Linear Regression Models

Segment 6 – Advanced Topics in Linear Regression

Topic 2 – Bayesian Linear Regression

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. Probabilistic View of Linear Regression

2. Maximum Likelihood Estimate

3. Maximum A Posteriori Estimate

Probabilistic View of Linear Regression

Probabilistic View of Linear Regression

- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has

Probabilistic View of Linear Regression

- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has zero mean,

Probabilistic View of Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

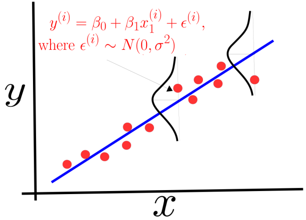
- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has zero mean, constant variance σ^2,

Probabilistic View of Linear Regression

- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has zero mean, constant variance σ^2, and is normally distributed.

Probabilistic View of Linear Regression

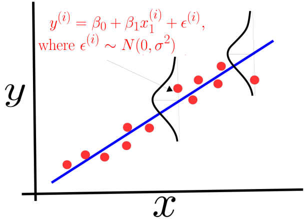
- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has zero mean, constant variance σ^2, and is normally distributed.



Probabilistic View of Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

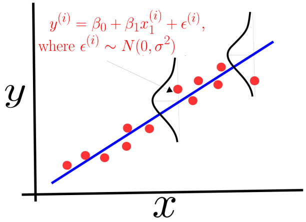
- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has zero mean, constant variance σ^2, and is normally distributed.



- None of these assumptions are needed to derive the formula for the coefficient estimates $\beta_0, \beta_1, \ldots, \beta_p$.

Probabilistic View of Linear Regression

- In the usual linear regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, we assumed that the random error term ϵ has zero mean, constant variance σ^2, and is normally distributed.



- None of these assumptions are needed to derive the formula for the coefficient estimates $\beta_0, \beta_1, \ldots, \beta_p$.
- However, this leads to a different and useful interpretation of the linear regression model.

Probabilistic View of Linear Regression – Continued

Probabilistic View of Linear Regression – Continued

- The probability density of $\epsilon^{(i)}$ is $P\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(\epsilon^{(i)}-0\right)^2}{2\sigma^2}}$.

Probabilistic View of Linear Regression – Continued

- The probability density of $\epsilon^{(i)}$ is $P\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(\epsilon^{(i)}-0\right)^2}{2\sigma^2}}$.
- This implies that

$$P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.$$

Probabilistic View of Linear Regression – Continued

- The probability density of $\epsilon^{(i)}$ is $P\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(\epsilon^{(i)}-0\right)^2}{2\sigma^2}}$.
- This implies that

$$P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.$$

- Note that the population parameters $\beta_0, \beta_1, \ldots, \beta_p$ are *fixed*.

Probabilistic View of Linear Regression – Continued

- The probability density of $\epsilon^{(i)}$ is $P\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(\epsilon^{(i)}-0\right)^2}{2\sigma^2}}$.
- This implies that

$$P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.$$

- Note that the population parameters $\beta_0, \beta_1, \ldots, \beta_p$ are *fixed*.
- For the ith sample, this means

$$y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p \sim N(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}, \sigma^2).$$

Probabilistic View of Linear Regression – Continued

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The probability density of $\epsilon^{(i)}$ is $P\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(\epsilon^{(i)}-0\right)^2}{2\sigma^2}}$.
- This implies that

$$P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.$$

- Note that the population parameters $\beta_0, \beta_1, \ldots, \beta_p$ are *fixed*.
- For the ith sample, this means
  $y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p \sim N(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}, \sigma^2)$.
- That is, the ith response value given the ith predictor values and parametrized by $\beta_0, \beta_1, \ldots, \beta_p$ is normally distributed with mean $\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}$ and variance σ^2.

Maximum Likelihood Estimate

Maximum Likelihood Estimate

- The probabilistic way of looking at linear regression leads to a different way of estimating the model parameters known as the *maximum likelihood estimation (MLE)*.

