Gaussian Processes for Natural Language Processing

Trevor Cohn
Computing and Information Systems
The University of Melbourne
trevor.cohn@gmail.com

Daniel Preoţiuc-Pietro and **Neil Lawrence**
Department of Computer Science
The University of Sheffield
{daniel,n.lawrence}@dcs.shef.ac.uk

1 Introduction

Gaussian Processes (GPs) are a powerful modelling framework incorporating kernels and Bayesian inference, and are recognised as state-of-the-art for many machine learning tasks. Despite this, GPs have seen few applications in natural language processing (notwithstanding several recent papers by the authors). We argue that the GP framework offers many benefits over commonly used machine learning frameworks, such as linear models (logistic regression, least squares regression) and support vector machines. Moreover, GPs are extremely flexible and can be incorporated into larger graphical models, forming an important additional tool for probabilistic inference. Notably, GPs are one of the few models which support analytic Bayesian inference, avoiding the many approximation errors that plague approximate inference techniques in common use for Bayesian models (e.g. MCMC, variational Bayes).[1] GPs accurately model not just the underlying task, but also the uncertainty in the predictions, such that uncertainty can be propagated through pipelines of probabilistic components. Overall, GPs provide an elegant, flexible and simple means of probabilistic inference and are well overdue for consideration of the NLP community.

This tutorial will focus primarily on regression and classification, both fundamental techniques of wide-spread use in the NLP community. Within NLP, linear models are near ubiquitous, because they provide good results for many tasks, support efficient inference (including dynamic programming in structured prediction) and support simple parameter interpretation. However, linear models are inherently limited in the types of relationships between variables they can model. Often non-linear methods are required for better understanding and improved performance. Currently, kernel methods such as Support Vector Machines (SVM) represent a popular choice for non-linear modelling. These suffer from lack of interoperability with down-stream processing as part of a larger model, and inflexibility in terms of parameterisation and associated high cost of hyperparameter optimisation. GPs appear similar to SVMs, in that they incorporate kernels, however their probabilistic formulation allows for much wider applicability in larger graphical models. Moreover, several properties of Gaussian distributions (closure under integration and Gaussian-Gaussian conjugacy) means that GP (regression) supports analytic formulations for the posterior and predictive inference.

This tutorial will cover the basic motivation, ideas and theory of Gaussian Processes and several applications to natural language processing tasks. GPs have been actively researched since the early 2000s, and are now reaching maturity: the fundamental theory and practice is well understood, and now research is focused into their applications, and improve inference algorithms, e.g., for scaling inference to large and high-dimensional datasets. Several open-source packages (e.g. GPy and GPML) have been developed which allow for GPs to be easily used for many applications. This tutorial aims to promote GPs, emphasising their potential for widespread application across many NLP tasks.

2 Overview

Our goal is to present the main ideas and theory behind Gaussian Processes in order to increase awareness within the NLP community. The first part of the tutorial will focus on the basics of Gaussian Processes in the context of regression. The Gaussian Process defines a prior over functions which applied at each input point gives a response

[1] This holds for GP regression, but note that approximate inference is needed for non-Gaussian likelihoods.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Tutorials, pages 1–3,
Baltimore, Maryland, USA, 22 June 2014.

value. Given data, we can analytically infer the posterior distribution of these functions assuming Gaussian noise.

This tutorial will contrast two main applications settings for regression: interpolation and extrapolation. Interpolation suits the use of simple radial basis function kernels which bias towards smooth latent functions. For extrapolation, however, the choice of the kernel is paramount, encoding our prior belief about the type of function wish to learn. We present several different kernels, including non-stationary and kernels for structured data (string and tree kernels). One of the main issues for kernel methods is setting the hyperparameters, which is often done in the support vector literature using grid search on held-out validation data. In the GP framework, we can compute the probability of the data given the model which involves the integral over the parameter space. This marginal likelihood or *Bayesian evidence* can be used for model selection using only training data, where by model selection we refer either to choosing from a set of given covariance kernels or choosing from different model hyperparameters (kernel parameters). We will present the key algorithms for type-II maximum likelihood estimation with respect to the hyper-parameters, using gradient ascent on the marginal likelihood.

