COMP4620/8620: Advanced Topics in AI
Foundations of Artificial Intelligence

Marcus Hutter
Australian National University
Canberra, ACT, 0200, Australia
http://www.hutter1.net/

ANU

5 MINIMUM DESCRIPTION LENGTH

- MDL as Approximation of Solomonoff's M
- The Minimum Description Length Principle
- Application: Sequence Prediction
- Application: Regression / Polynomial Fitting
- Summary

Minimum Description Length: Abstract

The Minimum Description/Message Length principle is one of the most important concepts in Machine Learning, and serves as a scientific guide, in general. The motivation is as follows: To make predictions involves finding regularities in past data, regularities in data allows for compression, hence short descriptions of data should help in making predictions. In this lecture series we approach MDL from a Bayesian perspective and relate it to a MAP (maximum a posteriori) model choice. The Bayesian prior is chosen in accordance with Occam and Epicurus and the posterior is approximated by the MAP solution. We reconsider (un)fair coin flips and compare the M(D)L to Bayes-Laplace's solution, and similarly for general sequence prediction tasks. Finally I present an application to regression / polynomial fitting.

From Compression to Prediction

The better you can compress, the better you can predict.

Being able to predict (the env.) well is key for being able to act well.

Simple Example: Consider "14159...[990 more digits]...01989".

- If it looks random to you, you can neither compress it nor can you predict the 1001st digit.
- If you realize that they are the first 1000 digits of π, you can compress the sequence and predict the next digit.

Practical Example: The quality of natural language models is typically judged by its perplexity, which is essentially a compression ratio.

Later: Sequential decision theory tells you how to exploit such models for optimal rational actions.

MDL as Approximation of Solomonoff's M

- Approximation of Solomonoff, since M incomputable:
- $M(x) \approx 2^{-Km(x)}$ (excellent approximation)
- $Km(x) \equiv Km_U(x) \approx Km_T(x)$
  (approximation quality depends on T and x)
- Predict y of highest $M(y|x)$ is approximately same as
- MDL: Predict y of smallest complexity $Km_T(xy)$.
- Examples for x: Daily weather or stock market data.
- Example for T: Lempel-Ziv decompressor.
- Prediction $\widehat{=}$ finding regularities $\widehat{=}$ compression $\widehat{=}$ MDL.
- Improved compressors lead to improved predictors.

Human Knowledge Compression Contest

- compression = finding regularities ⇒ prediction ≈ intelligence
  [hard file size numbers] [slippery concept]
- Many researchers analyze data and find compact models.
- Compressors beating the current compressors need to be smart(er).
- "universal" corpus of data ⇒ "universally" smart compressors.
- Wikipedia seems a good snapshot of the Human World Knowledge.
- The ultimate compressor of Wikipedia will "understand" all human knowledge, i.e. be really smart.
- Contest: Compress Wikipedia better than the current record.
- Prize: 50'000 Euro × the relative improvement to previous record.

[http://prize.hutter1.net]

The Minimum Description Length Principle

Identification of probabilistic model "best" describing data:

Probabilistic model(=hypothesis) H_ν with $\nu \in \mathcal{M}$ and data D.

Most probable model is $\nu^{\text{MDL}} = \arg\max_{\nu\in\mathcal{M}} p(H_\nu|D)$.

Bayes' rule: $p(H_\nu|D) = p(D|H_\nu)\cdot p(H_\nu)/p(D)$.

Occam's razor: $p(H_\nu) = 2^{-Kw(\nu)}$.

By definition: $p(D|H_\nu) = \nu(x)$, $D = x$ =data-seq., $p(D)$ =const.

Take logarithm:

Definition 5.1 (MDL) $\nu^{\text{MDL}} = \arg\min_{\nu\in\mathcal{M}}\{K\nu(x) + Kw(\nu)\}$

$K\nu(x) := -\log\nu(x)$ = length of Shannon-Fano code of x given H_ν.

$Kw(\nu)$ = length of model H_ν.

