Message Length Estimators, Probabilistic Sampling and Optimal Prediction

Ian Davidson, Ke Yin, SUNY – Albany, 1400 Washington Ave., Albany, NY 12222, davidson@cs.albany.edu

Abstract

The Rissanen (MDL) and Wallace (MML) formulations of learning by compact encoding only provide a decision criterion to choose between two or more models, they do not provide any guidance on how to search through the model space. Typically, deterministic search techniques such as the expectation maximization (EM) algorithm have been used extensively with the MML/MDL principles to find the single shortest model. However, the probabilistic nature of the MML and MDL approaches makes Markov chain Monte Carlo (MCMC) sampling readily applicable. Sampling involves creating a stochastic process that visits each model in the model space with a chance equal to its posterior probability and has many benefits. We show that for MML estimators using mixture modeling that sampling can find shorter models than deterministic EM search. Samplers can be used to perform optimal Bayesian prediction (OBP), also known as Bayesian model averaging which involves making predictions by calculating the expectation of the predictor with respect to the posterior over all models. We show that for prediction, OBP can outperform even the shortest model and discuss the implications of basing predictions from a collection of models rather than the shortest model. Furthermore, since MML/MDL effectively discretizes the parameter space attaching probability estimates to each region this makes possible sampling across model spaces of varying dimension/complexity.

Introduction

The process of inductive learning essentially abstracts, generalizes or compresses the data into a model from which predictions of the future can be made. This was first formally noted by Solomonoff [1] and Chaitin [2] but it was not until the Rissanen (MDL) [3] and Wallace (MML) [4] formulations of learning by compact encoding using Shannon's information theory that a computable approach became available. However, the MML and MDL approaches only provide a decision criterion to choose between two or more models, they do not provide any guidance on how to search through the collection of possible models in the model space. Though the complexity oriented Levin's optimal universal search [5] approach for classes of inversion problems exists, its application for probabilistically formulated MDL/MML problems seems difficult. Typically deterministic search techniques such as the expectation maximization (EM) algorithm have been used extensively [6] with the MML/MDL principles to find the single *best model* that results in the shortest total encoding of model and data given the model. However, the Bayesian nature of the MML and MDL approaches[1] means approaches in the field of Markov chain Monte Carlo (MCMC) sampling are readily applicable. Sampling involves creating a stochastic process so as to visit each model with a chance equal to its posterior probability and has several benefits over trying to converge to the best model. Furthermore, MML/MDL effectively discretizes the parameter space attaching probability estimates to each region making possible sampling across model spaces of varying dimension/complexity. We show that for MML estimators using mixture modeling that sampling can outperform deterministic EM search and can be used to perform optimal

[1] $P(\theta).P(D|\theta) = 2^{-Length(\theta)+Length(D|\theta)}$ when the lengths are measured in bits.

Bayesian prediction (OBP), also known as Bayesian model averaging, that outperforms even the best model. We briefly discuss the implications of using OBP instead of basing predictions from the best model.

MML Estimators

MDL/MML inference involves constructing a two-part string to be transmitted between a sender and receiver: the model or theory of the observations and the observations encoded with respect to the model. The best model has the shortest total (sum of both parts) message length [7]. A particularly desirable property of the principle is that it discretizes a continuous parameter space into regions attaching a probability *estimate* to each. This enables comparing models of different complexity, such as a three class and five class clustering model, as we have converted both models to the same units of measure, bits of information. Techniques such as maximum likelihood estimation compute probability *densities* making comparisons of models with different complexities analogous to comparing models whose goodness is measured in different units.

The the various formulations of the MML principle are effectively different ways to calculate the dimensions of each region. For example, the 1968 MML Gaussian formulation sub-optimally solved for the height and width *separately* to obtain $width_{\mu} = \bar{s}\sqrt{12/N}, height_{\sigma} = \bar{s}\sqrt{6/(N-1)}, \bar{s}$: sample standard deviation, N: number of instances. Later formulations of MML and MDL make use of the Fisher information to solve for all the region dimensions simultaneously. The MML formulation being $AOPV_{\theta} = \sqrt{12/F(\theta)}, F(\theta)$ is the expected Fisher information. Each region has a representative model that given the data is indistinguishable from all other models in the region. The MML estimate for a given induction problem is the representative model for the most probable region. Some highly probable regions for the simple univariate case are shown in Figure 1.

