Bayesian Analysis on the Fixed Acidity In Wine

A Report submitted in

Partial Fulfilment of the Requirements for the

course of Bayesian Theory and Computation

भारतीय प्रौद्योगिकी संस्थान हैदराबाद

Indian Institute of Technology Hyderabad

Submitted to,

Dr. Arunabha Majumdar,

Assistant Professor,

Department of Mathematics.

Submitted By,

ISHITA AGRAWAL (MA24RESCH11006)

RIMA RATI PATRA (MA23MSCST11015)

Table of Contents

Data Overview

We have taken the data from UCI Machine Learning Repository;

https://archive.ics.uci.edu/static/public/186/wine+quality.zip. This data incorporates wine qualities like acidity, sugar, chlorides, sulphate, pH, alcohol etc. in white wine. We are interested in calculating the average acidity level by taking the prior and likelihood as normal distribution.

Model Specification: Likelihood and Prior

- The two key elements we need to specify in a Bayesian model are the **prior distribution** and the **likelihood function**.
- In the data we have wine's physicochemical properties like acidity, sugar, chlorides, sulphate, pH, alcohol etc. But interested in calculating average acidity level of the wine.
- So, we are splitting our dataset into two parts where first 61% of the data is considered as prior and rest as likelihood.

Likelihood

- We choose $X\,|\mu\sim$ Normal(μ,σ^2)where, μ is the average alcohol content that we have to estimate, and σ^2 is the variance that is known from the likelihood data.
- Then, if the random variable representing the alcohol content is denoted by X, the Normal distribution is given by:

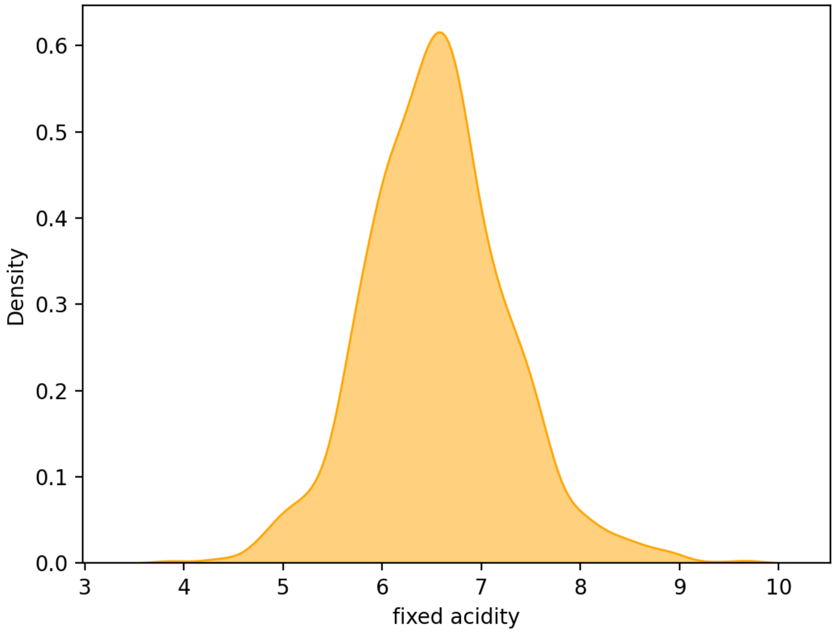
$$f(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where , $\sigma^2 > 0 \; and \; \mu \in R$

- Using MLE we have found the variance to be

$$S^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$$

Here $\sigma^2 \approx S^2$ = 0.522 and data mean $\bar{X}$= 6.552



Likelihood plot

Prior

- To finish the model specification, we need to define a prior distribution
- We consider our prior distribution to be **Normal** distribution with hyper parameters μ_0 and ${\sigma_0}^2$, and the distribution is given by:

$$f(\mu) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(\mu-\mu_0)^2}{2{\sigma_0}^2}}$$

