ML&MEA (2024)
Lecture 3 - Frequentist vs Bayesian

Quanying Liu

BME, SUSTech

2024.2.27

Content

1. Recap
2. Review Probability Theory
3. Frequentist vs Bayesian
4. Rethink linear regression
5. Summary

1 Recap

2 Review Probability Theory

3 Frequentist vs Bayesian

4 Rethink linear regression

5 Summary

Recap Lecture 2 - Linear regression

Simple linear regression

1. Data: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^1$, $y^{(i)} \in \mathbb{R}^1$

Recap Lecture 2 - Linear regression

Simple linear regression

1. Data: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^1$, $y^{(i)} \in \mathbb{R}^1$
2. Model: $y = \beta x$ with single parameter β;
   Loss function $\mathcal{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \beta x^{(i)}\right)^2$

Recap Lecture 2 - Linear regression

Simple linear regression

1. Data: $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^1$, $y^{(i)} \in \mathbb{R}^1$
2. Model: $y = \beta x$ with single parameter β;
   Loss function $\mathcal{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \beta x^{(i)}\right)^2$
3. Optimization algorithms

$$\text{Gradient: } \nabla\mathcal{L} = \frac{d\mathcal{L}(\beta)}{d\beta} = -\frac{2}{n}\sum_i x_i \left(y^{(i)} - \beta x^{(i)}\right)$$

$$\text{Analytical solution: } \hat{\beta} = \frac{\sum_{i=1}^{n} x^{(i)} y^{(i)}}{\sum_{i=1}^{n}\left(x^{(i)}\right)^2} = \frac{\mathbf{X}^T\mathbf{Y}}{\mathbf{X}^T\mathbf{X}}$$

$$\text{Gradient descent: } \beta^{(j+1)} \longleftarrow \beta^{(j)} - \eta\nabla\mathcal{L}$$

Recap Lecture 2 - Linear regression

Multiple linear regression

1. Data: $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}, \quad \mathbf{x}^{(i)} = [x_1 \; x_2 \ldots x_p]^T$

Recap Lecture 2 - Linear regression

Multiple linear regression

1. Data: $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}, \quad \mathbf{x}^{(i)} = [x_1 \; x_2 \ldots x_p]^T$
2. Model: $\mathbf{y} = \mathbf{X}\beta$;
   Loss function $\mathcal{L}(\beta) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\beta\|^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$

Recap Lecture 2 - Linear regression

Multiple linear regression

1. Data: $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}, \quad \mathbf{x}^{(i)} = [x_1 \ x_2 \dots x_p]^T$
2. Model: $\mathbf{y} = \mathbf{X}\beta$;
   Loss function $\mathcal{L}(\beta) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$
3. Optimization algorithms

$$\text{Gradient: } \nabla\mathcal{L} = -\frac{2}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\text{Analytical solution: } \hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

$$\text{Gradient descent: } \beta^{(j+1)} \longleftarrow \beta^{(j)} - \eta\nabla\mathcal{L}$$

Recap Lecture 2 - Linear regression

Derive the gradient of loss function using matrix calculus:

$$\begin{aligned}\mathcal{L}(\beta) &= \frac{1}{n}\left(\mathbf{y}-\mathbf{X}\beta\right)^T\left(\mathbf{y}-\mathbf{X}\beta\right)\\ &= \frac{1}{n}(\mathbf{y}^T-(\mathbf{X}\beta)^T)(\mathbf{y}-\mathbf{X}\beta)\\ &= \frac{1}{n}\left(\mathbf{y}^T\mathbf{y}-\mathbf{y}^T\mathbf{X}\beta-\beta^T\mathbf{X}^T\mathbf{y}+\beta^T\mathbf{X}^T\mathbf{X}\beta\right)\end{aligned}$$

From Eq(69) in the *Matrix Cookbook* : $\frac{\partial(\beta^T\mathbf{a})}{\partial\beta} = \frac{\partial(\mathbf{a}^T\beta)}{\partial\beta} = \mathbf{a}$

Gradient: $\nabla\mathcal{L}$
$$\begin{aligned}&=\frac{1}{n}\left(-(\mathbf{y}^T\mathbf{X})^T-(\mathbf{X}^T\mathbf{y})+\mathbf{X}^T\mathbf{X}\beta+(\beta^T\mathbf{X}^T\mathbf{X})^T\right)\\ &=-\frac{2}{n}\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\beta\right)\end{aligned}$$

