Implementation of MLE, MAP, Bayesian Estimation

Jayant Sharma, Rakesh Gorrepati, Sreeharsha Mourila,Saurabh Vemuri

November 26, 2017

Outline

- Introduction
- MLE vs MAP
- Bayesian Estimation
- Extending to classes
- Simulation Results

Introduction

- Several estimation Techniques are used to estimate unknown parameters from some observations
- MLE is one of the best practically realizable estimator which is sub optimal in nature
- However MLE doesn't allow us to include any prior knowledge on the unknown parameter.
- This fundamental drawback is solved by using assuming the unknown parameter to be a random variable
- And modelling this our prior information of in some prior pdf $P(\theta)$and calculating posterior after observing the data samples $P(\theta/D)$

MAP and Bayesian Estimation

- The probability density function is our entire knowledge on the unknown parameter θ after observing the data
- Minimizing the Bayesian MMSE error will gives the estimate to be the mean of $P(\theta/D)$.
- Finding the θ values that maximizes the posterior probability that is $P(\theta/D)$ will give us the map estimate of the parameter
- For Gaussian both the Mean and mode of the posterior turns out to be same
- To easily compute $P(\theta/D)$ we will choose some conjugate priors so that we know before that $P(\theta/D)$ will be in the same form as the prior.
- $P(\theta|D) = \dfrac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$

MLE and Map

The value of θ that maximizes the log-likelihood is the MLE estimate

This Likelihood estimate is calculated by solving $\frac{\partial}{\partial\theta} lnP(D|\theta) = 0$

For Gaussian source the MLE estimate for the mean and variance turns out to be sample mean and sample variance

For Exponential distribution the MLE estimate of rate parameter is $\frac{1}{\sum_{i=1}^{N} x_i}$

MAP Estimate can be derived as $P(y_i|x) = \dfrac{P(x|y)P(y)}{\sum_{i=1}^{K} P(x|y_i)P(y_i)}$ where y denotes the class label and x denotes the data observed

so MLE is special case of MAP when all classes are equally probable

Bayesian Parameter Estimation

example of the Bayesian technique:

$$P(\frac{X}{\mu}) \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

$$P(\mu) \sim \mathcal{N}(\mu_o, {\sigma_o}^2) \tag{2}$$

$$P(\mu) = \frac{1}{\sqrt{2\pi{\sigma_o}^2}} e^{-\frac{(\mu-\mu_o)^2}{2{\sigma_o}^2}} \tag{3}$$

$$P(\frac{\mu}{D}) = \frac{P(\frac{D}{\mu})P(\mu)}{\int P(\frac{D}{\mu})P(\mu)d\mu} \tag{4}$$

$$= \alpha \prod_{1}^{n} P(\frac{x_k}{\mu})P(\mu) \tag{5}$$

$$where \;\; \alpha = \frac{1}{\int P(\frac{D}{\mu})P(\mu)d\mu} \tag{6}$$

Bayesian Parameter Estimation Contd ..

$$= \alpha' \, exp[\frac{-1}{2}(\frac{n}{\sigma_o^2} + \frac{1}{\sigma_o^2})\mu^2 - \frac{1}{2}(\frac{1}{\sigma_o^2}\sum_{k=1}^{n} x_k + \frac{\mu_o}{\sigma_o^2})\mu] \tag{12}$$

$$P(\frac{\mu}{D}) = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}} \tag{13}$$

$$P(\frac{\mu}{D}) \sim \mathcal{N}(\mu_n, \sigma_n^2) \tag{14}$$

$$P(\frac{\mu}{D}) = \alpha exp(-\frac{\mu^2}{2\sigma_n^2} + \frac{\mu_n^2}{2\sigma_n^2} - 2\frac{\mu\mu_n}{2\sigma_n^2}) \tag{15}$$

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_o^2} \tag{16}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{n\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_o}{\sigma_o^2} \tag{17}$$

Bayesian Parameter Estimation Contd ..

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{n\sigma^2} \sum_{k=1}^{n} x_k + \frac{\mu_o}{\sigma_o^2} \tag{18}$$

$$\bar{x}_n = \frac{\sum_{k=1}^{n} x_k}{n} \tag{19}$$

$$\mu_n = \frac{n\sigma_o^2}{n\sigma_o^2 + \sigma^2} \bar{x}_n + \frac{\sigma^2}{n\sigma_o^2 + \sigma^2} \mu_o \tag{20}$$

$$\sigma_n^2 = \frac{\sigma_o^2 \sigma^2}{n\sigma_o^2 + \sigma^2} \tag{21}$$

Exponential pdf

$$P(x|\theta) = \begin{cases} \theta e^{-\theta x} & : x > 0 \\ x^{\alpha-1}\ \theta e^{-\theta x} & : \alpha = 1 \\ x^{\alpha-1}\beta e^{-\beta x} & : \beta = 1 \ and \ \theta = \beta \end{cases}$$

$\Gamma(\alpha) = x^{\alpha-1}\beta e^{-\beta x}$

Exponential pdf is a subspace of Gamma pdf.

