Predictive MDL & Bayes

Marcus Hutter

Canberra, ACT, 0200, Australia

http://www.hutter1.net/

ANU

RSISE

NICTA

Contents

Abstract

The minimum description length (MDL) principle recommends to use, among competing models, the one that allows to compress the data+model most. The better the compression, the more regularity has been detected, hence the better will predictions be. The MDL principle can be regarded as a formalization of Ockham's razor. In Bayesian learning, rather than selecting a single model, one takes a weighted mixture over all models. This takes justice of Epicurus' principle of multiple explanations. I show that for a countable class of models, MDL and Bayes predictions are close to the true distribution in total variation distance. The result is completely general. No independence, ergodicity, stationarity, identifiability, or other assumption on the model class need to be made. Implications for non-i.i.d. domains like time-series forecasting, discriminative learning, and reinforcement learning are discussed.

Discrete MDL Predicts in Total Variation

Marcus Hutter, ANU&NICTA, Canberra, Australia, www.hutter1.net

Main result informal: For *any* countable class of models $\mathcal{M}=\{Q_1,Q_2,...\}$ containing the *unknown* true sampling distribution P, MDL predictions converge to the true distribution in total variation distance. Formally ...

- Given $x\equiv x_1...x_\ell$, the Q-prob. of $z\equiv x_{\ell+1}x_{\ell+2}...$ is $Q(z|x)=\dfrac{Q(xz)}{Q(x)}$
- Use $Q=$ Bayes or $Q=$ MDL instead of P for prediction
- Total variation distance: $d_\infty(P,Q|x):=\sup_{A\subseteq\mathcal{X}^\infty}\big|Q[A|x]-P[A|x]\big|$
- $\text{Bayes}(x):=\sum_{Q\in\mathcal{M}}Q(x)w_Q$, $\quad[w_Q>0\,\forall Q\in\mathcal{M}$ and $\sum_{Q\in\mathcal{M}}w_Q=1]$
- MDL selects Q which leads to minimal code length for x:
  $\text{MDL}^x:=\arg\min_{Q\in\mathcal{M}}\{-\log Q(x)+K(Q)\}$, $\quad[\sum_{Q\in\mathcal{M}}2^{-K(Q)}\leq 1]$

Theorem 1 (Discrete Bayes&MDL Predict in Total Variation)

$$\left.\begin{array}{l} d_\infty(P,\text{Bayes}|x)\to 0 \\ d_\infty(P,\text{MDL}^x|x)\to 0 \end{array}\right\}\ \begin{array}{l}\text{almost surely}\\ \text{for}\ \ \ell(x)\to\infty\end{array}$$

[Blackwell&Dubins 1962]
[Hutter NIPS 2009]

No *independence*, *ergodicity*, *stationarity*, *identifiability*, or other assumption

PART I: SETUP AND MAIN RESULTS

- Multistep Lookahead Sequential Prediction
- Model Class and Distance
- Bayes & MDL – Informal "Derivation"
- Key Convergence Results
- Motivation
- Countable Class Assumption

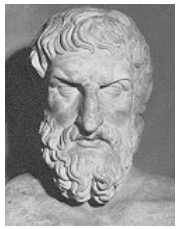
Bayes & MDL - Informally

Ockhams' razor (simplicity) principle (1285?–1349?)
Entities should not be multiplied beyond necessity.

Quantification:

Minimum description length (MDL) principle (1978)
Select the model that allows to compress the data+model most.

Justification: The simpler a model for the data, the more regularity has been detected, regularity tends to be stable over time, hence the better will the model predict.



Epicurus' principle of multiple explanations (342?–270? B.C.)
Keep all models consistent with the observations.

Quantification:

Bayes' rule for conditional probabilities (1763)
Take a (weighted) mixture over all models.

