Bayesian Linear Models

March 26, 2019

Bayesian Linear Models

- This example is a precursor/derivation for Theorem 2 on slide 52.
- We know that the minimum MSE estimator in the Bayesian casse is given by the average value of the posterior distribution:

$$\mathbb{E}_{\theta|x}[\theta|x]$$

- So, as long as we know the posterior distribution, we can gain an estimate.
- However, analytic solutions don't usually exist, especially as models and priors get more complicated/realistic.
- For linear models with gaussian priors we can find a solution!

Setting up the model

- Let $\mathbf{x} = [x[1]...x[n]]^T$ be a vector of samples modeled by :

$$\mathbf{x} = \mathbf{H}\theta + \mathbf{w}$$

  where $\mathbf{w} = [w[1]...w[n]]$ are i.i.d. samples of $N(0, C_w)$, and θ is a vector of parameters to be estimated..
- Then the likelihood function is $\mathrm{p}(\boldsymbol{x}|\theta) \sim N(\mathbf{H}\theta, C_w)$
- Let the prior distribution be $\pi(\theta) \sim N(\mu_\theta, C_\theta)$ be independent of $\mathbf{w}$.
- Then the normal bayesian approach is

$$\mathrm{p}(\theta|x) \propto \mathrm{p}(x|\theta)\pi(\theta)$$

- We know from "l4.pdf" pages 11-16 that multiplying gaussian likelihood with a gaussian prior should yield a gaussian result. However, this is a messy computation, so we will us a different approach to find the posterior.

Using Independence

- We know that $\mathbf{w}$ and θ are independent of each other, and because they are gaussian, this means that their joint distribution is also gaussian. Define:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \theta \end{bmatrix} = \begin{bmatrix} \mathbf{H} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \theta \\ \mathbf{w} \end{bmatrix}$$

- Then the expectations are:

$$\mathbb{E}(\mathbf{z}) = \mathbb{E}(\begin{bmatrix} \mathbf{x} \\ \theta \end{bmatrix}) = \begin{bmatrix} \mathbb{E}(\mathbf{H}\theta + \mathbf{w}) \\ \mathbb{E}(\theta) \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbb{E}(\theta) + 0 \\ \mu_\theta \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mu_\theta \\ \mu_\theta \end{bmatrix}$$

- The joint distribution is given by $\mathrm{p}(\mathbf{x}, \theta) \sim N(\begin{bmatrix} \mathbf{H}\mu_\theta \\ \mu_\theta \end{bmatrix}, \begin{bmatrix} C_{xx} & C_{x\theta} \\ C_{x\theta} & C_{\theta\theta} \end{bmatrix})$

Covariance matrices

- $C_{\theta\theta}$ is easy, since that is just the covariance of theta C_θ.
- The covariance of $\mathbf{x}$ is influenced by the prior pdf:

$$
\begin{aligned}
C_{xx} &= \mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x})(\mathbf{x} - \mathbb{E}(\mathbf{x}))^T] \\
&= \mathbb{E}(\mathbf{H}\theta + \mathbf{w} - \mathbf{H}\mu_\theta)(\mathbf{H}\theta + \mathbf{w} - \mathbf{H}\mu_\theta)^T \\
&= \mathbb{E}(\mathbf{H}(\theta - \mu_\theta) + \mathbf{w})(\mathbf{H}(\theta - \mu_\theta) + \mathbf{w})^T \\
&= \mathbf{H}\mathbb{E}[(\theta - \mu_\theta)(\theta - \mu_\theta)^T]\mathbf{H}^\mathbf{T} + 0 + 0 + \mathbb{E}(\mathbf{w}\mathbf{w}^T) \\
&= \mathbf{H}C_\theta\mathbf{H}^T + C_w
\end{aligned}
$$

- The cross covariance is given by

$$
\begin{aligned}
C_{x\theta} &= \mathbb{E}[(\theta - \mathbb{E}(\theta)(\mathbf{x} - \mathbb{E}(\mathbf{x}))^T \\
&= \mathbb{E}(\theta - \mu_\theta)(\mathbf{H}\theta + \mathbf{w} - \mathbf{H}\mu_\theta)^T \\
&= \mathbb{E}(\theta - \mu_\theta)(\mathbf{H}(\theta - \mu_\theta) + \mathbf{w})^T \\
&= \mathbb{E}(\theta - \mu_\theta)(\mathbf{H}(\theta - \mu_\theta))^T + 0 \\
&= C_\theta\mathbf{H}^T
\end{aligned}
$$