Names: Two-part MDL or MAP or MML ($\exists$ "slight" differences)

Predict with Best Model

- Use best model from class of models $\mathcal{M}$ for prediction:
- Predict y with probability $\nu^{\text{MDL}}(y|x) = \frac{\nu^{\text{MDL}}(xy)}{\nu^{\text{MDL}}(x)}$ (3 variants)
- $y^{\text{MDL}} = \arg\max_y \{\nu^{\text{MDL}}(y|x)\}$ is most likely continuation of x
- Special case: $Kw(\nu) =$const.
  $\Longrightarrow$ MDL $\rightsquigarrow$ ML:=Maximum likelihood principle.
- Example: $H_\theta =$Bernoulli(θ) with $\theta \in [0,1]$ and $Kw(\theta)$:=const. and $\nu(x_{1:n}) = \theta^{n_1}(1-\theta)^{n_0}$ with $n_1 = x_1 + ... + x_n = n - n_0$.
  $\Rightarrow \theta^{\text{MDL}} = \arg\min_\theta \{-\log\theta^{n_1}(1-\theta)^{n_0} + Kw(\theta)\} = \frac{n_1}{n} = \nu^{\text{MDL}}(1|x)$
  $=$ ML frequency estimate. (overconfident, e.g. $n_1 = 0$)
- Compare with Laplace' rule based on Bayes' rule: $\theta^{Laplace} = \frac{n_1+1}{n+2}$.

Application: Sequence Prediction

Instead of Bayes mixture $\xi(x) = \sum_\nu w_\nu \nu(x)$, consider MAP/MDL

$$\nu^{\text{MDL}}(x) = \max\{w_\nu \nu(x) : \nu \in \mathcal{M}\} = \arg\min_{\nu\in\mathcal{M}}\{K\nu(x) + Kw(\nu)\}.$$

Theorem 5.2 (MDL bound)

$$\sum_{t=1}^{\infty} \mathbb{E}\Big[\sum_{x_t}(\mu(x_t|x_{<t}) - \nu^{\text{MDL}}(x_t|x_{<t}))^2\Big] \leq 8w_\mu^{-1}$$

No log as for ξ $\qquad w_\mu \widehat{=} 2^{-K(\mu)}$ $\qquad$ Proof: [PH05]

$\Rightarrow$ MDL converges, but speed can be exp. worse than Bayes&Solomonoff

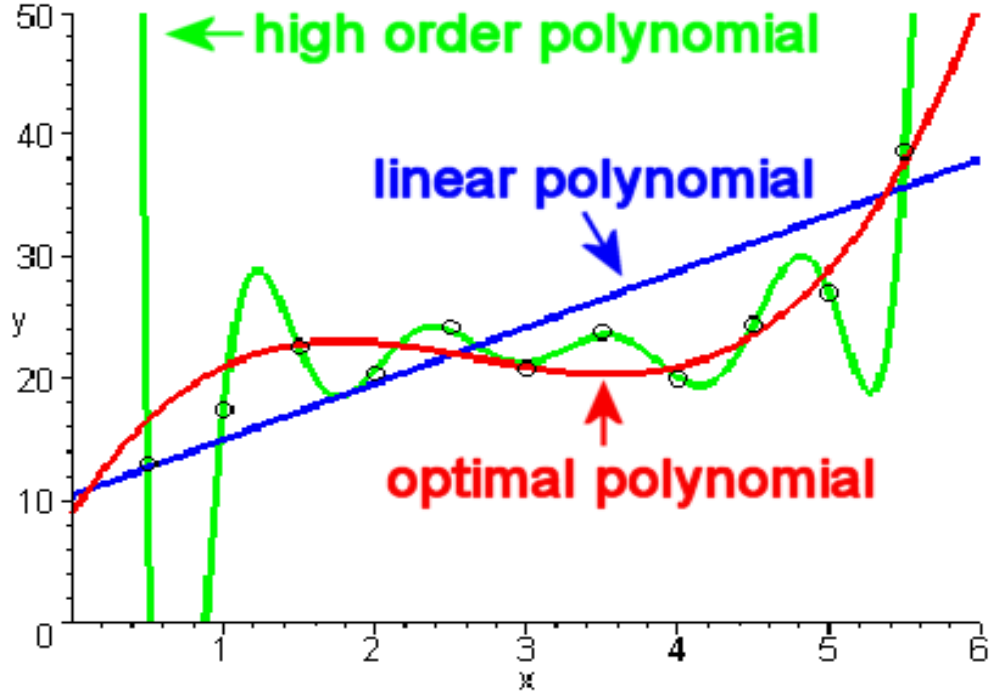
$\Rightarrow$ be careful (bound is tight).

For continuous smooth model class $\mathcal{M}$ and prior w_ν,

MDL is as good as Bayes.

