UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Spring 2025)

Machine Learning

Lecture 4: Review of Probability and Statistics

Chong Liu

Department of Computer Science

Feb 3, 2025

Announcement

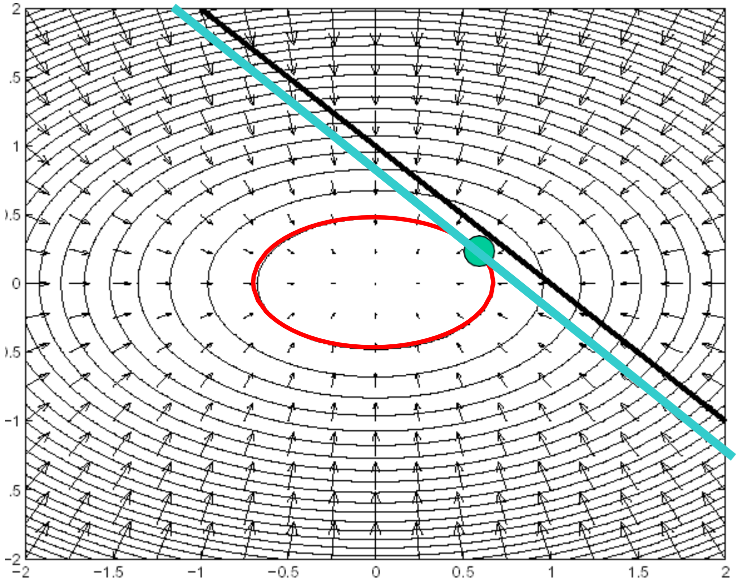
- Homework 1 has been released.
- Study group registration due today.
    - Sign up in waitlist if you don’t find a group.
- Project registration due this Wednesday.
    - [New Feature] You can also show your project interest in waitlist.

Recap: calculus and optimization review

- Multi-variate calculus
  - Partial derivative and gradient
  - Chain rule
  - Multiple integrals
  - Jacobian matrix and Hessian matrix
- Optimization
  - Convex set and convex function
  - Optimization problem formulation
  - Properties of convex optimization
  - Lagrange Multipliers

Equality constrained problem

- $\min f(x, y) = x^2 + 2y^2 - 2$
- s.t. $x + y = 1$



Equality constrained problem

- $\min f(x,y) = x^2 + 2y^2 - 2$
- s.t. $x + y = 1$

Introduce Lagrangian multiplier λ and form

- Solution:

$$L(x,y,\lambda) = x^2 + 2y^2 - 2 \;\; - \lambda(x + y - 1)$$

Then, differentiate with respect to x, y, λ : and set derivative to 0.

$$\left.\begin{array}{lll} \dfrac{\partial L}{\partial x} = 2x - \lambda = 0 & \Rightarrow & \lambda = 2x \\[2ex] \dfrac{\partial L}{\partial y} = 2y - \lambda = 0 & \Rightarrow & \lambda = 4y \\[2ex] \dfrac{\partial L}{\partial \lambda} = -x - y + 1 = 0 & \Rightarrow & -x - y + 1 = 0 \end{array}\right\} \quad \begin{array}{l} \lambda = \dfrac{4}{3} \\[2ex] x = \dfrac{2}{3} \\[2ex] y = \dfrac{1}{3} \end{array}$$

Today’s agenda

- Probability
    - Basic concepts
    - Probability properties
    - Random variable and distribution
    - Expectation and variance
    - Independence
    - Bernoulli distribution and Gaussian distribution
- Statistics
    - Maximum likelihood estimation

Basic concepts

- Experiment:
  - An action or process that leads to one or more possible outcomes.
- Outcome:
  - A single possible result of an experiment.
- Sample space:
  - The set of all possible outcomes of an experiment.
- Event:
  - A subset of the sample space. It is a collection of outcomes that share a common property.

Types of events

- Simple event:
  - An event that consists of exactly one outcome.
- Compound event:
  - An event that consists of more than one outcome.
- Mutually exclusive events:
  - Events that cannot occur simultaneously.
- Independent events:
  - Events where the occurrence of one event does not affect the occurrence of another.
- Complementary events:
  - If event A occurs, then the complement event A' does not occur, and vice versa.

