Bayesian nonparametrics in machine learning

Rémi Bardenet [1]

[1]CNRS & CRIStAL, Univ. Lille, France

Outline

Outline

Outline

Supervised learning with Gaussian Processes

Density estimation and clustering with Dirichlet processes

Outline

Supervised learning with Gaussian Processes

- From linear regression to GPs [2, Chapter 2]
- Basic theory: Existence and conditioning for GPs [1, Chapter 4]
- Modeling with GPs [2, Chapter 4]
- Inferring GPs [2, Chapter 5]
- Some more applications
- References and open issues

Density estimation and clustering with Dirichlet processes

- From finite mixture models to DPs [1, Chapter 2]
- Basic theory: Existence and conditioning for DPs [1, Chapter 2]
- Modeling with DPs [1, Chapter 2]
- Fitting a DP mixture model [1, Chapter 2]
- Some applications
- References and open issues

Commonly-used kernels

covariance function	expression	S	ND
constant	σ_0^2	$\surd$	
linear	$\sum_{d=1}^{D} \sigma_d^2 x_d x'_d$		
polynomial	$(\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p$		
squared exponential	$\exp(-\frac{r^2}{2\ell^2})$	$\surd$	$\surd$
Matérn	$\frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} r\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} r\right)$	$\surd$	$\surd$
exponential	$\exp(-\frac{r}{\ell})$	$\surd$	$\surd$
γ-exponential	$\exp\left(-(\frac{r}{\ell})^\gamma\right)$	$\surd$	$\surd$
rational quadratic	$(1 + \frac{r^2}{2\alpha\ell^2})^{-\alpha}$	$\surd$	$\surd$
neural network	$\sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1+2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}})(1+2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}')}}\right)$		$\surd$

Table 4.1: Summary of several commonly-used covariance functions. The covariances are written either as a function of $\mathbf{x}$ and $\mathbf{x}'$, or as a function of $r = |\mathbf{x} - \mathbf{x}'|$. Two columns marked 'S' and 'ND' indicate whether the covariance functions are stationary and nondegenerate respectively. Degenerate covariance functions have finite rank, see section 4.3 for more discussion of this issue.

Outline

Supervised learning with Gaussian Processes

Density estimation and clustering with Dirichlet processes

Outline

Emulators of expensive models

PLOS | ONE

RESEARCH ARTICLE

Bayesian Sensitivity Analysis of a Cardiac Cell Model Using a Gaussian Process Emulator

Eugene TY Chang[1,2], **Mark Strong**[3], **Richard H Clayton**[1,2]*

1 Insigneo Institute for *in-silico* Medicine, University of Sheffield, Sheffield, United Kingdom, **2** Department of Computer Science University of Sheffield, Sheffield, United Kingdom, **3** School of Health and Related Research, University of Sheffield, Sheffield, United Kingdom

* r.h.clayton@sheffield.ac.uk

CrossMark
click for updates

OPEN ACCESS

Citation: Chang ETY, Strong M, Clayton RH (2015)
Bayesian Sensitivity Analysis of a Cardiac Cell Model

Abstract

Models of electrical activity in cardiac cells have become important research tools as they can provide a quantitative description of detailed and integrative physiology. However, cardiac cell models have many parameters, and how uncertainties in these parameters affect the model output is difficult to assess without undertaking large numbers of model runs. In this study we show that a surrogate statistical model of a cardiac cell model (the Luo-Rudy 1991 model) can be built using Gaussian process (GP) emulators. Using this approach we

Nonparametric fits

Gaussian Process Cosmography

Arman Shafieloo[1], Alex G. Kim[2], Eric V. Linder[1,2,3]
[1] *Institute for the Early Universe WCU, Ewha Womans University, Seoul, Korea*
[2] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and*
[3] *University of California, Berkeley, CA 94720, USA*

(Dated: July 11, 2012)

Gaussian processes provide a method for extracting cosmological information from observations without assuming a cosmological model. We carry out cosmography – mapping the time evolution of the cosmic expansion – in a model-independent manner using kinematic variables and a geometric probe of cosmology. Using the state of the art supernova distance data from the Union2.1 compilation, we constrain, without any assumptions about dark energy parametrization or matter density, the Hubble parameter and deceleration parameter as a function of redshift. Extraction of these relations is tested successfully against models with features on various coherence scales, subject to certain statistical cautions.

