Active Bayesian Learning For Mixture Models

Ian Davidson
Silicon Graphics
1300 Crittenden Ln, MS 876
Mountain View, CA 94587
inpd@hotmail.com

Abstract

Traditionally, Bayesian inductive learning involves finding the most probable model from the entire data set. The induction algorithm is a passive recipient of the data. Active learning is when the learner starts with a small sample of data and specifies additional data points to collect to better understand the domain. Our active learning approach involves collecting alternative explanations (highly probable models) of the data and choosing new data points where the models make the most different predictions. This general idea has been qualitatively mentioned in the Bayesian literature as a method to reduce uncertainty amongst the competing models. We show that the approach is the optimum way to reduce entropy amongst the competing models' posterior probabilities and provide empirical results showing its effectiveness for active learning.

1 Introduction and Overview

Traditional inductive inference assumes that the learning algorithm passively receives the available observations and finds the best model in the available model space. In a Bayesian context, this is the model with the maximum a posteriori (MAP) estimate.

Active learning studies closed-loop learners that make queries that influence what data are added to their training set [1]. The hope is that good data choice will allow the selection of the "best" model using fewer observations. Three situations where active learning may be used are; when data is expensive, hard or dangerous to collect, when the learner cannot efficiently process all the data or to increase the rate of learning [2].

Consider the omnipotent situation where we have available all the observations, D, from a domain and know that θ_{TRUE} is the best model. Assume the model space to search, Θ, contains θ_{TRUE} ($\theta_{TRUE} \in \Theta$).

We can state the aim of active Bayesian learning is to select θ_{TRUE} as the MAP estimate using the smallest number of observations.

This assumes the learner is a consistent estimator such as the Minimum Message Length (MML) estimators [3][4]. An inconsistent estimator's best model would change with the size of the training set.

As D is too large or too expensive to collect we cannot give it to the learner to process. Instead the initial data set is a random selection of observations, D_0, $D_0 \subset$

D. We regularly add to D_0, batches of new observations, D_i, *purposefully* chosen from *D* (without replacement). At any instant, the observations currently available to the learner is, D^*, where $D^* = D_0 \cup D_1 \cup D_2 \cup D_3 \ldots \cup D_b$ where *b* is the number of batches chosen so far. We hope that θ_{TRUE} will be the most probable model using D^* where $D^* \subset D$ and $|D^*| << |D|$. The process of actually collecting a new batch of observations can be quite complicated in certain environments. In this paper we will assume all the observations in *D* are readily available and will focus on specifying a criterion identifying which observations to collect.

We postulate fast convergence to the true model will result when collecting observations where the models make the most different predictions.

1.1 Overview of the Paper

The following sections describe our active learning approach in more detail. We show that our approach is the optimum way to reduce the entropy of the competing models' posterior probabilities. We then attempt to show the usefulness of the approach for active learning by describing our results for multi-variate mixture modeling. Finally we conclude by describing other statistical and probabilistic active learning work, our contribution and future work.

2 Active Bayesian learning from Multiple Models.

From the initial data set and after each addition of data points to the training set the learner will find the alternative explanations (highly probable models). This is an active research question touching on how to identify distinct models [5][6] and efficient mixing between modes in the posterior [7]. In this paper we will focus on mixture modeling and use a Gibbs sampler [8] to find the *alternative* explanations/models in the data. For the remainder of this paper, we will assume our learner can find the alternative explanations/models in the model space.

Let the alternative explanations/models be $h_1, h_2, h_3, \ldots h_n$. The data collection focus is where the models make the most different predictions. However, how is it possible to measure the differences of two models' predictions? For specific applications such as mixture modeling key aspects of the model could be used such as the number of classes. We choose to use a more general approach. In a Bayesian inference setting the learnt models provide density estimations over the instance space. Two models' predictions differ by their respective estimate of the probability of an observation occurring in some instance space region. Consider the simple univariate mixture model situation of two competing Gaussian models. The models' parameter estimates are N(0,1) and N(3,1). The models' predictions will differ significantly at the intervals surrounding 0 and 3. In each case one model predicts an abundance of observations which the other does not.

