Model Selection Criteria for Learning Belief Nets: An Empirical Comparison

Tim Van Allen VANALLEN@CS.UALBERTA.CA
Russ Greiner GREINER@CS.UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2H1 Canada

Abstract

Learning the dependency structure of a (Bayesian) belief net involves a trade-off between simplicity and goodness of fit to the training data. We describe the results of an empirical comparison of three standard model selection criteria — viz., a Minimum Description Length criterion (MDL), Akaike's Information Criterion (AIC) and a Cross-Validation criterion (XV) — applied to this problem. Our results suggest that AIC and XV are both good criteria for avoiding overfitting, but MDL does not work well in this context.

This report focuses on the challenge of learning the (Bayesian) belief net *BN* [Pea88] that has minimum KL-divergence [KL51] from the true distribution, $\mathcal{D}$ over a set of discrete variables X — *i.e.*, the network that minimizes[1]

$$\text{info}(\, BN;\, \mathcal{D}\,) \;=\; -\sum_x \mathrm{P}_{\mathcal{D}}(\, X = x\,) \log \mathrm{P}_{BN}(\, X = x\,)$$

from a fixed training sample s drawn iid from $\mathcal{D}$. As it is easy to find the optimal parameter values (*i.e.*, "CPtable entries") for a given structure [CH92, Hec95], we focus further on selecting the best network structure – *i.e.*, on "model selection".

We let h vary over network structures, and $h(s)$ be the instantiated network formed by using the sample s to fill in h's parameters. It is tempting to simply find the structure h whose instantiation $h(s)$ minimizes the "training error"

$$\text{info}(\, h(s);\, s\,) \quad = \quad -\frac{1}{|s|}\sum_{x\in S} \log \mathrm{P}_{h(s)}(\, X = x\,)$$

Unfortunately, this will typically "overfit" the data, and not produce the best info$(\, h(s);\, \mathcal{D}\,)$ score. We therefore need a more sophisticated approach to find the best structure h.

Algorithms for model selection involve two components: a criterion for comparing models, and a search algorithm, for finding the best model in a given class (based on that criterion). Handling the bias-variance trade-off is primarily a matter of choosing the criterion to be applied. One approach to defining a criterion is to add a complexity penalty to the training error so that more complex models have to fit the data considerably better than smaller models, in order to outscore them. Two standard criteria are

Minimum Description Length **(MDL)** which seeks the h that minimizes

$$\text{MDL}(h; s) \quad = \quad \text{info}(\, h(s);\, s\,) \;+\; \frac{k \log |s|}{2\,|s|}$$

where k is the number of parameters of h [Ris87]

Akaike's Information Criterion **(AIC)** which uses

$$\text{MDL}(h; s) \quad = \quad \text{info}(\, h(s);\, s\,) \;+\; \frac{k \log e}{|s|}$$

where the $\log e$ is simply to convert from nats to bits [Boz87].

Another approach — called *Cross-Validation* (**XV**) — uses only part of the sample $s_1 \subset s$ to set the parameters, and uses the rest of the sample $s_2 = s - s_1$ to get an unbiased estimate of the true error [Sto74]; *i.e.*, evaluates a model h using

$$\text{XV}(h;\; s) \quad = \quad \text{info}(\, h(s_1);\, s_2\,)$$

We empirically compared these three model selection criteria, in our context of learning belief nets, over a range of training sample sizes and the true distributions. Because the space of network structures is huge for even a modest number of variables, a sys-

[1] Here, x ranges over all possible assignments to X. Also $\mathrm{P}_\omega(\, X = x\,)$ is the probability that the distribution ω assigns to x; we will later view an empirical sample s as a distribution. Finally, the true KL-divergence is actually this info$(\, BN;\, \mathcal{D}\,)$ term plus the entropy of the distribution entropy$(\mathcal{D}) \;=\; \sum_x \mathrm{P}_{\mathcal{D}}(\, X = x\,) \log \mathrm{P}_{\mathcal{D}}(\, X = x\,)$; we ignore that term as it is independent of the hypothesis being considered.

