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 $y_t = \boldsymbol{f}(x_t) + \epsilon_t$. In this work we restrict ourselves to the case where H is a Normal-Inverse Gamma distribution to parameterize the mean and variance of a normal mixture component.

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: $z_t \sim \text{CRP}(\alpha)$, $\theta_n \sim H$
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 $(\mu_n, \sigma_n^2) = \theta_n$.

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

$$p(z_t | \boldsymbol{K}, \boldsymbol{y}, \boldsymbol{z}_{\neg t}, \boldsymbol{\theta}, \alpha) \propto p(z_t | \boldsymbol{z}_{\neg t}, \alpha) p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{K}, \boldsymbol{\theta}). \tag{3}$$

We can compute $p(z_t | \boldsymbol{z}_{\neg t}$ easily using the standard conditional probability for a CRP. Given the noise parameters for every CRP partition are fixed, we can compute the marginal likelihood $p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{K}, \boldsymbol{\theta})$ analytically. (Equation 2.30 in Carl's book).

NOTE: we need to say something about sampling a new z_t, algorithm 8 from Radford Neal's paper should be able to make this happen.

Sampling $\boldsymbol{\theta}$ The condition distribution of $\boldsymbol{\theta}$ can be written as follows

$$p(\boldsymbol{\theta} | \boldsymbol{K}, \boldsymbol{y}, \boldsymbol{z}, H) \propto p(\boldsymbol{\theta} | H) p(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{K}. \boldsymbol{\theta}) \tag{4}$$

Unfortunately $\boldsymbol{\theta}$ influences the marginal likelihood in a complicated way which implies that an analytical expression for the condition distribution on $\boldsymbol{\theta}$ is impossible.

Metropolis-Hastings The simplest and probably least efficient method to sample form the conditional distribution of $\boldsymbol{\theta}$ is to use Metropolis-Hastings sampling. We create a proposal distribution around $\boldsymbol{\theta}$ and can then evaluate the conditional probability up to a proportionality constant by computing the marginal likelihood of the observations.

Auxiliary Variable An alternative method to sample from the conditional distribution on $\boldsymbol{\theta}$ is to introduce a latent variable $\boldsymbol{f}$ for the Gaussian Process. First, we sample

$$p(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{K}, \boldsymbol{y}, \boldsymbol{z}, H), \tag{5}$$

from a multivariate normal distribution.

Given the latent variable $\boldsymbol{f}$, each CRP partition becomes independent, and we can sample the $\boldsymbol{\theta}$.

$$p(\boldsymbol{\theta}|\boldsymbol{f}, \boldsymbol{K}, \boldsymbol{y}, \boldsymbol{z}, H) \propto \prod_k p(\theta_k) \prod_n p(y_n|f_n, \theta_{z_n}). \tag{6}$$

NOTE: Can we analytically compute the posterior of μ, σ? It's learning a number of Normal-Inverse Gamma distributions where the observations are Gaussians (the GP).

3.1. Software and Data

Acknowledgments