Application: Regression / Polynomial Fitting

- Data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$
- Fit polynomial $f_d(x) := a_0 + a_1 x + a_2 x^2 + ... + a_d x^d$ of degree d through points D
- Measure of error: $SQ(a_0 ... a_d) = \sum_{i=1}^{n} (y_i - f_d(x_i))^2$
- Given d, minimize $SQ(a_{0:d})$ w.r.t. parameters $a_0 ... a_d$.
- This classical approach does not tell us how to choose d? ($d \geq n - 1$ gives perfect fit)



MDL Solution to Polynomial Fitting

Assume y given x is Gaussian with variance σ^2 and mean $f_d(x)$, i.e.

$$P((x,y)|f_d) := P(y|x,f_d) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-f_d(x))^2}{2\sigma^2})$$

$$\Longrightarrow \quad P(D|f_d) = \prod_{i=1}^{d} P((x_i,y_i)|f_d) = \frac{e^{-SQ(a_{0:d})/2\sigma^2}}{(2\pi\sigma^2)^{n/2}}$$

The larger the error SQ, the less likely the data.

Occam: $P(f_d) = 2^{-Kw(f_d)}$. Simple coding: $Kw(f_d) \approx (d+1)\cdot C$, where C is the description length=accuracy of each coefficient a_k in bits $\Longrightarrow$

$$f^{\text{MDL}} = \operatorname*{argmin}_f\{-\log P(D|f)+Kw(f)\} = \operatorname*{argmin}_{d,a_{0:d}}\Big\{\frac{SQ(a_{0:d})}{2\sigma^2\ln 2}+(d+1)C\Big\}$$

Fixed d $\quad\Rightarrow\quad a_{0:d}^{\text{ML}} = \arg\min_{a_{0:d}} SQ(a_{0:d}) =$ classical solution

(by linear invariance of $\arg\min$)

MDL Polynomial Fitting: Determine Degree d

Determine d ($\min_f = \min_d \min_{f_d}$):

$$d = \arg\min_d \{ \frac{1}{2\sigma^2 \ln 2} \underbrace{SQ(a_{0:d}^{\mathsf{ML}})}_{\text{least square fit}} + \underbrace{\frac{n}{2}\log(2\pi\sigma^2)}_{\text{“constant”}} + \underbrace{(d+1)C}_{\text{complexity penalty}} \}$$

Interpretation: Tradeoff between SQuare error and complexity penalty

Minimization w.r.t. σ leads to $n\sigma^2 = SQ(d) := SQ(a_{0:d}^{\mathsf{ML}})$, hence

$$d = \arg\min_d \{ \tfrac{n}{2} \ln SQ(d) + (d+1)C \}.$$

With subtle arguments one can derive $C \stackrel{+}{=} \frac{1}{2} \ln n$.

Numerically find minimum of r.h.s.

Minimum Description Length: Summary

- Probability axioms give no guidance of how to choose the prior.
- Occam's razor is the only general (always applicable) principle for determining priors, especially in complex domains typical for AI.
- Prior $= 2^{-\text{descr.length}}$ — Universal prior $= 2^{-\text{Kolmogorov complexity}}$.
- Prediction $\widehat{=}$ finding regularities $\widehat{=}$ compression $\widehat{=}$ MDL.
- MDL principle: from a model class, a model is chosen that: minimizes the joint description length of the model and the data observed so far given the model.
- Similar to (Bayesian) Maximum a Posteriori (MAP) principle.
- MDL often as good as Bayes but not always.

Exercises

1. [C15] Determine an explicit expression for the $a_{0:d}^{\mathsf{ML}}$ estimates.
2. [C25] Use some artificial data by sampling from a polynomial with Gaussian or other noise. Use the MDL estimator to fit polynomials through the data points. Is the poly-degree correctly estimated?
3. [C20] Derive similar M(D)L estimators for other function classes like fourier decompositions. Use $C = \frac{1}{2}\ln n$ also for them.
4. [C25] Search for some real data. If other regression curves are available, compare them with your MDL results.

Literature

[Ris89] J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[Wal05] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Springer, Berlin, 2005.

[Gru05] P. D. Grünwald. *Introduction and Tutorial.* In Advances in MDL, Chapters 1 and 2. MIT Press, 2005. http://www.cwi.nl/~pdg/ftp/mdlintro.pdf

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