Multistep Lookahead Sequential Prediction

Setup: For $\ell \equiv \ell(x) = 0, 1, 2, ...,$ having

- observed sequence $x \equiv (x_1, x_2, ..., x_\ell) \equiv x_{1:\ell}$,
- predict $z \equiv (x_{\ell+1}, ..., x_{\ell+h}) \equiv x_{\ell+1:\ell+h}$, then
- observe $x_{\ell+1} \in \mathcal{X}$.

How far do we want to predict into the future?

- Classical prediction: $h = 1$,
- Multi-step prediction: $1 < h < \infty$,
- Total prediction: $h = \infty$.

Example 1: Having observed ℓ black ravens, what is the likelihood that the next / the next 100 / all ravens are black.

Example 2: What is the weather tomorrow / in 3 days.

Model Class and Distance

- Let $\mathcal{M} = \{Q_1, Q_2, ...\}$ be a countable class of models=theories=hypotheses=probabilities over sequences $\mathcal{X}^\infty$.
- Unknown true sampling distribution $P \in \mathcal{M}$.
- $Q(x)$ is Q-probability (or density) that a sequence starts with x.
- Given x, the true predictive prob. of z is $P(z|x) = P(xz)/P(x)$.
- Problem: P unknown.
- Approach: Use $Q(z|x) = Q(xz)/Q(x)$ for some Q for prediction.
- Question: How close is Q to P. Use distance measure:
- h-step predictive distance: $d_h(P,Q|x) := \sum_{z\in\mathcal{X}^h} \big|P(z|x) - Q(z|x)\big|$
- Total variation distance: $\frac{1}{2} d_\infty = \sup_{A\subseteq\mathcal{X}^\infty} \big|Q[A|x] - P[A|x]\big|$
- Property: $0 \leq d_1 \leq d_h \leq d_{h+1} \leq d_\infty \leq 2$

Bayes – Derivation

- Let w_Q be a prior weight (=believe=probability) for Q:
  $w_Q > 0 \forall Q \in \mathcal{M}$ and $\sum_{Q\in\mathcal{M}} w_Q = 1$.
- Bayesian mixture: $\boxed{\text{Bayes}(x) := \sum_{Q\in\mathcal{M}} Q(x) w_Q}$
- Bayesians use $\text{Bayes}(z|x) = \text{Bayes}(xz)/\text{Bayes}(x)$ for prediction.
- A natural choice is $w_Q \propto 2^{-K(Q)}$,
  where $K(Q)$ measures the complexity of Q.
- Simple choice: $w_{Q_i} = 1/i/(i+1)$.

MDL – Derivation: Two-Part Code

- Let $K(Q)$ be a prefix complexity=codelength of Q.
- Kraft inequality: $\sum_{Q\in\mathcal{M}} 2^{-K(Q)} \leq 1$.
- Examples: $K(Q_i) = \log_2[i(i+1)]$ or $K(Q_i) = \log_2 w_Q^{-1}$.
- Huffman coding: It is possible to code x in $\log P(x)^{-1}$ bits.
- Since x is sampled from P, this code is optimal (shortest among all prefix codes).
- Since we do not know P, we could select the $Q \in \mathcal{M}$ that leads to the shortest code on the observed data x.
- In order to be able to reconstruct x from the code, we need to know which Q has been chosen, so we also need to code Q, which takes $K(Q)$ bits.

$\Longrightarrow$ x can be coded in $\{-\log Q(x) + K(Q)\}$ bits (any Q).

MDL Criterion

- MDL selects Q which leads to minimal code length for x:
  $$\text{MDL}^x := \arg\min_{Q\in\mathcal{M}}\{-\log Q(x) + K(Q)\}$$
- Use $\text{MDL}^x(z|x) := \text{MDL}^x(xz)/\text{MDL}^x(x)$ for prediction.
- If $K(Q) = \log_2 w_Q^{-1}$ is chosen as complexity,
  then the maximum a posteriori estimate
  $\text{MAP}^x := \arg\max_{Q\in\mathcal{M}}\{\text{Pr}(Q|x)\} \equiv \text{MDL}^x$
  (proof by using Bayes rule: $\text{Pr}(Q|x) = Q(x)w_Q/\text{Bayes}(x)$)

$\Longrightarrow$ MDL results also apply to MAP.