$P(D|\theta) = \prod_{i=1}^{\infty} \theta e^{-\theta} x_i$ for $D = \{x_1, x_2, ..., x_N\}$

$= \theta^N e^{-\theta \sum_{i=1}^{N} x_i}$ for$x_i > 0$

Let $\sum_{i=1}^{N} x_i = s$ then,

$P(D|\theta) = \theta^N e^{-\theta s}$ where $p(\theta) = \lambda e^{-\lambda\theta}$ for $\theta > 0$

According to Bayes' Theorem,

$P(\theta|D) = \dfrac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)\ d\theta}$

Exponential pdf Contd

$$P(\theta|D) = \frac{\theta^N e^{-\theta s} \lambda e^{-\lambda\theta}}{\int_0^\infty \theta^N e^{-\theta s} \lambda e^{-\lambda\theta} d\theta}$$

$$P(\theta|D) = \frac{\theta^N e^{-\theta(s+\lambda)}}{\int_0^\infty \theta^N e^{-\theta(s+\lambda)} d\theta}$$

Let $\theta(s+\lambda) = t$

$$d\theta = \frac{dt}{s+\lambda} \text{ and } \theta^N = \frac{t^N}{(s+\lambda)^N}$$

$$P(\theta|D) = \frac{\theta^N e^(s+\lambda)\theta}{\frac{1}{(s+\lambda)^{N+1}} \int_0^\infty t^N e^{-t} dt}$$

By definition, $\Gamma(N+1) = \int_0^\infty x^N e^{-x} dx$

So, $P(\theta|D) = \dfrac{\theta^N e^{(s+\lambda)\theta} (\lambda+s)^{N+1}}{\Gamma(N+1)}$

Comparision with Gamma pdf

$$f(\theta, N, (s+\lambda)) = \frac{\theta^{(N+1)-1}(\lambda+s)^{N+1}e^{(s+\lambda)\theta}}{\Gamma(N+1)}$$

$$P(x|D) = \int_0^\infty P(x|\theta)P(\theta|D)d\theta$$

$$P(x|D) = \int_0^\infty \frac{\theta e^{-\theta x}\theta^N e^{-(s+\lambda)}(\lambda+s)^{N+1}d\theta}{\Gamma(N+1)}$$

$$P(x|D) = \frac{(\lambda+s)^{N+1}}{\Gamma(N+1)}\int_0^\infty \theta^{(N+1)}e^{-\theta(x+s+\lambda)}d\theta$$

Let $\theta(x+s+\lambda) = t$, then $d\theta = \frac{dt}{x+s+\lambda}$

$$\theta^{N+1} = \frac{t^{N+1}}{(x+s+\lambda)^N+1}$$

$$P(x|D) = \frac{(\lambda+s)^{N+1}}{\Gamma(N+1)}\int_0^\infty \frac{t^{N+1}e^{-t}dt}{(\lambda+s+x)^{N+2}}$$

$$P(x|D) = \frac{(\lambda+s)^{N+1}(\lambda+s+x)^{N+2}}{\Gamma(N+1)}\int_0^\infty t^{N+1}e^{-t}dt$$

By definition we have $t^{N+1}e^{-t} = \Gamma(N+2)$

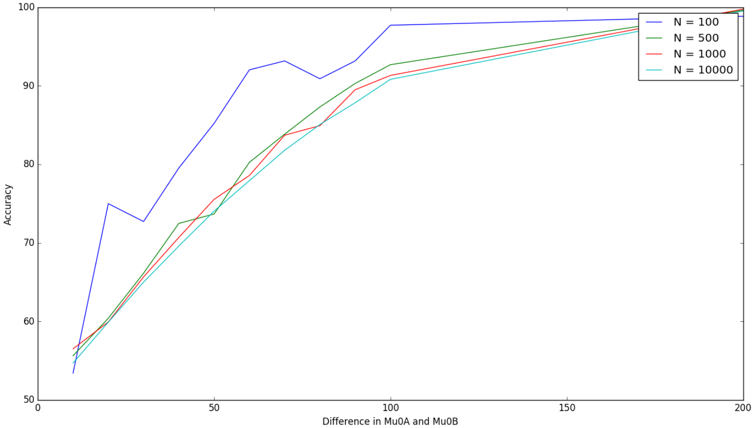
$$P(x|D) = \frac{(\lambda+s)^{N+1}}{(\lambda+s+x)^{N+2}} \frac{\Gamma(N+2)}{\Gamma(N+1)}$$

Also for integers $\Gamma(N) = (N-1)!$ So, $P(x|D) = \dfrac{(\lambda+s)^{N+1}(N+1)}{(\lambda+s+x)^{N+2}}$

The testing rule for D_1 and D_2 is :