Maximum Likelihood Estimate

- The probabilistic way of looking at linear regression leads to a different way of estimating the model parameters known as the *maximum likelihood estimation (MLE)*.

- Recall Bayes theorem:

Maximum Likelihood Estimate

- The probabilistic way of looking at linear regression leads to a different way of estimating the model parameters known as the *maximum likelihood estimation (MLE)*.

- Recall Bayes theorem: $\underbrace{P\left(model \,|\, data\right)}_{\text{posterior}} = \dfrac{\underbrace{P\left(data \,|\, model\right)}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}.$

Maximum Likelihood Estimate

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The probabilistic way of looking at linear regression leads to a different way of estimating the model parameters known as the *maximum likelihood estimation (MLE)*.
- Recall Bayes theorem: $\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \dfrac{\overbrace{\underbrace{P\left(data \mid model\right)}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}}{P(data)}.$
- The likelihood in linear regression is $P\left(y^{(1)}, \ldots, y^{(n)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; \beta_0, \beta_1, \ldots, \beta_p\right)$.

Maximum Likelihood Estimate

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- The probabilistic way of looking at linear regression leads to a different way of estimating the model parameters known as the *maximum likelihood estimation (MLE)*.
- Recall Bayes theorem: $\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \dfrac{\underbrace{P\left(data \mid model\right)}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}.$
- The likelihood in linear regression is $P\left(y^{(1)}, \ldots, y^{(n)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; \beta_0, \beta_1, \ldots, \beta_p\right).$
- Assuming independence between samples, the likelihood can be written as a function of the model parameters:

Maximum Likelihood Estimate

- The probabilistic way of looking at linear regression leads to a different way of estimating the model parameters known as the *maximum likelihood estimation (MLE).*
- Recall Bayes theorem: $\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \dfrac{\underbrace{P\left(data \mid model\right)}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}.$
- The likelihood in linear regression is $P\left(y^{(1)}, \ldots, y^{(n)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; \beta_0, \beta_1, \ldots, \beta_p\right).$
- Assuming independence between samples, the likelihood can be written as a function of the model parameters: $L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right).$

Maximum Likelihood Estimate – Continued

Maximum Likelihood Estimate – Continued

- The likelihood function

Maximum Likelihood Estimate – Continued

- The likelihood function

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} P\left(y^{(i)} \,|\, \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right),$$

Maximum Likelihood Estimate – Continued

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The likelihood function

$$
\begin{aligned}
L(\beta_0, \beta_1, \ldots, \beta_p) &= \prod_{i=1}^{n} P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right), \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.
\end{aligned}
$$

Maximum Likelihood Estimate – Continued

- The likelihood function

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} P\left(y^{(i)} \,|\, \mathbf{x}^{(i)}; \beta_0, \beta_1, \ldots, \beta_p\right),$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.$$

- The goal is now to find the optimal model parameters $\beta_0, \beta_1, \ldots, \beta_p$ such that the likelihood function is maximized.

Maximum Likelihood Estimate – Continued

- The likelihood function

$$L(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^{n} P\left(y^{(i)} \mid \mathbf{x}^{(i)}; \beta_0, \beta_1, \dots, \beta_p\right),$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-\frac{\left(y^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)^2}{2\sigma^2}}.$$

- The goal is now to find the optimal model parameters $\beta_0, \beta_1, \dots, \beta_p$ such that the likelihood function is maximized.
- To achieve that, we take the logarithm of the likelihood function which does not affect the optimal set of parameters.

