

CS4641 Spring 2025

Probability and Statistics Revisit

Bo Dai
School of CSE, Georgia Tech
bodai@cc.gatech.edu

Course Homepage

<https://bo-dai.github.io/CS4641-spring2025/>

Syllabus

Cover a number of most commonly used machine learning algorithms in sufficient amount of details on their mechanisms.

Organization

- *Background knowledge*
 - Linear Algebra, Probability and Statistics, Optimization
- *Supervised learning*
 - Learning with labels
- *Unsupervised learning*
 - Learning without labels
- *Advanced Topics*
 - Foundation Models / Large Language Models

Grading

- Homework (30%)
- Project (40%)
- Exam (30%)
- Participation bonus (5%)

Homework

- There will be **three** assignments, each account for 10% towards your final score.
- **Late policy:** Assignments are due at 11:59PM of the due date. You will be allowed 2 total late days (48 hours) without penalty for the entire semester (for homework only, not applicable to exams or projects). Once those days are used, you will be penalized according to the following policy:
 - Homework is worth full credit before the due time.
 - It is worth 75% credit for the next 24 hours.
 - It is worth 50% credit for the second next 24 hours.
 - It is worth zero credit after that.

Homework

You are required to use Latex ([OverLeaf Latex Example in the Video](#)), or a word processing software to generate your solutions to the written questions.

Handwritten solutions WILL NOT BE ACCEPTED. You can easily export your Jupyter Notebook to a Python file and import that to your desired python IDE to debug your code for assignments.

Project

Team Size

Each project must be completed in a team of 3-5. Once you have formed your group, please send one email per team to the class instructor list with the names of all team members. If you have trouble forming a group, please send us an email and we will help you find project partners.

The team formation email will be due at **11:59 PM on Feb 10th**.

Project

Project Topics:

- Reproduce classic papers, include but not limited to:
 - [Deep Residual Learning for Image Recognition](#)
 - Auto-Encoding Variational Bayes.
 - A Simple Framework for Contrastive Learning of Visual Representations.
 - [Sequence to Sequence Learning with Neural Networks](#)
 - etc
- You may also refer to the <https://cs231n.stanford.edu/project.html>.

Project

Deliverables:

- Presentation (15%)
- Final Report (25%): *All write-ups should use the [NeurIPS style](#).*

Your final report is expected to be 5 pages excluding references. It should have roughly the following format:

- *Introduction: problem definition and motivation*
- *Background & Related Work: background info and literature survey ([optional](#))*
- *Methods – Overview of your proposed method – Intuition on why should it be better than the state of the art – Details of models and algorithms that you developed*
- *Experiments – Description of your testbed and a list of questions your experiments are designed to answer – Details of the experiments and results*
- *Conclusion: discussion and future work*

The project final report will be due at **11:59 PM on April 28th**

Project

Criteria:

- 30% for proposed method (soundness and originality)
- 30% for correctness, completeness, and difficulty of experiments and figures
- 20% for empirical and theoretical analysis of results and methods
- 20% for quality of writing (clarity, organization, flow, etc.)

Exam

One exam will be held on [March 12](#) in lieu of the regular class:

- It will be a closed-book exam, so no notes or communication with peers is allowed.
- There will be no make-up exams, so be sure to attend on the scheduled date. Missing the exam will result in zero credit.
- It will only cover the content introduced before March 12.

Participation Bonus

We will be awarding, on a case-by-case basis, up to 5% in extra credit to the top Ed contributors based on the number of (meaningful) instructor-endorsed answers or other significant contributions that assist the teaching staff or other students in the course.

The most helpful contributor will receive the greatest amount of extra credit, and other students with significant contributions will receive a percentage of that.

Probability and Statistics

Revist

Basic Probability Concepts

- A **sample space** S is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite.)
 - E.g., S may be the set of all possible outcomes of a dice roll: S
(1 2 3 4 5 6)
 - E.g., S may be the set of all possible nucleotides of a DNA site: S
(A C G T)
 - E.g., S may be the set of all possible time-space positions of an aircraft on a radar screen.
- An **Event** A is any subset of S
 - Seeing "1" or "6" in a dice roll; observing a "G" at a site; UA007 in space-time interval



Discrete Probability Distribution

- A probability distribution P defined on a discrete sample space S is an assignment of a non-negative real number $P(s)$ to each sample $s \in S$:
 - Probability Mass Function (PMF): $p_x(x_i) = P[X = x_i]$
 - Properties: $p_x(x_i) \geq 0$ and $\sum_i p_x(x_i) = 1$
- Examples:

