Probability in Computing

LECTURE 8: CENTRAL LIMIT THEOREMS

Agenda

- Quick look at (30 min)
  - Law of large numbers
  - Normal distribution
  - Central limit theorem
- Midterm (60 min)

Law of large numbers

- The **law of large numbers** (**LLN**) is a theorem that describes the result of performing the same experiment a large number of times:
  - The average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.
  - if a large number of dice are rolled, the average of their values (sometimes called the sample mean) is likely to be close to 3.5, with the accuracy increasing as more dice are rolled.

LLN's importance

- The LLN is important because it "guarantees" stable long-term results for random events.
  - For example, while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins.
  - Any winning streak by a player will eventually be overcome by the parameters of the game. It is important to remember that the LLN only applies (as the name indicates) when a *large number* of observations are considered.

Basic idea of LLN

- With virtual certainty -- the sample average

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

converges to the expected value

  - where X_1, X_2, ... is an infinite sequence of i.i.d. (independent and identically distributed) random variables with finite expected value

    $E(X_1) = E(X_2) = \ldots = \mu < \infty$.

- The strong law and the weak law:
  - The two versions are concerned with the mode of convergence being asserted.

Weak law of large numbers

- The sample average converges in probability towards the expected value

$$\overline{X}_n \xrightarrow{p} \mu \qquad \text{when } n \to \infty.$$

- That is to say that for any positive number ε,

$$\lim_{n \to \infty} \Pr\left(|\overline{X}_n - \mu| < \varepsilon \right) = 1.$$

  - the weak law essentially states that for any nonzero margin specified, no matter how small, with a sufficiently large sample there will be a very high probability that the average of the observations will be close to the expected value, that is, within the margin.

Strong Law

- The **strong law of large numbers** states that the sample average converges almost surely to the expected value

$$\overline{X}_n \xrightarrow{a.s.} \mu \qquad \text{when } n \to \infty.$$

- That is

$$\Pr\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

The normal distribution

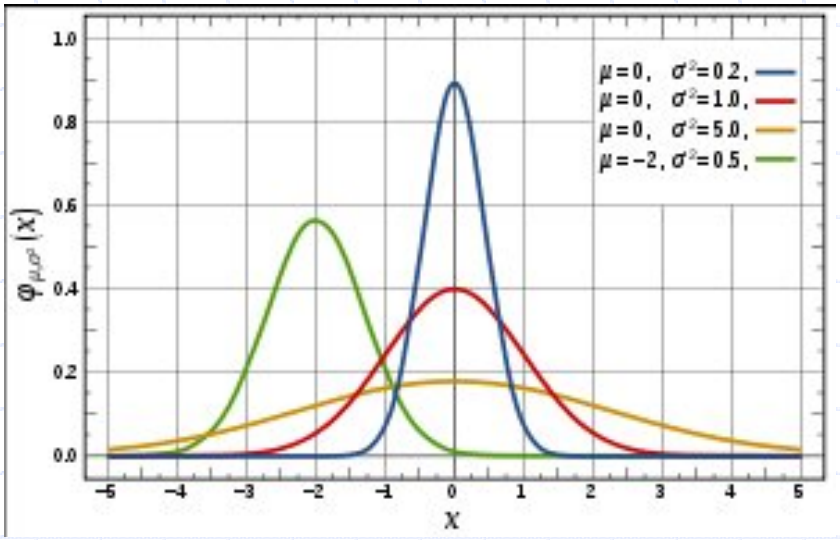
- The **normal distribution** or **Gaussian distribution**
  - A continuous probability distribution that often gives a good description of data that cluster around the mean
  - The graph of the associated probability density function is bell-shaped, with a peak at the mean, and is known as the **bell curve**.
  - The simplest case of a normal distribution is known as the **standard normal distribution**, described by pdf

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

  - More general with pdf $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$
  - Thus when a random variable X is distributed normally with mean μ and variance σ^2, we write $X \sim \mathcal{N}(\mu, \sigma^2).$

The normal distribution

- The **normal distribution** or **Gaussian distribution** is often used to describe, at least approximately, any variable that tends to cluster around the mean.
  - E.g., the heights of adult males in the United States are roughly normally distributed, with a mean of about 70 inches (1.8 m). Most men have a height close to the mean, though a small number of outliers have a height significantly above or below the mean.