I. INTRODUCTION

Cosmic acceleration is a fundamental mystery of great interest and importance to understanding cosmology, gravitation, and high energy physics. The cosmic expansion rate is slowed down by gravitationally attractive matter and sped up by some other, unknown contribution to the dynamical equations. While great effort is being put into identifying the source of this extra dark energy contribution, the overall expansion behavior also holds important clues to origin, evolution, and present

ing procedures have been suggested, e.g. [6], but tend to induce bias in the function reconstruction due to parametric restriction of the behavior or to have poor error control. Using a general orthonormal basis or principal component analysis is another approach, to describe the distance-redshift relation (e.g. [7]) or the deceleration parameter [8], or using a correlated prior for smoothness on the dark energy equation of state [9], but in practice a finite (and small) number of modes is significant beyond the prior, essentially reducing to a parametric approach. Gaussian processes [10] offer an interesting possibility for

Natural language processing

Using Gaussian Processes for Rumour Stance Classification in Social Media

MICHAL LUKASIK, University of Sheffield
KALINA BONTCHEVA, University of Sheffield
TREVOR COHN, University of Melbourne
ARKAITZ ZUBIAGA, University of Warwick
MARIA LIAKATA, University of Warwick
ROB PROCTER, University of Warwick

Social media tend to be rife with rumours while new reports are released piecemeal during breaking news. Interestingly, one can mine multiple reactions expressed by social media users in those situations, exploring their stance towards rumours, ultimately enabling the flagging of highly disputed rumours as being potentially false. In this work, we set out to develop an automated, supervised classifier that uses multi-task learning to classify the stance expressed in each individual tweet in a rumourous conversation as either supporting, denying or questioning the rumour. Using a classifier based on Gaussian Processes, and exploring its effectiveness on two datasets with very different characteristics and varying distributions of stances, we show that our approach consistently outperforms competitive baseline classifiers. Our classifier is especially effective in estimating the distribution of different types of stance associated with a given rumour, which we set forth as a desired characteristic for a rumour-tracking system that will warn both ordinary users of Twitter and professional news practitioners when a rumour is being rebutted.

1. INTRODUCTION

There is an increasing need to interpret and act upon rumours spreading quickly through social media during breaking news, where new reports are released piecemeal and often have an unverified

Bayesian optimization for hyperparameter tuning

Algorithms for Hyper-Parameter Optimization

James Bergstra
The Rowland Institute
Harvard University
bergstra@rowland.harvard.edu

Rémi Bardenet
Laboratoire de Recherche en Informatique
Université Paris-Sud
bardenet@lri.fr

Yoshua Bengio
Dépt. d'Informatique et Recherche Opérationelle
Université de Montréal
yoshua.bengio@umontreal.ca

Balázs Kégl
Linear Accelerator Laboratory
Université Paris-Sud, CNRS
balazs.kegl@gmail.com

Abstract

Several recent advances to the state of the art in image classification benchmarks have come from better configurations of existing techniques rather than novel approaches to feature learning. Traditionally, hyper-parameter optimization has been the job of humans because they can be very efficient in regimes where only a few

Outline

References

- Textbook by C. Rasmussen and C. Williams [2],
  - great for understanding, methods, pointers to ML and stats.
- Similar videolecture by C. Rasmussen.
- lecture notes by P. Orbanz [1].
  - mathematically clean, without losing the focus on ML.

Some open issues

- Fully Bayesian scalable approaches!
- Natural approaches to constrained GPs.
- Links with other models based on Gaussians and geometry.

Outline

The Dirichlet distribution

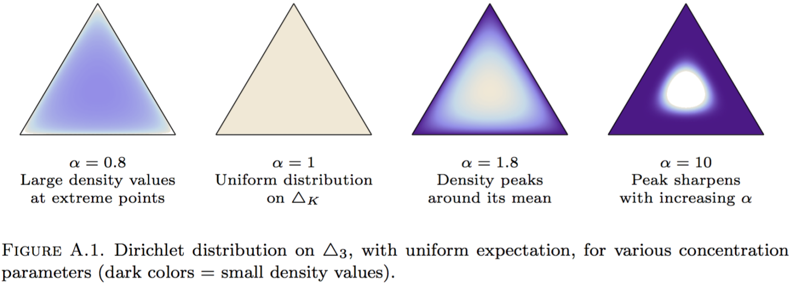


Figure: Taken from [1, Appendix 4]

Outline

Outline

Outline

Outline

Nonparametric latent Dirichlet allocation (HDP-LDA)

Hierarchical Dirichlet Processes

Yee Whye Teh tehyw@comp.nus.edu.sg
Department of Computer Science, National University of Singapore,
Singapore 117543

Michael I. Jordan jordan@eecs.berkeley.edu
Computer Science Division and Department of Statistics,
University of California at Berkeley, Berkeley CA 94720-1776, USA

Matthew J. Beal mbeal@cse.buffalo.edu
Department of Computer Science & Engineering,
State University of New York at Buffalo, Buffalo NY 14260-2000, USA

David M. Blei blei@eecs.berkeley.edu
Department of Computer Science, Princeton University,
Princeton, NJ 08544, USA

November 15, 2005

Nonparametric latent Dirichlet allocation (HDP-LDA)

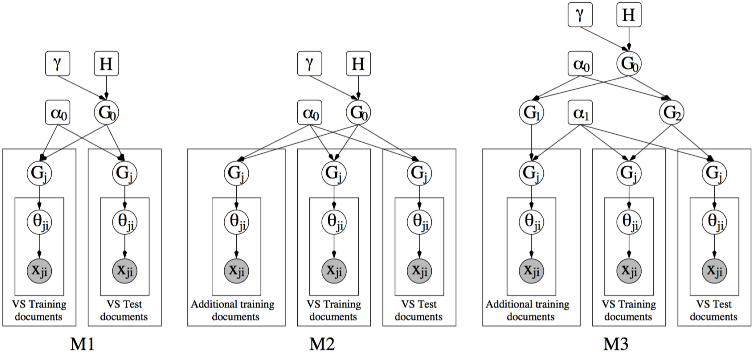


Figure 4: Three models for the NIPS data. From left to right: M1, M2 and M3.