$$\sigma^2 > 0 \; and \; \mu \in R$$

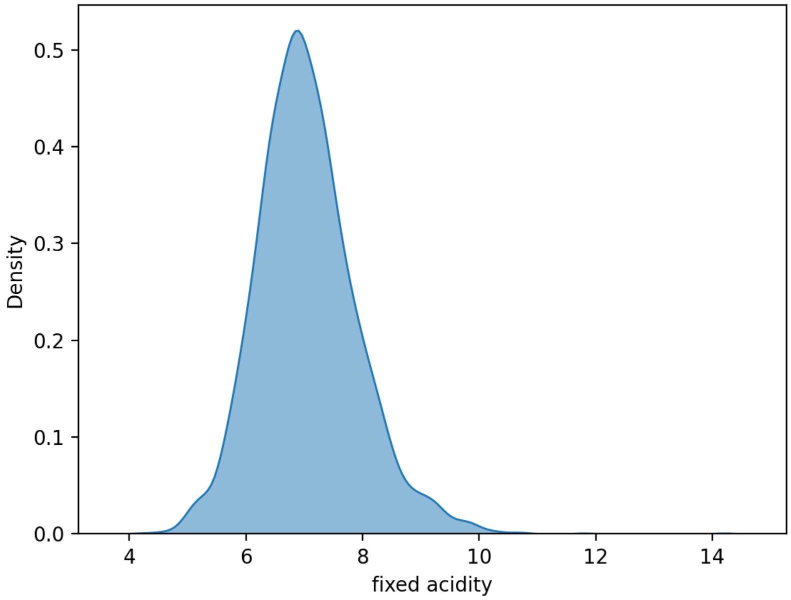
- Using MLE on the prior data

$$S^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$$

$$\overline{X} = \frac{\sum_1^n x_i}{n}$$

μ_0= 7.046
${\sigma_0}^2$= 0.738

- Thus μ~Normal(7.046,0.738)



Prior data plot

RESULT-1

We find the posterior distribution using the prior data and the likelihood function.

If, Prior : μ ~Normal(μ_0, ${\sigma_0}^2$), and Data-Likelihood :

X | μ ~Normal(μ, σ^2) then,

Posterior : μ |(X=$\overline{x}$) ~ Normal(μ_1, ${\sigma_1}^2$)

Where ${\sigma_1}^2 = \frac{1}{\frac{1}{\sigma^2}+\frac{1}{{\sigma_0}^2}}$, $\mu_1 = (\frac{x}{\sigma^2} + \frac{\mu_0}{{\sigma_0}^2}){\sigma_1}^2$

Posterior

Hence, the posterior distribution is given by:

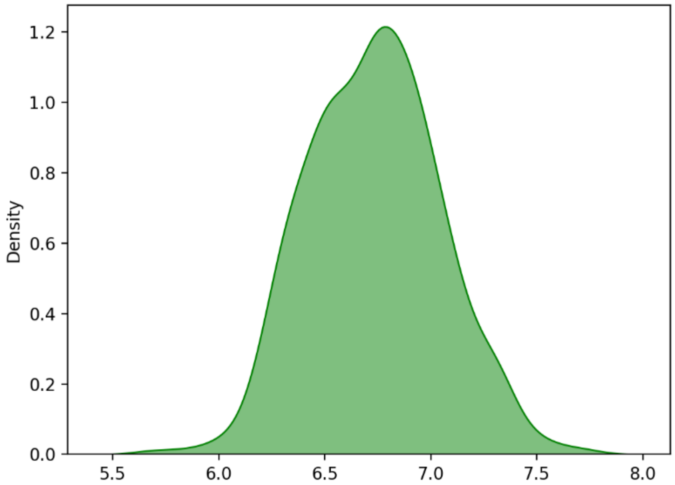
$$f(\mu|x) \propto \exp\left(-\frac{(\mu-\mu_1)^2}{2{\sigma_1}^2}\right)$$

Where,

$\mu_1 = \frac{\left(\frac{\overline{x}}{\sigma^2}+\frac{\mu_0}{{\sigma_0}^2}\right)}{\left(\frac{1}{\sigma^2}+\frac{1}{{\sigma_0}^2}\right)} = 6.74$ and

$${\sigma_1}^2 = \frac{1}{\left(\frac{1}{\sigma^2}+\frac{1}{{\sigma_0}^2}\right)} = 0.305$$