Maximum Likelihood Estimate – Continued

Maximum Likelihood Estimate – Continued

- The log-likelihood function

Maximum Likelihood Estimate – Continued

- The log-likelihood function

$$LL(\beta_0, \ldots, \beta_p) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \cdots + \beta_p x_p^{(i)}\right)\right)^2$$

Maximum Likelihood Estimate – Continued

- The log-likelihood function

$$LL(\color{red}{\beta_0, \ldots, \beta_p}) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \cdots + \beta_p x_p^{(i)}\right)\right)^2$$

- The optimal solution (parameters) for *maximizing* the log-likelihood function is equivalent to the optimal solution for *minimizing* the *negative* of the log-likelihood function;

Maximum Likelihood Estimate – Continued

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The log-likelihood function

$$LL(\beta_0, \ldots, \beta_p) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \cdots + \beta_p x_p^{(i)}\right)\right)^2$$

- The optimal solution (parameters) for *maximizing* the log-likelihood function is equivalent to the optimal solution for *minimizing* the *negative* of the log-likelihood function; this is usually referred to as the *loss* in the context of machine learning.

Maximum Likelihood Estimate – Continued

- The log-likelihood function

$$LL(\beta_0, \ldots, \beta_p) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y^{(i)} - \left(\beta_0 + \cdots + \beta_p x_p^{(i)}\right)\right)^2$$

- The optimal solution (parameters) for *maximizing* the log-likelihood function is equivalent to the optimal solution for *minimizing* the *negative* of the log-likelihood function; this is usually referred to as the *loss* in the context of machine learning.
- After a few steps of basic calculus, we arrive at the exact same solution for the model parameters that was shown during the matrix-formulation of the multiple linear regression model.

Maximum A Posteriori Estimate

Maximum A Posteriori Estimate

- In Bayes formula, we now focus on both the *likelihood* and *prior*

  terms:

Maximum A Posteriori Estimate

- In Bayes formula, we now focus on both the *likelihood* and *prior* terms:

$$\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \frac{\overbrace{P\left(data \mid model\right)}^{}\!\!\underbrace{}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}.$$

Maximum A Posteriori Estimate

- In Bayes formula, we now focus on both the <u>*likelihood*</u> and <u>*prior*</u> terms: $\underbrace{P(model \mid data)}_{\text{posterior}} = \dfrac{\overbrace{P(data \mid model)}^{\text{likelihood}} \times \overbrace{P(model)}^{\text{prior}}}{P(data)}$.
- This powerful approach called *maximum a posteriori estimation (MAP)* lets us calculate model parameters with <u>*prior*</u> uncertainty built into them;

Maximum A Posteriori Estimate

- In Bayes formula, we now focus on both the *likelihood* and *prior* terms: $\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \dfrac{\overbrace{P\left(data \mid model\right)}^{}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}$.

$$\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \frac{\underbrace{P\left(data \mid model\right)}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}.$$

- This powerful approach called *maximum a posteriori estimation (MAP)* lets us calculate model parameters with *prior* uncertainty built into them; that is, the model parameters are not assumed to be fixed but are random.

Maximum A Posteriori Estimate

- In Bayes formula, we now focus on both the <u>*likelihood*</u> and <u>*prior*</u> terms: $$\underbrace{P\left(model \mid data\right)}_{\text{posterior}} = \frac{\underbrace{P\left(data \mid model\right)}_{\text{likelihood}} \times \underbrace{P(model)}_{\text{prior}}}{P(data)}.$$
- This powerful approach called *maximum a posteriori estimation (MAP)* lets us calculate model parameters with <u>*prior*</u> uncertainty built into them; that is, the model parameters are not assumed to be fixed but are random.
- For example, we can assume that the model parameters $\beta_1, \ldots, \beta_p$, put together as the model parameters vector $\boldsymbol{\beta}$ have a *prior* joint-normal distribution with a constant variance;

Maximum A Posteriori Estimate

- In Bayes formula, we now focus on both the *likelihood* and *prior* terms: $\underbrace{P\left(\textcolor{red}{model} \,|\, \textcolor{blue}{data}\right)}_{\text{posterior}} = \dfrac{\overbrace{P\left(\textcolor{blue}{data} \,|\, \textcolor{red}{model}\right)}_{\text{likelihood}} \times \overbrace{P(\textcolor{red}{model})}_{\text{prior}}}{P(\textcolor{blue}{data})}.$
- This powerful approach called *maximum a posteriori estimation (MAP)* lets us calculate model parameters with *prior* uncertainty built into them; that is, the model parameters are not assumed to be fixed but are random.
- For example, we can assume that the model parameters $\beta_1, \ldots, \beta_p$, put together as the model parameters vector $\boldsymbol{\beta}$ have a *prior* joint-normal distribution with a constant variance; note that we are not including the intercept parameter β_0.

