Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Journal of Computational and Applied Mathematics 175 (2005) 231–243

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

www.elsevier.com/locate/cam

Bayesian analysis of binary sequences

David C. Torney*

Los Alamos National Laboratory, T-10, MS K710, Los Alamos, NM 87545, USA

Received 28 January 2003; received in revised form 18 April 2004

Dedicated to Professor Micah Dembo, Boston University, as it traces its origins to his clockwork exposition of Bayesian theory

Abstract

This manuscript details Bayesian methodology for "learning by example", with binary n-sequences encoding the objects under consideration. Priors prove influential; conformable priors are described. Laplace approximation of Bayes integrals yields posterior likelihoods for all n-sequences. This involves the optimization of a definite function over a convex domain—efficiently effectuated by the sequential application of the quadratic program.

MSC: primary: 62F15; secondary: 60C05; 65K05

Keywords: Concave; Convex; Cut polytope; Geometric probability; Laplace approximation; Machine learning; Moments; Nonlinear optimization; Polytope; Posterior likelihoods; Probability monomials; Quadratic program; Semidefinite

1. Introduction

This manuscript self-containedly describes new methodology for the well-rooted desideratum of "learning by example". To precise this aim, let binary n-sequences represent the elements of the universe of possible examples. Then, the desideratum may be paraphrased as "characterizing a distribution on binary n-sequences, based upon a sample from this distribution" because, for instance, the *sine qua non* of prediction is making intelligent use of the examples for the classification of supplementary elements. This aim compasses many applications, e.g., the screening of digital images or of biological sequences—to find elements resembling those from a data set.

* Tel.: +1-505-667-7510; fax: +1-505-665-3493.
E-mail address: dtorney@earthlink.net (D.C. Torney).

0377-0427/$ - see front matter
doi:10.1016/j.cam.2004.05.010

Bayesian methods constitute a peerless framework for data-based inference [4]. They engender a posterior distribution (what is inferred, based on the examples) from a prior distribution (what is assumed, at the outset, before the examination of the examples). As the desideratum is a distribution, the respective posterior distribution is a distribution on distributions – accentuating those distributions in greatest accordance with the examples, as described in Section 2. The reader who seeks an introduction to Bayesian statistics cannot do better than to consult [4], and, alternatively, the uninitiated may take its present instantiation (cf. (2)) as a postulate. The burgeoning of this application of Bayes' rule bided the impediment of a decided, unwelcome dependence of the posterior distribution upon the prior distribution. As will be seen in Section 2, a uniform (unprejudiced) prior is unavailing, and the new methodology comprises the detailed specification of priors. Insights sufficing for the selection of nonuniform priors may originate, for instance, from sample estimates of key parameters of probability distributions.

Parameterizations for distributions on binary n-sequences usually base their parameters on marginal distributions for subsets of digits [2,5,17,24], and sampling of these distributions is feasible using Markov chain Monte Carlo methods. The moment parameterization, reviewed in Section 3, is used to engineer the new methodology. Prior distributions are, herein, taken to be uniform over all distributions having specified vanishing moments (although fixing moments at nonzero values is also accommodated in the present framework). Section 4 introduces this prior, gives consideration to distributions exhibiting vanishing moments and digresses upon dialectics. Applications corroborate the felicitousness of this prior (viz. Section 11). The linearity of the moment parameterization is its greatest boon; whence, for example, each fixed-moment prior educes a convex polytope: an intersection of half spaces [11] constitutes the domain of admissible collections of (nonfixed) moment parameters. In detail, there is a half-space pertaining to each n-sequence, ensuring the nonnegativity of the respective linear combination of the moments expressing its probability.

As described in Section 4, according to Bayes' rule, given a prior uniform over admissible distributions, the sought posterior distributions result from integrals of a *probability monomial* over a respective polytope, with this monomial denoting the product of positive powers of sequences' probabilities, with the powers being the respective sequence multiplicities in the collection of examples. Probability monomials are multilinear functions of the (nonfixed) moments.

Explicit restriction to these polytopes, in theory, quells the "moment problem". However, as illustrated in Section 5, perplexity may persist through their intricacy, which may retard the implementation of this restriction. Section 6 is the starting point for the derivation of the main results, which are based upon a more detailed formulation given therein.

The logarithm of probability monomials is shown to be negative semidefinite in Section 7. Orthogonal projections may be employed, as necessary, to educe definiteness and, hence, a function whose unique local optimum is also its global maximum, as described in Section 8. The towering of this maximum, for samples of respectable size, triggers Laplace approximations for the Bayes integrals, described in Section 9. The integrand's maximum point parameterizes a distribution comprising the *posterior likelihoods*: the expected posterior probabilities (viz. Section 2). As a consequence of definiteness, convergence thither is efficiently effected by maximizing each of a sequence of quadratic approximations to the logarithm of a probability monomial, the latter by means of a quadratic program comprising the aforementioned half-space constraints—whose number is typically exponential in n; as detailed in Section 10 [19,22]. Implementations of the quadratic program exhibit low-degree-polynomial complexity in the number of optimized parameters.