For our data, $\mu_1 = 6.74$ and ${\sigma_1}^2 = 0.305$



Posterior distribution of the parameter μ

Expectation of Posterior Distribution

The expected value of posterior distribution of μ is given by:

$$\mu_1 = \frac{\left(\frac{\overline{x}}{\sigma^2} + \frac{\mu_0}{{\sigma_0}^2}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)}$$

$$\mu_1 = \frac{\left(\frac{\overline{x}}{\sigma^2} + \frac{\mu_0}{{\sigma_0}^2}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)}$$

$$= \frac{\frac{1}{\sigma^2}}{\left(\frac{1}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)}\,\overline{x} + \frac{\frac{1}{{\sigma_0}^2}}{\left(\frac{1}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)}\mu_0$$

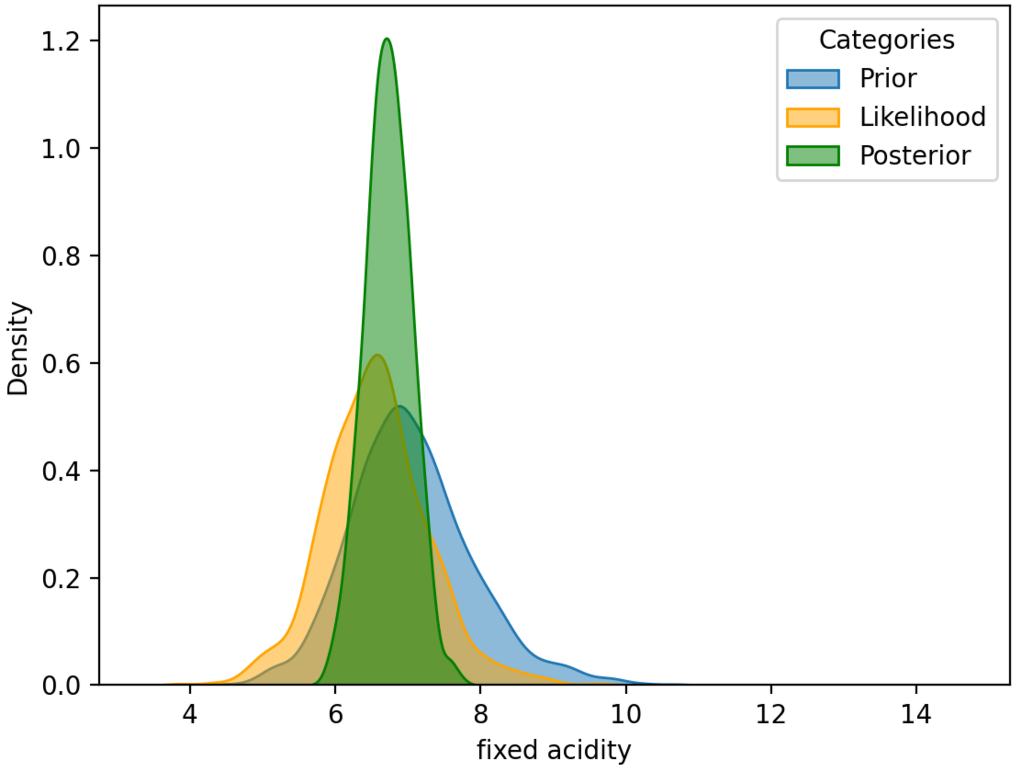
$$= w_1\,\overline{x} + \ w_2\,\mu_0$$

where, *w*1 = $\frac{\frac{1}{\sigma^2}}{\left(\frac{1}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)}$, *w*2 = $\frac{\frac{1}{{\sigma_0}^2}}{\left(\frac{1}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)}$, and *w*1 + *w*2 = 1.

Therefore, we have shown that the posterior mean is the linear combination of the data mean i.e. $\overline{X} = 6.55$ and the prior mean i.e. $\mu_0 = 7.046$

Observations

The posterior distribution represents a balance between the prior belief and the data likelihood.



Posterior distribution of the parameter μ

Model	Mean	Variance	SD
Prior	7.046	0.738	0.86
Posterior	6.74	0.522	0.72

Table: Summary of Prior and Posterior Distributions

<u>Inference:</u>

- The posterior mean of 6.74 suggests a shift from the prior mean (7.046) towards the observed data mean, reflecting the influence of the observed data on our belief about the true acidity level.
- The posterior standard deviation of 0.72, being lower than the prior standard deviation (0.86), indicates increased certainty in our estimate of the true acidity level due to the influence of the observed data.