

DDA2001: Introduction to Data Science

Midterm Review

Baicheng Chen

SDS, CUHK-Shenzhen

baichengchen@link.cuhk.edu.cn

Part 1: Probability (Include Lecture 1~9)

1. Mutually exclusive (Disjoint) \neq Independent

Disjoint: $P(A \cup B) = P(A) + P(B)$, otherwise,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Independent: $P(A \cap B) = P(A)P(B)$

2. Zero-probability event ($P(E) = 0$) \neq Impossible event ($E = \emptyset$)

Impossible Event \Rightarrow Zero-Probability Event, but the inverse is not true.

Useful example:

- {A: randomly pick a point from [0.1, 0.2]}
- {A': randomly pick a point from [0.1, 0.2]}
- {B: randomly pick a point from [0.2, 0.3]}

3. De Morgan's law: $(A \cup B)' = A' \cap B'$, $(A \cap B)' = A' \cup B'$

4. Functions:

(1) PMF: For discrete R.V., $f(x_i) = P(X = x_i)$, $\sum_{i=1}^n f(x_i) = 1$.

(2) PDF: For continuous R.V., $P(a \leq x \leq b) = \int_a^b f(x)dx$, $\int_{-\infty}^{\infty} f(x)dx = 1$.

$$\begin{cases} f(x) > 0, & \text{if } x \in S \\ f(x) = 0, & \text{if } x \notin S \end{cases}$$

(3) CDF: $F(x) = P(X \leq x) = \sum_{\tilde{x} \leq x} f(\tilde{x})$, $F(x) = P(X \leq x) = \int_{-\infty}^x f(u)du$. We can

consider CDF are areas, and use them to do some calculations.

Use CDF to calculate PMF: $P(X = x) = F(x) - F(x^-)$.

5. Discrete and Continuous:

- A sample space is discrete if it consists of a finite or countable infinite set of outcomes.
- A sample space is continuous if it contains an interval or a union of multiple intervals of real numbers.

6. Mean: $E[X] = \sum_x x f(x)$, $E[X] = \int_{-\infty}^{\infty} x f(x)dx$

Variance: $Var[X] = \sum_x (x - E[X])^2 f(x) = E[(X - E[X])^2] = E[X^2] - E[X]^2$

7. Linearity: $E[\sum_i C_i X_i] = \sum_i C_i E[X_i]$

Expectation of a function of X: $E[g(X)] = \sum_x g(x)P(X = x) = \sum_x g(x)f(x)$

8. Distributions:

Discrete R.V. distributions:

(1) Bernoulli Distribution: If $X \sim \text{Bernoulli}(p)$,

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

$$E(X) = p, \text{ Var}X = p(1 - p).$$

- (2) Binomial Distribution: If $X \sim \text{Binomial}(N, p)$, X denotes the **number** of success/failures during the first N experiments,

$$f(x) = p^{x_i}(1-p)^{1-x_i}, \quad x_i = 0/1$$

$$E(X) = Np, \text{Var}X = Np(1-p).$$

- (3) Geometric Distribution: If $X \sim \text{Geometric}(p)$, the X -th sample is the first success,

$$f(x) = P(X = k) = (1-p)^{k-1} \cdot p$$

$$E(X) = \frac{1}{p}, \text{Var}X = \frac{1-p}{p^2}. \text{ (Remember how to calculate them!)}$$

- (4) Discrete Uniform Distribution: If $X \sim U(n)$,

$$f(x) = \frac{1}{n}$$

$$E(X) = \frac{n+1}{2}, \text{Var}X = \frac{n^2-1}{12}.$$

Continuous R.V. distributions:

- (5) Uniform Distribution: If $X \sim \text{Uniform}(a, b)$,

$$f(x) = \frac{1}{b-a}$$

$$E(X) = \frac{a+b}{2}, \text{Var}X = \frac{(b-a)^2}{12}.$$

- (6) Normal Distribution: If $X \sim \text{Normal}(\mu, \sigma)$,

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu, \text{Var}X = \sigma^2.$$

Empirical Rule (68-95-99.7 Rule): Within one std (standard deviation of the mean), 68%, two \rightarrow 95%, three \rightarrow 99.7%.

Standard Normal Distribution: $X \sim \text{Normal}(0,1)$,

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

If an integral contains both e and x^2 , we can use the standard normal distribution to calculate the integral.

