Introduction to Gaussian Processes- Regression
DSA2019 Addis Ababa

Charles I. Saidu $^1$, Michael Mayhew $^2$

African University of Science and Technology, Nigeria and Baze University, Nigeria $^1$
Inflammatrix, USA $^2$

$^{isaidu}at^{aust}dot^{edu}dot^{ng}^1$
$^{mmayhew}at^{inflammatix}dot^{com}^2$

June 5, 2019
1 Data Modelling (Pre-introduction to Gaussian Processes-regression): Regression problem

2 Parametric Models: Motivation

3 Non-Parametric Models, and the Gaussian Process

4 Gaussian Processes

5 Additional Resources
Data Modelling: Regression

- Let say we have data $\mathbf{D} = \{X, y\}$
- We are interested in finding the function $f$, such that $y = f(x) + e$ describes the behaviour of data $\mathbf{D}$ with some error bars $e$
Possible approaches towards finding $f$ will be:

- **Parametric Approach**: Neural Networks family,
- **Non-Parametric Approach**: KNN, SVMs, Gaussian Processes
We have Data $D = \{X, y\}$

Parametric Approach:
- Choose a function class $f(x)$ or a mapping - With **FIXED NUMBER** of parameters $\Theta$
- Learn the **PARAMETERS** $\Theta^*$ of the model $f(x)$.

Parameter Estimations
- **Maximum Likelihood Estimation (MLE)**: Find the optimal value $\Theta^*$
- **Maximum A-Posterior Distribution (MAP)** → learn a point estimate (**Mode** of posterior distribution) of the **FIXED** $\Theta^*$, using a prior over $\Theta^*$. **Robust to overfitting**
- **Full Bayesian Methods**: Learn plausible/feasible distribution over **parameters** $\Theta^*$. Using approximation methods: Variational Methods, MCMC, Laplace
- Consider the data scatter plot below
- How would you fit this model $f(x)$?
Linear Model $f(x) = \theta_1 x + \theta_2$ - fitted using MLE or MAP
Optimal parameters $\Theta^* = \{\theta_1, \theta_2\}$
NOTE: Fixed Number of Parameters (2-parameters in this example)
Linear Model $f(x) = \theta_1 x + \theta_2$

Full Bayesian - Distribution of the parameters $\Theta^*$,

Optimal parameters $\Theta^* = \theta_1, \theta_2$

NOTE: Still Fixed Number of Parameters (2-parameters in this example)
Data Modelling: Graphical Intuition - Motivation: Quadratic Parametric models - MLE

- Quadratic Model $f(x) = \theta_1 x^2 + \theta_2 x + \theta_3$ - fitted using MLE or MAP
- Optimal parameters $\Theta^* = \{\theta_1, \theta_2, \theta_3\}$ NOTE: Fixed Number of Parameters (3-parameters in this example)
Data Modelling: Graphical Intuition - Motivation:
Quadratic Parametric models - MAP

- Quadratic Model \( f(x) = \theta_1 x^2 + \theta_2 x + \theta_3 \)
- **Full Bayesian** - Distribution of the parameters \( \Theta^* \)
- Optimal parameters \( \Theta^* = \{\theta_1, \theta_2, \theta_3\} \)
- **NOTE**: Still Fixed Number of Parameters (3-parameters in this example)
Data Modelling: Graphical Intuition - Motivation: Cubic Parametric models - MLE

- Quadratic Model $f(x) = \theta_1 x^3 + \theta_2 x^2 + \theta_3 x + \theta_4$ - fitted using MLE or MAP
- Optimal parameters $\Theta^* = \{\theta_1, \theta_2, \theta_3, \theta_4\}$
- **NOTE**: Fixed Number of Parameters (4-parameters in this example)
Data Modelling: Graphical Intuition - Motivation: Cubic Parametric models - MAP

- Quadratic Model $f(x) = \theta_1 x^3 + \theta_1 x^2 + \theta_2 x + \theta_3$
- **Full Bayesian.** Distribution of the parameters $\Theta^*$
- Optimal parameters $\Theta^* = \{\theta_1, \theta_2, \theta_3, \theta_4\}$
- NOTE: Fixed Number of Parameters (4-parameters in this example)
What if we don't want to specify the number of parameters upfront in our model?