Key Convergence Results

Theorem 2 (Bayes&MDL Predict in Total Variation ($h = \infty$))

$$\left.\begin{array}{l} d_\infty(P, \text{Bayes}|x) \to 0 \\ d_\infty(P, \text{MDL}^x|x) \to 0 \end{array}\right\} \begin{array}{l} \text{almost surely} \\ \text{for } \ \ell(x) \to \infty \end{array}$$

[Blackwell&Dubins 1962]
[Hutter NIPS 2009]

For $h < \infty$, an explicit bound for the # prediction errors is possible:

Theorem 3 (Bayes&MDL #Errors for $h < \infty$)

$\sum_{\ell=0}^{\infty} \mathbf{E}[d_h(P, \text{Bayes}|x_{1:\ell})] \leq h \cdot \ln w_P^{-1}$ [Solomonoff 1978, Hutter 2003]

$\sum_{\ell=0}^{\infty} \mathbf{E}[d_h(P, \text{MDL}^x|x_{1:\ell})] \leq 42\, h \cdot 2^{K(P)-1}$ [Poland&Hutter 2005]

where the expectation $\mathbf{E}$ is w.r.t. $P[\cdot|x]$.

This implies rapid convergence of $d_h \to 0$ almost surely (for $h < \infty$).

Motivation

The results hold for completely arbitrary countable model classes $\mathcal{M}$.

No *independence*, *ergodicity*, *stationarity*, *identifiability*, or other assumption need to be made.

Examples: Time-series prediction problems like weather forecasting and stock market prediction, the Universe, life.

Too much green house gases, a massive volcanic eruption, an asteroid impact, or another world war change the climate/economy irreversibly.

Life is not ergodic: one inattentive second in a car can have irreversible consequences.

Extensive games and multi-agent learning have to deal with non-stationary environments.

Identifiability: Data (asymptotically almost surely) uniquely reveals true distribution.

Countable Class Assumption

Countable class: Strong assumption but:

- Reduce semi-parametric to countable model by Bayes or NML, *or*
- Consider only countable dense class of (computable) parameters, *or*
- Reject non-computable parameters on philosophical grounds.

PART II: FACTS, INSIGHTS, PROBLEMS, PROOFS

- Deterministic MDL & Bayes
- Comparison: deterministic↔probabilistic and MDL↔Bayes
- Consistency of MDL for I.I.D and Stationary-Ergodic Sources
- Trouble Makers for General $\mathcal{M}$
- Predictive Bayes&MDL Avoid the Trouble
- Proof for Bayes&MDL for $h = \infty$

Deterministic MDL ($h = 1$)

= Gold style learning by elimination

- Each $Q \in \mathcal{M}$ is a model for one sequence $x^Q_{1:\infty}$, i.e. $Q(x^Q) = 1$.

$\Longrightarrow$ MDL selects the simplest Q consistent with true $x \equiv x^P_{1:\ell}$.

- For $h = 1$, a Q becomes (forever) inconsistent *iff* its prediction $x^Q_{\ell+1}$ is wrong ($\neq x^P_{\ell+1}$).

- Since elimination occurs in order of increasing complexity=codelength $K(Q_i) \approx \log_2 i$ (say), and $P = Q_m$ (say) never makes any error,

$\Longrightarrow$ MDL makes at most $m - 1$ prediction errors.

Deterministic MDL ($h > 1$)

= Gold style learning by elimination

For $1 < h < \infty$, prediction $x^Q_{\ell+1:\ell+h}$ may be wrong only on $x^Q_{\ell+h}$, which causes h wrong predictions before the error is revealed, since at time ℓ only x^P_ℓ is revealed.