Many problems in NLP involve learning from a range of different tasks. We present multi-task learning models by representing intra-task transfer simply and explicitly as a part of a parameterised kernel function. GPs are an extremely flexible probabilistic framework and have been successfully adapted for multi-task learning, by modelling multiple correlated output variables (Alvarez et al., 2011). This literature develops early work from geostatistics (*kriging* and *co-kriging*), on learning latent continuous spatio-temporal models from sparse point measurements, a problem setting that has clear parallels to transfer learning (including domain adaptation).

In the application section, we start by presenting an open-source software package for GP modelling in Python: GPy.[2] The first application we approach the regression task of predicting user influence on Twitter based on a range or profile and word features (Lampos et al., 2014). We exemplify how to identifying which features are best for predicting user impact by optimising the hyperparameters (e.g. RBF kernel length-scales) using Automatic Relevance Determination (ARD). This basically gives a ranking in importance of the features, allowing interpretability of the models. Switching to a multi-task regression setting, we present an application to Machine Translation Quality Estimation. Our method shows large improvements over previous state-of-the-art (Cohn and Specia, 2013). Concepts in automatic kernel selection are exemplified in an extrapolation regression setting, where we model word time series in Social Media using different kernels (Preoţiuc-Pietro and Cohn, 2013). The Bayesian evidence helps to select the most suitable kernel, thus giving an implicit classification of time series.

In the final section of the tutorial we give a brief overview of advanced topics in the field of GPs. First, we look at non-conjugate likelihoods for modelling classification, count and rank data. This is harder than regression, as Bayesian posterior inference can no longer be solved analytically. We will outline strategies for non-conjugate inference, such as expectation propagation and the Laplace approximation. Second, we will outline recent work on scaling GPs to big data using variational inference to induce sparse kernel matrices (Hensman et al., 2013). Finally – time permitting – we will finish with unsupervised learning in GPs using the latent variable model (Lawrence, 2004), a non-linear Bayesian analogue of principle component analysis.

3 Outline

1. GP Regression (60 mins)
   (a) Weight space view
   (b) Function space view
   (c) Kernels
2. NLP Applications (60 mins)
   (a) Sparse GPs: Predicting user impact
   (b) Multi-output GPs: Modelling multi-annotator data
   (c) Model selection: Identifying temporal patterns in word frequencies
3. Further topics (45 mins)
   (a) Non-congjugate likelihoods: classification, counts and ranking
   (b) Scaling GPs to big data: Sparse GPs and stochastic variational inference

[2] `http://github.com/SheffieldML/GPy`

(c) Unsupervised inference with the GP-LVM

4 Instructors

Trevor Cohn[3] is a Senior Lecturer and ARC Future Fellow at the University of Melbourne. His research deals with probabilistic machine learning models, particularly structured prediction and non-parametric Bayesian models. He has recently published several seminal papers on Gaussian Process models for NLP with applications ranging from translation evaluation to temporal dynamics in social media.

Daniel Preoţiuc-Pietro[4] is a final year PhD student in Natural Language Processing at the University of Sheffield. His research deals with applying Machine Learning models to model large volumes of data, mostly coming from Social Media. Applications include forecasting future behaviours of text, users or real world quantities (e.g. political voting intention), user geo-location and impact.

Neil Lawrence[5] is a Professor at the University of Sheffield. He is one of the foremost experts on Gaussian Processes and non-parametric Bayesian inference, with a long history of publications and innovations in the field, including their application to multi-output scenarios, unsupervised learning, deep networks and scaling to big data. He has been programme chair for top machine learning conferences (NIPS, AISTATS), and has run several past tutorials on Gaussian Processes.

References

Mauricio A. Alvarez, Lorenzo Rosasco, and Neil D. Lawrence. 2011. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: an application to machine translation quality estimation. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, ACL.

James Hensman, Nicolo Fusi, and Neil D. Lawrence. 2013. Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, UAI.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterising user impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL.

Neil D. Lawrence. 2004. Gaussian process latent variable models for visualisation of high dimensional data. *NIPS*, 16(329-336):3.

Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. A temporal model of text periodicities using Gaussian Processes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Kashif Shah, Trevor Cohn, and Lucia Specia. 2013. An investigation on the effectiveness of features for translation quality estimation. In *Proceedings of the Machine Translation Summit*.

[3] http://staffwww.dcs.shef.ac.uk/people/T.Cohn

[4] http://www.preotiuc.ro

[5] http://staffwww.dcs.shef.ac.uk/people/N.Lawrence