Composite likelihood: a family of proper scoring rules for probabilistic clustering

Niko Brümmer

July 2018

1 Introduction

1.1 Notation for partitions

Let $\mathcal{R}_{[n]} = \{R_1, \dots, R_n\}$ denote a set of n elements. Let $[n] = \{1, \dots, n\}$ denote the set of integer indices that identify the members of $\mathcal{R}_{[n]}$. Let $\mathcal{P}_n$ denote the set of all possible ways to partition $[n]$. If $L \in \mathcal{P}_n$, we say that L is a *partition* of $[n]$ and by association also of $\mathcal{R}_{[n]}$. A partition, L, can be specified as a collection of one or more mutually exclusive and exhaustive subsets of $[n]$. We shall refer to the subsets as *clusters*. The cardinality of $\mathcal{P}_n$, is the *Bell number*, B_n.

Let $\mathbb{P}_n$ denote the probability simplex, with B_n vertices, in which probability distributions over $\mathcal{P}_n$ live.

1.2 Notation for probability distributions

We shall concern ourselves with probability distributions over sets of partitions, of the form:

$$P(L \mid \mathcal{P}) \equiv P(L \mid L \in \mathcal{P}), \tag{1}$$

where the LHS is a more compact form for the RHS and where $\mathcal{P} \subseteq \mathcal{P}_n$. In cases where it is implicit that we condition on the full set, $\mathcal{P}_n$, we shall write simply $P(L)$.

We shall refer to $P(L \mid \mathcal{P})$ as *prior* probability distributions, while distributions conditioned also on (subsets of) $\mathcal{R}_{[n]}$, will be referred to as *posteriors*. Such prior distributions can be defined for example by a Chinese restaurant process (CRP) and as such would be conditioned on one or two parameters.

If those parameters are not of immediate interest, we don't show them in the notation.

We are also very interested in *partition likelihoods*, of the form:

$$\mathcal{L}(L \mid \mathcal{R}, \theta) \propto P(\mathcal{R} \mid L, \theta) \tag{2}$$

where $\mathcal{R} \subseteq \mathcal{R}_{[n]}$ and where θ is a set of parameters for a model that computes the likelihoods. Our interest will be in formulating tractable criteria for the evaluation of the goodness of the likelihood model and its parametrization. Such criteria can be used as optimization objective in training θ and also to compare the goodness of different likelihood models.

Given priors and likelihoods, *posteriors* are of course defined via Bayes' rule as:

$$P(L \mid \mathcal{P}, \mathcal{R}, \theta) = \frac{P(L \mid \mathcal{P})\mathcal{L}(L \mid \mathcal{R}, \theta)}{\sum_{L' \in \mathcal{P}} P(L' \mid \mathcal{P})\mathcal{L}(L' \mid \mathcal{R}, \theta)} \tag{3}$$

We shall assume that the posterior can be computed if the cardinality of $\mathcal{P}$ is small enough. In cases where $\mathcal{P} = \mathcal{P}_n$, the cardinality is B_n, which is intractable, except for very small n.