Mathematical techniques in data science

Lecture 2: Recap on Python and Probability

General information

- Classes:

  MW 5:00pm-6:15pm, Ewing Hall 207
- Office hours (starting from the 2nd week):
  - Tuesdays 5:00pm-6:00pm, via Zoom
  - Wednesdays 3:30pm-4:30pm, Ewing Hall 312
  - By appointments
- Instructor: Vu Dinh
- Website:

  `https://vucdinh.github.io/m637s23`

Evaluation

- Homework (theoretical + programming problems): 60%
- Final project: 40% (10% presentation, 30% final report) with a possible **+5% bonus**
- Grading system:

$\geq$ 94%	At least A
$\geq$ 90%	At least A-
$\geq$ 80%	At least B-
$\geq$ 70%	At least C-
$\geq$ 60%	At least D-
$<$ 60%	F

Platforms

- We will use Python during the course (there will be sessions to review the language). Specifically, we will use Google Colab for coding and programming assignments:

  `https://colab.research.google.com`

- We will use LaTeX to write the final report. The easiest way to use it collaboratively is to register an Overleaf account:

  `https://www.overleaf.com`

Final project

- Group project: 5-6 people (sign up on Canvas)
- The groups should be formed by the end of Week 4
- Data-oriented projects
  - Pick a practical learning problem with a dataset
  - Analyze the dataset
  - Write a report (in the form of a 4-page IEEE conference paper)
  - Present the project (last week of the semester)

Tentative schedule

- Introduction to (supervised) machine learning (4 weeks)
- Mathematical techniques in data science (8 weeks)
- Final project presentations (1 week)

Review: Probability

Topics

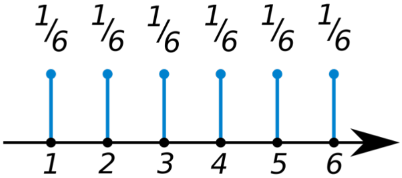
- Notations and definitions
- Basic probability rules
- Normal distribution

Random variables

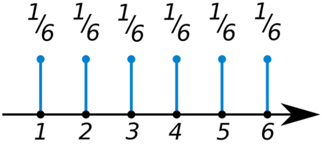
- Random variable X: used to describe random quantities
  Example: X $=$ number we get when rolling a dice
- Sample space: set of all possible outcomes of X
  Example: sample space $= \{1, 2, 3, 4, 5, 6\}$
- Event: a subset of sample space
  Example: event that X is even $= \{2, 4, 6\}$

Discrete random variable

- Sample space is discrete
- Probability mass function (pmf):
  - Assign a probability value to each outcome in sample space
  - Example: $P(1) = P(2) = \ldots = P(6) = 1/6$



Discrete random variable



- Probability of an event A:

$$P(A) = \sum_{x \in A} P(x)$$

  Example: $P(\{\text{X is even}\}) = P(2) + P(4) + P(6) = 1/2$

- Sometimes we write $P(X = x)$ for $P(x)$, for example, $P(X = 2) = P(2)$.

Continuous random variable

- Sample space is continuous (real values)
- Characterized by a density function P:
  - $P(x) \geq 0$ for all $x \in \mathbb{R}$
  - $\int_{-\infty}^{\infty} P(x)\ dx = 1$
  - For any fixed constant a, b,

$$P(a \leq X \leq b) = \int_a^b P(x)\ dx$$

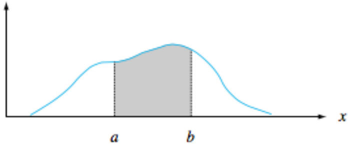


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

Joint probability distribution

- Random variables X and Y
- Sample space of $X = \{x_1, x_2, \ldots, x_n\}$
- Sample space of $Y = \{y_1, y_2, \ldots, y_m\}$
- Joint probability distribution of X and Y: assigns probability to each combination of values of X and Y.
- $P(X = x, Y = y) = P(x, y)$: probability that X has value x and Y has value y

Marginal distribution

- Marginal distribution of X:

$$P(x) = \sum_y P(x, y) = \sum_{i=1}^{m} P(x, y_i)$$

- Can be extended to more than two random variables:

$$P(z) = \sum_x \sum_y P(x, y, z)$$

- For continuous random variables

$$P(x) = \int_y P(x, y)\ dy$$

Conditional probability distribution

- Probability of X = x given Y = y:

$$P(y|x) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

- Product rule:

$$P(x, y) = P(x)P(y|x)$$

- Bayes' rule: very important in machine learning; allow us to reverse the order of conditional probabilities

