CSCI3230 (ESTR3108)
Fundamentals of Artificial Intelligence

Tutorial 4:
Understanding Linear Regression from Numerical and Probabilistic Perspectives

Tao Huang

Email: thuang22@cse.cuhk.edu.hk
Office: Room 1026, 10/F, SHB

Dept. of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

Part 1. A Numerical Perspective on Linear Regression

Part 2. A Probabilistic Perspective on Linear Regression

Part 1. A Numerical Perspective on Linear Regression

Linear Regression

- We denote $\mathbf{X} \in \mathbb{R}^{m \times n}$ as the data matrix, of which rows represent samples, columns represent features; $\Theta \in \mathbb{R}^n$ as the variables:

$$\mathbf{X} = \begin{pmatrix} {X^{(1)}}^T \\ \vdots \\ {X^{(m)}}^T \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(m)} & \dots & x_n^{(m)} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

Linear Regression

- We denote $\mathbf{X} \in \mathbb{R}^{m \times n}$ as the data matrix, of which rows represent samples, columns represent features; $\Theta \in \mathbb{R}^n$ as the variables:

$$\mathbf{X} = \begin{pmatrix} {X^{(1)}}^T \\ \vdots \\ {X^{(m)}}^T \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & \dots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(m)} & \dots & x_n^{(m)} \end{pmatrix}, \quad \Theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$$

- We aim to find the optimal solution by minimizing the following MSE objective:

$$J(\Theta) = \|\mathbf{X}\Theta - \mathbf{Y}\|_2,$$

where $\mathbf{Y}$ is the ground-truth target.

Analytic Optimal Solution

Find the global minimum of the convex objective $J(\Theta)$:

- $J(\Theta)$ is convex $\Rightarrow$ $\Theta^\star$ is the global minimum iff:

$$\nabla J(\Theta^\star) = 0, \quad \nabla^2 J(\Theta^\star) \succeq 0,$$

Analytic Optimal Solution

Find the global minimum of the convex objective $J(\Theta)$:

- $J(\Theta)$ is convex $\Rightarrow \Theta^\star$ is the global minimum iff:

  $$\nabla J(\Theta^\star) = 0, \quad \nabla^2 J(\Theta^\star) \succeq 0,$$

  where the notation $\succeq 0$ represents Positive Semidefinite (PSD):

  $$\text{Symmetric } \mathbf{V} \in \mathbb{R}^{n \times n} \text{ is PSD} \quad \Leftrightarrow \quad \mathbf{x}^T \mathbf{V} \mathbf{x} \geq 0 \text{ for } \forall \mathbf{x} \in \mathbb{R}^n.$$

Analytic Optimal Solution

Find the global minimum of the convex objective $J(\Theta)$:

- $J(\Theta)$ is convex $\Rightarrow \Theta^\star$ is the global minimum iff:

$$\nabla J(\Theta^\star) = 0, \quad \nabla^2 J(\Theta^\star) \succeq 0,$$

  where the notation $\succeq 0$ represents Positive Semidefinite (PSD):

$$\text{Symmetric } \mathbf{V} \in \mathbb{R}^{n\times n} \text{ is PSD} \quad \Leftrightarrow \quad \mathbf{x}^T\mathbf{V}\mathbf{x} \geq 0 \text{ for } \forall \mathbf{x} \in \mathbb{R}^n.$$

- Analytical solution:

$$\nabla J(\Theta^\star) = 2\mathbf{X}^T(\mathbf{X}\Theta - \mathbf{Y}) = 0 \quad \Rightarrow \quad \Theta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

Analytic Optimal Solution

Find the global minimum of the convex objective $J(\Theta)$:

- $J(\Theta)$ is convex $\Rightarrow$ $\Theta^\star$ is the global minimum iff:

$$\nabla J(\Theta^\star) = 0, \quad \nabla^2 J(\Theta^\star) \succeq 0,$$

  where the notation $\succeq 0$ represents Positive Semidefinite (PSD):

$$\text{Symmetric } \mathbf{V} \in \mathbb{R}^{n\times n} \text{ is PSD} \quad \Leftrightarrow \quad \mathbf{x}^T\mathbf{V}\mathbf{x} \geq 0 \text{ for } \forall \mathbf{x} \in \mathbb{R}^n.$$

- Analytical solution:

$$\nabla J(\Theta^\star) = 2\mathbf{X}^T(\mathbf{X}\Theta - \mathbf{Y}) = 0 \quad \Rightarrow \quad \Theta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},$$

  and the second order condition is satisfied:

$$\nabla^2 J(\Theta^\star) = 2\mathbf{X}^T\mathbf{X} \succeq 0. \quad \text{(Why?)}$$

L2 Regularization from a Numerical Perspective

However, we cannot ensure that $\mathbf{X}^T\mathbf{X}$ is invertible (e.g., zero matrix).