- (7) Exponential Distribution: If $X \sim \text{Exponential}(\beta)$,

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & 0 < x < \infty, \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \beta, \text{Var}X = \beta^2.$$

9. Approximate Problem: How to approximate $\int_a^b h(x)dx$?

Answer:

Step 1: If $X \sim \text{Uniform}(a, b)$, then the pdf of X will be ... ($f(x)$)

Step 2: I can prove that $E[(b-a)h(X)] = \int_a^b (b-a)h(x)f(x)dx = \int_a^b h(x)dx$

Step 3: Thus, by Lindeberg-Lévy CLT, I can approximate the integral by

- Draw n samples of $X \sim \text{Uniform}(a, b) = x_1, x_2, x_3, \dots, x_n$

- Calculate $\frac{\sum_i (b-a)h(x_i)}{n}$

Example: How to approximate π ?

Step 1: Draw a two-dimensional point from the square

- $X \sim \text{Uniform}(-1,1)$
- $Y \sim \text{Uniform}(-1,1)$

Then with the same chance (X, Y) is any point in the square

Step 2: Let $g(X, Y) = 1$ if $X^2 + Y^2 \leq 1$ and 0 otherwise. Then as the area represents probability, $E[g(X, Y)] = \frac{\pi}{4}$

Step 3: Thus, by Lindeberg-Lévy CLT, I can approximate π by

- Draw n samples of $X \sim \text{Uniform}(a, b) = x_1, x_2, x_3, \dots, x_n$
- Draw n samples of $Y \sim \text{Uniform}(a, b) = y_1, y_2, y_3, \dots, y_n$
- Calculate the proportion of $x_i^2 + y_i^2 \leq 1$

Step 4: Then π is approximated by the proportion calculated in step 3, multiplied by 4.

10. Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Part 2: Statistics (Include Lecture 10~11)

1. What is statistics? (Knowing) samples \rightarrow (estimate) true model
2. Need: Samples, possible models, criterion for quantifying model performance
3. Notation: The point estimator used to estimate a parameter θ is usually denoted as $\hat{\theta}$.
4. $\hat{\theta}$ is a function of samples (X_1, X_2, \dots, X_n) called statistic.
5. How to choose the statistic? In our course, use MLE (Maximum likelihood estimate)!
Why? Maximize the probability.
Given a model, the probability of generating such samples is called likelihood, notated as $L(\theta) = P(X_1, X_2, \dots, X_n | \theta)$. So, we want to find a parameter θ to maximize $L(\theta)$.
 \rightarrow We define $l(\theta) = \log L(\theta)$, called log-likelihood. (In our course, $\log = \ln$).
 \rightarrow To maximize $L(\theta)$ is to maximize $l(\theta)$, we can easily calculate $l'(\theta) = 0$ to find $\hat{\theta}$.
6. Likelihood Function:
Given a model with an unknown parameter θ . Given samples: X_1, X_2, \dots, X_n .
Continuous RV model (f is PDF):
 - Likelihood: $L(\theta) = \prod_i f(X_i | \theta)$
 - Log-likelihood: $l(\theta) = \sum_i \log f(X_i | \theta)$Discrete RV model (P is PMF):
 - Likelihood: $L(\theta) = \prod_i P(X_i | \theta)$
 - Log-likelihood: $l(\theta) = \sum_i \log P(X_i | \theta)$
7. Solve problem through MLE:
Step 1: Judge the type of distribution, writing the PDF/PMF.
Step 2: Use the formula in 6 to write the $L(\theta)$ and $l(\theta)$.
Step 3: Find $l'(\theta)$, let it equals to 0, find $\hat{\theta}$.
Examples are in the slides.
8. Linear Regression: Use a line to find the relationship between X and Y .
We use the model $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$ with normal distribution (samples are centralized) to represent. So, we should then use MLE to find the best β_0 , β_1 and σ^2 .

After least square regression and partial derivatives, we finally get the best parameters.

$$\begin{cases} \widehat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} \\ \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \end{cases}$$

Linear regression assumes:

- (1) The relationship between X and Y is linear.
 - (2) The variance of $Y - \beta_0 - \beta_1 X$ at every value of X is the same (the homogeneity of variances).
 - (3) Different observations are independent of each other.
9. Residual Analysis: To check the two assumptions of linear regression. They are linear and constant variance.

$$e_i := Y - \widehat{\beta}_0 - \widehat{\beta}_1 X$$