GAUSSIAN PROCESS

Chu Duc Thang
Nguyen Ngoc Tuan

OUTLINE

1. Prerequisite
2. Background
3. Requirements, assumptions and settings
4. Gaussian Process
5. Intuition, advantages and disadvantages
6. Kernel
7. Summary
8. Citation and Demo

1. MULTIVARIATE GAUSSIAN (PRE-REQUISITE)

Setting $x = \begin{bmatrix} x_A \\ x_B \end{bmatrix}$ $\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}$ $\Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$

$$p(x_A) = \int_{x_B} p(x_A, x_B; \mu, \Sigma) dx_B$$
$$p(x_B) = \int_{x_A} p(x_A, x_B; \mu, \Sigma) dx_A$$

$$p(x_A \mid x_B) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A} p(x_A, x_B; \mu, \Sigma) dx_A}$$
$$p(x_B \mid x_A) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B} p(x_A, x_B; \mu, \Sigma) dx_B}$$

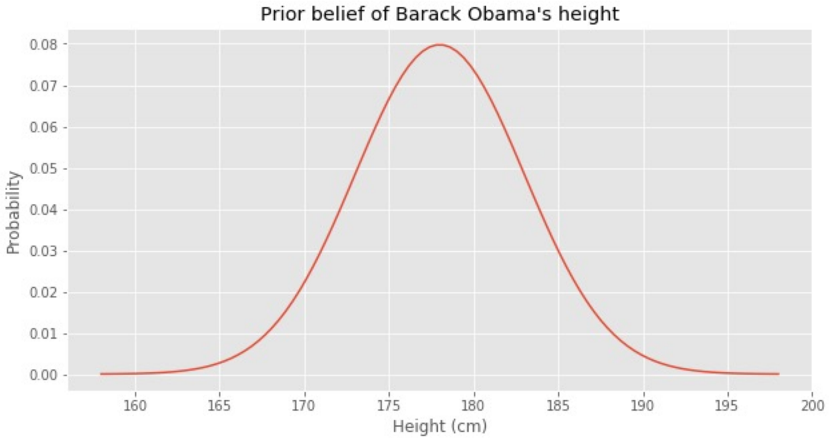
Marginalization $x_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$
$x_B \sim \mathcal{N}(\mu_B, \Sigma_{BB}).$

Conditional $x_A \mid x_B \sim \mathcal{N}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA})$
$x_B \mid x_A \sim \mathcal{N}(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}).$

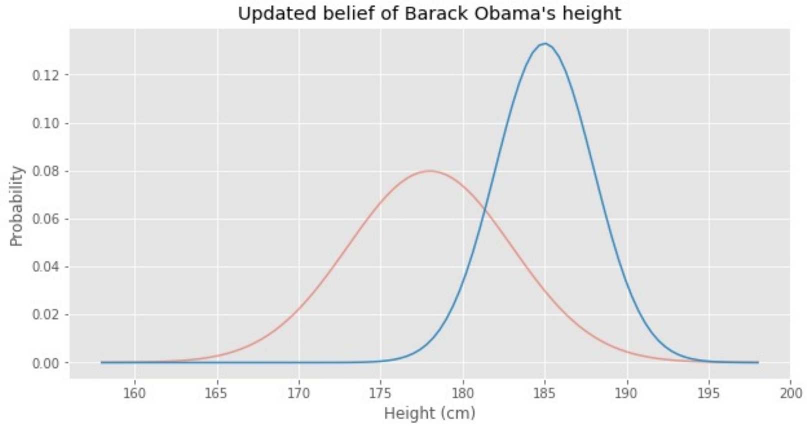
Note: Once Gaussian, always Gaussian

Importance: Central Limit Theorem

WHAT IS OBAMA’S HEIGHT



Prior belief of Barack Obama's height



Updated belief of Barack Obama's height

2. BACKGROUND

- Purpose: Learn underlying distribution from training data

- Parametric vs Nonparametric method
  - Given set of training points, there are potentially infinitely many functions that fit the data

- Bayesian Inference:
  - $p(w|D) = \frac{p(Data|w)p(w)}{p(Data)}$ or $p(w|D) \propto p(Data|w)p(w)$
  - Goal: $p(y^*|x, x^* \ , y)$ – with y*: posterior function, X*: test data, X: training data, y: prior function

3. REQUIREMENTS, ASSUMPTIONS AND SETTINGS

- Regression problem:

