Python for Data Science
Introduction to Bayesian Approaches in Learning from Data

Heru Praptono

heru.pra@cs.ui.ac.id

Salemba, December 2018

Outline

Insight

What we want to obtain?

Bayesian Learning from Data

Introd to Bayesian Linear Regression

Introd to Gaussian Processes

Insight: What we want to obtain?

- Get the idea at a glance on the differences between Frequentist and Bayesian approaches in learning from data.
- Implement in a very simple way how Python comes up with those stuff.

Revisit Linear Regression

The concept of linear regression is very interesting, because it gives the central idea of so many various machine learning models, as the function for approximations.

Bayesian Linear Regression

Revisit Linear Regression

In linear (linear in parameter) regression:

$$y_i = \beta x_i + c + \epsilon_i \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The estimation scenario is to find β by utilising e.g. least square error. The task is to minimise total residual,[1] S

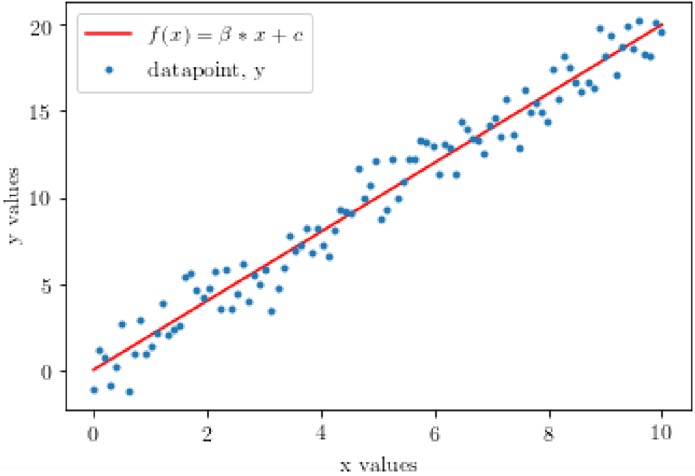
$$S = \sum_{i=1}^{N} r_i^2 \tag{2}$$

where $r_i = y_i - f(x_i, \beta)$

[1] Here we note residual, as the estimation from sample, rather than error which is the differences between the observed and the population

Bayesian Linear Regression

Revisit Linear Regression



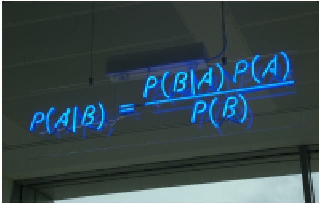
Bayesian Linear Regression

Revisit Linear Regression

- Usually we use Maximum Likelihood Estimation (MLE) to find the parameter β. The term "MLE" is just a fancy terminology from the simple one: "estimating the parameter by seeing available (sample) data"
- This approach is usually called as "frequentist" approach.
- Fitting function procedure includes e.g. *least square estimates*, given data (`np.linalg.lstsq` or `np.linalg.solve`). Usually, we are interested in a set of unique solution.
- We refer this as point-estimate approach, rather than expected value from a distribution.

Bayesian Linear Regression

Bayes Rule



Bayesian Linear Regression

Bayes Rule and Our parameter estimation approach

We have had Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

That is

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Normalisation}} \tag{4}$$

The `Normalisation` constant is needed in order to make a valid distribution (that is, the sum of the area under the curve must be 1). Our parameter estimation thus becomes:

$$p(\beta|y,x) = \frac{p(y|\beta,x)p(\beta|x)}{p(y|x)} \tag{5}$$

Bayesian Linear Regression

- For simplicity, through this slide the parts related to regression model, we will assume that all of the distributions (prior and likelihood) are Gaussian distribution.
- So that prior and posterior will have the same family distribution. This is called as conjugate prior concept.