Probability properties

- Non-negativity:
  - $P(A) \geq 0$ for any event A.
- Normalization:
  - $P(S) = 1$ for the whole sample space.
- Joint probability:
  - $P(A, B)$
- Marginal probability:
  - $P(A) = \sum_{b \in B} P(A, B)$
- Conditional probability:
  - $P(\, A \mid B \,) = P(A, B)/P(B)$

Probability of events

- For a finite sample space with equally likely outcomes,
    - $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes in sample space}}$
    - Example: a fair die with 6 outcomes

- Bayes' Theorem:
    - Find the probability of an event based on prior knowledge of conditions related to the event:
    - $P(A \mid B) = \frac{P(B|A) P(A)}{P(B)}$

In-class exercise: Bayes' theorem

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

- Suppose you have two coins:
  - Coin A is a fair coin (50% heads, 50% tails).
  - Coin B is biased, with a 70% chance of landing heads and 30% chance of landing tails.
- You randomly choose one of the two coins (with equal probability) and flip it. The result is heads. What is the probability that you chose the biased coin (Coin B)?

Random variable and distribution

- A random variable X is a numerical outcome of a random experiment

- The distribution of a random variable is the collection of possible outcomes along with their probabilities:
  - Discrete: $p(X = x) = p(x)$
  - Continuous: $p(a \leq X \leq b) = \int_a^b p(x)\,\mathrm{d}x$

Expectation

- Discrete case:
  - For a random variable $X \sim p(X = x)$, its expectation is
  - $E[X] = \sum_x xp(X = x)$
    - In an empirical sample, $x_1, x_2, \ldots x_N, E[X] = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Continuous case:
  - $E[X] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$

Properties of expectation

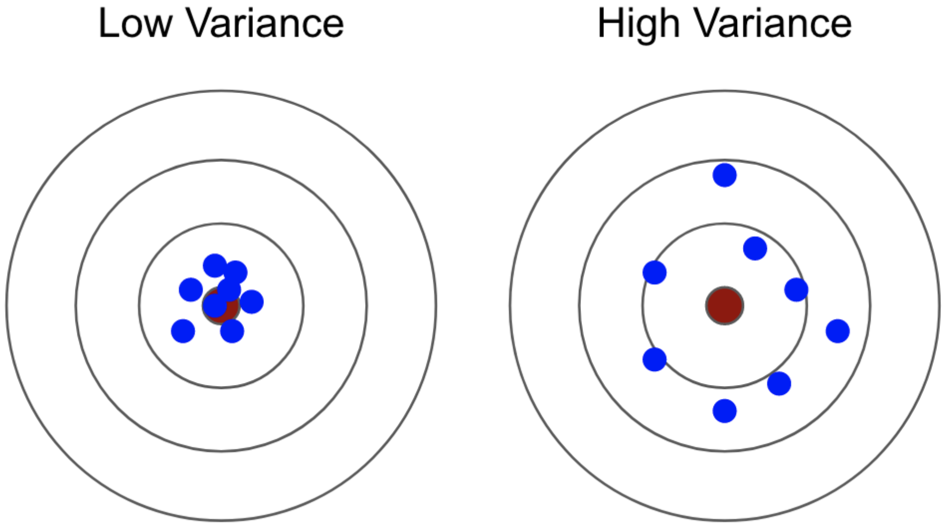
- Non-negativity:
  - If $X \geq 0$, then $E[X] \geq 0$.
- Linearity:
  - $E[X + Y] = E[X] + E[Y]$
  - $E[aX] = aE[X]$

- Discussion: expectation of $f(x)$, a function of random variable x?
  - $E[x] = \int f(X)p(x)dx$
  - $E[x] = \sum f(x)p(x)$

Variance

- Variance of a random variable X is the expected value of the squared deviation from the mean:
  - $\mathrm{Var}[X] = E[(X - E[X])^2]$

  - Mean $E[X]$
  - Deviation $X - E[X]$
  - Squared deviation $(X - E[X])^2$



https://towardsdatascience.com/what-bias-variance-bulls-eye-diagram-really-represent-ff6fb9670993

In-class exercise

- Use Markov's inequality to prove Chebyshev's inequality.
    - Markov's inequality:
        - For a *nonnegative* random variable X and any positive number a,
        - $P(X \geq a) \leq \frac{E[X]}{a}$
    - Chebyshev's inequality:
        - For a *nonnegative* random variable X and any positive number a,
        - $P(|X - E[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$

(Statistical) Independence

- Not the same as linear independence in linear algebra!