L2 Regularization from a Numerical Perspective

However, we cannot ensure that $\mathbf{X}^T\mathbf{X}$ is invertible (e.g., zero matrix).

- But we know that $\mathbf{X}^T\mathbf{X}$ is PSD. Consider the following propositions:

Non-negativity of Eigenvalue of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$, all its eigenvalues are non-negative, where the eigenvalue λ is defined as:

$$\mathbf{V}\mathbf{x} = \lambda\mathbf{x}, \quad \text{for some } \mathbf{x} \in \mathbb{R}^n$$

L2 Regularization from a Numerical Perspective

However, we cannot ensure that $\mathbf{X}^T\mathbf{X}$ is invertible (e.g., zero matrix).

- But we know that $\mathbf{X}^T\mathbf{X}$ is PSD. Consider the following propositions:

Non-negativity of Eigenvalue of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$, all its eigenvalues are non-negative, where the eigenvalue λ is defined as:

$$\mathbf{V}\mathbf{x} = \lambda\mathbf{x}, \quad \text{for some } \mathbf{x} \in \mathbb{R}^n$$

Invertible Condition of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$, it is invertible iff all its eigenvalues are positive,

L2 Regularization from a Numerical Perspective

However, we cannot ensure that $\mathbf{X}^T\mathbf{X}$ is invertible (e.g., zero matrix).

- But we know that $\mathbf{X}^T\mathbf{X}$ is PSD. Consider the following propositions:

Non-negativity of Eigenvalue of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$, all its eigenvalues are non-negative, where the eigenvalue λ is defined as:

$$\mathbf{V}\mathbf{x} = \lambda\mathbf{x}, \quad \text{for some } \mathbf{x} \in \mathbb{R}^n$$

Invertible Condition of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$, it is invertible iff all its eigenvalues are positive,

- Any way to approximate $\mathbf{X}^T\mathbf{X}$ such all eigenvalues are positive?

L2 Regularization from a Numerical Perspective

However, we cannot ensure that $\mathbf{X}^T\mathbf{X}$ is invertible (e.g., zero matrix).

- But we know that $\mathbf{X}^T\mathbf{X}$ is PSD. Consider the following propositions:

Non-negativity of Eigenvalue of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$, all its eigenvalues are non-negative, where the eigenvalue λ is defined as:

$$\mathbf{V}\mathbf{x} = \lambda\mathbf{x}, \quad \text{for some } \mathbf{x} \in \mathbb{R}^n$$

Invertible Condition of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$, it is invertible iff all its eigenvalues are positive,

- Any way to approximate $\mathbf{X}^T\mathbf{X}$ such all eigenvalues are positive?
- Yes! Consider the $\tilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}$ with a tiny $\alpha > 0$.

L2 Regularization from a Numerical Perspective

However, we cannot ensure that $\mathbf{X}^T\mathbf{X}$ is invertible (e.g., zero matrix).

- But we know that $\mathbf{X}^T\mathbf{X}$ is PSD. Consider the following propositions:

Non-negativity of Eigenvalue of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$, all its eigenvalues are non-negative, where the eigenvalue λ is defined as:

$$\mathbf{V}\mathbf{x} = \lambda\mathbf{x}, \quad \text{for some } \mathbf{x} \in \mathbb{R}^n$$

Invertible Condition of PSD Matrices

Given a PSD matrix $\mathbf{V} \in \mathbb{R}^{n\times n}$, it is invertible iff all its eigenvalues are positive,

- Any way to approximate $\mathbf{X}^T\mathbf{X}$ such all eigenvalues are positive?
- Yes! Consider the $\tilde{\mathbf{X}} = \mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}$ with a tiny $\alpha > 0$.
- If λ is an eigenvalue of $\mathbf{X}^T\mathbf{X}$, then $\lambda + \alpha > 0$ must be the eigenvalue of $\tilde{\mathbf{X}}$. (Proof: $\mathbf{V}\mathbf{x} = \lambda\mathbf{x} \Rightarrow (\mathbf{V} + \alpha\mathbf{I})\mathbf{x} = (\lambda + \alpha)\mathbf{x}$)

L2 Regularization from a Numerical Perspective

From the numerical perspective, we add L2 regularization on the original objective to approximate the optimal solution with numerical feasibility.