Bayesian Linear Regression

Point based estimate vs distribution based estimate

Our original model

$$y = \boldsymbol{\beta}^T \mathbf{x} + \epsilon \tag{6}$$

where $\boldsymbol{\beta} = \arg\max_{\boldsymbol{\beta}} p(y|\boldsymbol{\beta}, \mathbf{x})$ (Here I eliminate notation c (intercept) just to simplify, turning into $\boldsymbol{\beta} = (\beta_1, \beta_0)^T; \mathbf{x} = (x, 1)^T$). Our prediction y is rather a point estimate. But now, with Bayesian approach our prediction of y would rather be interpreted as a distribution p_y, that is:

$$y \sim \mathcal{N}(\boldsymbol{\beta}^T \mathbf{x}, \sigma_y^2) \tag{7}$$

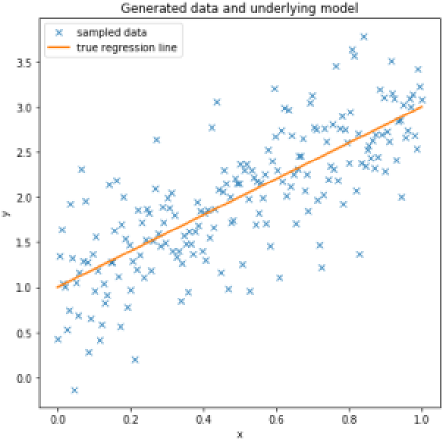
Recall that in learning process (or parameter estimation), our $\boldsymbol{\beta}$ comes from:

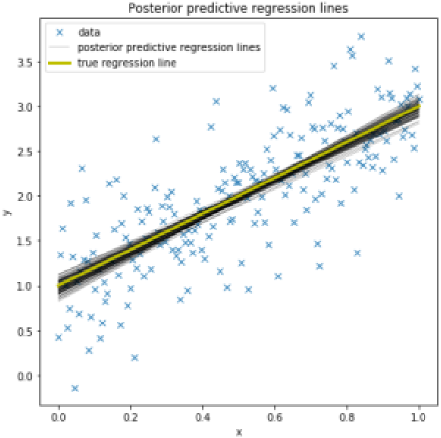
$$p(\boldsymbol{\beta}|y, \mathbf{x}) = \frac{p(y|\boldsymbol{\beta}, \mathbf{x})p(\boldsymbol{\beta}|\mathbf{x})}{p(y|\mathbf{x})} \tag{8}$$

So that our estimated y is represented as $\mathbb{E}(y) \approx \bar{y} = \frac{1}{N}\sum_i^N y_i$ where y_i is sampled from our posterior p_y

Bayesian Linear Regression

Frequentist vs Bayesian Approach in modelling linear regression.





Left: Fruequentist approach, **Right:** Bayesian approach.

Introduction to Gaussian Processes

In Gaussian process, our f is:

$$f \sim \mathcal{GP} \tag{9}$$

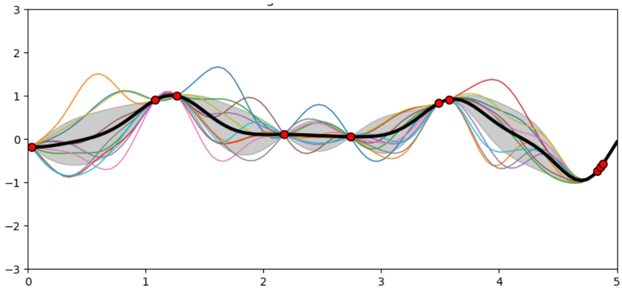
Given x, the function f is possibly any complex and expensive function to evaluate.

Notation

I will write the bold symbol to represent multivariate random variable,
$\mathbf{f} = (f_1, f_2, \ldots f_N)$.

We will demonstrate that f is of any regression function. That is, we refer this as Gaussian Process Regression.