1 Recap

2 Review Probability Theory

3 Frequentist vs Bayesian

4 Rethink linear regression

5 Summary

Basics of probability

- **Definition of *probability***:
  With each event X, one associates a number denoted by $P(X)$ and called the "probability of X".
  This number measures the likelihood of the event X to be realized a priori, before performing the experiment. It is chosen between 0 and 1, and the more likely the event is, the closer to 1 this number is. [Probability Essential, by Jean Jacod and Philip Protter]

Basics of probability

- **Definition of *probability***:
  With each event X, one associates a number denoted by $P(X)$ and called the "probability of X".
  This number measures the likelihood of the event X to be realized a priori, before performing the experiment. It is chosen between 0 and 1, and the more likely the event is, the closer to 1 this number is. [Probability Essential, by Jean Jacod and Philip Protter]
- **Rules of probability**:
  $0 \leq P(X) \leq 1$
  $P(\Omega) = 1$, where Ω is the full set of all possible events
  $P(A \cup B) = P(A) + P(B)$, if $A \cap B = \emptyset$

Random variable

- A *random variable* x is a quantity that is **uncertain**.
  It may be the result of experiment (*e.g.*, , draw a dice) or real-world measurement (*e.g.*, , measuring temperature).

Random variable

- A *random variable* x is a quantity that is **uncertain**.
It may be the result of experiment (*e.g.*, , draw a dice) or real-world measurement (*e.g.*, , measuring temperature).
- If observe x multiple times, we get different values.
Some values occur more than others; this information is captured by probability distribution $p(x)$.

Random variable

- A *random variable* x is a quantity that is **uncertain**.
  It may be the result of experiment (*e.g.*, , draw a dice) or real-world measurement (*e.g.*, , measuring temperature).
- If observe x multiple times, we get different values.
  Some values occur more than others; this information is captured by probability distribution $p(x)$.

If x is discrete, then $p(x)$ is probability mass function (PMF) with $\sum_x p(x) = 1$

If x is continuous, then $p(x)$ is probability density function (PDF) with $\int_x p(x)dx = 1$

PMF of discrete random variable

- A discrete random variable x can be fully described by a probability mass function (PMF).

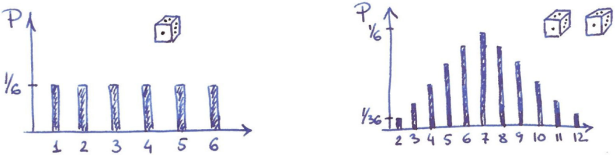


图 1: Two examples of PMF: (left) $p(x)$ where $x \in \{1, 2, \cdots, 6\}$. (right) $p(x)$ where $x \in \{1, 2, \cdots, 12\}$.

PMF of discrete random variable

- A discrete random variable x can be fully described by a probability mass function (PMF).

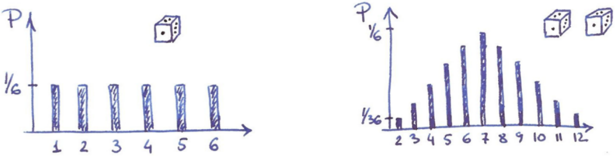


图 1: Two examples of PMF: (left) $p(x)$ where $x \in \{1, 2, \cdots, 6\}$. (right) $p(x)$ where $x \in \{1, 2, \cdots, 12\}$.

- We do an experiment (*e.g.*, draw a dice once), and then observe an event X, which means **sampling** from $p(x)$.

PDF of continuous random variable

- A continuous random variable x can be fully described by its probability density function (PDF).

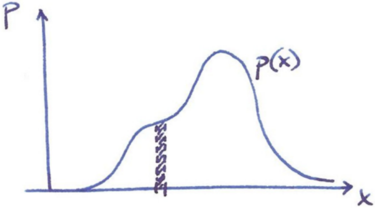


图 2: An example of PDF. $p(x)$ is the probability of an event X falling into the shadowed zone.

PDF of continuous random variable

- A continuous random variable x can be fully described by its probability density function (PDF).

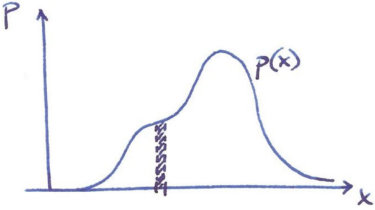


图 2: An example of PDF. $p(x)$ is the probability of an event X falling into the shadowed zone.

