Maximum A Posteriori Estimation

Nipun Batra

August 21, 2023

IIT Gandhinagar

Agenda

Revision

Coin Toss Problem

Univariate Normal Distribution

MAP for Linear Regression

MAP for Logistic Regression

Revision

Bayes Rule

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

- $P(\theta|D)$ is called the posterior
- $P(D|\theta)$ is called the likelihood
- $P(\theta)$ is called the prior
- $P(D)$ is called the evidence

Maximum Likelihood Estimation

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Given a dataset D, find the parameters θ that maximize the likelihood of the data.

$$\theta_{\text{MLE}} = \arg\max_\theta P(D|\theta)$$

For example, given a linear regression problem setup, we set the likelihood as normal distribution and find the parameters θ that maximize the likelihood of the data.

Maximum A Posteriori Estimation

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} = \frac{P(D|\theta) \cdot P(\theta)}{\int_\theta P(D|\theta) \cdot P(\theta) d\theta}$$

Given a dataset D, find the parameters θ that maximize the posterior of θ considering both the likelihood and the prior.

$$\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$$

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?
- **MAP**: Given N observations and prior knowledge, obtain best θ estimate (or θ_{MAP})

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?
- **MAP**: Given N observations and prior knowledge, obtain best θ estimate (or θ_{MAP})
- When do we need prior knowledge?

Maximum A Posteriori Estimation

- **MLE**: Given N observations, obtain best θ estimate (or θ_{MLE})
- What if we have prior knowledge about θ?
- **MAP**: Given N observations and prior knowledge, obtain best θ estimate (or θ_{MAP})
- When do we need prior knowledge?
    - When the dataset is not a good representation of the true distribution.
    - Can be a data quality and/or quantity issue.

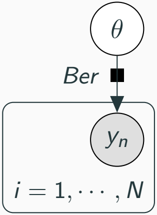
Coin Toss Problem

Coin Toss Problem

- Consider a sequence of independent N coin toss outcomes, $D = \{y_1, ..., y_N\}$ where each observation y_i is a binary random variable (Heads: 1, Tails: 0).

Coin Toss Problem

- Consider a sequence of independent N coin toss outcomes, $D = \{y_1, ..., y_N\}$ where each observation y_i is a binary random variable (Heads: 1, Tails: 0).
- Assuming $y_i \sim \text{Bernoulli}\,(\theta)$, $P(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$



Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial\theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H + n_T}$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H + n_T}$
- Suppose 10 tosses yield 9 heads and 1 tail. $\theta_{MLE} =$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial\theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H + n_T}$
- Suppose 10 tosses yield 9 heads and 1 tail. $\theta_{MLE} = 0.9$

Coin Toss Problem

- For the sequence: $P(D|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i}$
- **Recall**: $P(D|\theta) \longrightarrow$ Likelihood or $\mathcal{L}(\theta)$
- Log-Likelihood or $\mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta)$
- **Recall**: $\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$

$$\therefore \frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \implies \theta_{MLE} = \frac{\sum_{i=1}^{N} y_i}{N}$$

- Rewrite, $\theta_{MLE} = \frac{n_H}{n_H + n_T}$
- Suppose 10 tosses yield 9 heads and 1 tail. $\theta_{MLE} = 0.9$
- What if we have prior knowledge that the coin is fair?

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

$$\because P(\mathit{Head}) = \theta \in [0, 1]$$

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

$$\because P(\mathit{Head}) = \theta \in [0, 1]$$

- A resonable choice for prior is the Beta distribution.

Incorporating Prior Information

- We can incorporate prior information by assuming a prior distribution over θ.

$$\because P(Head) = \theta \in [0, 1]$$

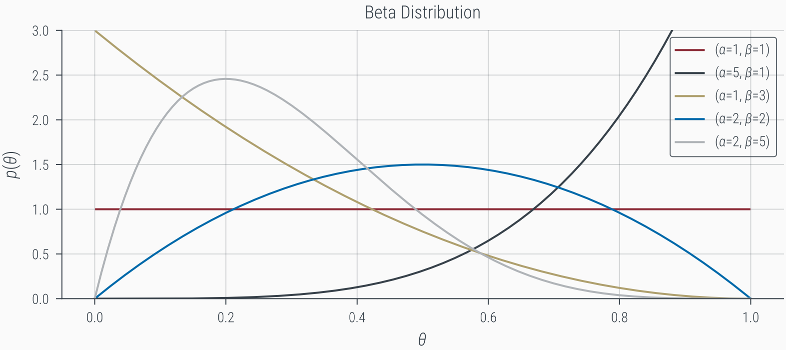
- A resonable choice for prior is the Beta distribution.