Introd to Gaussian Processes



Introd to Gaussian Processes

With Gaussian processes, we update our belief about our $\mathbf{f}$ every time we see the data $\mathbf{y}$

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{p(\mathbf{y})} \tag{10}$$

So posterior is proportional to prior and likelihood:

$$p(\mathbf{f}|\mathbf{y}) \propto (\mathbf{y}|\mathbf{f})p(\mathbf{f}) \tag{11}$$

Having conjugate prior, our posterior poses same family distribution as the prior.

$$p(\mathbf{f}|\mathbf{y}) = (\mathbf{y}|\mathbf{f})p(\mathbf{f}) \tag{12}$$

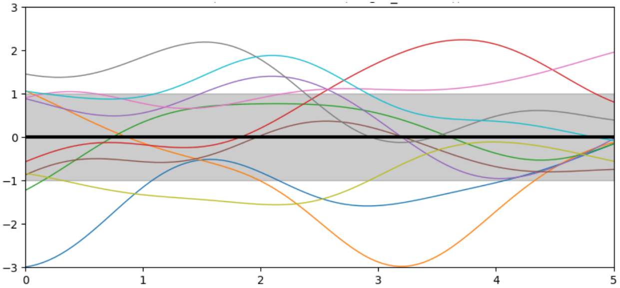
We know that $\mathbf{f}$ is a function of $\mathbf{x}$ to approximate $\mathbf{y}$. Thus, we need to model $\mathbf{x}, \mathbf{f}, \mathbf{y}$ for this.

Introd to Gaussian Processes

Prior

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{13}$$

where $\mathbf{K}$ is any kernel $K_{ij} = k(\mathbf{x}^i, \mathbf{x}^j)$ (usually we use Mercel kernel, in order to make sure that K is a positive definite matrix, so that it would be valid Gaussian).



Visualisation of Prior belief about f over x. Horizontal axes: x, vertical axes: f

Introd to Gaussian Processes

And now let us say we perform an experiment, so that we got $y^* = f(x)$. (Note: $y \sim \mathcal{N}(f, \sigma_y^2)$)
We then update our $\mathbf{f}$ become new one, $\mathbf{f}^*$. First, model the joint probability between $\mathbf{y} = (y_1 \dots y_n, y^*)$ and new $\mathbf{f}^*$.

$$p\left(\begin{bmatrix}\mathbf{y} \\ \mathbf{f}^*\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y} \\ \mathbf{f}^*\end{bmatrix}; \mathbf{0}, \begin{bmatrix}\mathbf{K}(X, X^*) + \sigma_y^2\mathbb{I} & \mathbf{K}(X, X^*) \\ \mathbf{K}(X^*, X) & \mathbf{K}(X^*, X^*)\end{bmatrix}\right) \tag{14}$$

Posterior Distribution

Our updated distribution of $\mathbf{f}^*$, that is posterior, becomes:

$$p(\mathbf{f}^*) = \mathcal{N}(\mathbf{f}^*; \mathrm{mean}(\mathbf{f}^*), \mathrm{cov}(\mathbf{f}^*)) \tag{15}$$

Where

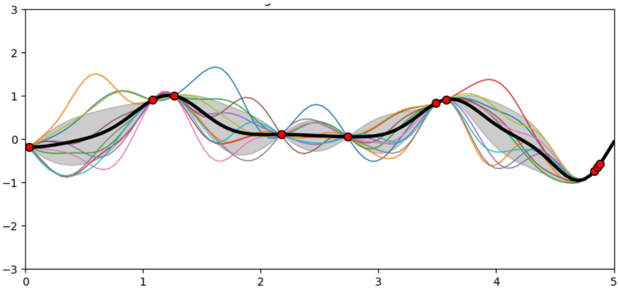
$$\mathrm{mean}(\mathbf{f}^*) = \mathbf{K}(X^*, X)(\mathbf{K}(X, X) + \sigma_y^2 \mathbb{I})^{-1}\mathbf{y} \tag{16}$$

and

$$\mathrm{cov}(\mathbf{f}^*) = \mathbf{K}(X^*, X^*) - \mathbf{K}(X^*, X)(\mathbf{K}(X, X) + \sigma_y^2 \mathbb{I})^{-1}\mathbf{K}(X, X^*) \tag{17}$$

Introd to Gaussian Processes

Our posterior..



Visualisation of Posterior belief about f over x. Horizontal axes: x, vertical axes: f

Introd to Gaussian Processes

As simple as that, but it is indeed a powerful method

Thank You