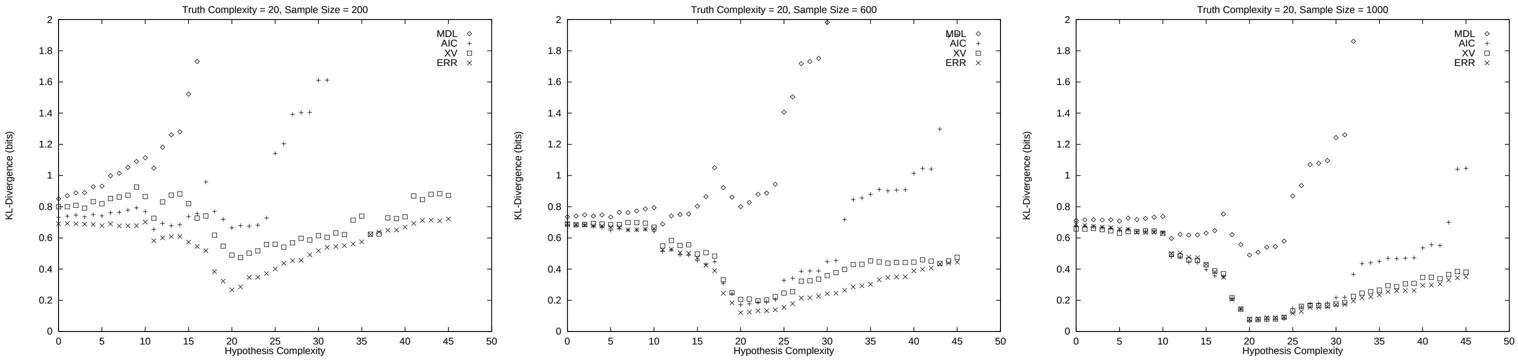

Figure 1. Case Study: (a) $m = 200$ (b) $m = 600$ (c) $m = 1000$

tematic exploration of that space was an unrealistic goal. Instead, we focussed on "trajectories" through that space; in particular, trajectories from the simplest to the most complex structure that include the true structure. We therefore performed many experiments of the following form:

Generate the "true model" BN_t

- Generate all $\binom{n}{2}$ edges over the n variables
- Randomly order these edges
- Pick a complexity value $k \in [0..\binom{n}{2}]$ for the true model
- Define BN_t's structure as having exactly the first k edges in the randomized list
- Generate random probabilities for BN_t's parameters. (Typically done uniformly, but excluding extreme values.)

Generate data sets

- Generate samples of various sizes $\{s_j\}$, each from the true model BN_t

Generate hypotheses

- For $i = 0$ to $\binom{n}{2}$, let hypothesis h_i be the BN-structure (over n variables) with exactly the first i edges in the randomized list.

Actual tests

- Apply each criterion to each hypothesis structure h_i and dataset s_j.

We experimentally found that $n = 10$ was sufficient to produce interesting results, and that our use of binary-valued variables had no qualitative impact on results.

We then observed the behaviour of each criterion across a spectrum of complexities and a range of sample sizes, and found a remarkably consistent pattern. **Figure 1** presents three "snapshots", taken at different sample sizes, of the results for one particular true model. These graphs show the criteria compared across the complexity spectrum (marked out in number of *dependencies*, not number of parameters) when they are evaluated on samples of size 200, 600 and 1000. Four values are plotted for each hypothesis structure: (ERR) the true error, which is the KL-divergence of the network with parameters estimated from the sample, (MDL) the MDL criterion, (AIC) the AIC criterion, and (XV) the Cross-Validation criterion. To scale everything, the true entropy of the distribution has been subtracted from each criterion. The true model is the one with $k = 20$ edges.

Given this set of hypotheses, an ideal learner using a criteria $\gamma(h; s)$ would pick a hypothesis with the lowest γ-value. So for the $m = 200$ graph, the MDL-based learner would pick h_0 (*i.e.*, the the 0-edge structure), the AIC-based learner would select h_{11} and the XV-based would pick h_{21}. While all are wrong (recall the true structure is h_{20}) note that the structure returned by MDL, h_0, is the worst, in that its KL-divergence is 0.65, while the answer returned by AIC has KL-divergence of 0.60 (h_{11}) and the answer for XV has KL-divergence of 0.23 (h_{21}). For $m = 600$, MDL picks h_{11} (KL of 0.55), AIC picks h_{20} (KL of 0.1) and XV picks h_{22} (KL of 0.1); and for $m = 1000$, all three (correctly) pick h_{20}. In all cases, we see that XV finds a structure that is close to optimal, while MDL does not, at least for small samples.