$$\implies P(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where,

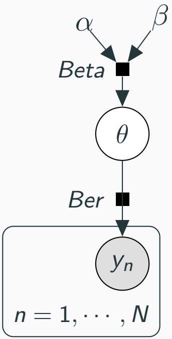
$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \ \text{(Gamma Function)}$$

Beta Distribution



Notebook

Coin Toss Problem with Prior



Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$

Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$
- The log-posterior for this coin-toss problem is given as,

Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$
- The log-posterior for this coin-toss problem is given as,

$$\log P(\theta|D) = \sum_{i=1}^{N} \log P(y_i|\theta) + \log P(\theta)$$

Deriving θ_{MAP}

- **Recall**: $\theta_{\text{MAP}} = \arg\max_\theta P(\theta|D) = \arg\max_\theta P(D|\theta) \cdot P(\theta)$
- The log-posterior for this coin-toss problem is given as,

$$\log P(\theta|D) = \sum_{i=1}^{N} \log P(y_i|\theta) + \log P(\theta)$$

$$\log P(\theta|D) = \sum_{i=1}^{N} y_i \log\theta + (1-y_i)\log(1-\theta) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

Deriving θ_{MAP}

$$\frac{\partial \log P(\theta|D)}{\partial \theta} = \frac{\sum_{i=1}^{N} y_i}{\theta} - \frac{\sum_{i=1}^{N}(1-y_i)}{1-\theta} + \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$$

$$\implies (1-\theta)\sum_{i=1}^{N} y_i + \theta\sum_{i=1}^{N}(1-y_i) + (1-\theta)(\alpha-1) - \theta(\beta-1) = 0$$

$$\implies \sum_{i=1}^{N} y_i - \theta\sum_{i=1}^{N} y_i - N\theta + \theta\sum_{i=1}^{N} y_i + \alpha - 1 - \theta\alpha + \theta - \theta\beta + \theta = 0$$

Deriving θ_{MAP}

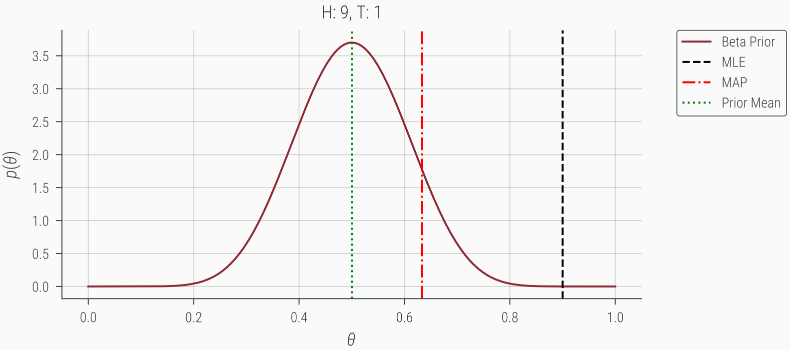
$$\implies \sum_{i=1}^{N} y_i + \alpha - 1 - \theta(N + \alpha + \beta - 2) = 0$$

$$\implies \theta_{MAP} = \frac{\sum_{i=1}^{N} y_i + \alpha - 1}{N + \alpha + \beta - 2}$$

Deriving θ_{MAP} (Coin toss context)

- Total number of tosses $= N$
- Number of heads $= n_H$
- Number of tails $= n_T$

Coin Toss Problem with Prior



Notebook

Univariate Normal Distribution

MAP for Normal Distribution

To estimate MAP for Normal Distribution, we can have the following 3 cases:

1. unknown μ, known σ^2
2. known μ, unknown σ^2
3. unknown μ, unknown σ^2

unknown μ, known σ^2

- Consider a sequence of independent N observations, $D = \{x_1, ..., x_N\}$ drawn from $\mathcal{N}(x_i|\mu, \sigma^2)$

unknown μ, known σ^2

- Consider a sequence of independent N observations, $D = \{x_1, ..., x_N\}$ drawn from $\mathcal{N}(x_i|\mu, \sigma^2)$
- Likelihood is given by (Note: only μ is a random variable, σ^2 is known and assumed fixed)
  $P(D|\mu, \sigma^2) = \mathcal{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

unknown μ, known σ^2

- Consider a sequence of independent N observations, $D = \{x_1, ..., x_N\}$ drawn from $\mathcal{N}(x_i|\mu, \sigma^2)$
- Likelihood is given by (Note: only μ is a random variable, σ^2 is known and assumed fixed)
  $P(D|\mu, \sigma^2) = \mathcal{L}(\mu) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$
- Log-Likelihood is given by

$$\log P(D|\mu, \sigma^2) = \mathcal{LL}(\mu) = \sum_{i=1}^{N} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\implies \mathcal{LL}(\mu) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

Obtaining μ_{MLE}

- For MLE for μ, we set

$$\frac{\partial \mathcal{LL}(\mu)}{\partial \mu} = 0 - \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(x_i - \mu) \right) = \frac{1}{\sigma^2} (\sum_{i=1}^{N} x_i - N\mu) = 0$$

or

$$\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Obtaining μ_{MLE}

- For MLE for μ, we set

$$\frac{\partial \mathcal{L}\mathcal{L}(\mu)}{\partial \mu} = 0 - \left(- \frac{1}{2\sigma^2} \sum_{i=1}^{N} 2(x_i - \mu) \right) = \frac{1}{\sigma^2} (\sum_{i=1}^{N} x_i - N\mu) = 0$$

or

$$\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- However, similar to Coin Toss problem, this is prone to overfit.

Incorporating Prior Information

- Since we need a prior over μ, we can choose $P(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$

Incorporating Prior Information

- Since we need a prior over μ, we can choose $P(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
- The Posterior for μ is given by

$$P(\mu|D) \propto P(D|\mu)P(\mu) \propto \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

Incorporating Prior Information

- Since we need a prior over μ, we can choose $P(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
- The Posterior for μ is given by

$$P(\mu|D) \propto P(D|\mu)P(\mu) \propto \prod_{i=1}^{N} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)$$

- Simplifying, we get

$$P(\mu|D) \propto \exp\left(-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right)$$

where,

$$(\mu_N, \sigma_N) = \left(\frac{\frac{\sigma^2}{N}}{\sigma_0 + \frac{\sigma^2}{N}} + \frac{\sigma_0^2}{\sigma_0 + \frac{\sigma^2}{N}} \frac{\sum_{i=1}^{N} x_i}{N}, \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}\right)^{-1}\right)$$

Obtaining MAP

- For MAP, we set

$$\frac{\partial \log P(\mu|D)}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} -2(x_i - \mu) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{N} 2(\mu - \mu_0) = 0$$

Obtaining MAP

- For MAP, we set

$$\frac{\partial \log P(\mu|D)}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} -2(x_i-\mu) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{N} 2(\mu-\mu_0) = 0$$

$$\implies \frac{1}{\sigma^2}(\sum_{i=1}^{N} x_i - N\mu) - \frac{N}{\sigma_0^2}(\mu-\mu_0) =$$

$$\mu\left(-\frac{N}{\sigma^2}-\frac{N}{\sigma_0^2}\right) + \frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{N\mu_0}{\sigma_0^2} = 0$$

Obtaining MAP

- For MAP, we set

$$\frac{\partial \log P(\mu|D)}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{N} -2(x_i - \mu) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{N} 2(\mu - \mu_0) = 0$$

$$\implies \frac{1}{\sigma^2}(\sum_{i=1}^{N} x_i - N\mu) - \frac{N}{\sigma_0^2}(\mu - \mu_0) =$$

$$\mu\left(-\frac{N}{\sigma^2} - \frac{N}{\sigma_0^2}\right) + \frac{\sum_{i=1}^{N} x_i}{\sigma^2} + \frac{N\mu_0}{\sigma_0^2} = 0$$

$$\mu_{MAP} = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{\frac{\sum_{i=1}^{N} x_i}{N}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}} = \frac{\sigma^2\mu_0 + \sigma_0^2\frac{\sum_{i=1}^{N} x_i}{N}}{\sigma_0^2 + \sigma^2}$$

known μ, unknown σ^2

Assuuming μ is known, the conjugate prior for σ^2 is Inverse Gamma(α_0, β_0) which gives,

$$P(\sigma^2|\alpha_0, \beta_0) \propto \frac{1}{(\sigma^2)^{\alpha_0+1}} \exp\left(-\frac{\beta_0}{\sigma^2}\right)$$

