10-708: Probabilistic Graphical Models 10-708, Spring 2018

24: Gaussian Process

Lecturer: Kayhan Batmanghelich *Scribes: Aditya Siddhant, Soham Ghosh, Chirag Nagpal*

1 Introduction

Many of the problems in Machine Learning such as Regression and Classification boils down to estimating a function and the goal of Gaussian process is to make this estimation non-parametric.

Non-parametric Approach The idea in non-parametric estimation is to integrate out all the parameters in a parametric model and focus directly on the joint distribution. For example in the equation (**??**), the goal is to directly model the conditional joint distribution on the L.H.S. without having exact mentioning of parameters θ. This is what Gaussian process tries to achieve.

$$p(y^*, \mathcal{Y}|x^*, \mathcal{X}) = \int_\theta p(y^*|x^*, \theta)p(\theta) \prod_n p(y^n|\theta, x^n)$$

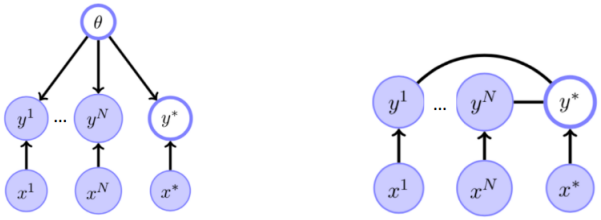

Figure 1: Parametric vs Non-parametric Approach

Revisiting Linear Regression Let us start with a simple regression example. We have a input vector x and feature space $\phi(x)$. Assuming noiseless prediction, we can write:

$$y = w^T \phi(x)$$

where

$$p(w) = \mathcal{N}(w|0, \Sigma_w)$$

Now, in non-parametric approach, when we integrate out w, it boils down to generic smoothness assumptions. If two inputs x and x' are close in input space, their output y and y' should be similar irrespective of the kernel ϕ. So, if the test data is close to train data, we have lower variance in prediction, while if the test data is far away from the train data, we expect higher variance in prediction.

Mathematically, integrating out $p(y^*, \mathcal{Y}|x^*, \mathcal{X})$ would give us something like

$$p(y|x) = \mathcal{N}(y|0, K)$$

where we specify k instead of ϕ directly

$$[K]_{n,n'} = \phi(x^n)^T \phi(x^{n'}) = k(x^n, x^{n'})$$

The idea of kernel k is to have some sort of similarity measure between data points. Some examples of $k(x, x')$ include:

- Linear kernel: $x^T x'$
- RBF kernel: $v_0 exp(-\frac{1}{2}\sum_l \lambda_l (x_l - x'_l)^2)$
- Matern: $||x - x'||^v K_v(||x - x'||)$

Some of these kernels require ϕ to be infinite dimensional and thus we model k directly rather than ϕ. It is important to note, however that the co-variance function controls the behaviour of prediction implicitly (via kernel). Potentially, the $\phi(x)$ can be viewed as an infinite dimensional basis function for the function $f(x)$ we are trying to estimate.

Stationary and Non-stationary Kernels Some kernels like linear and the RBF kernel depend on the distance between the two data points $(x - x')$ and thus are unaffected by starting point or origin, such kernels are called stationary kernels. While the ones like linear kernel are called non-stationary kernels.

Theorem 1 (Leove) *Kernel k corresponds to the covariance of a Gaussian Process iff k is a symmetric positive definite function.*

$$\int k(x, x') f(x) f(x') du(x) du(x') \geq 0$$

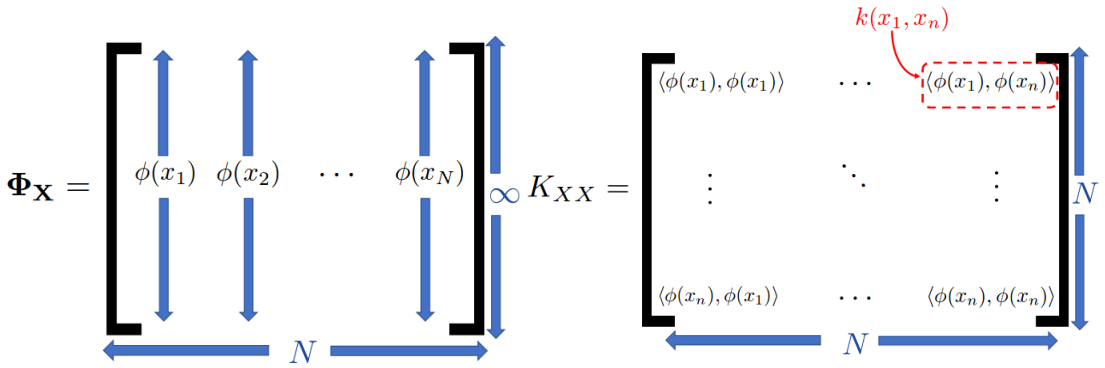

Figure 2: Dimensionality and Structure of Kernel and ϕ

2 Gaussian Process

Gaussian process(GP) can be seen as prior on functions. More formally, A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution. Covariance function and hyperparameters reflect the prior belief on function smoothness, length scales etc.