Also what if we want to consider a distribution over plausable functions that describe our data, such that these functions complexity/parameters scale with the data

Also we might want our model to be able to handle missing data: aka Generative Model

How???
Non Parametric models

What is a **Non-parametric model**?

- No! It does **NOT** mean the model has no parameters
- Simply means the models’s number of parameters is **NOT** fixed or determined upfront like in the previous examples - parametric models
- When you hear nonparametric, think models whose parameters scale with amount/complexity of data
Non Parametric models - GPs

- So we want a model whose parameters scale with data/complexity
- We also want to model plausible functions $f(x)$ that describes our data
- Consequently, we want is a distribution over these functions

Sounds Cool!!!
Non Parametric models -GPs

But How??
Non Parametric models - Gaussian Processes

- Let's define a vector of function values evaluated at $n$ points for $x_i \in \mathcal{X}$ as $\mathbf{f} = (f(x_1), f(x_2), \ldots, f(x_n))$
- Let's also assume the notion of smoothness of $\mathbf{f}$ to mean points $(f(x_i), f(x_{i+1}))$ that are closer in space are highly correlated.

Figure: Smoothness Assumption. Source: Neil Lawrence
**Definition:** Gaussian Process:

- Gaussian processes GPs assume neighbouring points $x_i, x_{i+1}$ are correlated and function values $f_i, f_{i+1}$ are distributed multivariate gaussian
- Hence, GPs are parameterized by $\mu(x)$ and covariance function or kernel $K(x_i, x_{i+1})$

\[
p(f_i, f_{i+1}) = \text{GP}(\mu, K) \tag{1}
\]

\[
\mu = \begin{bmatrix} \mu(x_i) \\ \mu(x_{i+1}) \end{bmatrix}, \quad K = \begin{bmatrix} K(x_i, x_i) & K(x_i, x_{i+1}) \\ K(x_{i+1}, x_i) & K(x_{i+1}, x_{i+1}) \end{bmatrix} \tag{2}
\]
Non Parametric models - Gaussian Processes

Similarly \( p(f) = p(f(x_1), f(x_2), \ldots, f(x_n)) \) is also multivariate gaussian given by

\[
p(f) = \mathcal{N}(\mu, K)
\]  

where

\[
\mu = \begin{bmatrix}
\mu(x_1) \\
\mu(x_1) \\
\vdots \\
\mu(x_n)
\end{bmatrix},
K = \begin{bmatrix}
K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_n) \\
K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_n) \\
\vdots & \vdots & \ddots & \vdots \\
K(x_n, x_1) & K(x_n, x_2) & \cdots & K(x_n, x_n)
\end{bmatrix}
\]

Note:

- function \( K \) generates the covariance matrix \( \Sigma \)
- \( \Sigma \) must be **positive definite** functions/matrices
- Note also that \( f \) could easily be infinite dimension as \( n \) tend to infinitinny
Brief Note on Multivariate Norm

Multivariate Normal - Statistic’s swiss army knife

\[ \overline{X} | \mu, \Sigma \sim \text{MVNorm}(\mu, \Sigma) \]

- A highly useful joint distribution for continuous, vector-valued observations
- Parameterized by mean vector \( \mu \) and covariance matrix \( \Sigma \)
Theorem
Suppose \( x = (x_1, x_2) \) is jointly Gaussian with parameters

\[
\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}
\]

Then the marginals are also Gaussians given by

\[
p(x_1) = \mathcal{N}(x_1 | \mu_1, \Sigma_{11})
\]
\[
p(x_2) = \mathcal{N}(x_2 | \mu_2, \Sigma_{22})
\]
Theorem - continues

The posterior is also gaussian given by

\[
p(x_1|x_2) = \mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2})
\]

\[
\begin{align*}
\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\
&= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(x_2 - \mu_2) \\
&= \Sigma_{1|2}(\Lambda_{11}\mu_1 - \Lambda_{12}(x_2 - \mu_2)) \\
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Lambda_{11}^{-1}
\end{align*}
\]
GPs **HAVE** parameters: they are parameterized by $\mu$ and class of kernel function $K(x_i, x_j)$:

- However, parameters scale with complexity/data
- An example of a **Kernel** function is

$$K(x_i, x_j|\Theta) = \theta_0 \exp \left[ - \frac{||x_i - x_j||^2}{2\ell^2} \right]$$ (5)

