An Introduction to the Minimum Message Length Principle

Mikko Koivisto

June 3, 2002

Seminar on algorithmics and machine learning
Department of Computer Science
University of Helsinki

Outline

1. The fundamental Bayesian approach
2. The Minimum Message Length Principle
3. From Strict MML to practical MML
4. Examples of MML
5. Discussion
6. References

Bayesianism and ‘optimal’ decisions

Principle: From a set of available actions, choose

$$\hat{a} = \arg\max_a E_\pi(\text{Utility}(a, w)),$$

where π is a probability distribution for possible ‘worlds’ w.

Assumptions:

- Utility function is well defined
- One can assign probabilities to ‘worlds’

Q: *How could I obtain a sensible π?*

A: *Approximate Your prior beliefs by a Bayesian model, and use the probability calculus!*

Model and calculus

1. Choose a set X_V of variables.
2. Specify the joint distribution $P(X_V)$.
3. Given $X_B = x_B$, put
   $\pi(w) = P(X_W = w \mid X_B = x_B)$.

Note:

- X_V together with $P(X_V)$ is called *Bayesian model*.
- $P(X_V)$ often constructed using the chain rule $P(X_V) = P(X_A)P(X_B \mid X_A)$, where $A \subseteq V$ indexes the *parameters* and $B = V \setminus A$ indexes the *observables*.
- Above we assumed that the possible 'worlds' relevant to the ultimate decision problem are represented by X_W.

Inference without utility

If no utility function is given, for Bayesians, the posterior distribution $P(X_W \mid X_B = x_B)$ *is the result of data-analysis. However, sometimes one searches for a good point estimate* $\hat{x}_W$*.*

Example: A physicist would like to know the estimated mass of an electron. The physicist will not necessarily know what uses the estimated mass will be put to.

Difficulty: For an estimate, should one choose, for instance,

- mean, (squared error)
- median, (absolute error)
- or mode? (the probability of larger error)

The Minimum Message Length (MML) inference gives an answer!?

MML principle – in words

Principle: *For each model, we consider the message consisting of a description of a model, followed by the data encoded under the assumption that the model is true. The preferred model will be the one with the shortest two part message length.*

Objectives: *The end product of inductive inference is the model that gives the shortest two part message length. This is the best explanation of the data. A model may consist of estimates of model parameters, perhaps representing real world quantities. Among other things, the model can be used for prediction.*

(adapted from Baxter & Oliver 1995)

Toward formulas – some definitions and notations

- Let $\mathcal{D}$ be a countable collection of data sets (or data strings).
- Let Θ be a set of hypotheses (parameter space).
- Let P denote a Bayesian probability model (in a generic way).

MML principle – in formulas

Discretization: Let $R = \{(\Theta_j, \theta_j) : j \in J\}$ be a countable *quantization* of Θ into *regions* Θ_j with *centres* θ_j.

Inference: Given R and a data set D, prefer

$$\arg\min_{\theta_j} Len_R(D, \theta_j),$$

among the centres. Here

$$\begin{aligned} Len_R(D, \theta_j) &\doteq Len_R(\theta_j) + Len(D \mid \theta_j) \\ Len(D \mid \theta_j) &\doteq -\log P(D \mid \theta_j) \\ Len_R(\theta_j) &\doteq -\log P(\Theta_j) \end{aligned}$$

But how to define and find a sensible quantization R in the first place?!

Optimal quantization

Naive idea: Try all possible quantizations and choose the one that gives the minimum message length.

Problem: The message will not be decodeable.

Better idea: Roughly: choose a quantization into regions such that no two centres (hypotheses) are so similar that the available volume of data cannot be expected to distinquish reliably between them.

Hierarchy of MML estimators

Strict MML (SMML) minimizes the expected message length under an optimal two part encoding.

Fairly Strict MML (FSMML) minimizes the expected message length under a specific form of two part encoding.

MML approximates FSMML under the assumptions the log-likelihood and log-prior are quadratic and that prior is locally uniform.