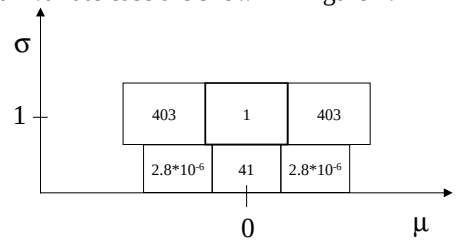

Figure 1: The posterior odds ratio of the MML regions containing and adjacent to the MML estimate for a 500 observation sample from the population μ=0, σ=1

Why Sample?

A good search algorithm will always choose the MML estimate, however a MCMC sampler stochastically chooses a region according to its posterior probability. The MML estimate will most often be chosen, but not always. But why sample instead of searching? The short answer is that it allows us more successfully to find the best model and to even make predictions better than this model.

When combined with approaches such as simulated annealing, sampling can consistently outperform deterministic greedy search algorithms such as EM at finding the shortest encoding [8][9] as they relax the gradient descent requirement. The posterior distribution for most interesting problem involve many local minima that a stochastic search algorithm can “escape” from. This is particularly true when multiple

model spaces of varying complexity are being searched. Our earlier work [8] shows that a MML mixture modeler where k (the number of clusters) is unknown can converge to a model whose Kullback-Leibler distance is closer to the generation mechanism (true model) than by using BIC (Bayesian information criterion) for model class selection and then EM to search the chosen model class.

OBP uses all available models in the model space and can out-perform any single model. However, using a sampler to perform OBP seems to be contradictory to the essence of choosing and using the model that results in the shortest encoding. Even though the model with the shortest encoding is the best given the available data and model space there is still uncertainty associated with this fact. This uncertainty maybe due to the intrinsic nature of the problem if it contains two or more alternative explanations of the data, model space selection or the amount of data available. The consistency of the MDL/MML estimators [10] means that more data or a better choice of model space will help to remove this uncertainty but for a fixed model space and set of data, averaging predictions over all models removes the uncertainty associated with stating that a particular model is the best. Since the "computational devices" used to model the data are rarely capable of universal computation the uncertainty due to the model space selection is not removed. Formally, consider a previously unseen instance that we must predict "+" or "–" for, a set of data D and the model space Θ. The OBP approach sums the belief that the prediction is "+" for each model weighted by its posterior probability. Formally:

$$
\begin{aligned}
P(+) &= \int_{\theta\in\Theta} P(+\mid\theta)P(\theta)P(D\mid\theta)\big/P(D)\,d\theta, \text{ + and } D \text{ are conditionally independent given } \theta \\
&= \int_{\theta\in\Theta} P(+,D\mid\theta)P(\theta)\big/P(D)\,d\theta \\
&= \int_{\theta\in\Theta} P(+,D,\theta)P(\theta)\big/P(D)P(\theta)\,d\theta, \text{ Cancelling terms and marginalizing} \\
&= P(+,D)\big/P(D) = P(+\mid D), c.f.\ P(+\mid\theta_{Shortest})
\end{aligned}
\qquad (1
$$

From equation (1) we see that OBC is effectively making a prediction from **all of the data** not just a single model so is effectively removing the model uncertainty. Such a classifier is optimal in the sense that it produces the minimal predictive risk for a 1-0 loss function for a given data set and model space, see [11] for details.