- X and Y are independent, i.e.,

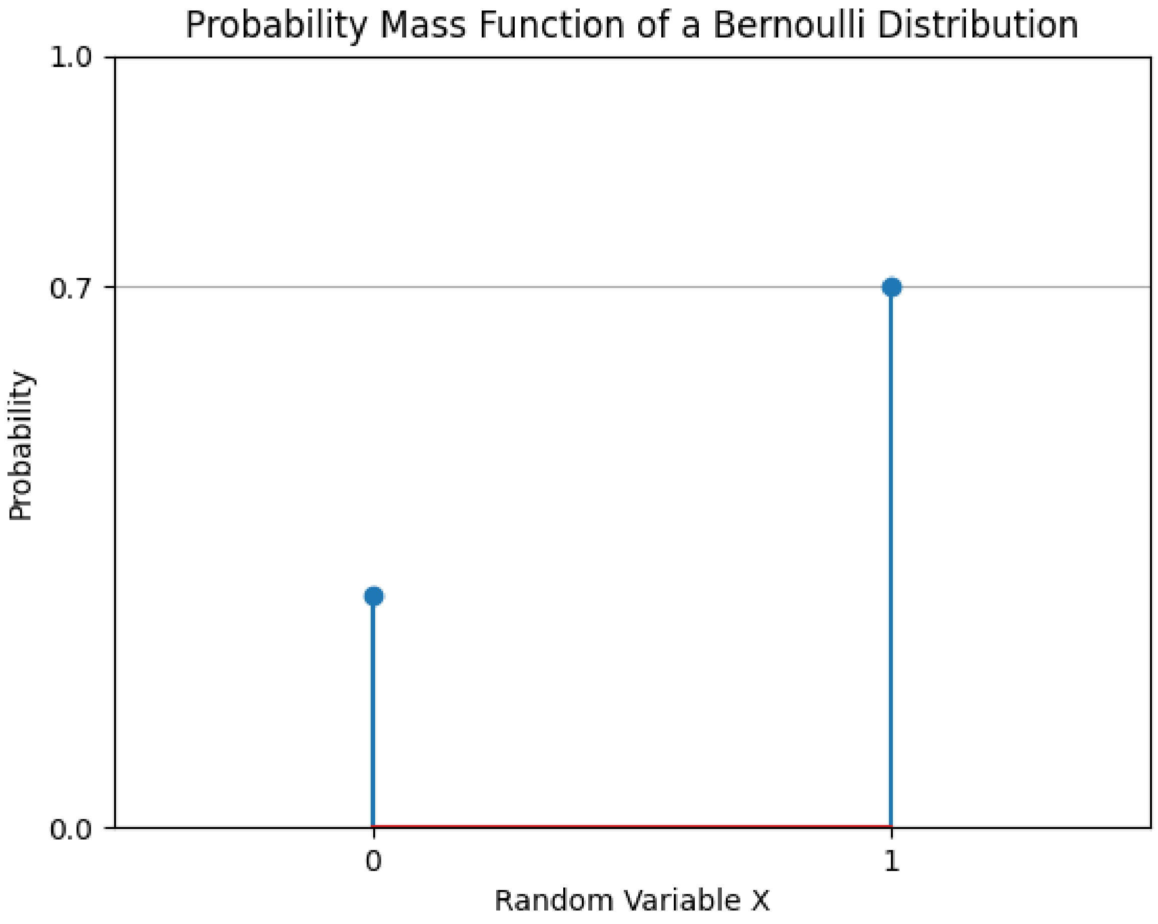
$$X \perp Y \text{ iff } P(X,Y) = P(X)P(Y) \text{ iff } P(X) = P(X|Y)$$