L2 Regularization from a Numerical Perspective

From the numerical perspective, we add L2 regularization on the original objective to approximate the optimal solution with numerical feasibility.

- We aim to minimize the following objective:

$$J(\Theta) = \|\mathbf{X}\Theta - \mathbf{Y}\|_2 + \alpha\|\Theta\|_2.$$

L2 Regularization from a Numerical Perspective

From the numerical perspective, we add L2 regularization on the original objective to approximate the optimal solution with numerical feasibility.

- We aim to minimize the following objective:

$$J(\Theta) = \|\mathbf{X}\Theta - \mathbf{Y}\|_2 + \alpha\|\Theta\|_2.$$

- Analytical solution:

$$\nabla J(\Theta^\star) = 2\mathbf{X}^T(\mathbf{X}\Theta - \mathbf{Y}) + 2\alpha\Theta = 0 \;\Rightarrow\; \Theta^\star = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y},$$

  and the second order condition is satisfied:

$$\nabla^2 J(\Theta^\star) = 2\mathbf{X}^T\mathbf{X} + 2\alpha\mathbf{I} \succ 0. \quad \text{(Positive definite)}$$

L2 Regularization from a Numerical Perspective

From the numerical perspective, we add L2 regularization on the original objective to approximate the optimal solution with numerical feasibility.

- We aim to minimize the following objective:

$$J(\Theta) = \|\mathbf{X}\Theta - \mathbf{Y}\|_2 + \alpha\|\Theta\|_2.$$

- Analytical solution:

$$\nabla J(\Theta^\star) = 2\mathbf{X}^T(\mathbf{X}\Theta - \mathbf{Y}) + 2\alpha\Theta = 0 \;\Rightarrow\; \Theta^\star = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y},$$

  and the second order condition is satisfied:

$$\nabla^2 J(\Theta^\star) = 2\mathbf{X}^T\mathbf{X} + 2\alpha\mathbf{I} \succ 0. \quad \text{(Positive definite)}$$

- Adding L2 regularization also enhances the numerical stability by reducing the noise sensitivity[1].

[1] https://www.cs.cornell.edu/~bindel/class/cs3220-s12/notes/lec11.pdf

Part 2. A Probabilistic Perspective on Linear Regression

The Principle of Maximum Likelihood Estimation

Maximum Likelihood Estimation[1]

Suppose we have a random sample of i.i.d. random variables $X_1, X_2, ..., X_n$ with a PMF or PDF $f_\theta(x)$ which depends on a parameter θ. The joint PMF/PDF (likelihood) is:

$$L(\theta) = f_\theta(x_1, x_2, ..., x_n) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

The Maximum Likelihood Estimator of θ (MLE) is the value $\hat{\theta}$ that maximizes the likelihood given the observed data $(x_1, x_2, ..., x_n)$.

[1] http://www2.stat.duke.edu/~vp58/sta111/lecture12.pdf

The Principle of Maximum Likelihood Estimation

Maximum Likelihood Estimation[1]

Suppose we have a random sample of i.i.d. random variables $X_1, X_2, ..., X_n$ with a PMF or PDF $f_\theta(x)$ which depends on a parameter θ. The joint PMF/PDF (likelihood) is:

$$L(\theta) = f_\theta(x_1, x_2, ..., x_n) = f_\theta(x_1) f_\theta(x_2) \cdots f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

The Maximum Likelihood Estimator of θ (MLE) is the value $\hat{\theta}$ that maximizes the likelihood given the observed data $(x_1, x_2, ..., x_n)$.

- Products are typically hard to maximize, so we usually take logarithms and maximize the log-likelihood $\ell(\theta) = \log L(\theta)$ instead.
- MLE finds the value of θ that makes the samples most probable.

[1] http://www2.stat.duke.edu/~vp58/sta111/lecture12.pdf

Example of MLE

Let $X_1, X_2, ..., X_n \sim^{iid}$ Bernoulli(p) with p unknown, and suppose that $x_1, x_2, ..., x_n$ have been observed (i.e., tossing a coin multiple times).

