

Multivariate Normal Distribution

Nipun Batra

August 26, 2023

IIT Gandhinagar

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.
- k is the dimensionality of the random vector $\boldsymbol{\theta}$ (number of variables).

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.
- k is the dimensionality of the random vector $\boldsymbol{\theta}$ (number of variables).
- $\boldsymbol{\Sigma}$ is the covariance matrix

Multivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

- $\boldsymbol{\theta}$ is the vector of random variables (observation) for which you want to calculate the PDF.
- k is the dimensionality of the random vector $\boldsymbol{\theta}$ (number of variables).
- $\boldsymbol{\Sigma}$ is the covariance matrix
- $\boldsymbol{\mu}$ is the mean vector.

Bivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

Bivariate Normal Distribution

$$\text{PDF}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)$$

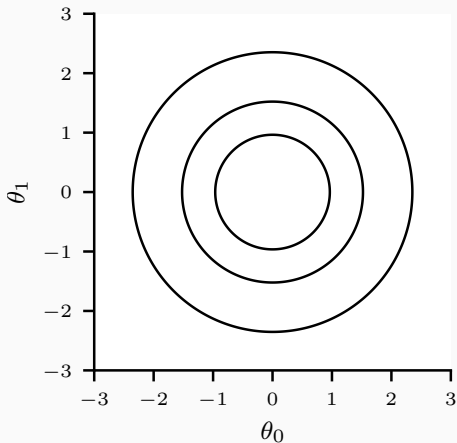
Slides heavily inspired from Richard Turner's slides

Bivariate Normal Distribution

Notebook ([visualise-normal.ipynb](#))

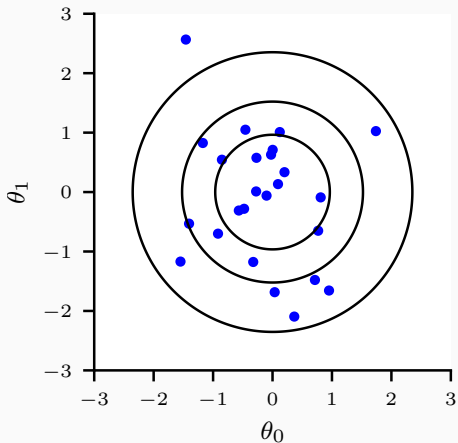
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \Sigma = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$



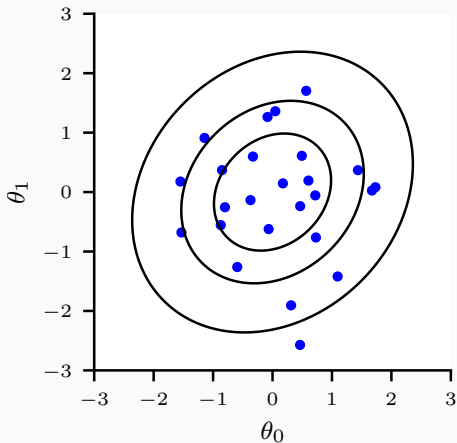
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \Sigma = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$



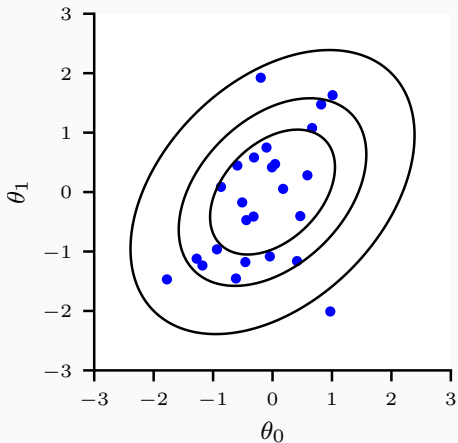
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \Sigma = \begin{bmatrix} 1.0 & 0.2 \\ 0.2 & 1.0 \end{bmatrix}$$



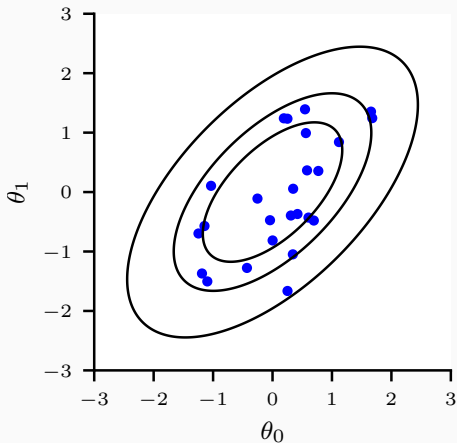
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \Sigma = \begin{bmatrix} 1.0 & 0.4 \\ 0.4 & 1.0 \end{bmatrix}$$