- Two properties of a PDF:

$$p(x) \geq 0, \quad \int_{x} p(x)dx = 1$$

Gaussian Distribution (or Normal Distribution)

- Gaussian Distribution is the most well-known and widely-used PDF, a special function with two parameters (mean μ, and variance σ^2),

$$p(x|\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{1}$$

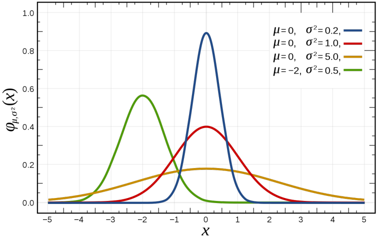


图 3: Four examples of Gaussian distribution.

Multivariate Gaussian Distribution

- To describe one random variable (1-dimension), we use the univariate Gaussian distribution:

$$p(x|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

Multivariate Gaussian Distribution

- To describe one random variable (1-dimension), we use the univariate Gaussian distribution:

$$p(x|\mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

- To describe high-dimensional random variables $\boldsymbol{x} \in \mathbb{R}^p$, we have to use multivariate Gaussian distribution.

Multivariate Gaussian Distribution

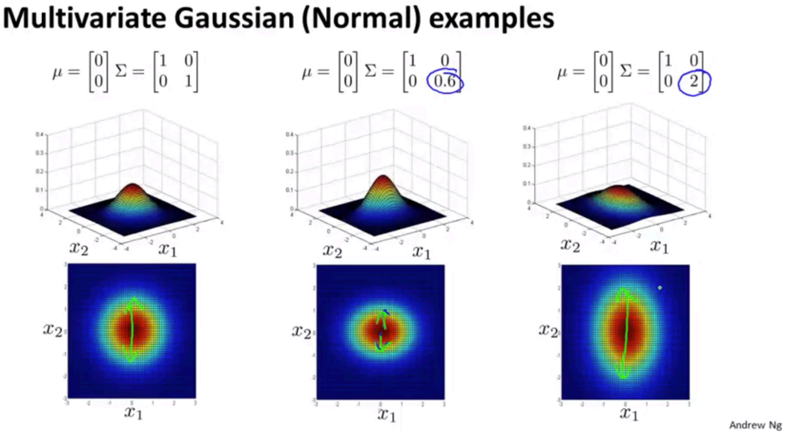
- To describe one random variable (1-dimension), we use the univariate Gaussian distribution:

$$p(x|\mu,\sigma^2) = \mathcal{N}(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

- To describe high-dimensional random variables $\boldsymbol{x} \in \mathbb{R}^p$, we have to use multivariate Gaussian distribution.
- $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \in \mathbb{R}^p$ with the mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ （协方差矩阵）

$$\begin{aligned} P(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ &= \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \end{aligned}$$

Visualize Multivariate Gaussian Distribution



Multivariate Gaussian (Normal) examples

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \boxed{0.6} \end{bmatrix}$

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \boxed{2} \end{bmatrix}$

Andrew Ng

图 4: Three examples of Multivariate Gaussian Distribution.

Motivation of Bayes' Theorem

- Sometime we have some prior information/knowledge about parameter θ. We want to **incorporate our prior knowledge of θ into the model**, such as the mean/variance/range of θ.
- For example, Let's denote the random variable x as Shenzhen's temperature.
  A simple model of x is $x \sim \mathcal{N}(\mu, \sigma^2)$.
- Our prior experiences tell us μ is close to $25°$C. We can assign a prior distribution to μ as $\mu \sim \mathcal{N}(25, 5^2)$.
  We just observed 5 days of temperature in Shenzhen is $\mathbf{X} = (12, 15, 13, 16, 17)^T$.
- Now our belief in μ is not $\mathcal{N}(25, 5^2)$ any more. We should update our belief in μ. How to update it based on our **prior knowledge** and our **current observations**?

Bayes' Theorem

Posterior is proportional to **likelihood** times **prior**.

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|\theta)P(\theta)}{\int_\theta P(\mathbf{X}|\theta)P(\theta)d\theta} \propto P(\mathbf{X}|\theta)P(\theta)$$

- **4 core components**
  1. **Prior**: $P(\theta)$ – what we know about θ BEFORE collecting any data $\mathbf{X}$
  2. **Likelihood**: $P(\mathbf{X}|\theta)$ – likelihood for observing a certain value of $\mathbf{X}$ given a certain value of θ
  3. **Posterior**: $P(\theta|\mathbf{X})$ – what we know about θ AFTER observing $\mathbf{X}$ just happened.
  4. **Evidence**: $P(\mathbf{X})$ – a constant, to ensure that the left hand side is a valid distribution

1. Recap
2. Review Probability Theory
3. Frequentist vs Bayesian
4. Rethink linear regression
5. Summary

Differences between Frequentist & Bayesian

- The fundamental difference between frequentist and Bayesian approaches lies in how they interpret probability and handle uncertainty, especially in the context of statistical inference and parameter estimation. Let's explore these differences and illustrate them with the example of estimating the mean temperature μ in Shenzhen based on observed temperatures.