$\Longrightarrow$ total #Errors $\leq h \cdot (m-1) \approx h \cdot 2^{K(Q_m)}$

For $h = \infty$, a wrong prediction gets eventually revealed

$\Longrightarrow$ each wrong Q_i $(i < m)$ gets eventually eliminated

$\Longrightarrow$ P gets eventually selected

$\Longrightarrow$ For $h = \infty$ the number of errors is finite, but

- no bound on the number of errors in terms of m only is possible.

Deterministic Bayes

= majority learning

- Models consistent with true observation $x_{1:\ell}^P$ have total weight W.
- Take weighted majority prediction (Bayes-optimal under 0-1 loss)
- For $h=1$, making a wrong prediction means that Q's contributing to at least half of the total weight W get eliminated.
- Since $P \equiv Q_m$ never gets eliminated, we have $w_P \leq W \leq 2^{-\#\text{Errors}}$

$\Longrightarrow$ $\#\text{Errors} \leq \log_2 w_P^{-1}$

- For proper probabilistic Bayesian prediction: $\#\text{Errors} \leq \ln w_P^{-1}$.
- $h>1$: Multiply bound by h.
- $h=\infty$: correct prediction eventually, but no explicit bound anymore.

Comparison

deterministic↔probabilistic and MDL↔Bayes

- Probabilistic and deterministic bounds are essentially the same.
- MDL bound is exponentially larger than Bayes bounds.
- In the deterministic case, the true P can be identified, but for probabilistic $\mathcal{M}$ in general not.
- In the probabilistic case, the proofs for the bounds are much harder, and much harder for MDL than for Bayes.

Consistency of MDL for I.I.D...

... and for stationary-ergodic sources

- For an i.i.d. class $\mathcal{M}$, the r.v. $Z_t := \log[P(x_t)/Q(x_t)]$ are i.i.d.
- KL-divergence: $\mathsf{KL}(P||Q) := \sum_{x_1} P(x_1) \log[P(x_1)/Q(x_1)]$
- Law of large numbers: $\frac{1}{\ell}\sum_{t=1}^{\ell} Z_t \to \mathsf{KL}(P||Q)$ with P-prob.1.
- Either KL=0, which is the case if and only if $P = Q$, or
- $\log P(x_{1:\ell}) - \log Q(x_{1:\ell}) \equiv \sum_{t=1}^{\ell} Z_\ell \sim \mathsf{KL}(P||Q)\ell \to \infty$, i.e. asymptotically MDL does not select Q.
- For countable $\mathcal{M}$, a refinement of this argument shows that MDL eventually selects P [Barron & Cover 1991].
- This reasoning can be extended to stationary-ergodic $\mathcal{M}$, but essentially not beyond.

Trouble Makers for General $\mathcal{M}$

asymptotically indistinguishable distributions

Asymptotically indistinguishable Bernoulli example:

P=Bernoulli(θ_0), but independent Q-probability that $x_t = 1$ is θ_t.

For a suitably converging but "oscillating" sequence $\theta_t \to \theta_0$ one can show that $\log[P(x_{1:t})/Q(x_{1:t})]$ converges to but oscillates around $K(Q) - K(P)$ w.p.1,

i.e. there are non-stationary distributions for which MDL does not converge (not even to a wrong distribution).

Trouble Makers for General $\mathcal{M}$

partitioning does not work in general

Idea: Partition $\mathcal{M}$ into asymptotically indistinguishable parts (like P and Q above) and ask MDL to identify a partition.

Problem: Asymptotic distinguishable can depends on the drawn sequence.

Example: Let P and Q be asymptotically (in)distinguishable *iff* $x_1 = (0)1$.

First observation can lead to totally different futures.

Predictive Bayes&MDL Avoid the Trouble

MDL and the Bayesian posterior do not need to converge to a single (true or other) distribution, in order for prediction to work.

At each time Bayes keeps a mixture and MDL selects a single distribution, but

Give up the idea that MDL/Bayes identify a single distribution asymptotically.

Just measure predictive success, and accept infinite oscillations.

Proof for Bayes and $h = \infty$

Let $(\Omega \equiv \mathcal{X}^\infty, \mathcal{F}, P)$ be the space of infinite sequences with natural filtration and product σ-field $\mathcal{F}$ and probability measure P.