- Bernoulli Distribution:

$$\begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

outcome of a coin

- Binomial Distribution:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Continuous Probability Distribution

- A continuous random variable X is defined on a continuous sample space: an interval on the real line, a region in a high dimensional space, etc.
 - It is meaningless to talk about the probability of the random variable assuming a particular value --- $P(x) = 0$
 - Instead, we talk about the probability of the random variable assuming a value within a given interval, half interval, or arbitrary Boolean combination of basic propositions.
 - Cumulative Distribution Function (CDF): $F_x(x) = P[X \leq x]$
 - Probability Density Function (PDF): $F_x(x) = \int_{-\infty}^x f_x(x)dx$ or $f_x(x) = \frac{dF_x(x)}{dx}$
 - Properties: $f_x(x) \geq 0$ and $\int_{-\infty}^{\infty} f_x(x)dx = 1$
 - Interpretation: $f_x(x) = P \left[X \in \frac{x, x+\Delta}{\Delta} \right]$

Continuous Probability Distribution

- Examples:
 - Uniform Density Function:

$$f_x(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

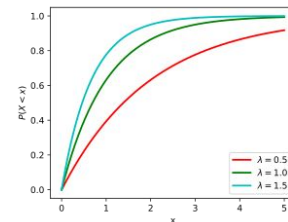
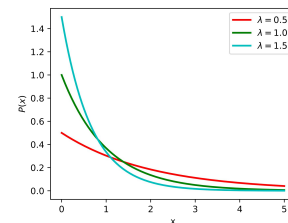
- Exponential Density Function:

$$f_x(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0$$

$$F_x(x) = 1 - e^{-\lambda x} \quad \text{for } x \geq 0$$

- Gaussian(Normal) Density Function

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

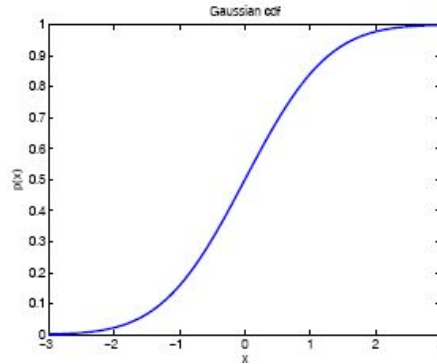
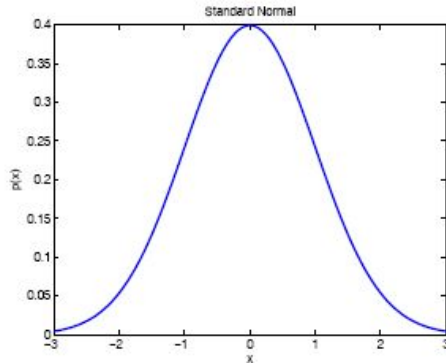


Continuous Probability Distribution

- Gaussian Distribution:
 - If $Z \sim N(0,1)$

$$F_x(x) = \Phi(x) = \int_{-\infty}^x f_x(z) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

- This has no closed form expression, but is built in most software packages.



Statistics

- Expectation: The mean value, center of mass, first moment:

$$E_X[g(X)] = \int_{-\infty}^{\infty} g(x)p_X(x)dx = \mu$$

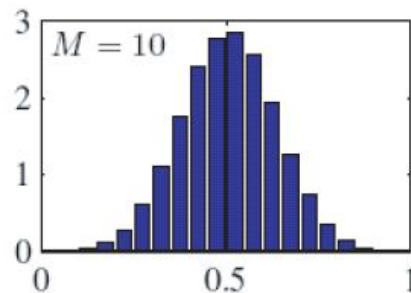
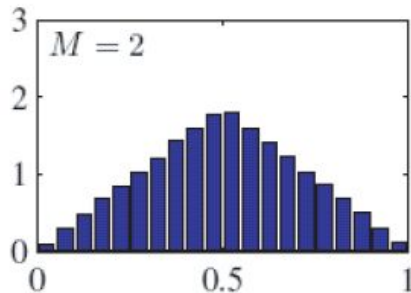
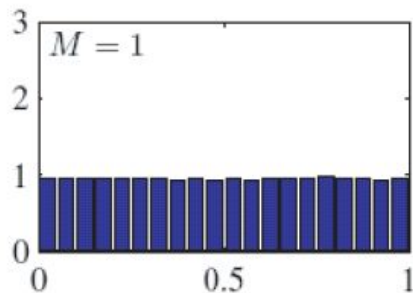
- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n$ $\mu = E_X[x]$
- Mean: $E_X[X] = \int_{-\infty}^{\infty} xp_X(x)dx$
 - $E[\alpha X] = \alpha E[X]$
 - $E[\alpha + X] = \alpha + E[X]$
- Variance(Second central moment): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
 - $Var(\alpha X) = \alpha^2 Var(X)$
 - $Var(\alpha + X) = Var(X)$