Differences between Frequentist & Bayesian

- The fundamental difference between frequentist and Bayesian approaches lies in how they interpret probability and handle uncertainty, especially in the context of statistical inference and parameter estimation. Let's explore these differences and illustrate them with the example of estimating the mean temperature μ in Shenzhen based on observed temperatures.
- **Interpretation of Probability**: In the frequentist view, probability is interpreted as the long-run frequency of events. It does not assign probabilities to hypotheses or fixed parameters. Thus, the true mean temperature μ is considered a fixed but unknown value.
  Bayesian probability is a measure of belief or certainty about events, including hypotheses or parameter values. It allows for the incorporation of prior knowledge and evidence to update beliefs.

Differences between Frequentist & Bayesian

- The fundamental difference between frequentist and Bayesian approaches lies in how they interpret probability and handle uncertainty, especially in the context of statistical inference and parameter estimation. Let's explore these differences and illustrate them with the example of estimating the mean temperature μ in Shenzhen based on observed temperatures.
- **Interpretation of Probability**: In the frequentist view, probability is interpreted as the long-run frequency of events. It does not assign probabilities to hypotheses or fixed parameters. Thus, the true mean temperature μ is considered a fixed but unknown value.
  Bayesian probability is a measure of belief or certainty about events, including hypotheses or parameter values. It allows for the incorporation of prior knowledge and evidence to update beliefs.
- **Parameter Estimation**: Frequentist methods use data to calculate a point estimate (e.g., the sample mean) or interval estimates (e.g., confidence intervals) for parameters without incorporating prior knowledge. Uncertainty in parameter estimates is expressed through confidence intervals and p-values.
  Bayesian methods combine prior beliefs about parameters (prior distributions) with observed data (likelihood) to update beliefs (posterior distributions). The result is a full probability distribution over the parameter, reflecting all known information.

Probabilistic model

- So far in this course, a *model* is referred to a parametric family of functions $f(x, \theta)$, where θ is the vector of parameters. For example, the simplest linear regression model, $f(x) = \beta x$.

Probabilistic model

- So far in this course, a *model* is referred to a parametric family of functions $f(x, \theta)$, where θ is the vector of parameters. For example, the simplest linear regression model, $f(x) = \beta x$.
- Now we will discuss *probabilistic models*, such as

$$f(x) = \beta x + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

where $\mathcal{N}(0, \sigma^2)$ denotes a Gaussian distribution with mean 0 and variance σ^2, referring to the distribution of error.

Probabilistic model

- So far in this course, a *model* is referred to a parametric family of functions $f(x, \theta)$, where θ is the vector of parameters.
  For example, the simplest linear regression model, $f(x) = \beta x$.
- Now we will discuss *probabilistic models*, such as

$$f(x) = \beta x + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

  where $\mathcal{N}(0, \sigma^2)$ denotes a Gaussian distribution with mean 0 and variance σ^2, referring to the distribution of error.
- Probabilistic models can be considered as *generative models*, where each data $\mathbf{x}$ is generated by sampling from the distribution $P(\mathbf{x}|\theta)$.
  Notice: Since $\mathbf{x} \in \mathbb{R}^{p\times 1}$, $P(\mathbf{x}|\theta)$ has p dimensions as well.

iid assumption

All data are independent and identically distributed (iid).

$$\mathbf{x}^{(i)} \overset{iid}{\sim} P(\mathbf{x}|\theta), \;\; i \in \{1, 2, \cdots, N\}$$

- Can you take some examples?

iid assumption

All data are independent and identically distributed (iid).

$$\mathbf{x}^{(i)} \overset{iid}{\sim} P(\mathbf{x}|\theta), \ \ i \in \{1, 2, \cdots, N\}$$

- Can you take some examples?
  1. Flip a coin N times.
     The probability the event 'head' happens is $1/2$ in each time.

iid assumption

All data are independent and identically distributed (iid).

$$\mathbf{x}^{(i)} \overset{iid}{\sim} P(\mathbf{x}|\theta), \;\; i \in \{1, 2, \cdots, N\}$$