$\therefore$ The posterior is given by,

$$P(\sigma^2|D; \alpha_0, \beta_0) \sim \text{Inverse Gamma}\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{\sum_{i=1}^{n}(x_i - \mu)}{2}\right)$$

Assuuming both μ and σ^2 are unknown, the conjugate prior for μ and σ^2 (or Precision $\tau = \frac{1}{\sigma^2}$) is as follows,

$$
\begin{aligned}
D|\mu, \tau &\sim \mathcal{N}(\mu, \tau^{-1}) \\
\mu|\tau &\sim \mathcal{N}(\mu_0, (\kappa_0 \tau)^{-1}) \\
\tau &\sim \text{Gamma}(\alpha_0, \beta_0)
\end{aligned}
$$

$\therefore$ The posterior is given by,

$$
\begin{aligned}
\mu|D, \tau &\sim \mathcal{N}\left(\frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_0 + n}, (\kappa_0 + n)^{-1}\right) \\
\tau|D &\sim \text{Gamma}\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{\kappa_0 n(\bar{x} - \mu_0)^2}{2(\kappa_0 + n)}\right)
\end{aligned}
$$

MAP for Linear Regression

MLE for Linear Regression

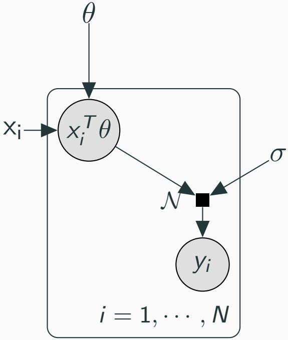
- Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

MLE for Linear Regression

- Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
- Suppose the data is generated from a linear model with additive Gaussian noise, i.e., $y_i = \theta^T x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

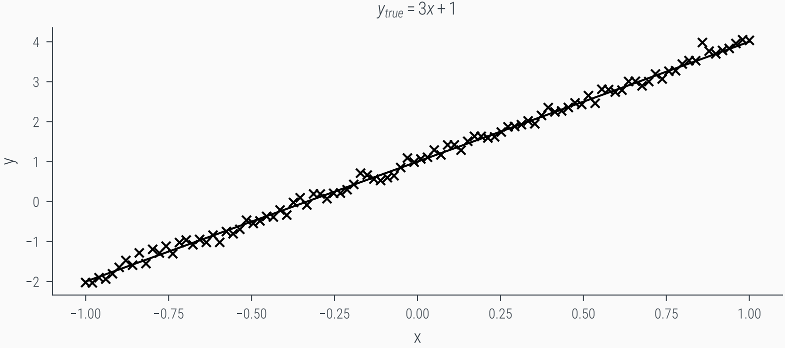
MLE for Linear Regression

- Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
- Suppose the data is generated from a linear model with additive Gaussian noise, i.e., $y_i = \theta^T x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.



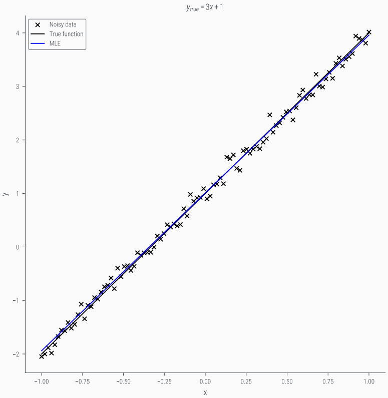
MLE for Linear Regression

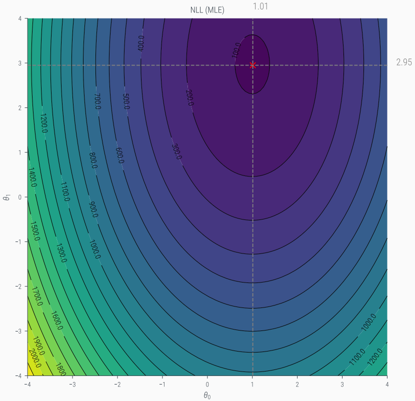
- Consider a dataset $D = \{(x_1, y_1) ... (x_N, y_N)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.
- Suppose the data is generated from a linear model with additive Gaussian noise, i.e., $y_i = \theta^T x_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.