Our uncertainty, *I*, of which is the true model for the current data set is simply the information content of the alternative models' normalized joint probabilities, as shown in equation (2.1).

$$I = -\sum_{i=1}^{n} \frac{P(h_i \cap D)}{\sum_{j=1}^{n} P(h_j \cap D)} \cdot \ln \frac{P(h_i \cap D)}{\sum_{j=1}^{n} P(h_j \cap D)} \qquad (2.1)$$

Uncertainty is maximized when each model is equally likely and is minimized when there is only one plausible model. New experiments can reduce uncertainty by generating observations that eliminate one of the models or make one model more probable than another. When a model's joint probability is not high enough to be

plausible the first situation occurs. Therefore, this is an extreme case of the second situation.

We will show the optimum way of reducing uncertainty is to collect observational data where the models make the most different predictions. We start with the simplest case of two alternative models (h_1, h_2) induced from a data set, to which, one additional observation will be added.

The initial set of observations is termed D_0 to which an additional observation, x, we add to D_0 to obtain D_1. We assume the joint probabilities of the data and model are normalized so that the sum for all plausible models is one. Our aim is to maximize the information gain, ΔI, by the addition of the observation.

Consider equation (2.2 **)**, the observation with the greatest information content to discriminate which is the better model will maximize equation (2.3). This occurs when the new observation maximizes the difference of the likelihood for the two models. One model may predict the observation will most likely occur whilst the other that it will not. We will call ΔI the information gain due to the addition of observation(s).

$$\Delta I = I_0 - I_1 = I_0 - -\sum_{i=1}^{2} \frac{P(h_i)P(D_0|h_i)P(x|h_i)}{\sum_{j=1}^{2} P(h_j \cap D_1)} \ln \frac{P(h_i)P(D_0|h_i)P(x|h_i)}{\sum_{j=1}^{2} P(h_j \cap D_1)} \qquad (2.2)$$

$$\text{if we treat } \sum_{j=1}^{2} P(h_j \cap D_1) \text{ as a normalizing constant then this expression is}$$

$$= I_0 - -\sum_{i=1}^{2} P(h_i)P(D_1|h_i).\ln(P(h_i)P(D_1|h_i)),$$

$$\text{this expression is maximized when } \left|P(h_1)P(D_1|h_1) - P(h_2)P(D_1|h_2)\right| \text{ is maximized} \qquad (2.3)$$

$$\text{If we assume uniform priors over the model space, } P(h_1) = P(h_2) \text{ and}$$

$$\text{as } P(D_1) = P(D_0)P(x) \text{ the expression to minimize is :}$$

$$-\sum_{i=1}^{2} P(x_i).\ln(P(x_i))$$

$$\text{This occurs when } \left|P(x|h_1) - P(x|h_0)\right| \text{ is maximized} \qquad (2.4)$$

We can generalize this finding for n models and m additional data points to determine the expression that maximizes ΔI. For n models and one additional data point this occurs when the data point maximizes the sum of differences in likelihood between every *combination* of model pairs, equation (2.5).

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left|P(D_0|h_i)P(x|h_i) - P(D_0|h_j)P(x|h_j)\right| \qquad (2.5)$$

Pragmatically, generalizing this for m data points involves using equation (2.5) to rank a set of observations in decreasing value. Then selecting m data points involves taking the top m observations. We can formally generalize for m data points ($x_1 \ldots x_m$) by choosing the m points that maximize equation (2.6).