- X and Y are independent implies

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

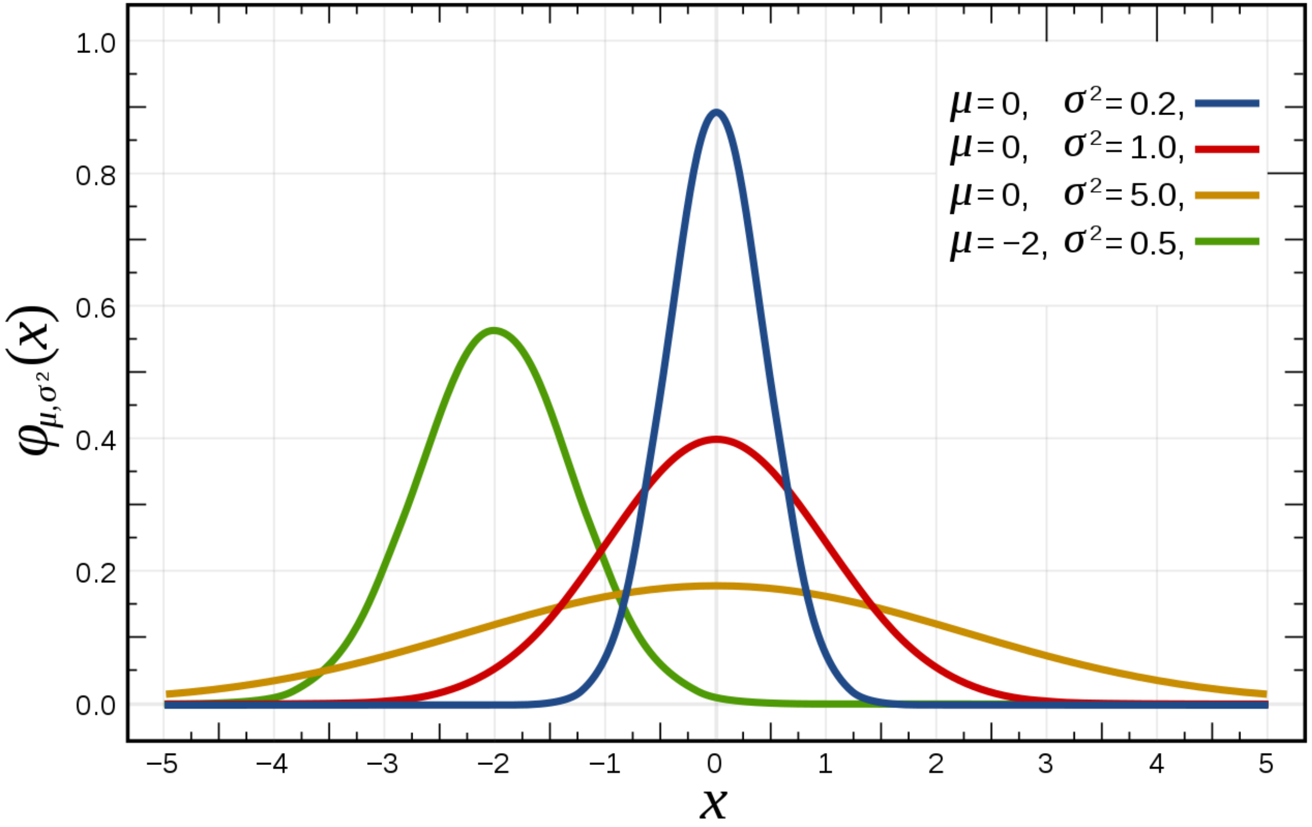
$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

Bernoulli distribution $X \sim \mathrm{Ber}(p)$



$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$



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

Statistics in one slide

- What is the difference between probability and statistics?
  - Statistics is the "science of data" --- it uses probability theory, but also other branches of mathematics and computational tools for making sense of data
- Typical problem: Statistical Estimation
  - Data: $X_1, X_2, \ldots, X_n \sim P$
  - Goal: estimate a statistical quantity θ of the distribution P
  - **Estimator** (really an algorithm): $\hat{\theta}$ that takes input data and output a guess of the true quantity θ
- Examples
  - Estimate the mean, variance, medians (and other quantiles).
  - Estimate the expected error of a given ML classifier using a holdout dataset.
  - Estimate the parameter θ of P if P is parameterized by θ, denoted by P_θ.

Examples of statistical estimation problem

- Example 1 (Biased coin): Toss a coin 100 times, observe the outcome “Head” or “Tail”. What is the probability of seeing “Head”?

- Example 2 (Average monthly precipitation in Albany, NY):
    - Observe data for Year 1960, 1961,…, 2024.
    - Each data point is a vector of 12 numbers measuring the number of inches of precipitation.
    - How to estimate the average?

Maximum likelihood estimation

- Used since Gauss, Laplace, …. Carefully analyzed by Ronald Fisher.
- Key idea:
  - Which distribution is more *likely* to have produced the data?
  - $\max_P f_{\text{Data}\sim P}(\text{Data})$

  - Example: $X_1, X_2, \ldots, X_n \sim D_\theta$
    - $\max P(X_1, X_2, \ldots, X_N|\theta)$

- Observation 1: If the data is i.i.d. then by independence the density factorizes
  - $P(X_1, X_2, \ldots, X_N|\theta) = P(X_1|\theta)P(X_1|\theta)\ldots P(X_1|\theta)$
- Observation 2: Taking log does not change the solution.
  - $\max P(X_1, X_2, \ldots, X_N|\theta) \leftrightarrow \max \log P(X_1, X_2, \ldots, X_N|\theta)$