MLE for Linear Regression

- The likelihood is given by, $P(y_i|x_i,\theta) = \mathcal{N}(y_i|\theta^T x_i, \sigma^2)$
- **Recall**: The negative log-likelihood is given by,
  $\mathcal{NLL}(\theta) = \frac{1}{2\sigma^2}(y - X\theta)^T (y - X\theta)$
- **Recall**: The MLE is given by,
  $\theta_{MLE} = \arg\min_\theta \mathcal{NLL}(\theta) = \left(X^T X\right)^{-1} X^T y$



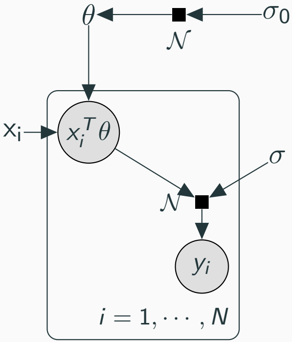


Using zero-mean Gaussian prior

Considering a zero-mean Gaussian prior on the weights, i.e., $P(\theta) = \mathcal{N}(\theta|0, \sigma_0^2)$, we have

$P(\theta|D) \propto P(D|\theta)P(\theta)$

$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$



Using zero-mean Gaussian prior

Rewrite

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

Using zero-mean Gaussian prior

Rewrite

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Using zero-mean Gaussian prior

Rewrite

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Question

What does this expression remind you of?

Using zero-mean Gaussian prior

Rewrite

$$\theta_{MAP} = \arg\min \log P(\theta|D) = \arg\min \mathcal{NLL}(\theta) + \log P(\theta)$$

We get

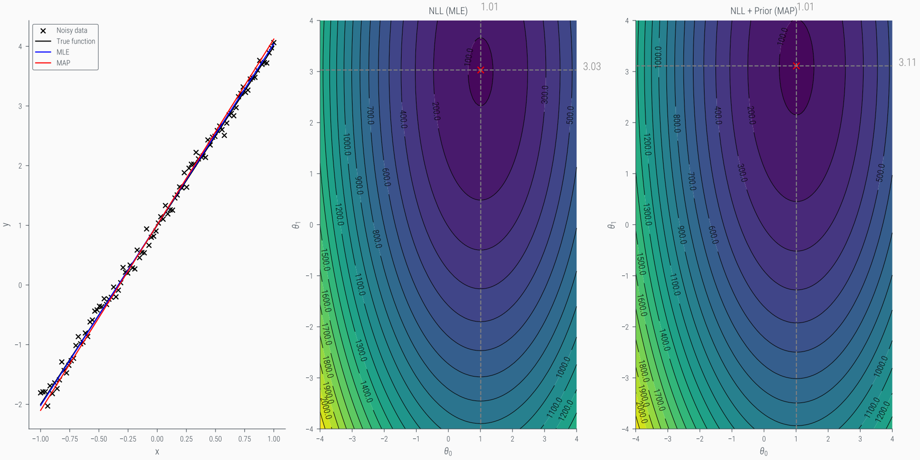
$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Question

What does this expression remind you of?

Answer: Ridge Regression

Using zero-mean Gaussian prior

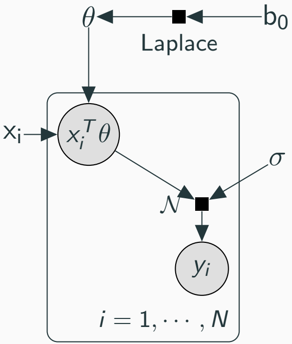


Notebook

Using Laplace prior

We can also use a Laplace prior on the weights, i.e.,

$$P(\theta) = \frac{1}{2b_0} \exp\left(-\frac{|x-\mu|}{b_0}\right)$$



Using Laplace prior

The MAP takes the form,

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{b_0}|\theta_i|$$

Using Laplace prior

The MAP takes the form,

$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{b_0}|\theta_i|$$

Question

What does this expression remind you of?