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left|P(D_0|h_i)\prod_{k=1}^{m} P(x_k|h_i) - P(D_0|h_j)\prod_{k=1}^{m} P(x_k|h_j)\right| \qquad (2.6)$$

We have described which data points will reduce the model uncertainty the most, how to collect them will be application specific. In our empirical mixture model trials (using uniform priors over the model space) we assume that all the data is available, but our mixture modeler would be too slow to process it all. We start with a small random subset of the data and rank order those observations not currently in

D^* according to equation (2.5) from largest to smallest. The observations at the top of the list are those that are predicted most differently by the alternative models.

2.1 Assumptions

Our aim is to reduce uncertainty by eliminating all but one model. Therefor phenomenon with two distinct and legitimate explanations will have only one explanation isolated. Other work in active learning implicitly make the assumption of the existence of only one posterior mode.

We assume that the learner can find alternative models of the data and is a consistent estimator. The assumption that one needs to find all the alternative models is a more considerable. Our empirical results show that not finding all the competing models still yields good results, but we plan to investigate this area more systematically in future work.

3 Experimental results with Mixture Modeling

We will use the MML mixture model Gibbs sampler defined in [5][8]. Each application of the sampler consists of 10,000 sweeps/iterations to find the best alternative models. We use a data set of six independent Gaussian variates consisting of six classes (generation mechanism) all with means of 0 and standard deviations of 0.5 except for class i whose i^{th} attribute has a mean of 1. That is, $\mu_{1...6,1..6}=0$ except $\mu_{i,i}=1$, $\sigma_{1..6,1..6}=0.5$. We start with a random sample of 900 data points but have many thousands of observations available to draw on. This data set contains many three, four, five and six class posterior modes as the classes overlap significantly.

We compare our active data selection approach against randomly chosen data. From the initial observations we find the four most distinct models, one each from the three, four, five and six class model spaces and calculate the entropy in their joint probabilities. We enforce this limitation to determine if the approach is viable if not *all* the alternative models are used. Our first strategy is to add fifty randomly chosen observations and repeat the search for the best models with the enforced limitation. Our second strategy is to purposefully select fifty observations that maximize equation (2.5) and repeat our best models search. For the remainder of this section we shall discuss average results for one hundred trials. We show in Table 1 that our approach maximizes ΔI.

Table 1: The posterior ratio of the most distinct models relative to the most probable model on a natural logarithm scale for: the original 900 observation sample, sample plus random observations and sample plus observation selected by guided experiments. Average results for 100 trials.

	Initial 900 observations	Initial 900 observations and 50 randomly chosen observations	Initial 900 observations and 50 purposely chosen observations
4 Class	1	1	8.03
5 Class	1.6	2.19	1
6 Class	8.14	6.57	12.73
3 Class	9.44	6.26	14.58
Entropy	0.28	0.21	0.12

We can measure how well a learner did on a data set by the Kullback-Leibler (KL) distances between the generation mechanism and the best parameter estimates found. When calculating the KL distances we arrange the component numbers to overcome the identifability problems that are common in mixture models.

The KL distance between the parameters of the generation mechanism and the parameter estimates of the six class models found from the initial sample, initial sample and fifty random data points and initial sample and fifty purposely chosen data points is 8, 5.03 and 3.82 respectively. As the KL distance is asymmetrical, this and other KL distances reported are the average result, for example $KL_{Average}(a,b)=(KL(a,b)+KL(b,a))/2$. For ease of writing we drop the "Average" label.

We show the KL distances between the various six class models in Table 2. We can see that after adding fifty observations, our active data selection approach is closest to the true model (first row). Interestingly the distance, $KL(\theta_{INITIAL}, \theta_{ACTIVE})$, is twice as large as $KL(\theta_{INITIAL}, \theta_{RANDOM})$ indicating the actively chosen data points were able to support a more different model than the one found from the initial data set.

Table 2: The Mean KL distances between the various six class models. Average results for 100 trials.