Maximum A Posteriori Estimate – Continued

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector:

$$\boldsymbol{\beta} \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$$, where $\mathbf{I}$ refers to the identity matrix.

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector: $\beta \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$, where $\mathbf{I}$ refers to the identity matrix.
- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}.$

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector:
  $$\boldsymbol{\beta} \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$$, where $\mathbf{I}$ refers to the identity matrix.
- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}.$
- This prior distribution assumption encourages the model parameters to be close to zero;

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector: $\beta \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$, where $\mathbf{I}$ refers to the identity matrix.
- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}.$
- This prior distribution assumption encourages the model parameters to be close to zero; similar to *ridge regularization*.

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector:
$$\beta \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$$, where $\mathbf{I}$ refers to the identity matrix.
- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}$.
- This prior distribution assumption encourages the model parameters to be close to zero; similar to *ridge regularization.*
- Now, we want to maximize the $\underbrace{likelihood \times prior}$ term in Bayes formula;

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector: $\beta \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$, where $\mathbf{I}$ refers to the identity matrix.
- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}$.
- This prior distribution assumption encourages the model parameters to be close to zero; similar to *ridge regularization.*
- Now, we want to maximize the $\underbrace{likelihood \times prior}$ term in Bayes formula; just as in maximum likelihood estimation, we assume independence between samples (which turns the product into summation),

Maximum A Posteriori Estimate – Continued

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- A normal prior distribution for model parameters vector: $\boldsymbol{\beta} \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$, where $\mathbf{I}$ refers to the identity matrix.
- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}.$
- This prior distribution assumption encourages the model parameters to be close to zero; similar to *ridge regularization*.
- Now, we want to maximize the $\underbrace{likelihood \times prior}$ term in Bayes formula; just as in maximum likelihood estimation, we assume independence between samples (which turns the product into summation), take logarithm (which does not affect the optimal solution),

Maximum A Posteriori Estimate – Continued

- A normal prior distribution for model parameters vector:

$$\boldsymbol{\beta} \sim N\left(\underbrace{\mathbf{0}}_{\text{zero mean vector}}, \underbrace{\lambda^{-1}\mathbf{I}}_{\text{constant diagonal covariance matrix}}\right)$$, where $\mathbf{I}$ refers to the identity matrix.

- $P(\boldsymbol{\beta}) = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\lambda^{-1}\mathbf{I})^{-1}(\boldsymbol{\beta}-\mathbf{0})} = \frac{1}{(2\pi\lambda^{-1})^{p/2}} e^{\frac{-\lambda}{2}\left(\beta_1^2+\beta_2^2+\cdots+\beta_p^2\right)}$.
- This prior distribution assumption encourages the model parameters to be close to zero; similar to *ridge regularization*.
- Now, we want to maximize the $\underbrace{likelihood \times prior}$ term in Bayes formula; just as in maximum likelihood estimation, we assume independence between samples (which turns the product into summation), take logarithm (which does not affect the optimal solution), and minimize the negative of the resulting expression.

Summary

Summary

- Linear regression from a probabilistic perspective.

Summary

- Linear regression from a probabilistic perspective.
- Maximum likelihood estimation approach for computing regression model parameters.

Summary

- Linear regression from a probabilistic perspective.
- Maximum likelihood estimation approach for computing regression model parameters.
- Maximum a posteriori estimation approach for computing regression model parameters and compare with regularization.