Using Laplace prior

The MAP takes the form,

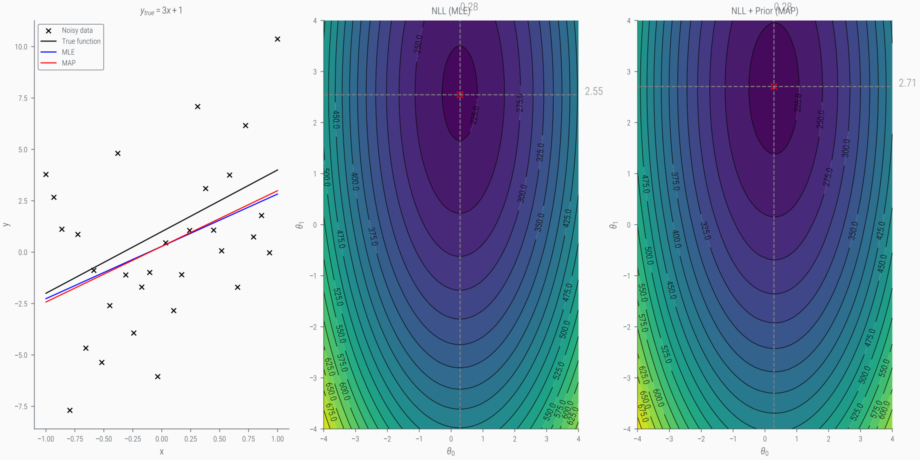
$$\theta_{MAP} = \arg\min \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta) + \frac{1}{b_0}|\theta_i|$$

Question

What does this expression remind you of?

Answer: Lasso Regression

Using Laplace prior



Notebook

MAP for Logistic Regression

MLE for Logistic Regression

Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ such that

$$P(y = 1|x) = \hat{y} = \frac{1}{1 + \exp(-X^T\theta)} = \sigma(X^T\theta)$$

Take $y \sim \text{Bernoulli}\left(\sigma(X^T\theta)\right)$

MLE for Logistic Regression

Consider a dataset $D = \{(x_1, y_1)...(x_N, y_N)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$ such that

$$P(y = 1|x) = \hat{y} = \frac{1}{1 + \exp(-X^T\theta)} = \sigma(X^T\theta)$$

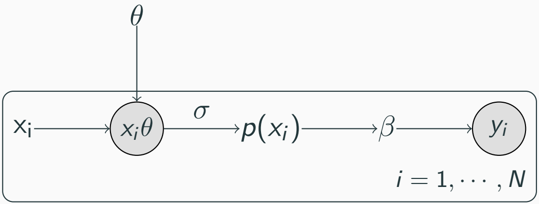
Take $y \sim \text{Bernoulli}\left(\sigma(X^T\theta)\right)$

The likelihood is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \hat{y}_i^{y_i}(i - \hat{y}_i)^{1-y_i}$$

$$\implies \mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

MLE for Logistic Regression



Binary Classification:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X)}}$$

$$\therefore \mathcal{LL}(\theta) = \sum_{i=1}^{N} y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))$$

Using zero-mean Gaussian prior

Considering a zero-mean Gaussian prior on the weights, i.e., $P(\theta) = \mathcal{N}(\theta|0, \sigma_0^2)$, the MAP is given by,

$$\theta_{MAP} = \arg\min \log(1 + \exp(-\theta^T X)) + \frac{1}{\sigma_0^2}\theta^T\theta$$

Using Laplace prior

Considering a Laplace prior on the weights, i.e., $P(\theta) = \prod_D \text{Laplace}(\theta_i | 0, b_0) \propto \prod_D \exp(-\frac{1}{b_0}|\theta_i|)$, the MAP is given by,

$$\theta_{MAP} = \arg\min \log(1 + \exp(-\theta^T X)) + \frac{1}{b_0}|\theta|$$

MAP for Logistic Regression

Self-Study: Modify the code for Linear Regression to implement MAP for Logistic Regression.