Bayesian ML: what, why, and how?

Part #1: what is Bayesian ML?

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France

Make sure you’re in the right class

These are more the applications we have in mind

PRL **119,** 141101 (2017) PHYSICAL REVIEW LETTERS week ending 6 OCTOBER 2017

GW170814: A Three-Detector Observation of Gravitational Waves from a Binary Black Hole Coalescence

B. P. Abbott *et al.*[*]

(LIGO Scientific Collaboration and Virgo Collaboration)

(Received 23 September 2017; published 6 October 2017)

On August 14, 2017 at 10:30:43 UTC, the Advanced Virgo detector and the two Advanced LIGO detectors coherently observed a transient gravitational-wave signal produced by the coalescence of two stellar mass black holes, with a false-alarm rate of $\lesssim 1$ in 27 000 years. The signal was observed with a three-detector network matched-filter signal-to-noise ratio of 18. The inferred masses of the initial black holes are $30.5^{+5.7}_{-3.0} M_\odot$ and $25.3^{+2.8}_{-4.2} M_\odot$ (at the 90% credible level). The luminosity distance of the source is 540^{+130}_{-210} Mpc, corresponding to a redshift of $z = 0.11^{+0.03}_{-0.04}$. A network of three detectors improves the sky localization of the source, reducing the area of the 90% credible region from 1160 deg^2 using only the two LIGO detectors to 60 deg^2 using all three detectors. For the first time, we can test the nature of gravitational-wave polarizations from the antenna response of the LIGO-Virgo network, thus enabling a new class of phenomenological tests of gravity.

DOI: 10.1103/PhysRevLett.119.141101

I. INTRODUCTION

The era of gravitational-wave (GW) astronomy began with the detection of binary black hole (BBH) mergers, by the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) detectors [1] during the first of the

waveform obtained from analysis of the LIGO detectors' data alone, we find that the probability, in 5000 s of data around the event, of a peak in SNR from Virgo data due to noise and as large as the one observed, within a time window determined by the maximum possible time of

These are more the applications we have in mind

just GWs but also broadband electromagnetic emission. LIGO and Virgo have been distributing low-latency alerts and localizations of GW events to a consortium now consisting of ground- and space-based facilities who are searching for gamma-ray, x-ray, optical, near-infrared, radio, and neutrino counterparts [57–59].

For the purpose of position reconstruction, the LIGO-Virgo GW detector network can be thought of as a phased array of antennas. Any single detector provides only minimal position information, its slowly varying antenna

due to the noise removal and final detector calibration, described in the previous section, that was applied for the full parameter estimation but not the rapid localization.

Incorporating Virgo data also reduces the luminosity distance uncertainty from 570^{+300}_{-230} Mpc (rapid localization) to 540^{+130}_{-210} Mpc (full parameter estimation). As with the previous paragraph, the three-dimensional credible volume and number of possible host galaxies also decreases by an order of magnitude [67–69], from 71×10^6 Mpc^3, to 3.4×10^6 Mpc^3, to 2.1×10^6 Mpc^3.

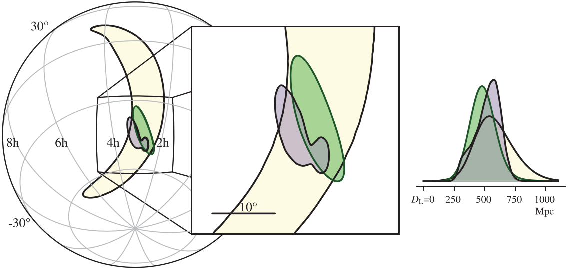


FIG. 3. Localization of GW170814. The rapid localization using data from the two LIGO sites is shown in yellow, with the inclusion of data from Virgo shown in green. The full Bayesian localization is shown in purple. The contours represent the 90% credible regions. The left panel is an orthographic projection and the inset in the center is a gnomonic projection; both are in equatorial coordinates. The inset on the right shows the posterior probability distribution for the luminosity distance, marginalized over the whole sky.

or that one

Journal of Machine Learning Research 14 (2013) 1303-1347 Submitted 6/12; Published 5/13

Stochastic Variational Inference

Matthew D. Hoffman MATHOFFM@ADOBE.COM
Adobe Research
Adobe Systems Incorporated
601 Townsend Street
San Francisco, CA 94103, USA

David M. Blei BLEI@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
35 Olden Street
Princeton, NJ 08540, USA

Chong Wang CHONGW@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Gates Hillman Centers, 8110
5000 Forbes Avenue
Pittsburgh, PA 15213, USA

John Paisley JPAISLEY@BERKELEY.EDU
Computer Science Division

Editor: Tommi Jaakkola

Abstract

We develop stochastic variational inference, a scalable algorithm for approximating posterior distributions. We develop this technique for a large class of probabilistic models and we demonstrate it with two probabilistic topic models, latent Dirichlet allocation and the hierarchical Dirichlet process topic model. Using stochastic variational inference, we analyze several large collections of documents: 300K articles from *Nature*, 1.8M articles from *The New York Times*, and 3.8M articles from *Wikipedia*. Stochastic inference can easily handle data sets of this size and outperforms traditional variational inference, which can only handle a smaller subset. (We also show that the Bayesian nonparametric topic model outperforms its parametric counterpart.) Stochastic variational inference lets us apply complex Bayesian models to massive data sets.

