MLE for Linear Regression

Nipun Batra

February 6, 2020

IIT Gandhinagar

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Notation: Let θ denote the parameters of the model and let $\mathcal{D}$ denote observed data. From Bayes Rule, we have

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

Bayes Rule

- Bayes Rule: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- Notation: Let θ denote the parameters of the model and let $\mathcal{D}$ denote observed data. From Bayes Rule, we have

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

- In the above equation $P(\theta|\mathcal{D})$ is called the posterior, $P(\mathcal{D}|\theta)$ is called the likelihood, $P(\theta)$ is called the prior and $P(\mathcal{D})$ is called the evidence.
- An example of a prior probability would be $\theta \sim \mathcal{N}(0, \mathcal{I}_n)$. The prior acts as a *regularizer as we will see.*

MLE for Linear Regression

- $\hat{\theta}_{LS} = \arg\min \epsilon^T \epsilon = \arg\min_\theta (y - X\theta)^T (y - X\theta)$.

MLE for Linear Regression

- $\hat{\theta}_{LS} = \arg\min \epsilon^T \epsilon = \arg\min_\theta (y - X\theta)^T (y - X\theta)$.
- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.

MLE for Linear Regression

- $\hat{\theta}_{LS} = \arg\min \epsilon^T \epsilon = \arg\min_\theta (y - X\theta)^T (y - X\theta)$.
- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.
- Let ϵ be independent across all the observations.

MLE for Linear Regression

- $\hat{\theta}_{LS} = \arg\min \epsilon^T \epsilon = \arg\min_\theta (y - X\theta)^T (y - X\theta)$.
- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.
- Let ϵ be independent across all the observations.
- $y \sim \mathcal{N}(X\theta, \sigma^2)$, since $\{X\theta + \mathcal{N}(0, \sigma^2)\}$

MLE for Linear Regression - Likelihood

- Likelihood $= P(\mathcal{D}|\theta)$. Data $= < x_i, y_i >$, Parameters $= \theta, \sigma$.

MLE for Linear Regression - Likelihood

- Likelihood $= P(\mathcal{D}|\theta)$. Data $= < x_i, y_i >$, Parameters $= \theta, \sigma$.
- $P(y|X, \theta, \sigma) = P(y_1|x_1, \theta, \sigma) \times P(y_2|x_2, \theta, \sigma) \dots$

MLE for Linear Regression - Likelihood

- Likelihood $= P(\mathcal{D}|\theta)$. Data $= < x_i, y_i >$, Parameters $= \theta, \sigma$.
- $P(y|X, \theta, \sigma) = P(y_1|x_1, \theta, \sigma) \times P(y_2|x_2, \theta, \sigma) \dots$
- $\implies$

$$\prod_{i=1}^{n} P(y|X, \theta, \sigma)$$

MLE for Linear Regression - Likelihood

- Likelihood $= P(\mathcal{D}|\theta)$. Data $= < x_i, y_i >$, Parameters $= \theta, \sigma$.
- $P(y|X, \theta, \sigma) = P(y_1|x_1, \theta, \sigma) \times P(y_2|x_2, \theta, \sigma) \dots$
- $\implies$
$$\prod_{i=1}^{n} P(y|X, \theta, \sigma)$$
- $\implies$
$$\prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y - x_i\theta)^2}$$

MLE for Linear Regression - Likelihood

- Likelihood $= P(\mathcal{D}|\theta)$. Data $= < x_i, y_i >$, Parameters $= \theta, \sigma$.
- $P(y|X, \theta, \sigma) = P(y_1|x_1, \theta, \sigma) \times P(y_2|x_2, \theta, \sigma) \ldots$
- $\implies$
$$\prod_{i=1}^{n} P(y|X, \theta, \sigma)$$
- $\implies$
$$\prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y - x_i\theta)^2}$$
- Log Likelihood :
$$\sum_{i=1}^{n} k \times (y_i - x_i\theta)^2 = k \times \sum_{i=1}^{n} \epsilon_i^2$$
where k is a constant

MLE for Linear Regression - Likelihood

- Likelihood $= P(\mathcal{D}|\theta)$. Data $= < x_i, y_i >$, Parameters $= \theta, \sigma$.
- $P(y|X, \theta, \sigma) = P(y_1|x_1, \theta, \sigma) \times P(y_2|x_2, \theta, \sigma) \ldots$
- $\implies$
$$\prod_{i=1}^{n} P(y|X, \theta, \sigma)$$
- $\implies$
$$\prod_{i=1}^{n} \sqrt{\frac{1}{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y - x_i\theta)^2}$$
- Log Likelihood :
$$\sum_{i=1}^{n} k \times (y_i - x_i\theta)^2 = k \times \sum_{i=1}^{n} \epsilon_i^2$$
where k is a constant
- $\hat{\theta}_{MLE} = \arg\max_\theta (y - X\theta)^T (y - X\theta) = \hat{\theta}_{LS}$, when the residues are normally distributed

MAP for Linear Regression

- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.
- Let ϵ be independent across all the observations.
- $y \sim \mathcal{N}(X\theta, \sigma^2)$, since $\{X\theta + \mathcal{N}(0, \sigma^2)\}$

MAP for Linear Regression

- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.
- Let ϵ be independent across all the observations.
- $y \sim \mathcal{N}(X\theta, \sigma^2)$, since $\{X\theta + \mathcal{N}(0, \sigma^2)\}$
- Suppose θ is a gaussian prior distribtuion, i.e., $\theta \sim \mathcal{N}(0, T^2\mathcal{I}_n)$

MAP for Linear Regression

- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.
- Let ϵ be independent across all the observations.
- $y \sim \mathcal{N}(X\theta, \sigma^2)$, since $\{X\theta + \mathcal{N}(0, \sigma^2)\}$
- Suppose θ is a gaussian prior distribution, i.e., $\theta \sim \mathcal{N}(0, T^2\mathcal{I}_n)$
- $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta) \times P(\theta) \implies log(P(\theta|\mathcal{D})) \propto log(P(\mathcal{D}|\theta)) + log(P(\theta))$

MAP for Linear Regression

- Let $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is known.
- Let ϵ be independent across all the observations.
- $y \sim \mathcal{N}(X\theta, \sigma^2)$, since $\{X\theta + \mathcal{N}(0, \sigma^2)\}$
- Suppose θ is a gaussian prior distribtuion, i.e., $\theta \sim \mathcal{N}(0, T^2\mathcal{I}_n)$
- $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta) \times P(\theta) \implies log(P(\theta|\mathcal{D})) \propto log(P(\mathcal{D}|\theta)) + log(P(\theta))$
- $\hat{\theta}_{MAP} = \arg\max_\theta \{log(P(\mathcal{D}|\theta)) + log(P(\theta))\}$

MLE for Linear Regression - Continued

- $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta) \times P(\theta) \implies log(P(\theta|\mathcal{D})) \propto log(P(\mathcal{D}|\theta)) + log(P(\theta))$
- $\hat{\theta}_{MAP} = \arg\max_\theta \{log(P(\mathcal{D}|\theta)) + log(P(\theta))\}$
  $= \arg\min_\theta (y - X\theta)^T (y - X\theta) + \lambda^2 \theta^T \theta$
- MAP with Gaussian Prior $\implies$ Ridge Regression

Priors and Regularization

- Prior leads to regularization
- $\theta \sim \mathcal{N}(0, \mathcal{I}_n)$: Ridge Regression
- $\theta \sim$ Laplace$(0, t)$: Lasso.