- Infinite Dimension : $f(x) \sim \mathcal{GP}(m, k)$
- Finite Dimension: $[f(x_1), ...f(x_N)] \sim \mathcal{N}(\mu, K)$ where $\mu_i = m(x_i), K_{ij} = k(x_i, x_j)$

2.1 Inference in GP

2.1.1 Regression

Given the observed noisy data $\mathcal{D} = ,$, the joint probability over the latent function values f and f^* given y is

$$p([f, f^*]|X, X^*, y, \theta_K, \sigma^2) \propto \mathcal{N}\left([f, f^*]|0, \begin{bmatrix} K_{X,X} & K_{X,X^*} \\ K_{X^*,X} & K_{X^*,X^*} \end{bmatrix}\right) \times \prod_{n=1}^{N} \mathcal{N}(y_n|f_n, \sigma^2)$$

where

$$f(X) = \begin{bmatrix} f(x_1) \\ f(x_2) \\ . \\ . \end{bmatrix}$$

$$f^*(X^*) = \begin{bmatrix} f^*(x_1^*) \\ f^*(x_2^*) \\ . \\ . \end{bmatrix}$$

We can write this down as

$$p([f, f^*]|X, X^*, y, \theta_K, \sigma^2) \propto \mathcal{N}\left([y, f^*]|0, \begin{bmatrix} K_{X,X} + \sigma^2\mathbb{I} & K_{X,X^*} \\ K_{X^*,X} & K_{X^*,X^*} \end{bmatrix}\right)$$

Finally, we integrating out y we have:

$$p(f^*|X, X^*, y, \theta_K, \sigma^2) = \mathcal{N}(f^*|\mu^*, \Sigma^*)$$
$$\mu^* = K_{X^*,X}[K_{X,X} + \sigma^2\mathbb{I}]^{-1}y$$
$$\Sigma^* = K_{X^*,X^*} - K_{X^*,X}[K_{X,X} + \sigma^2\mathbb{I}]^{-1}K_{X,X^*}$$

2.1.2 Classification

The problem becomes slightly difficult here as compared to regression because we have a likelihood that doesn't have a closed form and doesn't look like a Gaussian. For example in binary classification $y = \{0, 1\}$, likelihood is

$$p(y_n = 1|f_n) = \frac{1}{1 + exp(-f_n)}$$

When we have a likelihood like this, we cannot absorb it within our prior. We can however approximate this as a log of Gaussian (i.e quadratic) using Taylor Expansion of order 2.

Laplace Method If we consider any probability distribution $p(x)$, we can write it in terms of energy $E(x)$ as shown in the figure 3. Now, we can expand the term $E(x)$ using Taylor expansion as:

$$E(x) \approx E(x^*) + (x - x^*)^T \nabla E|_{x^*} + \frac{1}{2}(x - x^*)^T H(x - x^*)$$

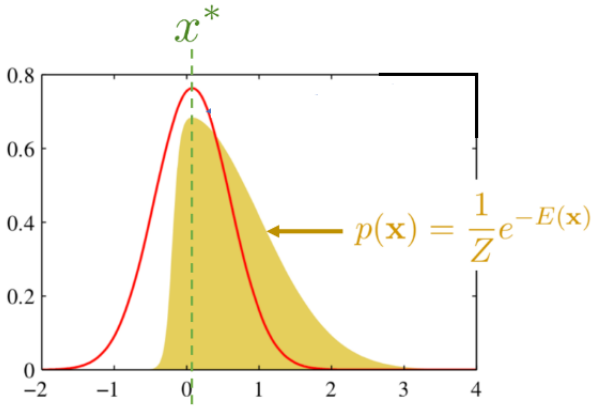

Figure 3: Approximating Gaussian using Taylor Expansion

2.2 Learning

We can use maximum likelihood to learn the parameters of the kernel (θ). Consider the likelihood function $p_\theta(\mathbf{y}|\mathbf{X})$ which can be found by integrating over all possible functions $\mathbf{f}$.

$$p_\theta(\mathbf{y}|\mathbf{X}) = \int p_\theta(\mathbf{y}|\mathbf{f}, \mathbf{X}) p_\theta(\mathbf{f}|\mathbf{X})$$

where $p_\theta(\mathbf{f}|\mathbf{X})$ is the prior and $p_\theta(\mathbf{y}|\mathbf{fX})$ is the likelihood. We consider Gaussian likelihood and Gaussian prior in GP, giving us a closed form solution for the posterior

