Bayesian Statistics and Machine Learning

Gan Luan

Department of Mathematical Sciences
New Jersey Institute of Technology

October, 23, 2020

A Simple Game

• Assume there are two types of coins: one is the regular fair coin (R) and the other is the special coin (S) with both sides are heads. If we toss a coin, we we got consecutive heads. How many consecutive heads do you want to see before you are willing to bet that this is a special coin?

- ◇ 3 consecutive heads? $\mathbb{P}(3H|R) = (\frac{1}{2})^3 = 1/8.$
- ◇ 5 consecutive heads? $\mathbb{P}(5H|R) = (\frac{1}{2})^5 = 1/32.$
- ◇ 10 consecutive heads? $\mathbb{P}(10H|R) = (\frac{1}{2})^{10} = 1/1024.$

• What if you were told that the coin were picked up from a bag of 1000 coins in total and 999 of them are regular and 1 of them is the special kind? Will you still bet the coin is a special one when you see 3, 5, or 10 consecutive heads?

• Prior knowledge about the coin matters!

Bayes' Rule

- What we really cares is the probability that the coin is a regular one when we see say 10 consecutive heads? i.e. $\mathbb{P}(R|10H)$.

Bayes' Rule

Let $C_1, C_2, \cdots, C_k$ form a partition of $\mathcal{C}$, and B be another random event with $P(B) \neq 0$, then

$$\mathbb{P}(C_j|B) = \frac{\mathbb{P}(C_j \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|C_j)\mathbb{P}(C_j)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|C_j)\mathbb{P}(C_j)}{\sum_i^k \mathbb{P}(B|C_i)\mathbb{P}(C_i)}$$

Bayes' Rule

- What we really cares is the probability that the coin is a regular one when we see say 10 consecutive heads? i.e. $\mathbb{P}(R|10H)$.
- First case,

$$\mathbb{P}(R|10H) = \frac{\mathbb{P}(10H|R)\mathbb{P}(R)}{\mathbb{P}(10H|R)\mathbb{P}(R) + \mathbb{P}(10H|S)\mathbb{P}(S)}$$
$$= \frac{(1/2)^{10} \cdot 1/2}{(1/2)^{10} \cdot 1/2 + 1 \cdot 1/2} \approx 0.001$$

- Second case,

$$\mathbb{P}(R|10H) = \frac{\mathbb{P}(10H|R)\mathbb{P}(R)}{\mathbb{P}(10H|R)\mathbb{P}(R) + \mathbb{P}(10H|S)\mathbb{P}(S)}$$
$$= \frac{(1/2)^{10} \cdot 999/1000}{(1/2)^{10} \cdot 999/1000 + 1 \cdot 1/1000} \approx 0.494$$

Frequentist versus Bayesian

- Frequentists treat parameters of interest as fixed value, while Bayesian treat parameters of interest as a random variable.

- For example, for a given coin, we are interested in the probability that it appears as head when toss it (let the probability be θ). To evaluate θ, we may toss the coin for N times and counted the number of heads, say y.

For frequentist, one common estimator of θ is $\hat{\theta} = y/N$.

For Bayesian, they first assign a prior distribution to θ, $\pi(\theta)$ and given θ, we have an likelihood $f(y|\theta)$ and then by Bayes' Theorem, the posterior distribution of θ is:

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)}, \quad f(\theta|y) \propto f(y|\theta)\pi(\theta)$$

where $f(y)$ is the marginal distribution and $f(y) = \int f(y|\theta)\pi(\theta)d\theta$.

- Difficulties with Bayesian approach

Coin example

Let the prior distribution of θ be $Beta(1,1)$ and clearly $y \sim Bino(N,\theta)$, so the likelihood is

$$f(y|\theta) = \binom{N}{y}\theta^{y}(1-\theta)^{N-y}$$

and it can be shown that the posterior distribution of θ also a Beta distribution, $Beta(y+1, N-y+1)$.

