

Gaussian Processes: Applications in Machine Learning

Abhishek Agarwal
(05329022)

Under the Guidance of **Prof. Sunita Sarawagi**
KReSIT, IIT Bombay

Seminar Presentation
March 29, 2006

- Introduction to Gaussian Processes(\mathcal{GP})
- Prior & Posterior Distributions
- \mathcal{GP} Models: Regression
- \mathcal{GP} Models: Binary Classification
- Covariance Functions
- Conclusion.

- Supervised Learning
- Gaussian Processes
 - Defines distribution over functions.
 - Collection of random variables, any finite number of which have joint Gaussian distributions.[1] [2]

$$f \sim \mathcal{GP}(m, k)$$

- Hyperparameters and Covariance function.
- Predictions

Prior Distribution

- Represents our belief about the function distribution, which we pass through parameters
- Example: $\mathcal{GP}(m, k)$

$$m(x) = \frac{1}{4}x^2, \quad k(x, x') = \exp(-\frac{1}{2}(x - x')^2).$$

- To draw sample from the distribution:
 - Pick some data points.
 - Find distribution parameters at each point.

$$\mu_i = m(x_i) \quad \& \quad \Sigma_{ij} = k(x_i, x_j) \quad i, j = 1, \dots, n$$

- Pick the function values from each individual distribution.

Prior Distribution(contd.)

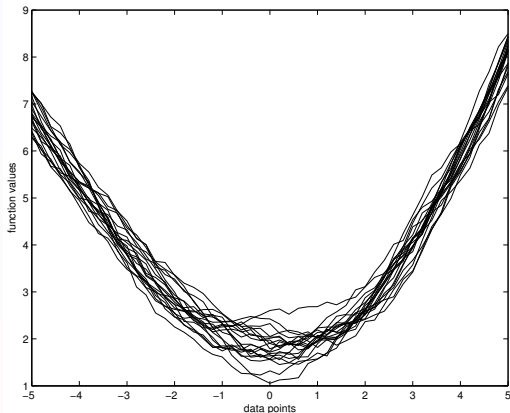


Figure: *Prior distribution over function using Gaussian Process*

Posterior Distribution

- Distribution changes in presence of Training data $\mathcal{D}(x, y)$.
- Functions which satisfy \mathcal{D} are given higher probability.

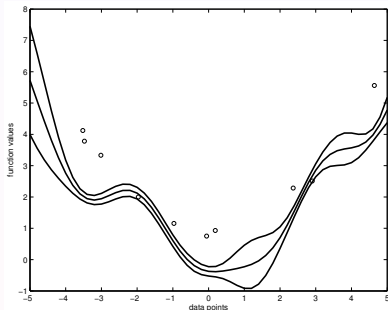


Figure: *Posterior distribution over functions using Gaussian Processes*

Posterior Distribution (contd.)

- Prediction for unlabeled data \mathbf{x}_*
 - \mathcal{GP} outputs the function distribution at x_*
 - Let \mathbf{f} be the distribution at data points in \mathcal{D} and \mathbf{f}_* at x_*
 - \mathbf{f} and \mathbf{f}_* will have a *joint Gaussian distribution*, represented as:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \left(\begin{bmatrix} \mu \\ \mu_* \end{bmatrix}, \begin{bmatrix} \Sigma & \Sigma_* \\ \Sigma_*^T & \Sigma_{**} \end{bmatrix} \right)$$

- Conditional distribution of \mathbf{f}_* given \mathbf{f} can be expressed as:

$$\mathbf{f}_* | \mathbf{f} \sim \mathcal{N}(\mu_* + \Sigma_*^T \Sigma^{-1}(\mathbf{f} - \mu), \Sigma_{**} - \Sigma_*^T \Sigma^{-1} \Sigma_*) \quad (1)$$

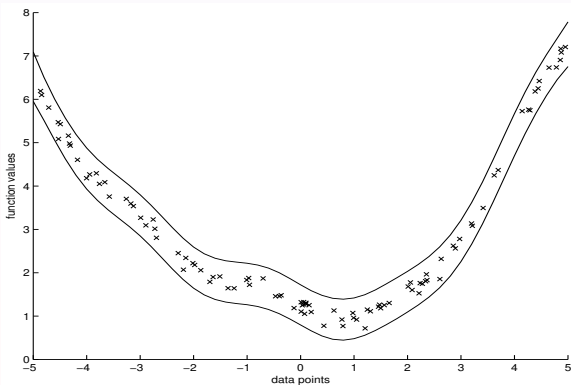
Posterior Distribution (contd.)

- Parameters of the posterior in Eq. 1 are:

$$f_* | \mathcal{D} \sim \mathcal{GP}(m_{\mathcal{D}}, k_{\mathcal{D}}),$$

$$\text{where } m_{\mathcal{D}}(x) = m(x) + \Sigma(X, x)^T \Sigma^{-1}(\mathbf{f} - \mathbf{m})$$

$$k_{\mathcal{D}}(x, x') = k(x, x') - \Sigma(X, x)^T \Sigma^{-1} \Sigma(X, x')$$



- GP can be directly applied to Bayesian Linear Regression model like:
 - $f(x) = \phi(x)^T \mathbf{w}$ with prior $\mathbf{w} \sim \mathcal{N}(0, \Sigma)$
 - Parameters for this distribution will be:

$$\begin{aligned}\mathbb{E}[f(x)] &= \phi(x)^T \mathbb{E}[\mathbf{w}] = 0, \\ \mathbb{E}[f(x)f(x')] &= \phi(x)^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi(x') = \phi(x)^T \Sigma_p \phi(x')\end{aligned}$$

- So, $f(x)$ and $f(x')$ are jointly Gaussian with zero mean and covariance $\phi(x)^T \Sigma_p \phi(x')$.