$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}(0, \sigma^2 I)$$
$$\mathbf{f}|\mathbf{X} \sim \mathcal{N}(0, \mathbb{K})$$
$$\log p(\mathbf{y}|\mathbf{X}) = \log \mathcal{N}(0, \mathbb{K}_y)$$

where $\mathbf{K}_y$ is $\mathbf{K} + \sigma^2 I$. We can learn the parameters θ by taking the gradient of the log-likelihood, which is given by

$$\frac{1}{2}\mathbf{y}^T \mathbf{K}_y^- 1 \frac{\partial \mathbf{K}_y}{\partial \theta_i} \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \operatorname{tr}(\mathbf{K}_y^{-1} \frac{\partial \mathbf{K}_y}{\partial \theta_i})$$

A challenge of these gradients is that there is a significant computational challenge as inverting the $\mathbf{K}_y$ is $O(N^3)$ and also storage of the $O(N^2)$ matrix might be expensive. One way of combating this is by using conjugate gradients. For the latter, we can keep some 'important' points, and discard the rest.

2.3 Choosing $k(x, x')$

The choice of a kernel affects the performance of our model on a certain dataset. For an example we consider the CO2 concentration dataset.

To generate meaningful predictions on this dataset, we need to

- Capture the general increasing trend in emissions
- Capture the miniature local oscillations in the trend

The choice of kernel restricts the space of functions searched upon to fit the data. This can lead to poor modeling of the data. For example for RBF and RBF+Quadratic kernel, poor results are produced as shown in figure 4.

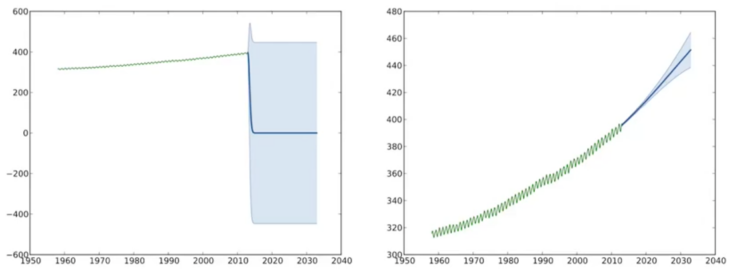

Figure 4: Bad kernel choice

We can instead consider a combination of kernels (which is also a kernel). In this case adding a quadratic kernel, a periodic kernel to capture oscillations, we obtain good results. Adding two kernels together models the data as a superposition of independent functions. Multiplying by a RBF kernel locally smooths the predictions of the first kernel.

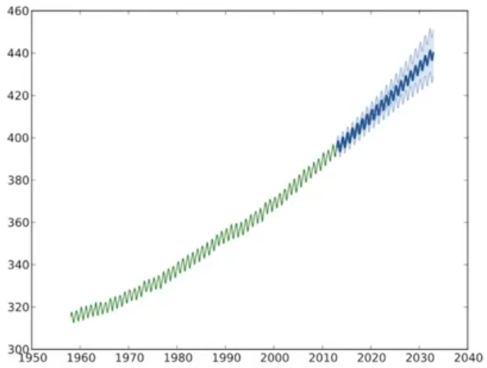

Figure 5: Combination of kernels searches a larger space of functions

2.4 Application to Graphical Models

The main advantage of Gaussian processes lies in going from explicit representation of features to high (or infinite) dimensional representations of the features.

In GMs we model the CPTs wither by tables for discrete data or for continuous data using distributions. We can use GPs to obtain a non-paramateric form of the conditionals.

Recall we can represent a distribution using moments of the distribution as a vector μ_X. Just the first moment is no sufficient - we cannot disambiguate between Gaussian, Laplacian etc. We can add more moments, but this is not scalable.

However, consider an infinite dimensional vector space ϕ_X such that the representation of our distribution is $\mu_X = \mathbb{E}[\phi_X]$. μ_X is called the Hilbert Space embedding. This formulation yields parallel equivalents to the original space (figure 6). The original paper claims this approach can be used in two-sample tests, which are used for determining whether two sets of observations arise from the same distribution, covariate shift correction, local learning, measures of independence, and density estimation.

Original Space	RKHS Space
$\mathbb{P}(X, Y)$	$\mathcal{C}_{YX} = \mathbb{E}[\psi_Y \otimes \phi_X]$
$\text{Diag}[P(X)]$	$\mathcal{C}_{XX} = \mathbb{E}[\phi_X \otimes \phi_X]$
$\mathbb{P}[Y\|X] = \mathbb{P}[Y, X] \times \text{Diag}(\mathbb{P}[X])^{-1}$	$\boldsymbol{\mathcal{C}_{Y\|X} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}}$
$\mathbb{P}[X] = \int_Y \mathbb{P}[X, Y] = \int_Y \mathbb{P}[X\|Y]\mathbb{P}[Y]$	$\boldsymbol{\mu_X = \mathcal{C}_{X\|Y}\mu_Y}$
$\mathbb{P}[X, Y] = \mathbb{P}[X\|Y]\mathbb{P}[Y]$	$\boldsymbol{\mathcal{C}_{YX} = \mathcal{C}_{Y\|X}\mathcal{C}_{XX}}$

Figure 6: Embedding Distributions