Figure: Taken from [3]

Nonparametric latent Dirichlet allocation (HDP-LDA)

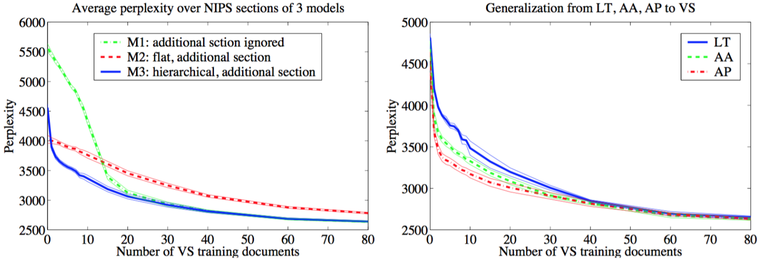


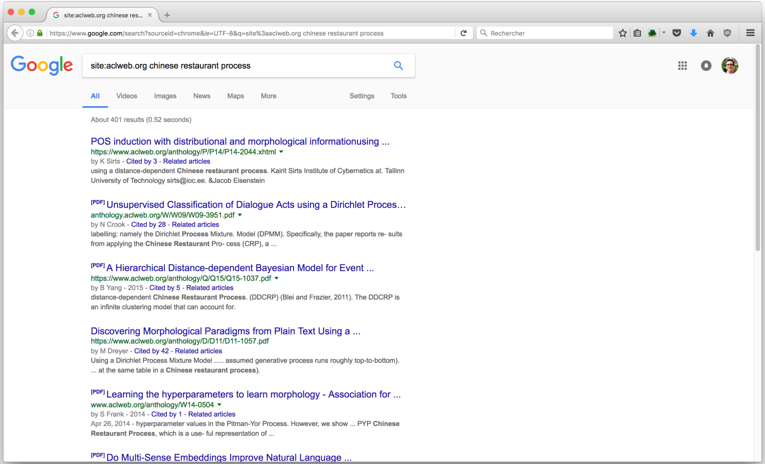
Figure: Taken from [3]

Nonparametric latent Dirichlet allocation (HDP-LDA)

CS	task representation pattern processing trained representations three process unit patterns examples concept similarity bayesian hypotheses generalization numbers positive classes hypothesis
NS	cells cell activity response neuron visual patterns pattern single fig visual cells cortical orientation receptive contrast spatial cortex stimulus tuning
LT	signal layer gaussian cells fig nonlinearity nonlinear rate eq cell large examples form point see parameter consider random small optimal
AA	algorithms test approach methods based point problems form large paper distance tangent image images transformation transformations pattern vectors convolution simard
IM	processing pattern approach architecture single shows simple based large control motion visual velocity flow target chip eye smooth direction optical
SP	visual images video language image pixel acoustic delta lowpass flow signals separation signal sources source matrix blind mixing gradient eq
AP	approach based trained test layer features table classification rate paper image images face similarity pixel visual database matching facial examples
CN	ii tree pomdp observable strategy class stochastic history strategies density policy optimal reinforcement control action states actions step problems goal

Figure: Taken from [3]

DP and CRP are also popular in Natural language processing (NLP)



Outline

References

- My primary source is P Orbanz' 2014 BNP lecture notes again, with a great overview of relevant theory.
- Chapters 1, 2, 5, 6 of the 2010 book edited by N. Hjort et al.
  - This book is a collection of dense independent chapters and not really a textbook, more of a state-of-the-art with a lot of pointers. There is ample material for those who want to read further on BNP.
- There is an upcoming book by YW Teh and M Jordan that should become the reference ML textbook that [2] is to GPs, keep an eye open!
  - Meanwhile, check out recent videolectures by the authors YW Teh and M Jordan.

Some open issues

- Lots of potential applications in other fields such as signal processing.
- Fast variational vs slow “exact” approaches.
- Characterize models for which efficient Gibbs is possible.
- Discover/build new models and name them after your favourite food source.

References I

[1] P. Orbanz.
Lecture notes on Bayesian nonparametrics, 2014.

[2] C. E. Rasmussen and C. K. I. Williams.
Gaussian Processes for Machine Learning.
MIT Press, 2006.

[3] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei.
Hierarchical dirichlet processes.
In *Advances in Neural Information Processing Systems (NIPS)*.