	True Model	**Initial Data (ID)**	**ID+ Random Data**	**ID + Active Selection**
True Model	0	8.00	5.03	3.82
Initial Data	8.00	0	1.08	2.54
ID+ Random Selection	5.03	1.08	0	2.55
ID+ Active Selection	3.82	2.54	2.55	0

We add six fifty-observation batches, re-running the mixture modeler after each addition. Figure 1 shows the mean KL Distance between the true model and the best six-class models found from the respective data sets. The parameter estimates of the models found from the actively selected data are always closer to the generation mechanisms' parameters. After the addition of all 300 observations our approach is nearly twice as close.

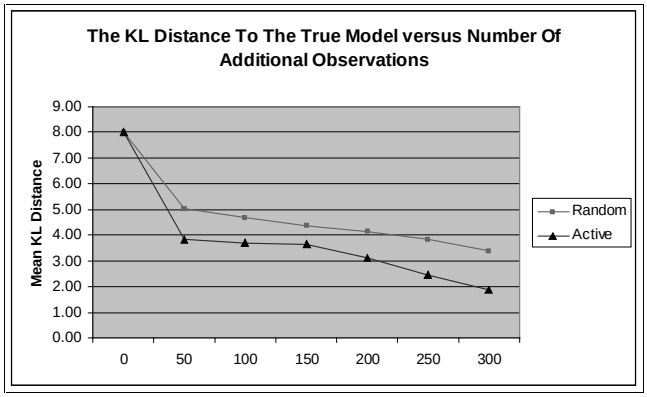

Figure 1: The Mean KL Distances, $KL(\theta_{TRUE}, \theta_{ACTIVE})$ and $KL(\theta_{TRUE}, \theta_{RANDOM})$ between the best six-class models found and the true model after the addition of batches of fifty observations. Average results for 100 independent executions. Standard deviation of results for active and random selection after the addition of 50 points is 0.42 and 0.56

respectively. Standard deviation of results for active and random selection after the addition of 300 points is 0.24 and 0.38 respectively. Average results for 100 trials.

4 Summary of Other Active Learning Work

Active learning has its roots in at least the three fields; machine learning [2], Bayesian inference [9] [10] and statistics [1]. I will limit my discussion to the work of Mackay and Cohn as their work is most relevant to the ideas in this paper. Most active learning work in machine learning [11] focuses on non-probabilistic approaches such as Windowing. The simplest form of Windowing [12] involves adding those test set observations that are mis-classified to the training set.

Our approach to active learning has been postulated generally in the Bayesian experimental design literature [13]. Sivia states that, "a model selection experiment will be optimised if most of the data are collected where the competing hypotheses give (drastically) different predictions". However, he gives only qualitative advice.

Mackay's work on active data selection has some similarity to our own [10]. He uses information theoretic objective functions to actively select data. His third problem definition discusses maximizing the information gain (or reducing the uncertainty) between two competing hypotheses by the addition of one data point to the original *N*. Equation (4.1) illustrates the objective function to maximize.

$$\Delta I = I_N - I_{N+1}, \text{ where } I = -\sum_{i=1}^{2} P(h_i).\log(P(h_i)) \qquad (4.1)$$

Mackay specifically addresses optimally reducing model uncertainty for Bayesian interpolation. Bayesian interpolation involves making inferences about a function *f(x)* from data derived from the function at coarse intervals [13]. The probability of a new observation, *x*, is $P(x|h_j)$ = Normal(μ_j,σ_j^2), these parameters are for the specific data point and are obtained from the interpolation model's best-fit parameters. He shows that the expected information gain from the addition of one data point is:

$$E(\Delta I) \approx \frac{P(h_1).P(h_2)}{2}\left[\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)(\mu_1-\mu_2)^2+\left(\frac{\sigma_1^2-\sigma_2^2}{\sigma_1\sigma_2}\right)^2\right] \qquad (4.2)$$

This expression is maximized when the observation's value chosen causes μ_1-μ_2 to be maximally separated and σ_1-σ_2 to "differ significantly from each other". This is precisely where the models should make the most different predictions. Our approach is similar to Mackay's work with respect to the objective function to minimize and we believe is consistent with and generalizes his findings for a general Bayesian setting.