- In Regression, posterior distribution over the weights, is given as (9):

$$\mathbf{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{marginal likelihood}}$$

- Both prior $p(\mathbf{f}|X)$ and likelihood $p(y|\mathbf{f}, X)$ are Gaussian:

$$\text{prior: } \mathbf{f}|X \sim \mathcal{N}(0, K) \quad (5)$$

$$\text{likelihood: } \mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma_n^2 \mathbf{I})$$

- **Marginal Likelihood** $p(y|X)$ is defined as (6):

$$p(y|X) = \int p(y|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f} \quad (2)$$

- Modeling Binary Classifier
 - Squash the output of a regression model using a response function, like sigmoid.
 - Ex: Linear logistic regression model:

$$p(C_1|x) = \lambda(x^T w), \quad \lambda(z) = \frac{1}{1 + \exp(-z)}$$

- Likelihood is expressed as (7):

$$p(y_i|x_i, w) = \sigma(y_i f_i), \\ f_i \sim f(x_i) = x_i^T w$$

and therefore its **non-Gaussian**.

- Distribution over latent function, after seeing the test data, is given as:

$$p(f_*|X, y, x_*) = \int p(f_*|X, x_*, \mathbf{f})p(\mathbf{f}|X, y)d\mathbf{f}, \quad (3)$$

where $p(\mathbf{f}|X, y) = p(y|\mathbf{f})p(\mathbf{f}|X)/p(y|X)$ is the posterior over the latent variable.

- Computation of the above integral is analytically intractable
 - Both, likelihood and posterior are non-Gaussian.
 - Need to use some analytic Approximation of integrals.

- Gaussian Approximation of $p(\mathbf{f}|X, y)$:
 - Using second order Taylor expansion, we obtain:

$$q(\mathbf{f}|X, y) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, A^{-1})$$

where where $\hat{\mathbf{f}} = \operatorname{argmax}_{\mathbf{f}} p(\mathbf{f}|X, y)$ and
 $A = -\nabla \nabla \log p(\mathbf{f}|X, y)|_{\mathbf{f}=\hat{\mathbf{f}}}$

- To find $\hat{\mathbf{f}}$, we use Newton's method, because of non-linearity of $\nabla \log p(\mathbf{f}|X, y)$ (9)
- Prediction is given as:

$$\pi_* = p(y_* = +1|X, y, \mathbf{x}_*) = \int \sigma(f_*)p(f_*|X, y, \mathbf{x}_*)df_*, \quad (4)$$

Covariance Function

- Encodes our belief about the prior distribution over function
- Some properties:
 - Stationary
 - Isotropic
 - Dot-Product Covariance
- Ex: Squared Exponential(SE) covariance function:

$$\text{cov}(f(x_p), f(x_q)) = \exp\left(-\frac{1}{2}|x_p - x_q|^2\right)$$





- Learned with other hyper-parameters.

Summary and Future Work

- Current Research:
 - Fast sparse approximation algorithm for matrix inversion.
 - Approximation algorithm for non-Gaussian likelihoods.
- \mathcal{GP} approach has outperformed traditional methods in many applications.
 - Gaussian Process based Positioning System (GPPS) [6]
 - Multi user Detection (MUD) in CDMA [7]

\mathcal{GP} models are more powerful and flexible than simple linear parametric models and less complex in comparison to other models like multi-layer perceptrons. [1]

-  Rasmussen and Williams. Gaussian Process for Machine Learning, The MIT Press, 2006.
-  Matthias Seeger. Gaussian Process for Machine Learning, 2004. International Journal of Neural Systems, 14(2):69-106, 2004.
-  Christopher Williams, Bayesian Classification with Gaussian Processes, In IEEE Trans. Pattern analysis and Machine Intelligence, 1998
-  Rasmussen and Williams, Gaussian Process for Regression. In Proceedings of NIPS' 1996.
-  Rasmussen, Evaluation of Gaussian Processes and Other Methods for Non-linear Regression. PhD thesis, Dept. of Computer Science, University of Toronto, 1996. Available from <http://www.cs.utoronto.ca/~carl/>

-  Anton Schwaighofer, et. al. GPPS: A Gaussian Process Positioning System for Cellular Networks, In proceedings of NIPS' 2003.
-  Murillo-Fuentes, et. al. Gaussian Processes for Multiuser Detection in CDMA receivers, Advances in Neural Information Processing System' 2005
-  David Mackay, Introduction to Gaussian Processes
-  C. Williams. Gaussian processes. In M. A. Arbib, editor, Handbook of Brain Theory and Neural Networks, pages 466-470. The MIT Press, second edition, 2002.

Thank You !!

Questions ??

- Prior:

$$\log p(\mathbf{f}|X) = -\frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \frac{1}{2}\log |K| - \frac{n}{2}\log 2\pi \quad (5)$$

- Marginal likelihood

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^T (K + \sigma_n^2 \mathcal{I})^{-1}\mathbf{y} - \frac{1}{2}\log |K + \sigma_n^2 \mathcal{I}| - \frac{n}{2}\log 2\pi \quad (6)$$

- Likelihood

$$p(y = +1|x, \mathbf{w}) = \sigma(x^T \mathbf{w}), \quad (7)$$

For symmetric likelihood $\sigma(-z) = 1 - \sigma(z)$.

$$p(y_i|x_i, \mathbf{w}) = \sigma(x_i^T \mathbf{w}), \quad (8)$$

- first derivative of posterior

$$\hat{\mathbf{f}} = K(\nabla \log p(\mathbf{f}|X, y))$$

- Prediction

$$p(w|y, X) = \frac{\mathbf{p}(y|\mathbf{X}, \mathbf{w}) * \mathbf{p}(\mathbf{w})}{p(y|X)}$$