Keywords: Bayesian inference, variational inference, stochastic optimization, topic models, Bayesian nonparametrics

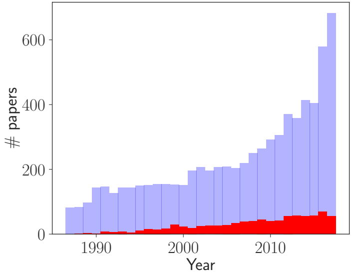
1. Introduction

Modern data analysis requires computation with massive data. As examples, consider the following. (1) We have an archive of the raw text of two million books, scanned and stored online. We want to discover the themes in the texts, organize the books by subject, and build a navigator for users

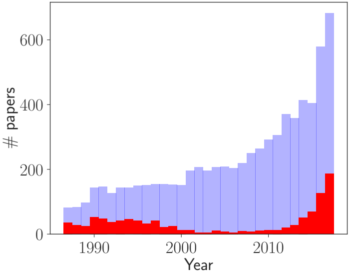
Outline

Outline

Bayesian keywords in NeurIPS abstracts, up to 2016



(a) “Bayesian” at NeurIPS



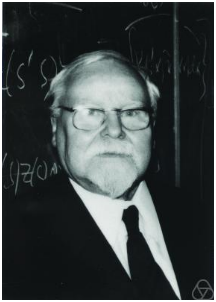
(b) “Neural net” at NeurIPS

Topics automatically extracted from 1000+ “Bayesian” abstracts

model models data process latent Bayesian Dirichlet hierarchical nonparametric inference
features learn problem different knowledge learning image object example examples
method neural Bayesian using linear state based kernel approach model
belief propagation nodes local tree posterior node nbsp given algorithm
learning data Bayesian model training classification performance selection prediction sets
inference Monte Carlo Markov sampling variational time algorithm MCMC approximate
function optimization algorithm optimal learning problem gradient methods bounds state
learning networks variables structure network Bayesian EM paper distribution algorithm
Bayesian gaussian prior regression non estimation likelihood sparse parameters matrix
model information Bayesian human visual task probability sensory prior concept

Figure: Topics extracted by stochastic variational latent Dirichlet allocation, using scikit-learn ().

Inference in regression models



Inference in regression models

Inference in regression models

Describing a decision problem under uncertainty (Wal50)

- A state space $\mathcal{S}$,
  Every quantity you need to consider to make your decision.
- Actions $\mathcal{A} \subset \mathcal{F}(\mathcal{S}, \mathcal{Z})$,
  Making a decision means picking one of the available actions.
- A reward space $\mathcal{Z}$,
  Encodes how you feel about having picked a particular action.
- A loss function $L : \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$.
  How much you would suffer from picking action a in state s. It is also customary to first define a utility $u : \mathcal{Z} \to \mathbb{R}_+$, and then let

$$L(a, s) = \sup_{a' \in \mathcal{A}} u\big(a'(s)\big) - u\big(a(s)\big) \in \mathbb{R}_+.$$

Estimation as a decision problem

- $\mathcal{S} =$
- $\mathcal{Z} =$
- $\mathcal{A} =$
-

Classification as a decision problem

- $\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \times \mathcal{Y}$, i.e. $s = (x_{1:n}, y_{1:n}, x, y)$.
- $\mathcal{Z} = \{0, 1\}$.
- $\mathcal{A} = \{a_g : s \mapsto 1_{y \neq g(x; x_{1:n}, y_{1:n})}, \quad g \in \mathcal{G}\}$.
- $L(a_g, s) = 1_{y \neq g(x; x_{1:n}, y_{1:n})}$.

PAC bounds; see e.g. (ShBe14)

Let $(x_{1:n}, y_{1:n}) \sim \mathbb{P}^{\otimes n}$, and independently $(x, y) \sim \mathbb{P}$, we want an algorithm $g(\cdot; x_{1:n}, y_{1:n}) \in \mathcal{G}$ such that if $n \geqslant n(\delta, \varepsilon)$,

$$\mathbb{P}^{\otimes n}\left[\mathbb{E}_{(x,y)\sim\mathbb{P}} L(a_g, s) \leqslant \varepsilon\right] \geqslant 1 - \delta.$$