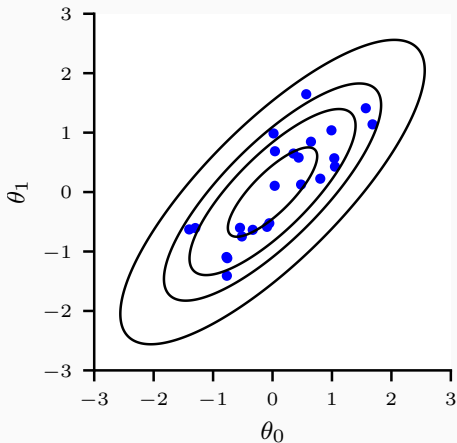
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \Sigma = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$$



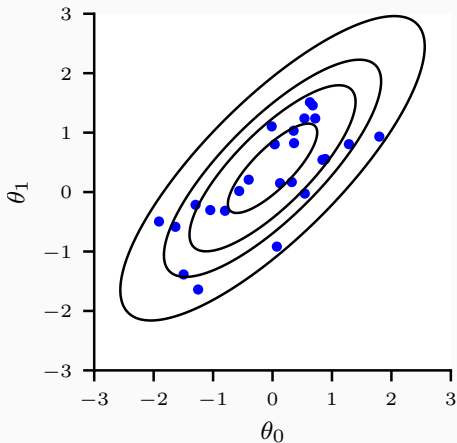
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \Sigma = \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix}$$



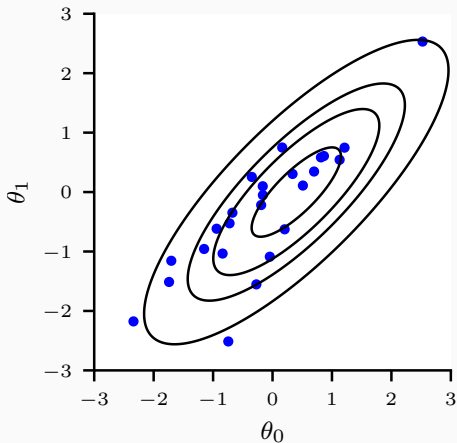
Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \mu = \begin{bmatrix} 0.0 \\ 0.4 \end{bmatrix}$$



Bivariate Normal Distribution

$$PDF(\mu, \Sigma) \propto \exp\left(-\frac{1}{2}(\theta - \mu)^\top \Sigma^{-1}(\theta - \mu)\right) \quad \mu = \begin{bmatrix} 0.4 \\ 0.0 \end{bmatrix}$$



Bayesian Linear Regression

Nipun Batra

August 26, 2023

IIT Gandhinagar

Bayesian Linear Regression

$$\boldsymbol{\theta}_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\theta_{\text{MLE}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

For θ_{MAP} estimation, we assume a Gaussian prior
 $p(\boldsymbol{\theta}) = \mathcal{N}(0, b^2 \mathbf{I})$

$$\theta_{\text{MLE}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

For θ_{MAP} estimation, we assume a Gaussian prior $p(\theta) = \mathcal{N}(0, b^2 \mathbf{I})$

$$\theta_{\text{MAP}} = \left(\mathbf{X}^T \mathbf{X} + \frac{\sigma^2}{b^2} \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

$$\theta_{\text{MLE}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

For θ_{MAP} estimation, we assume a Gaussian prior $p(\theta) = \mathcal{N}(0, b^2 \mathbf{I})$

$$\theta_{\text{MAP}} = \left(\mathbf{X}^T \mathbf{X} + \frac{\sigma^2}{b^2} \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

where \mathbf{X} is the feature matrix, \mathbf{y} is the corresponding ground truth values and σ is the standard deviation of Gaussian distribution in the MLE estimation.

