mixture among the set of all Gaussian mixtures with arbitrary number of components
  - Regression

Other Justifications

- Rissanen does not believe that true distributions or models exist. He thinks the goal of inductive inference should be to pick the model that 'captures the most regularity in the data'
  - i.e. best summarizes the data, give the meaningful information in the data
  - He tries to justify MDL in terms of the **Kolmogorov Minimal Sufficient Statistic** (based on lossy rather than lossless compression)

MDL and Bayes

- Heated debates galore!
- First insight:
  - Two tenets of Bayesian statistics:
    1. All uncertainty should be handled using probability
    2. All decisions should be done based on (expectations according to) prior/posterior
  - MDL sticks with 1, not 2 (NML code!)

How to use MDL in practical Model Selection Problems

In order of preference:

1. Try o(1)-universal models: NML distributions or non-informative Bayesian mixtures *or*
2. Use predictive MDL
   - with sequential Bayes-MAP estimates
3. Use asymptotic expansion (k/2 log n +…) (be careful!) *or*
4. Use two-part code MDL *or*
5. *Use another O(1)-universal model*

Overview – part II

- Prequential interpretation of MDL
- What if NML distribution undefined?
- MDL and Bayes; philosophy of MDL

Overview – part II

- Prequential interpretation of MDL
- What if NML distribution undefined?
- MDL and Bayes; philosophy of MDL

What if NML distribution undefined?

- In many interesting applications, NML distribution undefined
  - Examples: linear regression, normal distribution: P_{nml} should have density

$$\frac{f(x^n|\hat{\mu},\hat{\sigma}^2)}{\int_{x^n} f(x^n|\hat{\mu}(x^n),\hat{\sigma}^2(x^n))dx^n}$$

  - Undefined since complexity

$$\int_{x^n} f(x^n|\hat{\mu}(x^n),\hat{\sigma}^2(x^n))dx^n$$

diverges!

What if NML distribution undefined?

- In many interesting applications, NML distribution undefined
- In such cases also $\int \sqrt{I(\theta)} d\theta$ diverges
- Hence Jeffreys' prior improper
- However, integral typically remains small even if parameters get quite close to boundary of parameter space

Undefined NML, II

- Simplistic solution:
  - start with

$$f_{\mathrm{nml}}(x^n | \mathcal{M}_{[\tau,K]}) := \frac{f(x^n | \hat{\mu}, \hat{\sigma}^2)}{\int_{x^n} f(x^n | \hat{\mu}(x^n), \hat{\sigma}^2(x^n)) dx^n}$$

  where (ML) parameters are restricted to

$$-K \leq \mu \leq K \qquad\qquad \sigma^2 > \tau$$

  - this is finite for each pair of ‘hyperparameters’ K and τ

Undefined NML, II

- Explicitly encode hyperparameters by encoding integers a, b with

$$\tau = 2^{-a} \qquad K = 2^{b}$$

- We need

$$-\log f^{*}(x^{n} \mid \mathcal{M}) := \inf_{\tau, K} \; -\log f_{\mathrm{nml}}(x^{n} \mid \mathcal{M}_{[\tau, K]}) + L(\tau) + L(K)$$

  - Unless for outrageous data sets, need much less bits for hyperparameters than for ordinary parameters

Undefined NML, II

- Explicitly encode hyperparameters by encoding integers a, b with

$$\tau = 2^{-a} \qquad K = 2^{b}$$

- We get as our new code length:

$$-\log f^*(x^n|\mathcal{M}) := \inf_{\tau, K} \; -\log f_{\mathrm{nml}}(x^n|\mathcal{M}_{[\tau,K]}) + \underbrace{L(\tau)}_{2\log a} + \underbrace{L(K)}_{2\log b}$$

More sophisticated ideas

- Rissanen's Renormalization (2001)
- Barron and Liang's conditional minimax universal codes
  - Elegant solution for variable selection in regression
- Many others (hot topic!)