Regression as a decision problem

- $\mathcal{S} =$
- $\mathcal{Z} =$
- $\mathcal{A} =$
-

Model choice as a decision problem

- $\mathcal{S} =$
- $\mathcal{Z} =$
- $\mathcal{A} =$
-

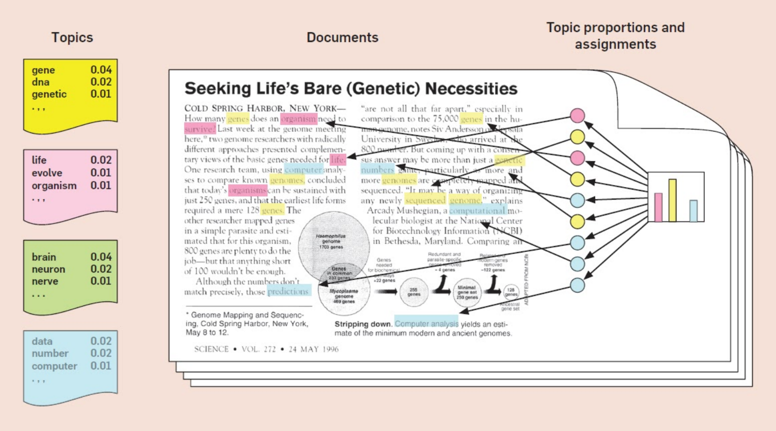


Figure: Topic modeling. *Credits to D. Blei?*

Topic modeling as a decision problem

- $\mathcal{S} =$
- $\mathcal{Z} =$
- $\mathcal{A} =$
-

SEU is what defines the Bayesian approach

The subjective expected utility principle

1. Choose $\mathcal{S}, \mathcal{Z}, \mathcal{A}$ and a loss function $L(a, s)$,
2. Choose a distribution p over $\mathcal{S}$,
3. Take the the corresponding Bayes action

$$a^\star \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s \sim p} L(a, s). \tag{1}$$

Corollary: minimize the posterior expected loss

If we partition $s = (s_o, s_\mathrm{u})$, then

$$a^\star \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s_o} \mathbb{E}_{s_\mathrm{u}|s_o} L(a, s).$$

Equivalently to (1), given s_o, we choose

$$a^\star = \delta(s_o) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s_\mathrm{u}|s_o} L(a, s).$$

A recap on probabilistic graphical models

- PGMs (aka "Bayesian" networks) represent the dependencies in a joint distribution $p(y)$ by a directed graph $G = (E, V)$.
- Two important properties:

$$p(y) = \prod_{v \in V} p(y|y_{\mathrm{pa(v)}}) \qquad \text{and} \qquad y_v \perp y_{nd(v)}|y_{pa(v)}.$$

- Also good to know how to determine whether $A \perp B|C$; see (**Mur12**).

Estimation as a decision problem: point estimates

- $\mathcal{S} = \mathcal{Y}^n \times \Theta$.
- $\mathcal{Z} = \Theta$.
- $\mathcal{A} = \{a_g : s \mapsto \theta - g(y_{1:n})\}$.
- $L(a_g, s) = \|\theta - g(y_{1:n})\|^2$.

Estimation as a decision problem: credible intervals

- $\mathcal{S} = \mathcal{Y}^n \times \Theta$.
- $\mathcal{Z} = \Theta$.
- $\mathcal{A} = \{a_g : s \mapsto (1_{\theta \in g(y_{1:n})}, |g(y_{1:n})|)\}$.
- $L(a_g, s) = 1_{\theta \in g(y_{1:n})} + \gamma |g(y_{1:n})|$.

Topic modelling as a decision problem

- $\mathcal{S} =$.
- $\mathcal{Z} =$.
- $\mathcal{A} =$.
- $L(a_g, s) =$.

Outline

Classification as a decision problem

- $\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \times \mathcal{Y}$, i.e. $s = (x_{1:n}, y_{1:n}, x, y)$.
- $\mathcal{Z} = \{0, 1\}$.
- $\mathcal{A} = \{a_g : s \mapsto 1_{y \neq g(x; x_{1:n}, y_{1:n})}\}$.
- $L(a_g, s) = 1_{y \neq g(x; x_{1:n}, y_{1:n})}$.

Classification as a decision problem

- $\mathcal{S} = \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \times \mathcal{Y}$, i.e. $s = (x_{1:n}, y_{1:n}, x, y)$.
- $\mathcal{Z} = \{0, 1\}$.
- $\mathcal{A} = \{a_g : s \mapsto 1_{y \neq g(x; x_{1:n}, y_{1:n})}\}$.
- $L(a_g, s) = \alpha 1_{y \neq g(x)} 1_{y=0} + \beta 1_{y \neq g(x)} 1_{y=1}$.

Prediction in regression as a decision problem

- $\mathcal{S} = \mathcal{X}^n \times \mathbb{R}^n \times \mathcal{X} \times \mathbb{R}$, i.e. $s = (x_{1:n}, y_{1:n}, x, y)$.
- $\mathcal{Z} = \mathbb{R}$.
- $\mathcal{A} = \{a_g : s \mapsto y - g(x; x_{1:n}, y_{1:n})\}$.
- $L(a_g, s) = \|y - g(x; x_{1:n}, y_{1:n})\|^2$.

Wrapping up

A Bayesian minimizes a posterior expected loss

$$a^\star = \delta(s_o) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{s_u|s_o} L(a, s).$$

- SEU allows to formalize most ML questions.
- Choosing L and π is often relatively natural.

Good's *46656 varieties of Bayesians*

- Bayesian subschools differ on how they justify, interpret, and implement that principle.
- Different interpretations lead to different degrees of freedom for the joint model.

References I