Central Limit Theorem

- If (X_1, X_2, \dots, X_n) are i.i.d. continuous random variables, then the joint distribution is $f(\bar{X})$
- CLT proves that $f(\bar{X})$ is Gaussian with mean $E[X_i]$ and $Var[X_i]$

$$\bar{X} = f(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{as } n \rightarrow \infty$$

- Somewhat of a justification for assuming Gaussian noise



Joint RVs and Marginal Densities

- Joint cumulative distribution:

$$F_{X,Y}(x, y) = P[\{X \leq x\} \cap \{Y \leq y\}] = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(\alpha, \beta) d\alpha d\beta$$

- Marginal densities:

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, \beta) d\beta$

- $p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j)$

- Expectation and Covariance:

- $E[X + Y] = E[X] + E[Y]$

- $cov(X, Y) = E[(X - E_X[X])(Y - E_Y[Y])] = E[XY] - E[X]E[Y]$

- $Var(X + Y) = Var(X) + 2cov(X, Y) + Var(Y)$

Conditional Probability

- $P(X | Y)$ = Fraction of the worlds in which X is true given that Y is also true.
- For example:
 - H = "Having a headache"
 - F = "Coming down with flu"
 - $P(\text{Headche} | \text{Flu})$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?
- Definition:

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

Corollary:

$$P(X, Y) = P(Y | X)P(X)$$

This is called **Bayes Rule**

Is the following statement TRUE or FALSE: $p(X = x|Y = y) = \frac{p(X=x)p(Y=y|X=x)}{\sum_x p(X=x')p(Y=y|X=x')}$

The Bayes Rule

$$P(\textit{Headache} | \textit{Flu}) = \frac{P(\textit{Headache}, \textit{Flu})}{P(\textit{Flu})} = \frac{P(\textit{Flu} | \textit{Headache})P(\textit{Headache})}{P(\textit{Flu})}$$

- Other cases:

- $$P(Y | X) = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y)+P(X|\neg Y)P(\neg Y)}$$

- $$P(Y = y_i | X) = \frac{P(X|Y)P(Y)}{\sum_{i \in S} P(X|Y=y_i)P(Y=y_i)}$$

- $$P(Y | X, Z) = \frac{P(X|Y,Z)P(Y,Z)}{P(X,Z)} = \frac{P(X|Y,Z)P(Y,Z)}{P(X|Y,Z)P(Y,Z)+P(X|\neg Y,Z)P(\neg Y,Z)}$$

Rules of Independence

- Recall that for events E and H , the probability of E given H , written as $P(E | H)$, is

$$P(E | H) = \frac{P(E, H)}{P(H)}$$

- E and H are (statistically) independent if

$$P(E, H) = P(E)P(H)$$

- Or equivalently

$$P(E) = P(E | H)$$

That means, the probability of E is true doesn't depend on whether H is true or not

- E and F are conditionally independent given H if

$$P(E | H, F) = P(E | H)$$

- Or equivalently

$$P(E, F | H) = P(E | H)P(F | H)$$

Suppose random variables Y, x and ϵ are related by $Y = \beta_0 + \beta_1 x + \epsilon$, with β_0 and β_1 are parameters and ϵ is assumed to independent of x and follow normal distribution with mean 0 and constant variance. Please calculate: (1) $E(\epsilon|x)$, and (2) $E(Y|x)$.

