DDA2001: Introduction to Data Science

Midterm Review

Baicheng Chen

<baichengchen@link.cuhk.edu.cn>

Part 1: Probability (Include Lecture 1~9)

1. Mean:

Variance:

1. Linearity:

Expectation of a function of X:

1. Distributions:

**Discrete R.V. distributions:**

1. Bernoulli Distribution: If ,

, .

1. Binomial Distribution: If , X denotes the **number** of success/failures,

, .

1. Geometric Distribution: If the X-th sample is the first success,

, .

**Continuous R.V. distributions:**

1. Uniform Distribution: If ,

, .

1. Normal Distribution: If ,

, .

1. Exponential Distribution: If ,

, .

1. Approximate Problem: How to approximate ?

Answer:

Step 1: If , then the pdf of X will be … ()

Step 2: I can prove that

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

* Draw n samples of =
* Calculate

Example: How to approximate π?

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

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

Step 2: Let if and 0 otherwise. Then as the area represents probability,

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

* Draw n samples of =
* Draw n samples of =
* Calculate the proportion of

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

Part 2: Statistics (Include Lecture10~11)

1. What is statistics? (Knowing) samples → (estimate) true model
2. Need: Samples, possible models, criterion for quantifying model performance
3. Notation: The point estimator used to estimate a parameter is usually denoted as .
4. is a function of samples 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 . So, we want to find a parameter to maximize .

→ We define , called log-likelihood. (In our course, log=ln).

→ To maximize is to maximize , we can easily calculate to find .

1. Likelihood Function:

Given a model with an unknown parameter . Given samples: .

Continuous RV model ( is PDF):

* Likelihood:
* Log-likelihood:

Discrete RV model ( is PMF):

* Likelihood:
* Log-likelihood:

1. 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 and .

Step 3: Find , let it equals to 0, find .

Examples are in the slides.

1. Linear Regression: Use a line to find the relationship between X and Y.

We use the model with normal distribution (samples are centralized) to represent. So, we should then use MLE to find the best , and . After least square regression and partial derivatives, we finally get the best parameters.

Linear regression assumes:

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