Learning Bayesian Networks under Equivalence Constraints (Abstract)

Tiansheng Yao, Arthur Choi and Adnan Darwiche
Computer Science Department
University of California, Los Angeles
{*tsyao,aychoi,darwiche*}*@cs.ucla.edu*

Introduction

Machine learning tasks typically assume that the examples of a given dataset are independent and identically distributed (i.i.d.). Yet, there are many domains and applications where this assumption does not strictly hold. Further, there may be additional information available that ties together the examples of a dataset, which we could exploit to learn more accurate models. For example, there are clustering tasks in the domain of semi-supervised learning where, for example, we have available side information that tells us that certain pairs of examples belong to the same cluster. To incorporate such information, *constrained* versions of k-means clustering (Wagstaff et al. 2001), Gaussian mixture models (Lu and Leen 2004; Shental et al. 2003) and a variety of other models and algorithms, have been proposed in the literature; see, e.g., the surveys (Davidson 2009; Han, Kamber, and Pei 2011).

We propose here to abstract such problems in more general terms, as a task of learning from datasets that are subject to *equivalence constraints*. We formalize the notion of learning a Bayesian network subject to equivalence constraints, introducing a notion of a *constrained dataset*, which implies a corresponding *constrained log likelihood*. The constrained log likelihood provides a simple and principled way to learn, for example, the parameters of a Bayesian network from a constrained dataset. The constrained log likelihood, however, is intractable in general, although we identify a special case where we can design practical algorithms for optimizing the constrained log likelihood. In particular, we propose, as an example, a *constrained* generalization of expectation maximization (EM), for a class of models that subsumes those for constrained clustering tasks as a special case.

Constrained Datasets

We introduce a particular type of dataset, called a *constrained dataset*, which is a traditional dataset that is further annotated with *equivalence constraints*. Suppose we are given an incomplete dataset, where we do now know the particular values of a hidden variable across examples in a dataset. Suppose, however, that no matter what that value is, we do know that it must be the same across certain examples in a dataset. For example, consider the incomplete dataset:

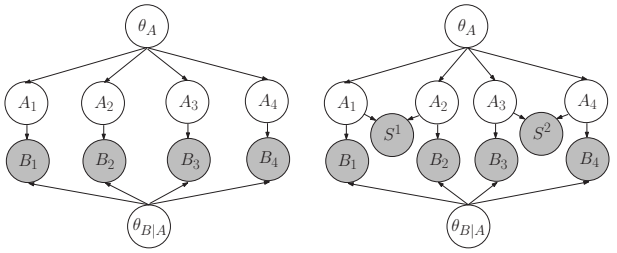


Figure 1: On the left, is a meta-network for a Bayesian network $A \rightarrow B$ with four examples, where A is hidden and B is observed. On the right is a meta-network with equivalence constraints: constraint S^1 on A_1 and A_2, and constraint S^2 on A_3 and A_4, are represented as observed variables.

example	A	B	C
1	a_1	b_1	c_2
2	?	b_2	?
3	a_2	?	c_1
4	?	b_2	c_1

While we do not know the specific value of variable A in examples 2 and 4, suppose that we happened to know that the value of A in these examples must be the same. For example, this value of A could represent the presence or absence of a disease in a patient that has visited two different doctors, who performed two independent sets of tests. While we do not know whether the disease is present or absent in the patient, we do know that it is either present in both cases, or absent in both cases. We view background knowledge such as this, as an equivalence constraint on a dataset. Ideally, we would like to take advantage of such information, in order to learn more accurate models. Our goal now is to formalize such learning tasks, which leads us to the notion of a *constrained log likelihood*.

Constrained Log Likelihoods

To formalize the notion of learning from a constrained dataset, we appeal to the notion of a meta-network, which is typically used to motivate (Bayesian) learning of parameters in Bayesian networks (Darwiche 2009); see Figure 1

(left) for an example. Suppose we are given a dataset $\mathcal{D}$ that is subject to equivalence constraints $\boldsymbol{\mathcal{S}}$. We can represent these equivalence constraints explicitly in a meta-network, as observed variables; see Figure 1 (right) for an example. As a meta-network induces a log likelihood, the constrained meta-network induces a *constrained log likelihood (CLL)*:

$$CLL(\theta \mid \mathcal{D}, \boldsymbol{\mathcal{S}}) = \log \mathcal{P}(\mathcal{D} \mid \boldsymbol{\mathcal{S}}, \theta)$$

where $\mathcal{P}$ denotes the (meta-)distribution induced by the constrained meta-network. To learn the parameters of a Bayesian network, subject to equivalence constraints, we can thus seek to obtain those estimates maximizing the above constrained log likelihood. Note that the constrained log likelihood reduces to the traditional log likelihood when there are no equivalence constraints.