Coin Example

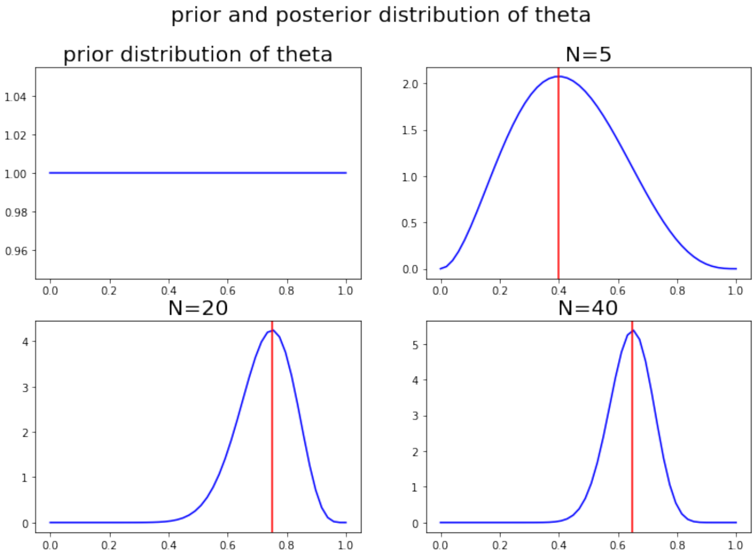


Figure: Plot of prior and posterior distribution of θ

Linear regression

Assume we have a linear regression $y = w_0 + w_1 x + \epsilon$ and $\epsilon \sim N(0, 1/\beta)$. We are interested in the unknown parameter $\boldsymbol{w} = (w_0, w_1)^T$.

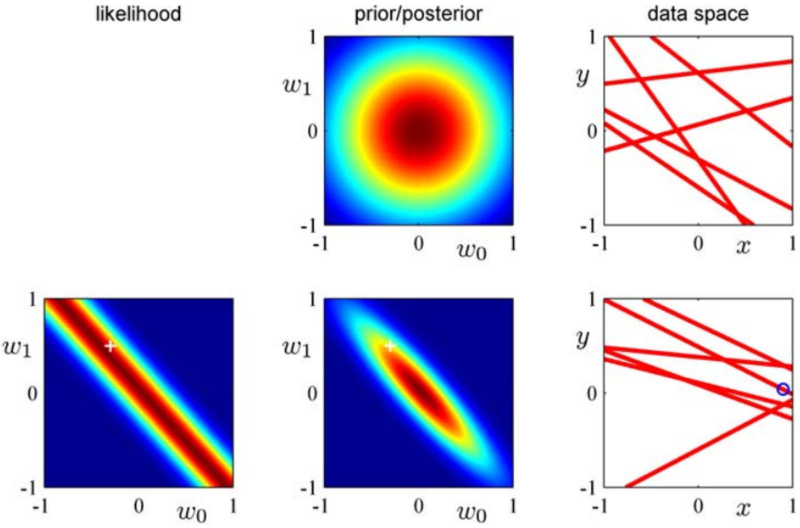
We generate synthetic data from the function $f(x, \boldsymbol{a}) = a_0 + a_1 x$ with $a_0 = -0.3$ and $a_1 = 0.5$. We first choosing values of x_n from the uniform distribution $U(x|-1, 1)$, and then evaluating $f(x_n, \boldsymbol{a})$ and finally adding Gaussian noise with standard deviation of 0.2 to obtain the target values t_n. From this data we are trying to recover the value of w_0 and w_1.

For frequentist, we could use ordinary least squares or maximum likelihood to estimate $\boldsymbol{w}$. We can also do this by Bayesian method. Assume the prior distribution of $\boldsymbol{w}$ is:

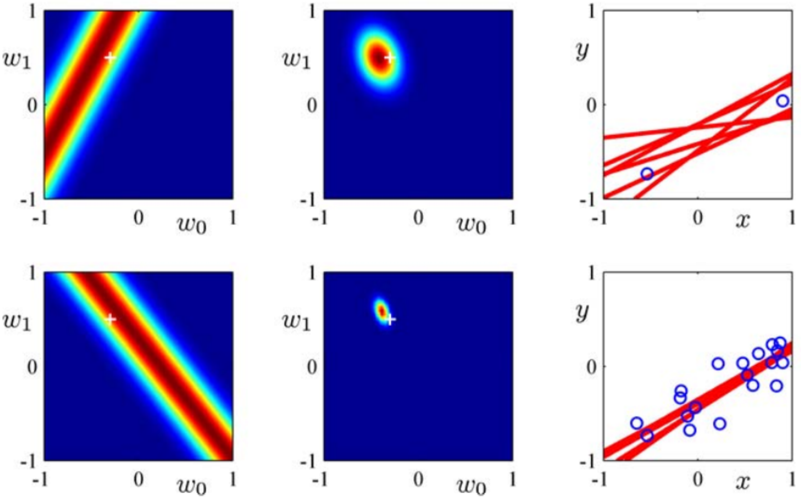
$$\boldsymbol{w} \sim N(0, 1/\alpha \boldsymbol{I})$$

The posterior distribution of $\boldsymbol{w}$ is also a Gaussian distribution.

Linear Regression with Bayesian Method



Linear Regression with Bayesian Method



Reference

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Carlin, Baradley Louis, Thomas (2008) Bayesian Methods for Data Analysis, third edition, CRC press

Thank You!