General Picture

- $\mathcal{M}$ such that there is no universal model $P_{\mathcal{M}}$ achieving uniform/minimax regret
- Then carve up $\mathcal{M}$ into subsets

  $$\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots$$

  and define P^* such that for each x^n,

  $$-\log P^*(x^n|\mathcal{M}) - [-\log P(x^n|\hat{\theta}(x^n))]$$

  is almost as small as the uniform/minimax regret of the *smallest* $\mathcal{M}_k$ containing $P(\cdot|\hat{\theta}(x^n))$

- P^* achieves ‘nearly’, ‘almost’ uniform regret

  $$-\log P_{\mathrm{nml}}(x^n|\mathcal{M}_j) + \mathrm{small}$$

General Principle

- We were doing exactly the same thing when trying to find the best order Markov chain among the class of all Markov chains
- 'Luckiness' idea:
  - Let $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$ be the union of 1st- and 2nd order MC models, and compare the NML distribution $P_{\text{nml}}(\cdot|\mathcal{M}_2)$ with the distribution

$$-\log P^*(x^n) = \inf_{k \in \{1,2\}} -\log P_{\text{nml}}(x^n|\mathcal{M}_k) + 1$$

which we implicitly used in model selection

‘Luckiness Idea’

- If you’re lucky, you need much less bits using the code P^* than the code $P_{\mathrm{nml}}(\cdot|\mathcal{M}_2)$
- If you’re not lucky, you need hardly any more bits (max. 1) using the code P^* than the code $P_{\mathrm{nml}}(\cdot|\mathcal{M}_2)$
  - Related to Luckiness principle in Computational Learning Theory (Herbrich and Williamson, 2001)

The MDL Principle

- First principle: try to be as ‘honest’ as possible, associating models (sets of distributions) with uniform/minimax regret universal models
- Second principle: if regret becomes too large, carve up your model into submodels and use a ‘quasi-uniform’ universal model
  - Never much worse than uniform regret model
  - If you’re lucky, considerably better than uniform regret model

Overview – part II

- Prequential interpretation of MDL
- What if NML distribution undefined?
- MDL and Bayes; philosophy of MDL

MDL and Bayes

- Heated debates galore!
- First insight:
  - Two tenets of Bayesian statistics:
    1. All uncertainty should be handled using probability
    2. All decisions should be done based on (expectations according to) prior/posterior
  - MDL sticks with 1, not 2 (NML code!)

Brands of Bayesian Statistics

‘modern’ Bayesian Statistics has (at least) three founding fathers, each with (quite) different ideas

L. Savage
The Foundations of Statistics (1954)

B. De Finetti
Theory of Probability (‘1937’,1974)

H. Jeffreys
Theory of Probability (1939, 1961)

MDL and Bayes, Philosophy

- MDL = Maximum Probability Principle, *not* Savage's \`Maximizing Expected Utility according to prior' principle
- In MDL priors used as a tool that do not have anything to do with 'degrees of belief'
- Indeed 'degree of belief' in a *hypothesis* is meaningless according to MDL
  - Naïve Bayes, speech recognition
  - some universal models do not have anything like 'prior' or 'posterior'

MDL and Bayes, II

- In MDL we certainly don't believe that a first-order Markov chain is much more likely to have generated the data, although we give individual 1st order Markov chains an infinitely higher prior density than individual 2nd order Markov chains

- Instead, we would like to select a 1st order chain as long as the sample is so small that the inferred chain is likely to lead to better predictions of future data

MDL and Bayes, Philosophy

- Nevertheless, MDL (that is, Rissanen) considers probabilities of data as *subjective* - probabilities are something to be used for prediction or description, the ‘true’ distribution does not exist other than as a mental construct

Rissanen: ‘We only have the data’

- … so, in the end, this *is* very similar to De Finetti’s ideas:

De Finetti: ‘Probabilities Do Not Exist’

MDL and Bayes in practice

- In practice ‘objective’ Bayesians do model selection in almost the same way as MDL (when applied with Bayesian universal model)
- Yet some differences remain:
  - MDL does not restrict type of universal model used (more freedom)
  - MDL **never** allows taking expectations over the prior (less freedom)
    - If prior that is good in minimizing worst-case code length assigns large probability to a set A, this certainly does not imply that A will indeed be realized
      - Both degree-of-belief and frequentist justification of taking expectation fail; expectation according to prior is meaningless!
    - difference to MML

Thank you for your attention!