Linear Regression using Basis Functions

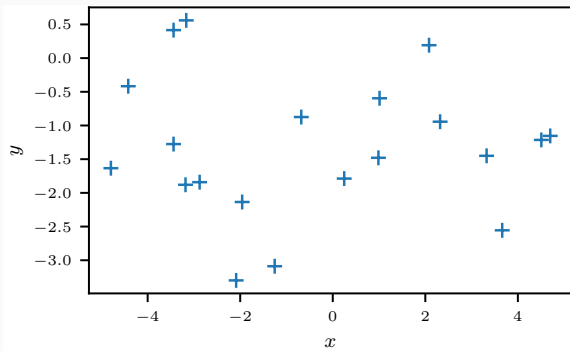


Figure 1: Data

Linear Regression using Basis Functions

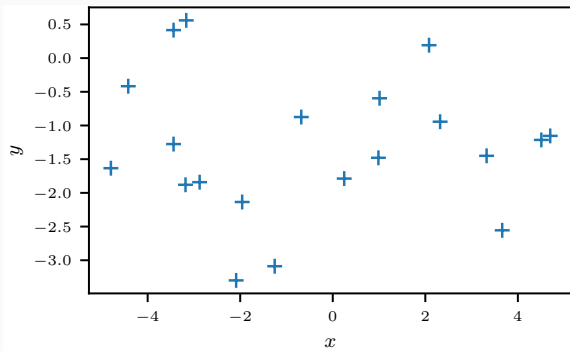


Figure 1: Data

We can use basis functions to fit a non-linear function to the data.

Linear Regression using Basis Functions

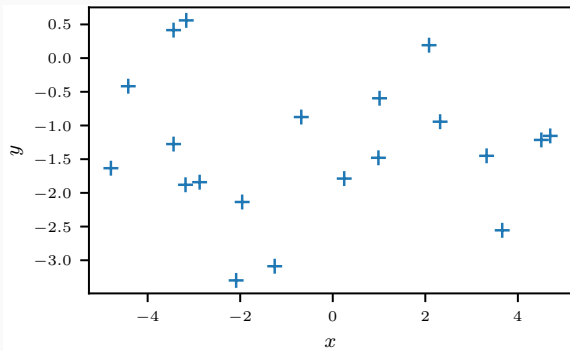


Figure 1: Data

We can use basis functions to fit a non-linear function to the data. For example we can use a polynomial basis function to fit a polynomial to the data, where $\phi_j(x) = x^j$.

MLE and MAP

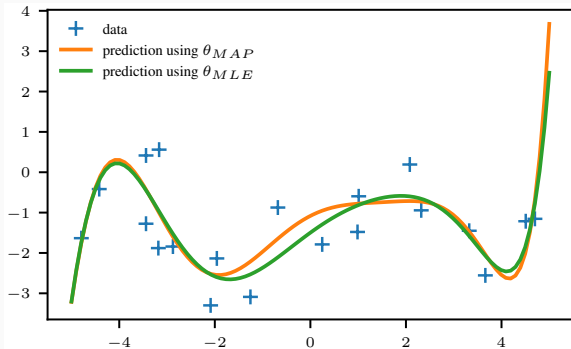


Figure 2: MLE and MAP

Bayesian Linear Regression

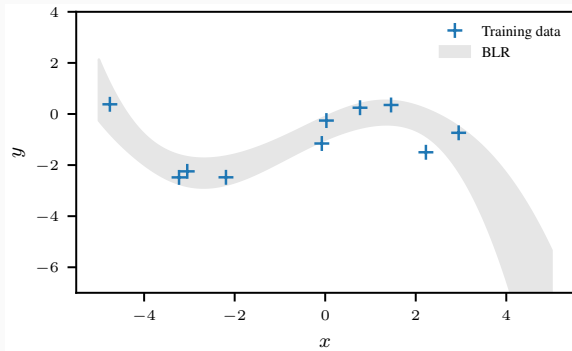
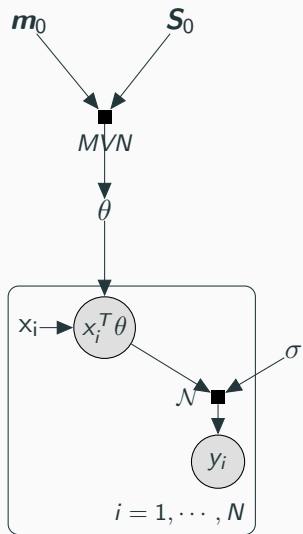