- Can you take some examples?
  1. Flip a coin N times.
     The probability the event 'head' happens is $1/2$ in each time.
  2. Draw a dice N times.
     The probability the event '1' happens is $1/6$ in each time.

iid assumption

All data are independent and identically distributed (iid).

$$\mathbf{x}^{(i)} \overset{iid}{\sim} P(\mathbf{x}|\theta), \ \ i \in \{1, 2, \cdots, N\}$$

- Can you take some examples?
  1. Flip a coin N times.
     The probability the event 'head' happens is $1/2$ in each time.
  2. Draw a dice N times.
     The probability the event '1' happens is $1/6$ in each time.
  3. Scores of students in our course are iid, $x \overset{iid}{\sim} \mathcal{N}(90, 5^2)$.

iid assumption

All data are independent and identically distributed (iid).

$$\mathbf{x}^{(i)} \overset{iid}{\sim} P(\mathbf{x}|\theta), \;\; i \in \{1, 2, \cdots, N\}$$

- Can you take some examples?
  1. Flip a coin N times.
     The probability the event 'head' happens is $1/2$ in each time.
  2. Draw a dice N times.
     The probability the event '1' happens is $1/6$ in each time.
  3. Scores of students in our course are iid, $x \overset{iid}{\sim} \mathcal{N}(90, 5^2)$.
- **Why *iid* assumption is important?**
  For a set of N observed samples, $\mathbf{X} = [\mathbf{x}^{(1)},\; \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(N)}]^T$, the probability that the event $\mathbf{X}$ happens is $P(\mathbf{X}|\theta) = \prod_{i=1}^{N} P(\mathbf{x}^{(i)}|\theta)$.

Frequentist vs Bayesian

Frequentist:

- **Core idea**: The parameters θ are constant.

Frequentist vs Bayesian

Frequentist:

- **Core idea**: The parameters θ are constant.
- Maximum log-likelihood estimation (*i.e.*, MLE) to estimate parameters θ:

$$
\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta} \log P(\mathbf{X}|\theta) && (3)\\
&= \arg\max_{\theta} \log \prod_{i=1}^{N} P(\mathbf{x}^{(i)}|\theta) && (4)\\
&= \arg\max_{\theta} \sum_{i=1}^{N} \log P(\mathbf{x}^{(i)}|\theta) && (5)
\end{aligned}
$$

Frequentist vs Bayesian

Bayesian:

- **Core idea**: The parameters θ are *not* constant, which have their prior distributions $P(\theta)$.

Frequentist vs Bayesian

Bayesian:

- **Core idea**: The parameters θ are *not* constant, which have their prior distributions $P(\theta)$.
- Recall Bayes' theorem:

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|\theta)P(\theta)}{\int_\theta P(\mathbf{X}|\theta)P(\theta)d\theta} \propto P(\mathbf{X}|\theta)P(\theta)$$

Frequentist vs Bayesian

Bayesian:

- **Core idea**: The parameters θ are *not* constant, which have their prior distributions $P(\theta)$.
- Recall Bayes' theorem:

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})} = \frac{P(\mathbf{X}|\theta)P(\theta)}{\int_\theta P(\mathbf{X}|\theta)P(\theta)d\theta} \propto P(\mathbf{X}|\theta)P(\theta)$$

- Maximum a posterior (*i.e.*, MAP) to estimate parameters θ:

$$\theta_{MAP} = \arg\max_\theta \log P(\theta|\mathbf{X}) \tag{6}$$

$$= \arg\max_\theta \sum_{i=1}^{N} [\log P(\mathbf{x}^{(i)}|\theta) + \log P(\theta)] \tag{7}$$

1. Recap
2. Review Probability Theory
3. Frequentist vs Bayesian
4. Rethink linear regression
5. Summary

Rethink linear regression as a Probabilistic model

- Probabilistic model for linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon = \beta^T \mathbf{x} + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Rethink linear regression as a Probabilistic model

- Probabilistic model for linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon = \beta^T \mathbf{x} + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

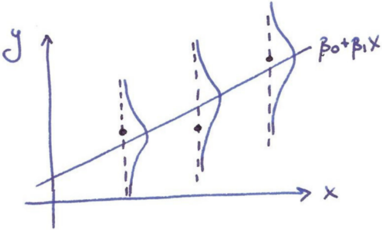


图 5: The error from each sample is assumed as iid Gaussian distribution.

