How an objective Bayesian integrates data

Jürgen Landes joint work with Jon Williamson

January 13, 2015

Keywords: Objective Bayesianism, data integration, Bayesian nets, inductive inference

Introduction

Computers have made it possible to collect and store large data sets. Reasoners would like to make use of as many data sets as possible which are as large as possible while still allowing for computationally feasible inferences. Ideally, one could simply combine all the available data sets. Unfortunately, the available data sets do not always all employ the same variables, mainly because the data sets have been collected by different persons/groups at different times with varying interests and resources. The challenge arises how to practically make sense of all this data.

The objective Bayesian approach to integrating data, as presented in [3], is roughly as follows

1) gather all relevant data sets which constitute one's evidence,
2) compute the set of probability functions consistent with the evidence, $\mathbb{E}$,
3) adopt the probability function in $\mathbb{E}$ with maximal entropy, $P^\dagger \in \mathbb{E}$.

A simple minded example

Let us see how this approach applies to an example with two data sets, DS_1 and DS_2. Let DS_i employ the variables $\vec{x} \cup \vec{y}^i$ where $\mathbf{x} = \{x_1, \ldots, x_k\}$ and $\mathbf{y}^i = \{y^i_1, \ldots, y^i_{l_i}\}$. L_i is then the finite propositional language generated by the variables in $\mathbf{x} \cup \mathbf{y}^i$ and L is the language generated by the variables in $\mathbf{x} \cup \mathbf{y}^1 \cup \mathbf{y}^2$. The set of probability functions on L is denoted by $\mathbb{P}$. A state of a language is the usual conjunction of negated or non-negated literals. A state of $L_1 \cap L_2$ is denoted by ω_x and a state of $L_i \setminus \{L_1 \cap L_2\}$ is denoted by ω_i.

We denote the observed frequencies in DS_i by P^*_i. If for all $\varphi \in SL_1 \cap SL_2$ it holds that $P^*_1(\varphi) = P^*_2(\varphi)$, we say that the data sets are *consistent* and drop the index i on P^*. That is, the observed frequencies in the data sets agree on all states ω_x.

In my talk, I will show that in the case of two consistent data sets the entropy maximiser $P^\dagger$ is given by

$$P^\dagger(\omega_x \wedge \omega_1 \wedge \omega_2) \cdot P^*(\omega_x) = P^*(\omega_x \wedge \omega_1) \cdot P^*(\omega_x \wedge \omega_2) \ .$$

This follows since the variables $\vec{x}$ screen off the variables in $\vec{y}^1$ from those in $\vec{y}^2$, [2] – as I will explain in some detail. That is, we can read-off $P^\dagger$ directly from the data, without performing any calculations.

Yes, but ...

Meaningful applications tend to be somewhat harder than the above example. The following problems arise in practice:

1. How to treat inconsistent data sets?

2. How to deal with more than two data sets?

3. Computing conditional probabilities of the form $P^\dagger(\varphi|\psi)$ is a computationally hard problem, even if $P^\dagger$ is known.

In this talk I shall address these difficulties.

Inconsistent Data Sets

Denote by N_i the number of observations in DS_i. Objective Bayesians want to match degrees of belief to observed frequencies[1], and thus it ought to hold that

$$P^\dagger(\omega_x) = \frac{N_1}{N_1+N_2} P_1^*(\omega_x) + \frac{N_2}{N_1+N_2} P_2^*(\omega_x) \quad . \tag{1}$$

The conditional observed frequencies of the form $P_i^*(\omega_{y^i}|\omega_x)$ do not depend on whether DS_1 and DS_2 are consistent. Hence, $P^\dagger(\omega_{y^i}|\omega_x) = P_i^*(\omega_{y^i}|\omega_x)$ holds.

We can now work out the constraints which determine $\mathbb{E}$ for the case of two inconsistent data sets and obtain after a short while (if at least one instance of ω_x has been observed)

$$P^\dagger(\omega_x \wedge \omega_1 \wedge \omega_2) = P_1^*(\omega_1|\omega_x) \cdot P_2^*(\omega_2|\omega_x) \cdot P^\dagger(\omega_x) \quad .$$

N nice data sets

The above arguments can be extended to $N \geq 3$ data sets, if these data sets are *nice*. A collection of data sets is *nice*, iff there exists one data set DS_I which contains all variables which are used in two or more other data sets. That is, $L_i \cap L_k \subseteq L_I$ for all $1 \leq i < k \leq N$. The variables contained in more than one data set are again denoted by $\mathbf{x}$. The variables used in DS_i which are unique to DS_i are the $\mathbf{y}^i$.

The entropy maximiser for a collection N nice data sets, which may be inconsistent, is

$$P^\dagger(\omega_x \wedge \bigwedge_{i=1}^{N} \omega_i) = P^\dagger(\omega_x) \cdot \prod_{i=1}^{N} P_i^*(\omega_i|\omega_x) \quad .$$

This holds for the same reason as in the case of two data sets: the variables in $\mathbf{x}$ screen off the variables in $\mathbf{y}^i$ from $\mathbf{y}^k$ for $i < k$; see [2] for details. $P^\dagger(\omega_x)$ can be computed from the P_i^* in a similar way as in (1). Again, we can now simply read off the entropy maximiser $P^\dagger$ from the observed P_i^*.

Inference

Time permitting, I will talk about the following inference problem. The analytical solution of $P^\dagger$ presented above is of little use for practical applications in which one wants to calculate conditional probabilities. If the number of variables used in the data sets is in the thousands, then the number of states is north of 2^{1000}. Thus, adding the probabilities of all states which satisfy certain conditions is computationally infeasible.

Fortunately, there exist efficient algorithms which we can employ to learn the P_i^* in a Bayesian network representation, [1]. I will show how we can then represent $P^\dagger$ as a Bayesian network, if the collection of data sets is nice. Collections of data sets which are not nice are currently under investigation.

References

[1] Ioannis Tsamardinos, LauraE. Brown, and ConstantinF. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

[2] Jon Williamson. *Bayesian Nets and Causality.* Oxford University Press, 2005.

[3] Jon Williamson. *In Defence of Objective Bayesianism.* Oxford University Press, 2010.

[1] at least if the number of observations is sufficiently large and there is no other evidence