In general, computing the constrained log likelihood is intractable. However, under a certain assumption on the equivalence constraints, the constrained log likelihood is no more difficult to compute than the traditional log likelihood. In particular, if only a single network variable is subject to equivalence constraints, the constrained log likelihood assumes a simple closed-form:

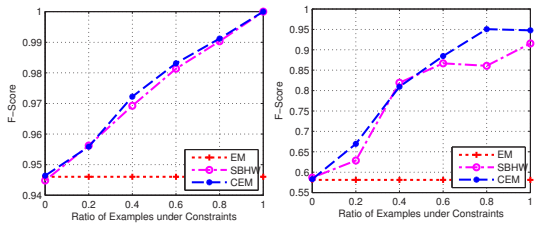
$$\begin{aligned} &CLL(\theta \mid \mathcal{D}, \boldsymbol{\mathcal{S}}) \\ &= LL(\theta \mid \mathcal{D}) + \log \mathcal{P}(\boldsymbol{\mathcal{S}} \mid \mathcal{D}, \theta) - \log \mathcal{P}(\boldsymbol{\mathcal{S}} \mid \theta) \end{aligned}$$

where $LL(\theta \mid \mathcal{D})$ is the traditional log likelihood. We see here that optimizing the constrained log likelihood balances between optimizing the traditional log likelihood and an additional term over equivalence constraints. Under our given assumption, these additional terms can be computed efficiently, as a by-product of computing the log likelihood, which we must compute anyways. We omit these details, however, in this abstract.

An Application in Semi-Supervised Learning

Now having formalized the notion of learning from a constrained dataset, we could appeal to off-the-shelf systems for optimizing the corresponding constrained log likelihood. Here, we consider an EM algorithm (Dempster, Laird, and Rubin 1977; Lauritzen 1995), but one that is adapted to learn the parameters of a Bayesian network from constrained datasets. In particular, we derived a *constrained* EM (CEM) algorithm for the simplified case where a single root variable is subject to equivalence constraints. Such an algorithm is sufficient for semi-supervised clustering tasks that have seen increasing interest in recent years.

We applied our CEM algorithm to learn both naive Bayes models and Gaussian mixture models (GMMs), from constrained datasets. In preliminary experiments on datasets from the UCI ML repository, we found algorithms that take advantage of side-information can exhibit much better clustering performance than vanilla EM, with smooth increase in performance as we provide more side-information. Further, our constrained EM algorithm is competitive with, and sometimes outperforming, more specialized algorithm specifically designed for this domain (Shental et al. 2003); for an example, see Figure 2.



Breast Ionosphere

Figure 2: In GMMs, we observe an increase in F-measure value (y-axis) as the amount of side-information is increased (x-axis).

On Generalizations

In initial experiments, we considered an EM algorithm adapted for the special case where a single root variable is subject to equivalence constraints. Similar assumptions are typically assumed in models used by more specialized algorithms for constrained clustering. However, generalizing to non-root variables, for example, is non-trivial in these cases.

In contrast, in our framework, it is not too difficult to perform this generalization. The constrained log likelihood is still as easy to evaluate in this case, and we further have an analogous EM algorithm to optimize it. There are less trivial generalizations that are correspondingly efficient, for constraints on multiple variables, further under the assumption that the number of constrained variables is bounded.

In general, with arbitrary equivalence constraints, the constrained log likelihood is intractable. However, our formulation further naturally admits a certain approximation to the constrained log likelihood, that leads to another EM-based algorithm that is as efficient as the one we considered in our preliminary experimental results. We plan to investigate such generalizations and approximations in future work.

References

Darwiche, A. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press.

Davidson, I. 2009. Clustering with constraints. In Liu, L., and Özsu, M. T., eds., *Encyclopedia of Database Systems*. Springer US. 393–396.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39:1–38.

Han, J.; Kamber, M.; and Pei, J. 2011. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Lauritzen, S. 1995. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* 19:191–201.

Lu, Z., and Leen, T. K. 2004. Semi-supervised learning with penalized probabilistic clustering. In *NIPS*.

Shental, N.; Bar-hillel, A.; Hertz, T.; and Weinshall, D. 2003. Computing Gaussian mixture models with EM using equivalence constraints. In *NIPS*. MIT Press.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schrödl, S. 2001. Constrained k-means clustering with background knowledge. In *ICML*, 577–584.