Example of MLE

Let $X_1, X_2, ..., X_n \sim^{iid}$ Bernoulli(p) with p unknown, and suppose that $x_1, x_2, ..., x_n$ have been observed (i.e., tossing a coin multiple times).

- The likelihood is:

$$L(p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{S_n}(1-p)^{(n-S_n)},$$

Example of MLE

Let $X_1, X_2, ..., X_n \sim^{iid}$ Bernoulli(p) with p unknown, and suppose that $x_1, x_2, ..., x_n$ have been observed (i.e., tossing a coin multiple times).

- The likelihood is:

$$L(p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{S_n}(1-p)^{(n-S_n)},$$

  where $S_n = \sum_{i=1}^{n} x_i$.
- The log-likelihood is:

$$\ell(p) = S_n \log p + (n - S_n)\log(1-p).$$

Example of MLE

Let $X_1, X_2, ..., X_n \sim^{iid}$ Bernoulli(p) with p unknown, and suppose that $x_1, x_2, ..., x_n$ have been observed (i.e., tossing a coin multiple times).

- The likelihood is:

$$L(p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{S_n}(1-p)^{(n-S_n)},$$

  where $S_n = \sum_{i=1}^{n} x_i$.
- The log-likelihood is:

$$\ell(p) = S_n \log p + (n - S_n) \log(1-p).$$

- Let $\ell'(p) = 0$:

$$p^\star = S_n/n = \bar{x}.$$

Example of MLE

Let $X_1, X_2, ..., X_n \sim^{iid}$ Bernoulli(p) with p unknown, and suppose that $x_1, x_2, ..., x_n$ have been observed (i.e., tossing a coin multiple times).

- The likelihood is:

$$L(p) = \prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{S_n}(1-p)^{(n-S_n)},$$

  where $S_n = \sum_{i=1}^{n} x_i$.
- The log-likelihood is:

$$\ell(p) = S_n \log p + (n - S_n)\log(1-p).$$

- Let $\ell'(p) = 0$:

$$p^\star = S_n/n = \bar{x}.$$

- We can say that the maximum likelihood estimator is the value of p that is "most likely" to have the generated data.

MLE for Linear Regression

Consider the linear regression model with data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$:

$$\hat{y} = \mathbf{x}^T\theta + \epsilon,$$

where $\theta = [\theta_1, \theta_2, ..., \theta_n]^T$. We assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.

MLE for Linear Regression

Consider the linear regression model with data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$:

$$\hat{y} = \mathbf{x}^T\theta + \epsilon,$$

where $\theta = [\theta_1, \theta_2, ..., \theta_n]^T$. We assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.

- The conditional distribution of Y_i given $X_i = \mathbf{x}_i$ is:

$$Pr(Y_i|\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i^T\theta, \sigma^2)$$

MLE for Linear Regression

Consider the linear regression model with data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$:

$$\hat{y} = \mathbf{x}^T\theta + \epsilon,$$

where $\theta = [\theta_1, \theta_2, ..., \theta_n]^T$. We assume $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise.

- The conditional distribution of Y_i given $X_i = \mathbf{x}_i$ is:

$$Pr(Y_i|\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i^T\theta, \sigma^2)$$

- We are given the log-likelihood of Gaussian distributions:

$$L_m(\theta) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(y_i - \mathbf{x}_i^T\theta)^2}{2\sigma^2}),$$

$$\ell_m(\theta) = \ln(\sigma^2 2\pi)^{-m/2} + (-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2)$$

MLE for Linear Regression

So the log MLE is given by:

$$\ell_m(\theta) = -\frac{m}{2}\ln(\sigma^2)^{-m/2} - \frac{m}{2}\ln(2\pi) + (-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2,$$

where $\mathbf{X}$ denotes $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)$, $\mathbf{Y}$ denotes $(y_1, y_2, ..., y_m)$.

MLE for Linear Regression

So the log MLE is given by:

$$\ell_m(\theta) = -\frac{m}{2}\ln(\sigma^2)^{-m/2} - \frac{m}{2}\ln(2\pi) + (-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2,$$

where $\mathbf{X}$ denotes $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)$, $\mathbf{Y}$ denotes $(y_1, y_2, ..., y_m)$.