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)} = \frac{P(x)P(y|x)}{\sum_{x'} P(x')P(y|x')}$$

Expectation of random variables

- Expectation (expected value or mean) of a discrete random variable X:

$$E[X] = \sum_{x} xP(x) = \sum_{i=1}^{n} x_i P(x_i)$$

- For continuous variables:

$$E[X] = \int_{x} xP(x)dx$$

- Can be used for functions:

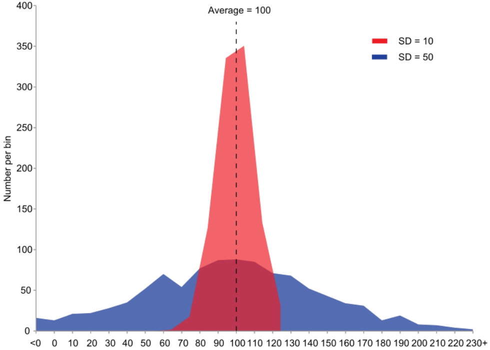
$$E[g(X)] = \sum_{x} g(x)P(x)$$

or

$$E[g(X)] = \int_{x} g(x)P(x)dx$$

Variance of random variables

- Measure the spread of values of a random variable around the mean:
$$Var(X) = E[(X - E(X))^2]$$
- Standard deviation: $sd(X) = \sqrt{Var(X)}$

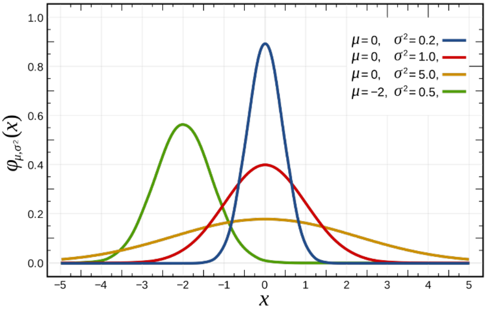


Normal distribution (Gaussian distribution)

- Notation: $\mathcal{N}(\mu, \sigma^2)$
- Continuous random variable with density

$$\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- $E(X) = \mu,\ Var(X) = \sigma^2$



Linear combination of random variables

Theorem

Let $X_1, X_2, \ldots, X_n$ be independent random variables (with possibly different means and/or variances). Define

$$T = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

then the mean and the standard deviation of T can be computed by

- $E(T) = a_1 E(X_1) + a_2 E(X_2) + \ldots + a_n E(X_n)$
- $\sigma_T^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma_{X_2}^2 + \ldots + a_n^2 \sigma_{X_n}^2$

Example

Let $X_1, X_2, \ldots, X_n$ be independent random sample from a distribution with μ and standard deviation σ.
Define

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

What are the mean and the standard deviation of $\bar{X}$?

Mean and variance of the sample mean

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

1. $E(\overline{X}) = \mu_{\overline{X}} = \mu$

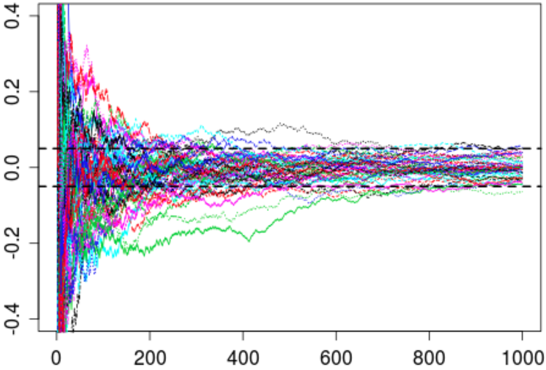
2. $V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n \quad \text{and} \quad \sigma_{\overline{X}} = \sigma/\sqrt{n}$

Law of large numbers

THEOREM If $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and variance σ^2, then $\overline{X}$ converges to μ

a. In mean square $E[(\overline{X} - \mu)^2] \rightarrow 0$ as $n \rightarrow \infty$

b. In probability $P(|\overline{X} - \mu| \geq \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$



The Central Limit Theorem

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then, in the limit when $n \to \infty$, the standardized version of $\bar{X}$ have the standard normal distribution

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z\right) = \mathbb{P}[Z \leq z] = \Phi(z)$$

Example

Problem

When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g.

If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity X is between 3.5 and 3.8 g?

Hint:

- First, compute $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$
- Note that

$$\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

  is (approximately) standard normal.