$A \subseteq \mathcal{F}$ measurable set of infinite sequences.

P is said to be absolutely continuous relative to Q, written

$$P \ll Q \quad :\Leftrightarrow \quad [Q[A] = 0 \text{ implies } P[A] = 0 \text{ for all } A \in \mathcal{F}]$$

The famous Blackwell&Dubins convergence result:

$$\text{If} \quad P \ll Q \quad \text{then} \quad d_\infty(P, Q|x) \to 0 \quad \text{w.p.1} \quad \text{for} \quad \ell(x) \to \infty$$

If $P \in \mathcal{M}$, then obviously $P \ll \text{Bayes}$, hence $d_\infty(P, \text{Bayes}|x) \to 0$ w.p.1

MDL Proof for Finite $\mathcal{M}$ and $h = \infty$

Key Lemma 4 (generalizes Blackwell&Dubins 1962)
$Q(x_{1:\ell})/P(x_{1:\ell}) \to 0$ or $d_\infty(P,Q|x) \to 0$ for $\ell(x) \to \infty$ w.p.1

MDL will asymptotically not select Q for which $Q(x)/P(x) \to 0$.

$\Longrightarrow$ For those Q potentially selected by MDL,
we have $d_\infty(P,Q|x) \to 0$ w.p.1 (by Lemma 4)

MDL Proof for Countable $\mathcal{M}$ and $h = \infty$

Key Lemma 5 (MDL avoids complex probability measures Q)
$P[Q(x_{1:\ell})/P(x_{1:\ell}) \geq c$ for infinitly many $\ell] \leq 1/c$.

$\Longrightarrow$ Prob. that MDL asymptotically selects any "complex" Q is small.

PART III: IMPLICATIONS

- Time-Series Forecasting
- Classification and Regression
- Discriminative MDL&Bayes
- Reinforcement Learning
- Variations
- References

Time-Series Forecasting

Online learning: Direct application of convergence results for $h = 1$.

Offline learning: Train predictor on $x_{1:\ell}$ for fixed ℓ in-house, and then sell and use the predictor on $x_{\ell+1:\infty}$ without further learning.

Convergence results for $h = \infty$ show that for enough training data, predictions "post-learning" will be good.

Classification and Regression

Simply replace $x_t \rightsquigarrow (x_t, y_t)$.

Assumes distributions over x and y.

Discriminative MDL&Bayes

- Replace $x_t \rightsquigarrow (x_t, y_t)$. Only assumes distributions over x *given* y.
- $\text{Bayes}(x|y) := \sum_{Q\in\mathcal{M}} Q(x|y) w_Q$,
  $\text{MDL}^{x|y} := \arg\min_{Q\in\mathcal{M}}\{-\log Q(x|y) + K(Q)\}$

$\Longrightarrow$ $\sup_A \left| \overset{\text{MDL}^{x|y}}{\underset{\text{Bayes}}{}}[A|x,y] - P[A|x,y] \right| \to 0$ for $\ell(x) \to \infty$,
$P[\cdot|y_{1:\infty}]$ almost surely, for every sequence $y_{1:\infty}$.

Intuition for finite $\mathcal{Y}$ and conditionally independent x:

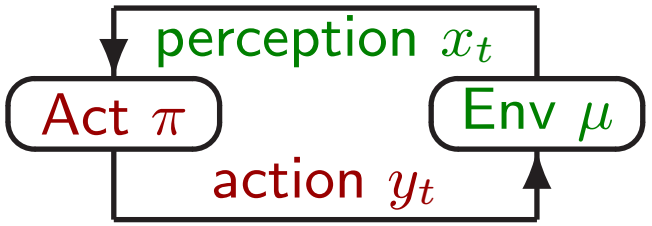
If $\dot{y}$ appears in $y_{1:\infty}$ only finitely often, it plays asymptotically no role;
if it appears infinitely often, then $P(\cdot|\dot{y})$ can be learned.