Figure 3: Bayesian linear regression

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

- $P(\theta|D)$ is called the posterior
- $P(D|\theta)$ is called the likelihood
- $P(\theta)$ is called the prior
- $P(D)$ is called the evidence

Bayesian Linear Regression



In Bayesian linear regression, we consider the model:

$$\text{prior : } p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$$

with \mathbf{m}_0 and \mathbf{S}_0 as the mean and covariance matrix and

$$\text{likelihood : } p(y | \mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y | \mathbf{x}^\top \boldsymbol{\theta}, \sigma^2)$$

Given a training set of inputs $\mathbf{x}_n \in \mathbb{R}^D$ and corresponding observations $y_n \in \mathbb{R}$, $n = 1, \dots, N$, we compute the posterior over the parameters using Bayes' theorem as

$$p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y} \mid \mathcal{X})}$$

where \mathcal{X} is the set of training inputs and \mathcal{Y} the collection of corresponding training targets.

We find the closed form solution of posterior $p(\boldsymbol{\theta} \mid \mathcal{X})$ to be a normal distribution with mean \mathbf{m}_N and covariance matrix \mathbf{S}_N

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) &= \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N) \\ \mathbf{S}_N &= \left(\mathbf{S}_0^{-1} + \sigma^{-2} \mathbf{X}^\top \mathbf{X} \right)^{-1} \\ \mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \sigma^{-2} \mathbf{X}^\top \mathbf{y} \right) \end{aligned}$$

where the subscript N indicates the size of the training set.

$$\text{Posterior : } p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{Y} | \mathcal{X})}$$

$$\text{Likelihood : } p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} | \mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I})$$

$$\text{Prior : } p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} | \mathbf{m}_0, \mathbf{S}_0)$$

The sum of the log-prior and the log-likelihood is

$$\begin{aligned} & \log \mathcal{N}(\mathbf{y} \mid \mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}) + \log \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_0, \mathbf{S}_0) \\ &= -\frac{1}{2} \left(\sigma^{-2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + (\boldsymbol{\theta} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\boldsymbol{\theta} - \mathbf{m}_0) \right) + \text{const} \end{aligned}$$

We ignore the constant term independent of θ . We now factorize, which yields

We ignore the constant term independent of θ . We now factorize, which yields

$$= -\frac{1}{2} \left(\sigma^{-2} \mathbf{y}^\top \mathbf{y} - 2\sigma^{-2} \mathbf{y}^\top \mathbf{X} \theta + \theta^\top \sigma^{-2} \mathbf{X}^\top \mathbf{X} \theta + \theta^\top \mathbf{S}_0^{-1} \theta - 2\mathbf{m}_0^\top \mathbf{S}_0^{-1} \theta + \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 \right)$$

We ignore the constant term independent of θ . We now factorize, which yields

$$= -\frac{1}{2} \left(\sigma^{-2} \mathbf{y}^\top \mathbf{y} - 2\sigma^{-2} \mathbf{y}^\top \mathbf{X} \theta + \theta^\top \sigma^{-2} \mathbf{X}^\top \mathbf{X} \theta + \theta^\top \mathbf{S}_0^{-1} \theta - 2\mathbf{m}_0^\top \mathbf{S}_0^{-1} \theta + \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 \right)$$

$$= -\frac{1}{2} \left(\theta^\top \left(\sigma^{-2} \mathbf{X}^\top \mathbf{X} + \mathbf{S}_0^{-1} \right) \theta - 2 \left(\sigma^{-2} \mathbf{X}^\top \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0 \right)^\top \theta \right) + \text{const}$$

Now, we evaluate the posterior distribution,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) &= \exp(\log p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y})) \propto \exp(\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})) \\ &\propto \exp\left(-\frac{1}{2} \left(\boldsymbol{\theta}^\top \left(\sigma^{-2} \mathbf{X}^\top \mathbf{X} + \mathbf{S}_0^{-1} \right) \boldsymbol{\theta} - 2 \left(\sigma^{-2} \mathbf{X}^\top \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0 \right)^\top \boldsymbol{\theta} \right)\right) \end{aligned}$$

Normalizing the posterior distribution

We now normalize this Gaussian distribution into the form that is proportional to $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N)$, i.e., we need to identify the mean \mathbf{m}_N and the covariance matrix \mathbf{S}_N .