Central limit theorem

- By the central limit theorem, under certain conditions the sum of a number of random variables with finite means and variances approaches a normal distribution as the number of variables increases.
  - For this reason, the normal distribution is commonly encountered in practice, and is used throughout statistics, natural science, and social science as a simple model for complex phenomena.
  - For example, the observational error in an experiment is usually assumed to follow a normal distribution
- The central limit theorem is also known as the second fundamental theorem of probability (first is LLN)

Central limit theorem

- Let X_1, X_2, X_3, …, X_n be a sequence of n i.i.d. random variables each having finite values of expectation μ and variance $\sigma^2 > 0$. The CLT states that
  - as the sample size n increases the distribution of the sample *average* of these random variables approaches the normal distribution with a mean μ and variance σ^2/n irrespective of the shape of the common distribution of the individual terms X_i.
  - More precisely, let S_n be the sum $S_n = X_1 + \cdots + X_n.$
  - Then, if we define new random variables $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}},$
  - then they will converge in distribution to the standard normal distribution $N(0,1)$ as n approaches infinity.

$$Z_n \xrightarrow{d} \mathcal{N}(0,1).$$

Midterm

- *Đề thi giữa kỳ Toán chuyên đề. (Người ra đề: TS. Nguyễn Khánh Văn)*
- *Thời gian làm bài **60** phút. Điền đầy đủ các thông tin dưới đây rồi **NỘP ĐỀ BÀI KÈM THEO BÀI LÀM**.*
  - *Có thể sử dụng tài liệu giấy nhưng nghiêm cấm dùng máy tính, điện thoại di động và các thiết bị có chức năng nhớ/liên lạc nói chung.*
  - ***Họ tên:***
  - ***MSSV (mã số sinh viên):*** ***Lớp:***
  - **X=** **Y=**
  - Trong đó gọi **X** là số xác định bởi hai chữ số cuối của MSSV. **Gọi Y = X mod 60 +5**
  - (Đồng thời ghi các thông tin trên trong bài làm)

Midterm

1. ***Randomized Min Cut***

Cho đồ thị G có n=**Y** đỉnh. Giả sử C là một min-cut của đồ thị G. Ta xem xét thuật toán Karger và phân tích xác suất sự kiện thuật toán cho kết quả là C.

a) Đánh giá xác xuất sự kiện F1: cạnh đầu tiên chọn để collapse không nằm trong C. Lý giải về cách tính.

b) Đánh giá xác xuất sự kiện F2: hai cạnh đầu tiên chọn để collapse không nằm trong C. Lý giải.

c) Nếu muốn xác suất chọn sai (min-cut đồ thị) không vượt quá 0.0001 ta phải tiến hành lặp thuật toán Karger bao nhiêu lần trở lên?

Midterm

2. Hashing

a) Để kiểm tra mật khẩu người dùng tạo ra có tốt hay không, người ta xây dựng một từ điển gồm 2^{32} mật khẩu nên tránh dùng bằng cách sử dụng 64-bit fingerprints của chúng để so sánh. Cho biết xác suất loại bỏ nhầm mật khẩu tốt trong hệ kiểm tra này (false positive probability).

- Một bảng băm có *n= Y*1000* vị trí (slots/bins). Sử dụng một hàm băm ngẫu nhiêm, người ta lần lượt đưa *m* khóa vào bảng băm này.

b) Cho biết sau khi đưa vào khoảng bao nhiêu khóa thì bảng băm hết ô trống? (tính kỳ vọng)

Midterm

2c. Trong giải pháp Open addressing/linear probing, người ta không dùng danh sách móc nối để lưu các khóa cùng giá trị băm mà tận dụng triệt để chỗ trống của bảng băm như sau:

- khi lưu một khóa mới nếu ô ứng với giá trị băm đã bị chiếm người ta tìm đến ô gần nhất ngay sau đó mà còn trống để lưu khóa này
- ngược lại, khi đi tìm khóa trong bảng băm trước hết kiểm tra ô có địa chỉ bằng giá trị băm, nếu không thấy khóa cần tìm thì lần lượt kiểm tra các ô kế tiếp cho đến khi thấy.
- Giả sử ta đã đưa vào bảng được m=n/2 khóa theo phương pháp này. Hỏi rằng tìm kiếm trên bảng đang có sẽ mất trung bình (kỳ vọng) bao nhiêu phép kiểm tra ô?

Midterm

3. Quicksort

Trong phân tích của thuật toán QuickSort, cho biết ý nghĩa của biến ngẫu nhiên X_{ij} và công dụng của nó. Lập luận lý giải chi tiết công thức $E[X_{ij}] = 2/ (j-i+1)$.

Gợi ý: Nếu một phần tử ở giữa y_i và y_j được chọn làm pivot thì X_{ij} sẽ có giá trị gì?