Whether purposeful data point choices bias our inferences in favoring the wrong model is an important question that Mackay addresses. As Bayesian inference is consistent with the likelihood principle that states we make our inferences from the data we collect not the data we don't collect, no bias is introduced.

Cohn, Ghahramani and Jordan [1] discuss how to select data to minimize the variance of the learning error for unbiased supervised learning. Their approach uses the variance of the learner as an objective function to minimize. For the example of mixture modeling, they use the EM algorithm to find the best parameter estimates. The adjusted variance of the learner is calculated using a Monte Carlo approximation from a set of sample reference points. Those points that minimize the variance are then added to the data set. As is noted by the authors, for high dimensional space, the number of reference points drawn may need to be large.

5 Conclusion

We have presented an approach to active Bayesian learning from multiple models. The approach involves using distinct highly probable models of the data to select observations whose chance of occurring are predicted differently by the models.

We compare our approach against adding observations that are randomly chosen for multivariate mixture modeling. We show formally and empirically that our approach is better at reducing the entropy amongst the competing model's posterior probabilities. We empirically show that applying this approach in an active learning context results in the best model's parameter estimates being closer to the generation mechanism's parameters. Our approach is similar to Mackay's work, as both our objective functions are to minimize the entropy of the competing models posterior probabilities. Mackay suggests an approach specifically for Bayesian interpolation whilst we focus on a general approach for Bayesian learning that we illustrate using mixture modeling. We believe that our approach is consistent with and generalizes Mackay's findings.

We plan further work to understand key aspects of the approach and their effect on the rate of uncertainty reduction. Examples of these aspects include sensitivity to finding all or some of the most probable models and the batch size of observations to add. We also wish to explore either formally or empirically the rate of convergence to the true model. A more long-term objective is comparison against active learners in which only one model is used [1].

References

[1] Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1995. Active learning with statistical models. In Tesauro, G.; Touretzky, D.; and Alspector, J., eds., Advances in Neural Information Processing, volume 7. Morgan Kaufmann.

[2] Blum A.L., and Langley, P., "Selection of Relevant Features and Examples in Machine Learning", Artificial Intelligence, Volume 1, Number 2, 1997, 245-271.

[3] Wallace, C.S., Boulton D.M., An Information Measure for Classification, Computer J Volume 11 No. 2, pp. 185-194, 1968

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

[5] Davidson, I., Markov Monte Carlo Sampling and The Minimum Message Length Principal : Application and Uses, PhD Thesis, Department of Computer Science, Monash University, 1998.

[6] Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62, 795--809

[7] Gilks, W, Richardson, S and Spiegelhalter, D. editors Markov Chain Monte Carlo In Practice, Chapman Hall Publishers, 1996.

[8] Davidson, I., Minimum Message Length Clustering Using Gibbs Sampling, Proceedings of the Uncertainty in Artificial Intelligence Conference 2000.

[9] Cheeseman, P., Chapter 4: On Finding The Most Probable Model, In Computational Models of Scientific Discovery and Theory Formation, Editors Langely, P. and Shrager, J., Publishers Morgan Kaufmann, 1990.

[10] Mackay, D., Information-based objective functions for active data selection, Neural Computation 4:4, pages 589-603, 1992

[11] Furnkranz, J., "Integrative Windowing", Journal of Artificial Intelligence Research, Volume 8, 1998, 129-164.

[12] Quinlan, J. R. (1983). Learning efficient classification procedures and their application to chess end games. In Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.), Machine Learning. An Artificial Intelligence Approach, pp. 463--482. Tioga, Palo Alto, CA.

[13] Sivia, D., Data Analysis A Bayesian Tutorial, Clarendon Press, 1996.