Intuition for infinite $\mathcal{Y}$ and deterministic $\mathcal{M}$:

Every $\dot{y}$ might appear only once, but probing enough
function values $x_t = f(y_t)$ allows to identify the function.

Stochastic Agent-Environment Setup

- In the agent framework, an agent interacts with an environment in cycles.



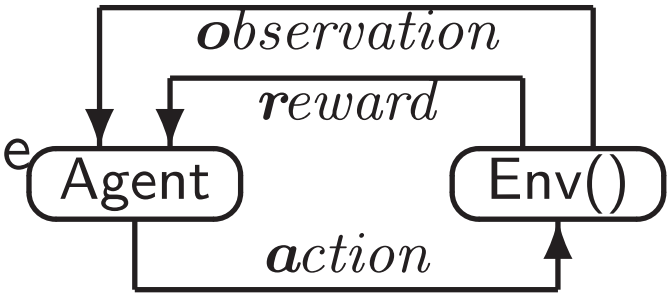
- At time t, agent chooses action y_t with probability $\boldsymbol{\pi}(y_t|x_{<t}y_{<t})$ based on past experience $x_{<t} \equiv (x_1, ..., x_{t-1})$ and actions $y_{<t}$. This leads to a new perception x_t with probability $\boldsymbol{\mu}(x_t|x_{<t}y_{1:t})$. Then cycle $t+1$ starts.

- Joint interaction probability: $P(xy) = \prod_{t=1}^{\ell} \mu(x_t|x_{<t}y_{1:t})\pi(y_t|x_{<t}y_{<t})$

- We make no (Markov, stationarity, ergodicity) assumption on μ,π. They may be partially observable MDPs or beyond.

Reinforcement Learning (RL)

- In reinforcement learning (RL), the perception $x_t := (o_t, r_t)$ consists of some regular observation o_t and a reward $r_t \in [0,1]$



- Goal is to find a policy which maximizes accrued reward in the long run.

- True Value of π := Future γ-discounted P-expected reward sum:
  $V_P[xy] := \mathbf{E}_{P[\cdot|xy]}[r_{\ell+1} + \gamma r_{\ell+2} + \gamma^2 r_{\ell+3} + ...]$
  Similarly define $V_Q[xy]$ for Q.

$\Longrightarrow$ The Bayes&MDLvalues converge to the true Value (for fixed π):

$V_{\text{Bayes\&MDL}^{x|y}}[xy] - V_P[xy] \to 0$ w.p.1. for any policy π [M.H.2005]

Variations

Incremental MDL:

$$\text{MDLI}(x_{1:\ell}) := \prod_{t=1}^{\ell} \text{MDL}^{x_{<t}}(x_t|x_{<t})$$
$$\text{MDLI}(z|x) = \text{MDLI}(xz)/\text{MDLI}(x).$$

Properties of MDLI:

+ should predict better than MDL^x (proof?)
+ defines a single measure over $\mathcal{X}^\infty$,
– is $\notin \mathcal{M}$.
– is harder to compute/use.

Thanks! THE END Questions?

- Want to work on this or other things ?
- Apply at ANU/NICTA/me for a PhD or PostDoc position !
- Canberra, ACT, 0200, Australia
  `http://www.hutter1.net/`


ANU


RSISE


NICTA

References

[BC91] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. on Information Theory*, 37:1034–1054, 1991.

[BD62] D. Blackwell and L. Dubins. Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33:882–887, 1962.

[Hut03] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Trans. on Information Theory*, 49(8):2061–2067, 2003.

[Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[Hut09] M. Hutter. Discrete MDL predicts in total variation. In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, 2009.

[PH05] J. Poland and M. Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Trans. on Information Theory*, 51(11):3780–3795, 2005.

[Sol78] R. J. Solomonoff. Complexity-based induction systems. *IEEE Trans. on Information Theory*, IT-24:422–432, 1978.

See http://www.hutter1.net/official/publ.htm for publications