- Extension version can be used for classification problem
- $y\ = f(x) + \varepsilon$

$$X = \begin{bmatrix} \text{—} & (x^{(1)})^T & \text{—} \\ \text{—} & (x^{(2)})^T & \text{—} \\ & \vdots & \\ \text{—} & (x^{(n)})^T & \text{—} \end{bmatrix} \in \mathbf{R}^{n \times d} \quad \vec{f} = \begin{bmatrix} f(x^{(1)}) \\ f(x^{(2)}) \\ \vdots \\ f(x^{(n)}) \end{bmatrix}, \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbf{R}^{n},$$

$$X_* = \begin{bmatrix} \text{—} & (x_*^{(1)})^T & \text{—} \\ \text{—} & (x_*^{(2)})^T & \text{—} \\ & \vdots & \\ \text{—} & (x_*^{(n_*)})^T & \text{—} \end{bmatrix} \in \mathbf{R}^{n_* \times d} \quad \vec{f}_* = \begin{bmatrix} f(x_*^{(1)}) \\ f(x_*^{(2)}) \\ \vdots \\ f(x_*^{(n_*)}) \end{bmatrix}, \quad \vec{\varepsilon}_* = \begin{bmatrix} \varepsilon_*^{(1)} \\ \varepsilon_*^{(2)} \\ \vdots \\ \varepsilon_*^{(n_*)} \end{bmatrix}, \quad \vec{y}_* = \begin{bmatrix} y_*^{(1)} \\ y_*^{(2)} \\ \vdots \\ y_*^{(n_*)} \end{bmatrix} \in \mathbf{R}^{n_*}.$$

$$\begin{bmatrix} \vec{\varepsilon} \\ \vec{\varepsilon}_* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \sigma^2 I & \vec{0} \\ \vec{0}^T & \sigma^2 I \end{bmatrix}\right).$$

4. GAUSSIAN PROCESS

Definition: A GP is a (potentially infinite) collection of random variables such that the joint distribution of every finite subset of random variables is multivariate Gaussian

Prior: Assuming the mean of function is 0 $\mathrm{p(y|x)} \sim \mathcal{N}(\mu = 0, K)$

Posterior: After observing $y^* \ and \ x^*$, we use the Bayesian rules as followed

4. GAUSSIAN PROCESS

Gaussian process: modelling probability distributions over functions $p(y^*|y, x^*, x)$

$$\begin{bmatrix} \vec{y} \\ \vec{y}_* \end{bmatrix} \Bigg| X, X_* = \begin{bmatrix} \vec{f} \\ \vec{f}_* \end{bmatrix} + \begin{bmatrix} \vec{\varepsilon} \\ \vec{\varepsilon}_* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} K(X,X) + \sigma^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) + \sigma^2 I \end{bmatrix}\right).$$

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_y^2 \mathbf{I})$$

$$\mu^* = K(X_*,X)\left(K(X,X) + \sigma^2 I\right)^{-1} \vec{y}$$

$$\Sigma^* = K(X_*,X_*) + \sigma^2 I - K(X_*,X)\left(K(X,X) + \sigma^2 I\right)^{-1} K(X,X_*).$$

5. INTUITION, ADVANTAGES AND DISADVANTAGES

- Intuition: Gaussian process versus classical linear regression
  - GP: The posterior distribution over the y* for new input x* reflect the uncertainty in our prediction y* due to the ε and choice of prior θ.
  - Classical linear regression models, estimate θ directly from the training data but provide no estimate of how reliable these learned parameters may be.

- Advantages:
  - Elegant solution
  - Assigning probabilities to each of these functions
  - Incorporate confidence in the prediction

- Disadvantages:
  - Computationally expensive

6. KERNEL

- Definition: Outputs the similarity between 2 given points
- Example: Squared Exponential Kernel

$$k_{SE}(x, x') = \quad \sigma^2 \exp\left(-\frac{||t-t'||^2}{2l^2}\right)$$

- Note: Relation between covariance and kernel
- Parameters:
  - σ^2: average distance away from the function’s mean
  - l : reach of influence on neighbors

7. SUMMARY

-GP: estimation of function, instead of parameters

- Bayesian inference

- Non-parametric method

8. CITATION

https://distill.pub/2019/visual-exploration-gaussian-processes/

http://cs229.stanford.edu/summer2020/gaussian_processes.pdf

https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote15.html