- To find the MLE of θ, we set the gradient to zero:

$$\frac{\partial}{\partial\theta}\ell_m(\theta) = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{X}\theta - \mathbf{X}^T\mathbf{Y}) = 0 \;\Rightarrow\; \theta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

MLE for Linear Regression

So the log MLE is given by:

$$\ell_m(\theta) = -\frac{m}{2}\ln(\sigma^2)^{-m/2} - \frac{m}{2}\ln(2\pi) + (-\frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2,$$

where $\mathbf{X}$ denotes $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m)$, $\mathbf{Y}$ denotes $(y_1, y_2, ..., y_m)$.

- To find the MLE of θ, we set the gradient to zero:

$$\frac{\partial}{\partial\theta}\ell_m(\theta) = \frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{X}\theta - \mathbf{X}^T\mathbf{Y}) = 0 \Rightarrow \theta^\star = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

- Conclusion: Under the previous probabilistic assumptions on the data, least-square regression corresponds to finding the maximum likelihood estimate of θ.

Mean Absolute Error

Consider the linear regression model with data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$:

$$\hat{y} = \mathbf{x}^T\theta + \epsilon,$$

- For a data point $(\mathbf{x}_i, y_i)$, we can also assume that the error ϵ follows a Laplacian distribution Laplace$(0, \sigma^2)$, i.e.,

$$Pr(y_i|\mathbf{x}_i, \theta) = \frac{1}{2b}\exp(-\frac{\|y_i - \mathbf{x}_i^T\theta\|}{b}).$$

Mean Absolute Error

Consider the linear regression model with data $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$:

$$\hat{y} = \mathbf{x}^T \theta + \epsilon,$$

- For a data point $(\mathbf{x}_i, y_i)$, we can also assume that the error ϵ follows a Laplacian distribution Laplace$(0, \sigma^2)$, i.e.,

$$Pr(y_i|\mathbf{x}_i, \theta) = \frac{1}{2b} \exp(-\frac{\|y_i - \mathbf{x}_i^T \theta\|}{b}).$$

- Through similar steps, the finding the MLE of θ is equivalent to solving the linear regression with mean absolute error:

$$\theta^\star = \arg\min_\theta \|\mathbf{X}\theta - \mathbf{Y}\|_1, \quad \text{(Prove it by yourself)}$$

Maximum A Posteriori Estimation[1]

- MLE objective:

$$\arg\max_\theta \log P(\mathbf{Y}|\mathbf{X},\theta) = \arg\min_\theta \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2.$$

- MLE solution:

$$\theta_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

[1] https://www.cse.iitk.ac.in/users/piyush/courses/pml_winter16/slides_lec4.pdf

Maximum A Posteriori Estimation[1]

- MLE objective:

$$\arg\max_{\theta} \log P(\mathbf{Y}|\mathbf{X},\theta) = \arg\min_{\theta} \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2.$$

- MLE solution:

$$\theta_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

- MAP objective:

$$\arg\max_{\theta} \log P(\theta|\mathbf{X},\mathbf{Y}) \propto \arg\max_{\theta} \underbrace{P(\mathbf{Y}|\mathbf{X},\theta)P(\theta)}_{\text{Bayes Theorem}} = \arg\min_{\theta}\sum_{i=1}^{m}(y_i - \mathbf{x}_i^T\theta)^2 + \beta\|\theta\|_2$$

[1] https://www.cse.iitk.ac.in/users/piyush/courses/pml_winter16/slides_lec4.pdf

Maximum A Posteriori Estimation[1]

- MLE objective:

$$\arg\max_{\theta} \log P(\mathbf{Y}|\mathbf{X}, \theta) = \arg\min_{\theta} \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mathbf{x}_i^T \theta)^2.$$

- MLE solution:

$$\theta_{MLE} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

- MAP objective:

$$\arg\max_{\theta} \log P(\theta|\mathbf{X}, \mathbf{Y}) \propto \arg\max_{\theta} \underbrace{P(\mathbf{Y}|\mathbf{X}, \theta) P(\theta)}_{\text{Bayes Theorem}} = \arg\min_{\theta} \sum_{i=1}^{m} (y_i - \mathbf{x}_i^T \theta)^2 + \beta \|\theta\|_2$$

- MAP solution:

$$\theta_{MAP} = (\mathbf{X}^T \mathbf{X} + \beta \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y},$$

[1] https://www.cse.iitk.ac.in/users/piyush/courses/pml_winter16/slides_lec4.pdf