MML and MCMC

Both Gibbs sampling and the Metropolis-Hastings algorithm are popular approaches to construct ergodic Markov chains for a specific stationary distribution [12], in our case the posterior defined by the message lengths. Gibbs sampling is possible if conditional probability estimates of the model parts exists which is the case for many latent variable models such as mixture and hidden Markov models. Consider the situation of model being represented by a number of random variables, $X_1^{(t)} \ldots X_n^{(t)}$. Gibbs sampling performs an *asynchronous* update of each random variable to derive the new value of the chain at time $t+1$. In theory there is no particular order of updating though usually each random variable is updated in sequence, a process known as a sweep. Updating a random variable stochastically assigns it a new value according to the distribution defined over all possible values conditional on all other random variable values. That is the conditional distribution for updating of the i^{th} random variable is $P(X_i^{(t+1)} \mid x_1^{(t)} \ldots x_{i-1}^{(t)}, x_{i+1}^{(t)} \ldots x_n^{(t)})$.

In this paper we discuss two types of MCMC samplers: approximate Gibbs (aGibbs) and exact Gibbs (eGibbs) sampling. Though the former is only an approximation to Gibbs sampling it is worth discussing as it can be easily implemented

for a number of MDL/MML estimators with only a minor change to the EM algorithm producing superior results. The eGibbs algorithm compares favorably with other approaches that sample across a model space of varying dimension such as reverse jump samplers and can perform OBP. We compare these two types of samplers against standard search and prediction techniques for mixture models, but the approaches are applicable to most problems.

A Graphical View of Mixture Models

Figure 2 shows the variables/parameters used in mixture models. The model (*M*) consists of the parameters for each cluster $\theta_1 \ldots \theta_k$, while $q_1 \ldots q_k$ is the cluster's relative frequency. The data *X* is a fixed collection (hence its shading) of observations/instances with the purpose of clustering to allocate each to a group, this allocation is represented by the latent *Z* variables. The arrows in the diagram represent the influence of one part on another and can be used to derive the conditional probability estimates.

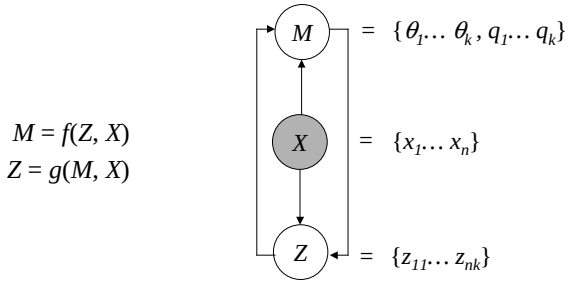

Figure 2. The parts of a clustering model and the dependencies between each part

This graphical representation of mixture modeling shows the central problem is to find the values of *M* and *Z* that minimize the message length. However, there exists a circular dependency of *M* depending partially on *Z* which in turn depends on *M*. The EM (expectation maximization) algorithm addresses this situation by iteratively calculating the *expected* value of *Z* and then choosing the values for *M* that are maximized given the *Z* values. Such iterative algorithms are common approaches in combinatorial optimization a common example being Newton's algorithm. In the MML context the expected *Z* values for each observation/instance are calculated from the message lengths $P(z_{ij})=2^{-\text{MessLen}(zij=1)} / \Sigma_k 2^{-\text{MessLen}(zik=1)}$, while the values for *M* are the MML estimates given the *Z* values. This approach of calculating the expected values for the latent (hidden) variables in the model makes the EM approach easily applicable to a variety of problems such as hidden Markov models where the gamma variables are latent in the Rabiner notation [13].

Approximate Gibbs Sampling (AGS)

MCMC sampling treats *M* and *Z* as being random variables aiming to create a stochastic process that produces a stationary distribution (the posterior) over them. Rather than calculating the expected values now only one of $z_{i,1} \ldots z_{i,k}$ is set to one, the remainder are set to zero. In aGibbs sampling we sample the *Z* values by randomly and exclusively assigning each instance to only one cluster according to the distribution given by the message lengths. The values of *M* are still chosen to be the MML estimate.

That is the *g* function is stochastic while the *f* function is deterministic. Our results using this approach against the EM algorithm with and without BIC for model class selection for a small 3-class clustering problem show its ability to find models that are shorter and better performing on unseen data. This small change is readily applicable to a variety of problems that have used EM to solve latent variables problems. We show some empirical results in Figure 3, see [9] for more empirical results.