Finding the actual estimator

- Now that we have the covariance matrices, we can find the posterior distribution using the conditional Gaussian formula from lecture 1!
- $p(\theta|\mathbf{x}) \sim N(\mu_\theta + C_{x\theta}C_{xx}^{-1}(\mathbf{x} - \mathbf{H}\mu_\theta), C_{\theta\theta} - C_{\mathbf{x}\theta}C_{xx}^{-1}C_{\mathbf{x}\theta})$
- So our MMSE estimator is:

$$\hat{\theta} = \mu_\theta + C_\theta \mathbf{H}^T(\mathbf{H}C_\theta\mathbf{H}^T + C_w)^{-1}(\mathbf{x} - \mathbf{H}\mu_\theta)$$

- Compare this to the Maximum Likelihood Estimation:

$$\hat{\theta}_{MLE} = (\mathbf{H^T H})^{-1}\mathbf{H^T x}$$

- The bayesian case looks more complicated, but sometimes $\mathbf{H^T H}$ is not invertible, and by choosing the right covariance matrices we can fix that issue.
- In this case C_θ "fixes" $\mathbf{H}$.

Example: Fourier Transform

- Recall the linear example about the fourier series:

$$x[n] = \Sigma_{k=1}^{M} a_k \cos(\frac{2\pi kn}{N}) + b_k \sin(\frac{2\pi kn}{N})_{+} w[n]$$

- We constructed a linear model by creating the matrix $\mathbf{H}$:

$$\mathbf{H} = \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \\ \cos(\frac{2\pi}{N}) & \dots & \cos(\frac{2\pi M}{N}) & \sin(\frac{2\pi}{N}) & \dots & \sin(\frac{2\pi M}{N}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cos(\frac{2\pi(N-1)}{N}) & \dots & \cos(\frac{2\pi M(N-1)}{N}) & \sin(\frac{2\pi(N_1)}{N}) & \dots & \sin(\frac{2\pi M(N-1)}{N}) \end{bmatrix}$$

- For simplicity set

$$\mu_\theta = \mathbf{0}$$

$$C_\theta = \sigma_\theta^2 \mathbb{I}$$

$$C_w = \sigma_w^2 \mathbb{I}$$

Bayesian Estimator of the Fourier Transform

- Using our formula, the bayesian least squares estimator is

$$\hat{\theta} = 0 + \sigma_\theta^2 \mathbf{H}^T(\sigma_\theta^2 \mathbf{H}\mathbf{H}^T + \sigma_w^2 \mathbb{I})^{-1}(\mathbf{x})$$

- But $\mathbf{H^T H} = \frac{N}{2}\mathbb{I}$, which means we can simplify further.

$$\hat{\theta} = \sigma_\theta^2 (\frac{\sigma_\theta^2 N}{2} + \sigma_w^2)\mathbb{I})^{-1}\mathbf{H}^T\mathbf{x}$$

$$= \frac{\sigma_\theta^2}{\frac{\sigma_\theta^2 N}{2} + \sigma_w^2}\mathbf{H^T x}$$

- From the last time we looked at this example, $\mathbf{H^T x} = \Sigma_{n=0}^{N-1} \cos(\frac{2\pi kn}{N})x[n]$ or $\Sigma_{n=0}^{N-1} \sin(\frac{2\pi kn}{N})x[n]$

Comments

- Our "almost" Fourier Coefficients are:

$$\hat{a_k} = \frac{\sigma_\theta^2}{\frac{\sigma_\theta^2 N}{2} + \sigma_w^2} \Sigma_{n=0}^{N-1} \cos(\frac{2\pi kn}{N}) x[n]$$

$$\hat{b_k} = \frac{\sigma_\theta^2}{\frac{\sigma_\theta^2 N}{2} + \sigma_w^2} \Sigma_{n=0}^{N-1} \sin(\frac{2\pi kn}{N}) x[n]$$

- These look like fourier transform coefficients.
- The prior variance here is important, as is it's relative size to the noise variance.
- This derivation was possible because we assumed white noise. It gets more complicated otherwise.