Normalizing the posterior distribution

We now normalize this Gaussian distribution into the form that is proportional to $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N)$, i.e., we need to identify the mean \mathbf{m}_N and the covariance matrix \mathbf{S}_N .

To do this, we use the concept of completing the squares. The desired log posterior is

Normalizing the posterior distribution

We now normalize this Gaussian distribution into the form that is proportional to $\mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N)$, i.e., we need to identify the mean \mathbf{m}_N and the covariance matrix \mathbf{S}_N .

To do this, we use the concept of completing the squares. The desired log posterior is

$$\begin{aligned}\log \mathcal{N}(\boldsymbol{\theta} \mid \mathbf{m}_N, \mathbf{S}_N) &= -\frac{1}{2} (\boldsymbol{\theta} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\boldsymbol{\theta} - \mathbf{m}_N) + \text{const} \\ &= -\frac{1}{2} \left(\boldsymbol{\theta}^\top \mathbf{S}_N^{-1} \boldsymbol{\theta} - 2\mathbf{m}_N^\top \mathbf{S}_N^{-1} \boldsymbol{\theta} + \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N \right).\end{aligned}$$

Normalizing the posterior distribution

We factorize the quadratic form $(\boldsymbol{\theta} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\boldsymbol{\theta} - \mathbf{m}_N)$ into a term that is quadratic in $\boldsymbol{\theta}$ alone, a term that is linear in $\boldsymbol{\theta}$, and a constant term. This allows us now to find \mathbf{S}_N and \mathbf{m}_N by matching the expressions, which yields

$$\begin{aligned}\mathbf{S}_N^{-1} &= \mathbf{X}^\top \sigma^{-2} \mathbf{I} \mathbf{X} + \mathbf{S}_0^{-1} \\ \implies \mathbf{S}_N &= \left(\sigma^{-2} \mathbf{X}^\top \mathbf{X} + \mathbf{S}_0^{-1} \right)^{-1}\end{aligned}$$

and

$$\begin{aligned}\mathbf{m}_N^\top \mathbf{S}_N^{-1} &= \left(\sigma^{-2} \mathbf{X}^\top \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0 \right)^\top \\ \implies \mathbf{m}_N &= \mathbf{S}_N \left(\sigma^{-2} \mathbf{X}^\top \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0 \right).\end{aligned}$$

Posterior Predictive Distribution

Goal: Find $p(y_* | \mathcal{X}, \mathcal{Y}, \mathbf{x}_*)$

$$\begin{aligned} p(y_* | \mathcal{X}, \mathcal{Y}, \mathbf{x}_*) &= \int p(y_* | \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) d\boldsymbol{\theta} \\ &= \int \mathcal{N}(y_* | \mathbf{x}_*^\top \boldsymbol{\theta}, \sigma^2) \mathcal{N}(\boldsymbol{\theta} | \mathbf{m}_N, \mathbf{S}_N) d\boldsymbol{\theta} \\ &= \mathcal{N}(y_* | \mathbf{x}_*^\top \mathbf{m}_N, \mathbf{x}_*^\top \mathbf{S}_N \mathbf{x}_* + \sigma^2) \end{aligned}$$

Posterior Predictive Distribution

Goal: Find $p(y_* | \mathcal{X}, \mathcal{Y}, \mathbf{x}_*)$

$$\begin{aligned} p(y_* | \mathcal{X}, \mathcal{Y}, \mathbf{x}_*) &= \int p(y_* | \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) d\boldsymbol{\theta} \\ &= \int \mathcal{N}(y_* | \mathbf{x}_*^\top \boldsymbol{\theta}, \sigma^2) \mathcal{N}(\boldsymbol{\theta} | \mathbf{m}_N, \mathbf{S}_N) d\boldsymbol{\theta} \\ &= \mathcal{N}(y_* | \mathbf{x}_*^\top \mathbf{m}_N, \mathbf{x}_*^\top \mathbf{S}_N \mathbf{x}_* + \sigma^2) \end{aligned}$$

Two kinds of uncertainty:

- **Aleatoric uncertainty:** Uncertainty in the data - given as σ^2
- **Epistemic uncertainty:** Uncertainty in the model - given as $\mathbf{x}_*^\top \mathbf{S}_N \mathbf{x}_*$

- TFP blog: Aleatoric v/s Epistemic Uncertainty
- MML book: Figure 9.4

Bishop book: Figure 3.7