Likelihood

- Probabilistic model for linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon = \beta^T \mathbf{x} + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Likelihood

- Probabilistic model for linear regression:

$$
\begin{aligned}
&y = \beta_0 + \beta_1 x + \epsilon = \beta^T \mathbf{x} + \epsilon \\
&\epsilon \sim \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

- For a given β and given $\mathbf{x}^{(i)}$, the likelihood $P(y|\mathbf{x}^{(i)}, \beta)$ is as follows,

$$
y^{(i)} \sim \mathcal{N}(\beta^T \mathbf{x}^{(i)}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y - \beta^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right] \quad (8)
$$

Likelihood

- Probabilistic model for linear regression:

$$y = \beta_0 + \beta_1 x + \epsilon = \beta^T \mathbf{x} + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

- For a given β and given $\mathbf{x}^{(i)}$, the likelihood $P(y|\mathbf{x}^{(i)}, \beta)$ is as follows,

$$y^{(i)} \sim \mathcal{N}(\beta^T \mathbf{x}^{(i)}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y - \beta^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right] \quad (8)$$

- For the entire training set $(\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$,

$$P(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y^{(i)} - \beta^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right] \quad (9)$$

Linear regression in frequentist view

- For the entire training set $(\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$,

$$P(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y^{(i)} - \beta^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right]$$

Linear regression in frequentist view

- For the entire training set $(\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$,

$$P(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y^{(i)} - \beta^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right]$$

- MLE to estimate the parameter β

$$\begin{aligned}\beta_{MLE} &= \arg\max_{\beta} \log P(\mathbf{y}|\mathbf{X}, \beta) \\ &= \arg\max_{\beta} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y^{(i)} - \beta^T \mathbf{x}^{(i)})^2}{2\sigma^2}\right]\right) \\ &= -\arg\max_{\beta} \sum_{i=1}^{N} \left(y^{(i)} - \beta^T \mathbf{x}^{(i)}\right)^2\end{aligned}$$

MLE = minimizing *negative log-likelihood* = minimizing MSE

- MLE is the same as minimizing the *negative log-likelihood*.

MLE = minimizing *negative log-likelihood* = minimizing MSE

- MLE is the same as minimizing the *negative log-likelihood*.
- MLE to estimate the parameter β

$$\begin{aligned}\beta_{MLE} &= -\arg\max_{\beta} \sum_{i=1}^{N} \left(y^{(i)} - \beta^T \mathbf{x}^{(i)}\right)^2 \\ &= \arg\min_{\beta} \sum_{i=1}^{N} \left(y^{(i)} - \beta^T \mathbf{x}^{(i)}\right)^2 \\ &= \arg\min_{\beta} ||y - \mathbf{X}\beta||^2\end{aligned}$$

MLE = minimizing *negative log-likelihood* = minimizing MSE

- MLE is the same as minimizing the *negative log-likelihood*.
- MLE to estimate the parameter β

$$\begin{aligned}\beta_{MLE} &= -\arg\max_{\beta} \sum_{i=1}^{N} \left(y^{(i)} - \beta^{T}\mathbf{x}^{(i)}\right)^2 \\ &= \arg\min_{\beta} \sum_{i=1}^{N} \left(y^{(i)} - \beta^{T}\mathbf{x}^{(i)}\right)^2 \\ &= \arg\min_{\beta} ||y - \mathbf{X}\beta||^2\end{aligned}$$

- Big surprise! It is the same as the mean-squared-error (MSE) loss in Lecture 2.
  Here, MLE is the same as minimizing MSE loss.

1. Recap

2. Review Probability Theory

3. Frequentist vs Bayesian

4. Rethink linear regression

5. Summary

Summary of Lecture 4

- **Probability theory**: random variable, PMF, PDF
  Gaussian distribution, Multivariate Gaussian distribution
- **Frequentist vs Bayesian**:
  1. Frequentist: The parameters θ are constant.
     Maximum (log-)likelihood estimation (MLE) to estimate θ:

     $$\theta_{MLE} = \arg\max_{\theta} \log P(\mathbf{X}|\theta)$$

  2. Bayesian: The parameters θ are not constant. We have some prior information of θ.
     Maximum a posterior (MAP) to estimate θ:

     $$\theta_{MAP} = \arg\max_{\theta} \log P(\theta|\mathbf{X}) = \arg\max_{\theta} \log P(\mathbf{X}|\theta) + \log P(\theta)$$

- **Linear regression in frequentist view**:
  MLE = minimizing *negative log-likelihood* = minimizing MSE