The extended paper [VG00] further compares the complexity penalties of MDL and AIC with the actual amount of overfitting measured, and observes that, while AIC does a reasonably good job of matching the overfitting until the network complexity gets too high, the MDL penalty is much larger than the amount of overfitting.

Table 1 summarizes the results of a more comprehensive study: For each $\langle$sample-size m, truth complexity k $\rangle$ combination we carried out 30 experiments of the type described above. For each experiment, for each criterion, we took the network that scored the

Table 1. Comprehensive Study.

	$k=0$			$k=10$			$k=20$			$k=30$		
m	MDL	AIC	XV	MDL	AIC	XV	MDL	AIC	XV	MDL	AIC	XV
200												
μ	**0.0015**	0.0074	0.0210	0.0618	**0.0138**	0.0258	0.3079	0.0483	**0.0377**	0.4705	0.2008	**0.0477**
M	**0.0000**	**0.0000**	0.0117	0.0044	**0.0000**	0.0096	0.3167	**0.0031**	**0.0031**	0.4419	0.1861	**0.0275**
400												
μ	**0.0000**	0.0020	0.0050	0.0277	**0.0036**	0.0141	0.1658	0.0181	**0.0064**	0.4965	0.0864	**0.0231**
M	**0.0000**	**0.0000**	0.0014	**0.0000**	**0.0000**	0.0038	0.1332	**0.0000**	**0.0000**	0.4884	0.0521	**0.0000**
600												
μ	**0.0000**	0.0017	0.0016	0.0111	**0.0010**	0.0049	0.0946	**0.0008**	0.0033	0.3601	0.0589	**0.0143**
M	**0.0000**	**0.0000**	0.0001	**0.0000**	**0.0000**	**0.0000**	0.0461	**0.0000**	**0.0000**	0.3143	0.0031	**0.0000**
800												
μ	**0.0003**	0.0022	0.0032	0.0023	**0.0014**	0.0047	0.0510	**0.0001**	0.0084	0.3684	0.0294	**0.0020**
M	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.0006	**0.0000**	**0.0000**	**0.0000**	0.3408	**0.0000**	**0.0000**
1000												
μ	**0.0000**	0.0023	0.0027	**0.0013**	0.0023	0.0036	0.0319	**0.0016**	0.0027	0.3150	0.0232	**0.0032**
M	**0.0000**	**0.0000**	0.0003	**0.0000**	**0.0000**	**0.0000**	0.0160	**0.0000**	**0.0000**	0.3504	**0.0000**	**0.0000**

best and subtracted its error from the lowest error attained by any network. We summarize these 30 values obtained by giving their mean and median, and use a large font to indicate the "winners" in each cell. Note, however, the differences are more important than distinguishing the best: Where MDL "won" (on left), the other methods also did quite well in attaining low error; but where MDL did poorly (on right), it did very poorly relative to the other criteria.

These empirical results show that optimizing for the MDL criterion can be a risky strategy for learning belief net structures. While MDL does seem to work for sufficiently large samples, it can be arbitrarily worse for even slightly smaller samples; therefore there is no guarantee of graceful degradation. Furthermore, there is no way to know *a priori* whether MDL has sufficient data to be effective. By contrast, we found XV to be a "safe bet", one which was never that bad.[2] (Table 1 shows that XV's average error never exceeded 0.1; and in fact, XV was the minimax over the three criteria.) AIC's performance was in-between, but closer to XV, in terms of its risk.

Based on our experience: for learning belief net structures, if there is no prior knowledge, we advise using XV; if there is a prior expectation of simplicity, we advise using AIC; and we advise against the use of MDL.

For more information, including a more complete description of our data and results, see `http://www.cs.ualberta.ca/~greiner/CRITERIA`.

[2] This is consistent with Cross-Validation's other name, "the jackknife" — *i.e.*, a jack of all trades, even if a master of none.

References

[Boz87] H. Bozdogan. Model selection and akaike's information criterion (aic): the general theory and its analytical extensions. *Psychometrica*, 52(2):345–370, 1987.

[CH92] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[Hec95] David E. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.

[KL51] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:76–86, 1951.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, 1988.

[Ris87] J. Rissanen. Stochastic complexity (with discussion). *J. Royal Statistical Society*, 49:223–239 and 253–265, 1987.

[Sto74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Society*, 36:41–47, 1974.

[VG00] Tim Van Allen and Russ Greiner. Model selection criteria for learning belief nets: An empirical comparison. Technical report, UofAlberta, Dept of CS, 2000.