bayes

Kaio Brasil

July 2024

1 Introdução

O objetivo deste arquivo é apresentar meu entendimento sobre redes bayesianas, principalmente em modelos preditores de regressão. O estudo se baseia no livro Doing Bayesian Data Analysis: A Tutorial with R and BUGS, a seguir será descito os topicos que achei mais importante no desenvolvimento do livro.

2 Regra de Bayes

Descreveremos a regra de bayes abaixo:

$$p(\theta|D) = \frac{p(D|\theta).p(\theta)}{p(D)} \tag{1}$$

Onde $p(D)$ é:

p(D)= $\int d\theta.p(D|\theta)p(\theta)$ (2)

A priori, $p(\theta)$, é a força da crença em θ sem os dados D. A posteriori, $p(\theta|D)$, é a força da crença em θ quando os dados foram tidos em conta.
A verossimilhança $p(D|\theta)$, é a probabilidade de os dados poderem ser gerados pelo modelo com valores de parâmetro θ. A evidência, $p(D)$, é a probabilidade dos dados de acordo com o modelo, determinada pela soma de todos os valores de parâmetros possíveis, ponderados pela força de crença nos valores dos parâmetros.

Abaixo encontramos a regra de bayes para duas variáveis:

$$p(\theta|D,M) = \frac{p(D|\theta,M).p(\theta|M)}{p(D|M)} \tag{3}$$

$p(\theta|D,M)$ é a função posteriori, $p(D|\theta,M)$ é a função de verossimilhança, $p(\theta|M)$ é a priori e $p(D|M)$ é a evidência dos dados.
A evidência dos dados pode ser expressa de outra forma que é bem mais comum ao longo do livro, onde $p(D|M)$ é igual a:

$$p(D|M) = \int d\theta\, p(D|\theta,M).p(\theta|M)$$

A evindência dos dados muitas vezes pode ser dificil de calcular, pois quando há muitos dados para aplicar a regra de bayes a integral na evidência tende a se complicar. Por isso, existe um metodo que faz o calculo dessas multiplas integrais o Markov Chain Monte Carlo .MCMC.
Há tambem outra questão sobre a formula de bayes e suas distribuições..
Na regra de bayes se a priori em θ segue uma distribuição normal então a posteriori em θ também segue uma distribuição normal, isto pode ser encontrado no capitulo 15 do livro sobre análise.

3 Regressão Linear

Veremos a seguir a aplicação da formúla de bayes na regressão linear.
No modelo de regressão linear usando o método de estimação OLS, há uma suposição de erro distribuído normal que é $\epsilon \sim N(0,\sigma^2)$.
Como o erro é normalmente distribuído, as variáveis $(Y|X,\beta,\sigma^2)$ também são normalmente distribuídas. Assim, as variáveis] $(Y|X,\beta,\sigma^2) \sim N(0,\sigma^2)$ e a função densidade de probabilidade (pdf) dessas variáveis são as seguintes.

$$p(Y|X,\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right] \tag{4}$$

Com base na função de densidade de probabilidade (pdf) acima, pode ser definida a função de verossimilhança dessas variáveis da seguinte forma:

$$p(Y|X,\beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right] \tag{5}$$

$$p(Y|X,\beta,\sigma^2) = (\sigma^2)^{-2/n} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right] \tag{6}$$

$$p(Y|X,\beta,\sigma^2) \propto (\sigma^2)^{\frac{v}{2}} \exp\left[-\frac{vs^2}{2\sigma^2}\right] \times \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right] \quad (7)$$

Existem várias distribuições anteriores que podem ser usadas na abordagem bayesiana do modelo de regressão linear, uma das quais é a distribuição do conjugado anterior.A estimação dos parâmetros do modelo de regressão com abordagem bayesiana pode ser feito iterando na marginal posterior. A distribuição posterior é calculada multiplicando a distribuição anterior e a função de verossimilhança.

$$Posterior \propto Likelihood \times Prior \quad (8)$$

$$p(\beta,\sigma^2|Y,X) \propto p(Y|X,\beta,\sigma^2)p(\sigma^2)p(\beta|\sigma^2) \quad (9)$$

$$p(\beta,\sigma^2|Y,X) \propto \quad (10)$$

$$\propto (\sigma^2)^{-n/2} \exp\left[-\tfrac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right] \times (\sigma^2)^{-\frac{v}{2}+1} \exp\left[-\tfrac{vs^2}{2\sigma^2}\right] \times (\sigma^2)^{-k/2} \exp\left[-\tfrac{1}{2\sigma^2}(\beta-\mu)^T(\beta-\mu)\right]$$

O processo de obtenção da estimação dos parâmetros do modelo de regressão com abordagem bayesiana pode ser feito usando o algoritmo MCMC (Markov Chain Monte Carlo). Um dos algoritmos comumente usados no MCMC é o Gibbs Sampling. Em seguida, é o processo de iteração para estimar os parâmetros até que as condições de burn-in sejam atendidas