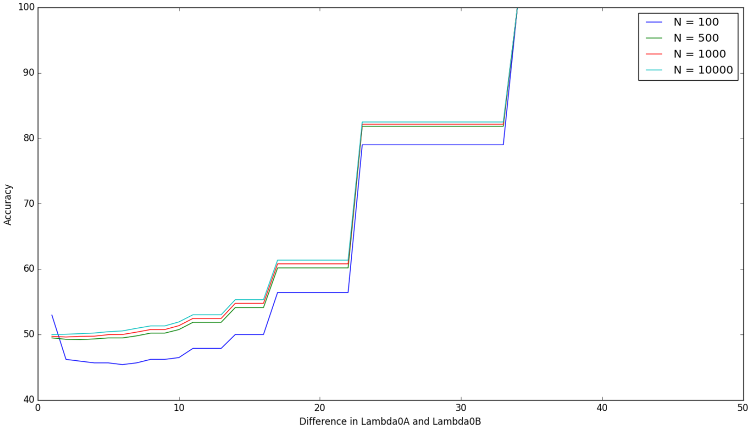
$$\frac{P(x|D_1)}{P(x|D_2)} = \left(\frac{\lambda_2+s_2+x}{\lambda_1+s_1+x}\right)^{N+1} \left(\frac{\lambda_1+s_1}{\lambda_2+s_2}\right)^{N+1}$$

Simulation



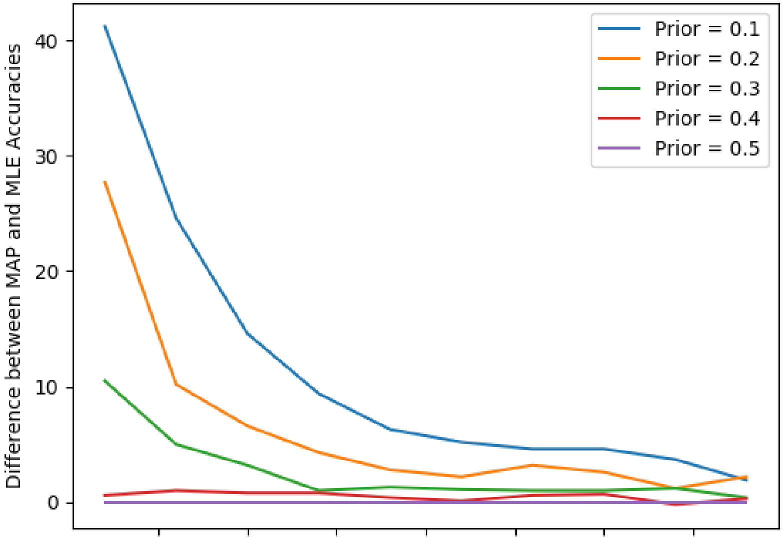
N = 100
N = 500
N = 1000
N = 10000
Accuracy
Difference in Mu0A and Mu0B
100
90
80
70
60
50
0
50
100
150
200

Simulation



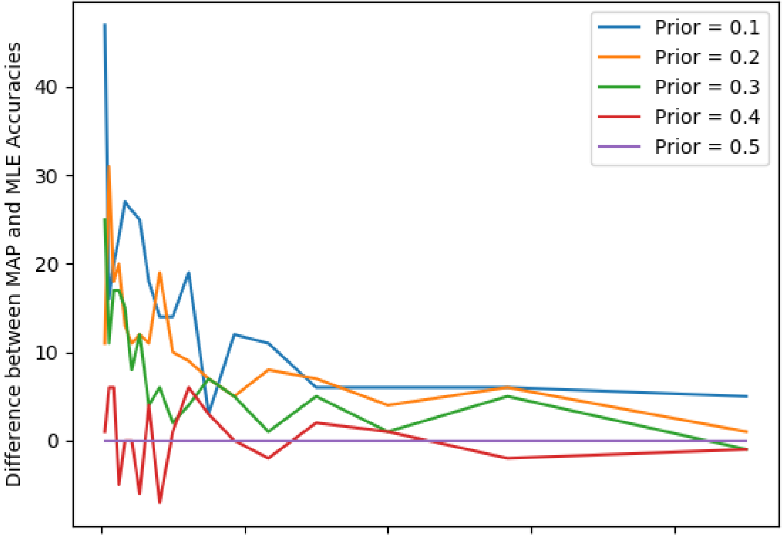
100
90
80
70
60
50
40
0
10
20
30
40
50
Accuracy
Difference in Lambda0A and Lambda0B
N = 100
N = 500
N = 1000
N = 10000

Simulation



Prior = 0.1
Prior = 0.2
Prior = 0.3
Prior = 0.4
Prior = 0.5
Difference between MAP and MLE Accuracies
0
10
20
30
40

Simulation



Difference between MAP and MLE Accuracies
40
30
20
10
0
Prior = 0.1
Prior = 0.2
Prior = 0.3
Prior = 0.4
Prior = 0.5