Strict MML

The SMML estimator minimizes the expected two-part message length,

$$SMML \doteq \arg\min_{m:\mathcal{D}\to\Theta} -\sum_{D\in\mathcal{D}} P(D)\log P(D, m(D)).$$

Remark: Each m defines a countable subset

$$\Theta(m) \doteq \{m(D) : D \in \mathcal{D}\} = \{\theta_j : j \in J\} \subseteq \Theta,$$

and hence a partition of $\mathcal{D}$,

$$B(m) \doteq \{m^{-1}(\theta_j) : j \in J\} = \{\mathcal{D}_j : j \in J\}.$$

We note that, for $m(D) = \theta_j$, we have

$$P(m(D)) = P(\mathcal{D}_j) = \sum_{D\in\mathcal{D}_j} P(D).$$

Besides the most simple cases, the search for an optimal m (and hence $SMML$) is extreamly difficult.

Fairly Strict MML

Optimal quantization is now defined as

$$R^* \doteq \arg\min_R \sum_{j \in J} \int_{\Theta_j} \sum_{D \in \mathcal{D}} P(D,\theta) Len_R(D,\theta_j) d\theta.$$

The FSMML estimator is defined by

$$FSMML(D) \doteq \arg\min_{\theta_j \text{ of } R} Len_{R^*}(D,\theta_j)$$

Remark: Note that

$$R^* = \arg\min_R \mathrm{E}\{Len_R(D,\theta_j)\},$$

where the expectation is taken over the joint distribution $P(D,\theta)$, and ignoring the dependence of θ_j on D through the inference step.

The search for an optimal R (and hence $FSMML$) becomes infeasible in multi-dimensional problems.

Approximation approach

1. Consider any $\theta \in \Theta$ as a putative center.
2. Assume the prior is locally uniform.
3. Consider just the *volume* $V(\theta)$ of the region of θ.
4. Try to minimize the expectation of

$$Len_V(D, \theta) \doteq -\log(P(\theta)V(\theta)) - \log P(D \mid \theta).$$

5. Find an analytical solution using Tailor's quadratic approximation to $\log P(D \mid \theta)$.

A practical approximation formula for MML

Some assumptions, approximations, algebra, and questionable arguments yield

$$\begin{aligned} Len_{V^*}(D,\theta) &= -\log P(\theta) - \log P(D \mid \theta) \\ &\quad + \frac{1}{2}\log|F(\theta)| + \frac{d}{2}(1 + \log \kappa_d) \end{aligned}$$

for an optimal volume mapping V^*.

- $|F(\theta)|$ is the determinant of the Fisher Information Matrix
- d is the dimension of Θ
- κ_d is d-dimensional optimal quantizing lattice constant

The MML estimator is defined by

$$MML(D) \doteq \arg\min_\theta Len_{V^*}(D,\theta).$$

Two examples of MML

Let us consider a coin flipping experiment with x observed tails in n trials. Let the prior of the unknown bias be uniform.

With the *binomial* model with fixed n,

$$f_{bin}(x \mid \theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

we have $F_{bin}(\theta) = \frac{n}{\theta(1-\theta)}$, and hence,

$$\hat{\theta}_{bin} = MML(x) = \frac{x + 0.5}{n + 1}.$$

With the *negative binomial* model with fixed x,

$$f_{negbin}(n \mid \theta, x) = \binom{n-1}{x-1} \theta^x (1-\theta)^{n-x},$$

we have $F_{negbin}(\theta) = \frac{x}{\theta^2(1-\theta)}$, and hence,

$$\hat{\theta}_{negbin} = MML(n) = \frac{x + 1}{n + 1.5}.$$

Note: the posterior mean is $\frac{x+1}{n+2}$ *in both cases.*

Properties of MML inference

Pros:

- Generality: allows model selection too
- Invariance under transformations of prior
- Efficiency: sensible discretization
- Consistency

Cons:

- Mysterious an messy derivation of the parctical MML
- Violates the Likelihood Principle: estimates may depend on the data representation. For instance, stopping rule (in binomial versus negative binomial experiment) is not without effect.

References

- Baxter RA. Minimum Message Length Inductive Inference - Theory and Application. PhD thesis, Dept. Computer Science, Monash University, 1996
- Baxter RA, Oliver JJ. MDL and MML: similarities and differences. Technical report TR 207, Dept. of Computer Science, Monash University, 1995.
- Oliver JJ, Baxter RA. MML and Bayesianism: similarities and differences. Technical report TR 206, Dept. of Computer Science, Monash University, 1994.
- Wallace CS, Boulton DM. An information measure for classification. *Computer Journal*, 11, 185-194, 1968.
- Special issue in *Computer Journal*, Volume 42, Issue 4, 1999. Papers and discussions by B. Clarke, A. P. Dawid, J. Rissanen, and C. S Wallace and D. L. Dowe