	Lowest Error	Mean Error	Shortest ML	Mean ML
Approximate Gibbs Sampler – k=3	0.05	0.08	6925	6944
EM Algorithm – k=3	0.07	0.11	6927	7027
Approximate Gibbs Sampler – k unknown	0.05	0.09	6925	6946
BIC and EM Algorithm – k unknown	0.07	0.14	6927	7037

Figure 3: Approximate Gibbs sampling against EM for the IRIS data set for 1000 trials from random restarts. The error is the percentage of observations that have labels different to the majority of observations within the cluster. We assign observations to their most probable cluster.

Exact Gibbs Sampling (EGS)

Functionally, exact Gibbs sampling differs from approximate Gibbs sampling only that we update all parts of the model stochastically and don't automatically pick the MML region containing the MML estimate. That is the *f* function is stochastic. Though this is a small functional change the amount of computation to achieve this change is substantially more. Rather than performing one message length calculation for the MML estimate containing region we need to now calculate the message lengths for each region. We propose a method where the message length of the Gaussian MML regions are viewed as a two-dimensional sequence and derive an efficient approach to calculate the probability of any region relative to the MML estimate containing region (details omitted in extended abstract). Empirically we show for the hand written digit recognition problem shown in Figure 5 that OBP using MML estimators outperforms the single best model found using either EM or eGibbs sampling. We present some of results in Figure 4.

	Lowest Error	Mean Error
Approximate Gibbs Sampler	12%	15%
EM Algorithm	13%	17%
OBP Using Exact Gibbs Sampler	9%	11%

Figure 4: Comparing OBP with the best model found using approximate Gibbs and EM for the Digit data set for 1000 random division of data into training (80%) and test (20%) data sets.

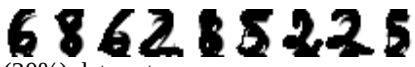

Figure 5: Examples of Differently Drawn Numbers from the UCI Digit Dataset.

To do ????

Maths for 68 estimators

Maths for 87 estimators

Results for 2 variate problem

Results for digit data set

References

[1] Solomonoff, R., A Formal Theory of Inductive Inference: Part 1, *IEEE Information Theory and Control*, 7, pp. 1-22, 1964.

[2] Chaitin, G., On The Difficulty of Computations, IEEE Information Theory, IT-16, pp. 5-9, 1970

[3] Rissanen, J.J, Modelling by Shortest Data Description, *Automatica*, 14, pp. 465-471, 1978.

[4] Wallace, C.S., Boulton D.M., An Information Measure for Classification, Computer J Volume 11 No. 2, pp. 185-94, 1968

[5] J. Schmidhuber. Discovering neural nets with low Kolmogorov complexity and high generalization capability. Neural Networks, 10(5):857-873, 1997 (123 K).

[6] Oliver, J., Baxter, R., and Wallace, C.S., Unsupervised Learning Using MML, International Conference on Machine Learning: Proceedings of the Thirteenth International Conference, 1996

[7] Wallace, C.S., Freeman, P.R., Estimation and Inference by Compact Encoding, *Journal Royal Statistical Society, Series B, Methodology*, volume 49 No 3, pp. 240-265, 1987

[8] Davidson, I., Minimum Message Length Clustering and Gibbs Sampling, 16th Uncertainty in A.I. Conference, 2000.

[9] Davidson, I., "Combining Probabilistic Search, Latent Variable Analysis and Classification Models", AAAI Workshop on Probabilistic Search, 2002

[10] Barron, A., Cover T., Minimum Complexity Density Estimation, *IEEE Transactions on Information Theory*, 37, pp. 1034-1054, 1991.

[11] Duda R., Hart and Stork, *Pattern Recognition*, Wiley, 2002.

[12] Neal, R., Probabilistic Inference Using Markov Chain Monte Carlo Method, *Technical Report CRG-TR-93-1*, Department of Computer Science University of Toronto, 1993.

[13] Rabiner, L. R. and Juang, B. H., An introduction to hidden Markov models, *IEEE ASSP Magazine*, 1986.