Hyper parameters $= [\theta_0, \ell]$ - parameter vector

- $\ell$ is the **lengthscale**
- $\theta_0$ is known as the amplitude
Some kernel functions

\[ \kappa = \exp\left(-\frac{||x-x'||^2}{2l^2}\right) \]
\[ \kappa = \min(x, x') \]
\[ \kappa = (x^T x' + c)^2 \]

**Figure:** Effect on choosing different kernels on the prior function distribution. Source: wikipedia
Once we design on our kernel function

Gaussian processes can thus be used for bayesian regression:

\[ p(f|D) = \frac{p(D|f)p(f)}{p(D)} \]  \hspace{1cm} (6)

Where \( p(f) \) represents our prior before of the functions
\( p(D|f) \) is our likelihood of the Data \( D \) given the functions
\( p(f|D) \) is our posterior after observing the data \( D \)
Recap: Bayes’ Theorem/What’s a likelihood?

\[ Pr(\theta|D) = \frac{Pr(D|\theta)Pr(\theta)}{Pr(D)} \]

\[ Pr(D|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \]

**NOTE:** A likelihood is your model for how the data was generated.
Non Parametric models - Where does the likelihood come from?

- All probability modeling starts with a preliminary analysis or visual inspection of the data
  - Called **Exploratory Data Analysis (EDA)**
  - Motivates choice/formulation of the likelihood

**NOTE**

- Carrying out EDA doesn’t violate spirit of prior specification unless the prior is engineered to look exactly like what’s in the data
- This is why we tend to
  - elicit priors from third-party experts
  - use *flat*, non-informative priors
We have training data $D = \{X, y\}$

We want to predict $y_*$ given points $X_*$

Our model is

- $y_n = f_n + e_n$
- $f \sim GP(0, K)$

Then we can make predictions by combining the likelihood and posterior theoretically as

$$p(y_*|X_*, D) = \int p(y_*|X_*, f, D)p(f|D)df$$  \hspace{1cm} (7)
If we assume Gaussian noise: \( y_n = f_n + e_n \), where \( e \sim N(0, \sigma^2) \)

**Likelihood** is gaussian: IID samples

Predictive distribution has Gaussian **Analytical solution** as

\[
p(y_*|X_*, D) \sim \mathcal{N}(f|\mu_*, \Sigma_*)
\]

\[
\mu_* = K_*^T K_y^{-1} y
\]

\[
\Sigma_* = K_{**} - K_y^{-1} K_*
\]
Gaussian Process

\[ p(y_*|X_*, D) \sim \mathcal{N}(f|\mu_*, \Sigma_*) \] (11)

\[ \mu_* = K_*^T K_y^{-1} y \] (12)

\[ \Sigma_* = K_{**} - K_y^{-1} K_* \] (13)

Where

- \( K_y = K + \sigma I \)
- \( K \) - is a kernel function covariance matrix of \( x_1, x_2, \ldots, x_n \)
- \( K_* \) - correlation between \( x_1, x_2, \ldots, x_n \) and \( X_* \) - the test points
- \( K_{**} \) - correlation between \( X_* \)
Non-Parametric Models - Gaussian Processes - Regression Prediction

How do we choose hyper-parameters

- Optimizations to find hyperparameters

What about NON-Gaussian Likelihood functions

- For **NON Gaussian Likelihood**, The posterior does **NOT** have analytical form. **NO SUMMARISING STATISTICS**. Hence, we obtain posterior via
  - Sampling
  - Analytic approximations
Why Gaussian Processes

What are they good for

- Good for time series data
- Directly captures model uncertainty
- Work very well no so large datasets
- Ability to be able to encode prior information of the model
- handles model complexity and scalability quite well

Some limitations

- Not so great for large dataset (time/space complexity). However, parallelization, Sparse GPs and other techniques tries to solve this
- May not be your number one go-to option for classification problems
- Notebook on coding GPs using the equations above using python numpy included (Just the intuition).
- Practical handson using GPy Coming up!

Lecture Notes Neil Lawrence - http://inverseprobability.com


Lecture Notes Nando de Freitas Video - http://www.cs.ox.ac.uk/people/nando.defreitas/

Gaussian processes website - http://www.gaussianprocess.org/
Thank you: Questions?