Nonparametric Noise models for the Gaussian Process

Ulrich Paquet ULRIPA@MICROSOFT.COM
Microsoft Research, Cambridge

Jurgen Van Gael JURGEN@RANGESPAN.COM
Rangespan Ltd., B131 MacMillan House, Paddington Station, London W2 1FT, UK

Abstract

Notes while developing a Gaussian Process with nonparametric noise model.

1. Introduction

It is sometimes really hard to tell what noise model we want to use for a GP. A specific example is quantile regression for product demand forecasting. We've got an underlying trend, perhaps with cyclical component, which we can easily model with a GP by encoding prior knowledge in the covariance matrix. Unfortunately sales data might have spike and other irregularities which make choosing a noise model quite tricky. One option is to use a robust noise model like student-t or Laplace. In this work, we learn a noise model by using a non-parametric mixture of Gaussians.

2. Model

Imagine we have a time series with observations y_t at times x_t with $t \in [0, T]$. We model this data by assuming a latent Gaussian process

$$\boldsymbol{f} \sim \mathcal{GP}\left(0, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\right) \tag{1}$$

We model the noise as a non-parametric mixture model using the Dirichlet process. Let

$$\boldsymbol{G} \sim \mathcal{DP}\left(\alpha, H\right) \tag{2}$$

be a Dirichlet process with concentration parameter α and base measure H. For each time t we introduce a noise variable $\epsilon_t \sim G$. We then model the observation $y_t = \boldsymbol{f}(x_t) + \epsilon_t$.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

Algorithm 1 Collapsed Gibbs Sampling

Input: data $\boldsymbol{x}, \boldsymbol{y}$ and kernel $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})$
Initialisation: $z_t \sim \mathrm{CRP}(\alpha)$, $\theta_n \sim H$
repeat
  Sample $z_t | \boldsymbol{K}, \boldsymbol{y}, \boldsymbol{z}_{\neg t}, \boldsymbol{\theta}$
  Sample $\theta_n | \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}$
until convergence

3. Inference

We can perform inference in this model using a collapsed Gibbs sampler. In order to work with the CRP representation of the Dirichlet process we introduce a new variable z_t which will represent CRP partition that datapoint t belongs to. For each CRP partition n we represent the cluster parameters using θ_n.

In algorithm 1 we integrate out the Gaussian process $\boldsymbol{f}$. In what follows we derive the resampling steps for z_t and θ_n.

Sampling z_t The conditional distribution of z_t can be written as follows

$$\begin{aligned} p(z_t | \boldsymbol{K}, \boldsymbol{y}, \boldsymbol{z}_{\neg t}, \boldsymbol{\theta}) &\propto \int p(\boldsymbol{y} | \boldsymbol{f}, \boldsymbol{z}_{\neg t}, z_t, \boldsymbol{\theta}) p(\boldsymbol{f} | \boldsymbol{K}) d\boldsymbol{f}, \\ &= \int p(y_t | f_t, \theta_{z_t}) p(f_t | \boldsymbol{y}, \boldsymbol{z}_{\neg t}, \boldsymbol{\theta}, \boldsymbol{K}) df_t. \end{aligned}$$

The key bit is that $p(f_t | \boldsymbol{y}, \boldsymbol{z}_{\neg t}, \boldsymbol{\theta}, \boldsymbol{K})$ is a Gaussian process where every datapoint contributes noise that is dependent on the CRP partition it belongs to.

$$\begin{bmatrix} \boldsymbol{y} \\ f_t \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{\Sigma} & \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x_t}) \\ \boldsymbol{K}(\boldsymbol{x_t}, \boldsymbol{x}) & \boldsymbol{K}(\boldsymbol{x_t}, \boldsymbol{x_t}) \end{bmatrix}\right) \tag{3}$$

Where $\boldsymbol{\Sigma}$ is a diagonal matrix with on the diagonal $\boldsymbol{\Sigma}_{ii} = \theta_{z_{ii}}$.

We know that (**?**) $p(\boldsymbol{f} | \boldsymbol{K})$ can be represented as a multivariate Gaussian distribution.

Sampling θ_n

3.1. Software and Data

Acknowledgments