Suppose random variables Y, x and ϵ are related by $Y = \beta_0 + \beta_1 x + \epsilon$, with β_0 and β_1 are parameters and ϵ is assumed to independent of x and follow normal distribution with mean 0 and constant variance. Please calculate: (1) $E(\epsilon|x)$, and (2) $E(Y|x)$.

$$E[\epsilon|x] = E[\epsilon] = 0$$

Suppose random variables Y, x and ϵ are related by $Y = \beta_0 + \beta_1 x + \epsilon$, with β_0 and β_1 are parameters and ϵ is assumed to independent of x and follow normal distribution with mean 0 and constant variance. Please calculate: (1) $E(\epsilon|x)$, and (2) $E(Y|x)$.

$$E[\epsilon|x] = E[\epsilon] = 0$$

$$E[Y|x] = E[\beta_0 + \beta_1 x + \epsilon|x] = \beta_0 + \beta_1 x + E[\epsilon|x] = \beta_0 + \beta_1 x$$

- $E[X + Y] = E[X] + E[Y]$

Rules of Independence

- Examples:

$$P(\text{Virus} \mid \text{Drink Beer}) = P(\text{Virus})$$

iff **Virus** is independent of **Drink Beer**

$$P(\text{Flu} \mid \text{Virus}; \text{Drink Beer}) = P(\text{Flu} \mid \text{Virus})$$

iff **Flu** is independent of **Drink Beer**, given **Virus**

$$P(\text{Headache} \mid \text{Flu}; \text{Virus}; \text{Drink Beer}) = P(\text{Headache} \mid \text{Flu}; \text{Drink Beer})$$

iff **Headache** is independent of **Virus**, given **Flu** and **Drink Beer**

Assume the above independence, we obtain:

$$P(\text{Headache}; \text{Flu}; \text{Virus}; \text{Drink Beer})$$

$$= P(\text{Headache} \mid \text{Flu}; \text{Virus}; \text{Drink Beer}) P(\text{Flu} \mid \text{Virus}; \text{Drink Beer})$$

$$P(\text{Virus} \mid \text{Drink Beer}) P(\text{Drink Beer})$$

$$= P(\text{Headache} \mid \text{Flu}; \text{Drink Beer}) P(\text{Flu} \mid \text{Virus}) P(\text{Virus}) P(\text{Drink Beer})$$

Multivariate Gaussian

$$p(x | \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}$$

- Moment Parameterization $\mu = E(X)$

$$\Sigma = Cov(X) = E[(X - \mu)(X - \mu)^\top]$$

- Mahalanobis Distance $\Delta^2 = (x - \mu)^\top \Sigma^{-1} (x - \mu)$
- Tons of applications (MoG, FA, PPCA, Kalman filter,...)

Multivariate Gaussian $P(X_1, X_2)$

- Joint Gaussian $P(X_1, X_2)$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- Marginal Gaussian

$$\mu_2^m = \mu_2 \quad \Sigma_2^m = \Sigma_{22}$$

- Conditional Gaussian $P(X_1 | X_2 = x_2)$

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Operations on Gaussian R.V.

- The **linear transform** of a Gaussian r.v. is a Gaussian. Remember that no matter how x is distributed

$$E(AX + b) = AE(X) + b$$

$$\text{Cov}(AX + b) = ACov(X)A^T$$

this means that for Gaussian distributed quantities:

$$X \sim N(\mu, \Sigma) \rightarrow AX + b \sim N(A\mu + b, A\Sigma A^T)$$

- The **sum** of two independent Gaussian r.v. is a Gaussian

$$Y = X_1 + X_2, X_1 \perp X_2 \rightarrow \mu_y = \mu_1 + \mu_2, \Sigma_y = \Sigma_1 + \Sigma_2$$

- The **multiplication** of two Gaussian functions is another Gaussian function (although no longer normalized)

$$N(a, A)N(b, B) \propto N(c, C)$$

$$\text{where } C = (A^{-1} + B^{-1})^{-1}, c = CA^{-1}a + CB^{-1}b$$

Maximum Log-Likelihood Estimation (MLE)

- Example: toss a coin N times with n head
- Objective function:

$$l(\theta; \text{Head}) = \log P(\text{Head} | \theta) = \log \theta^n (1 - \theta)^{N-n} = n \log \theta + (N - n) \log (1 - \theta)$$


Maximum Log-Likelihood Estimation (MLE)

- Example: toss a coin N times with n head
- Objective function:

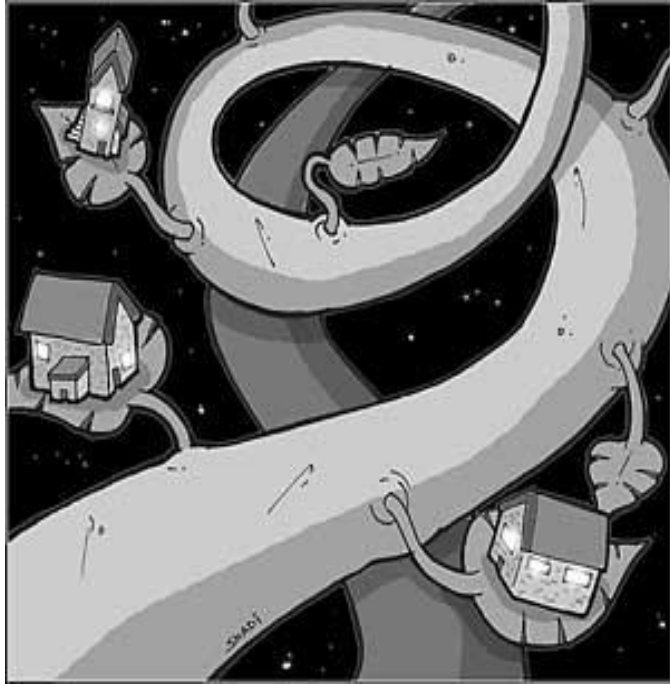
$$l(\theta; \text{Head}) = \log P(\text{Head} | \theta) = \log \theta^n (1 - \theta)^{N-n} = n \log \theta + (N - n) \log (1 - \theta)$$

- We need to maximize this w.r.t. θ
- Take derivatives w.r.t. θ

$$\frac{dl}{d\theta} = \frac{n}{\theta} - \frac{N-n}{1-\theta} = 0$$

 $\hat{\theta}_{MLE} = \frac{n}{N}$

Machine Learning for Apartment Hunting



- Suppose you are to move to Atlanta
- And you want to find the **most reasonably priced** apartment satisfying your **needs**:
monthly rent = $\theta_1(\text{living area}) + \theta_2(\# \text{ bedroom})$

Living area (ft ²)	# bedroom	Monthly rent (\$)
230	1	900
506	2	1800
433	2	1500
190	1	800
...		
150	1	?
270	1.5	?

Linear Regression Model

- Assume y is a linear function of x (features) plus noise ϵ
monthly rent = θ_1 (living area) + θ_2 (# bedroom)
$$y = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n + \epsilon$$

where ϵ is an error model as Gaussian $N(0, \sigma^2)$

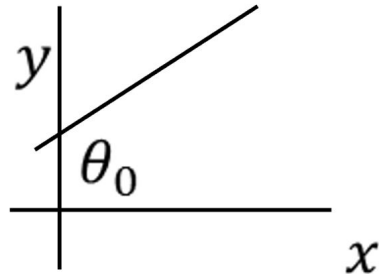
Probability

- Let $\theta = (\theta_0, \theta_1, \dots, \theta_n)^\top$, and augment data by one dimension

Linear algebra $x \leftarrow (1, x)^\top$

Then $y = \theta^\top x + \epsilon$

Linear algebra



Gaussian Likelihood

- Assume y is a linear in x plus noise ϵ

$$y = \theta^\top x + \epsilon$$

- Assume ϵ follows a Gaussian $N(0, \sigma)$

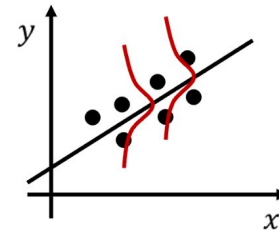
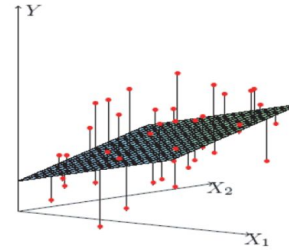
$$p(y^i | x^i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

- By independence assumption, likelihood is

$$L(\theta)$$

$$= \prod_i^m p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{\sum_i^m (y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

← Probability



MLE

$$L(\theta) = \prod_{i=1}^m p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^m \exp \left(- \frac{\sum_i^m (y^i - \theta^\top x^i)^2}{2\sigma^2} \right)$$

MLE

$$L(\theta) = \prod_{i=1}^m p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^m \exp \left(- \frac{\sum_{i=1}^m (y^i - \theta^\top x^i)^2}{2\sigma^2} \right)$$

$$\max_{\theta} \log L(\theta) = - \frac{1}{2\sigma^2} \sum_{i=1}^m (y^i - \theta^\top x^i)^2 - m \log(\sqrt{2\pi}\sigma)$$

Least Mean Square

Reference

- Chapter 2 in [Pattern Recognition and Machine Learning](#). Springer. 2006

Q&A