CX4240 Spring 2026
Probability and Statistics Revisit

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

Office Hours

Friday: 3:00-4:00pm, *Online Session*

Office Hours

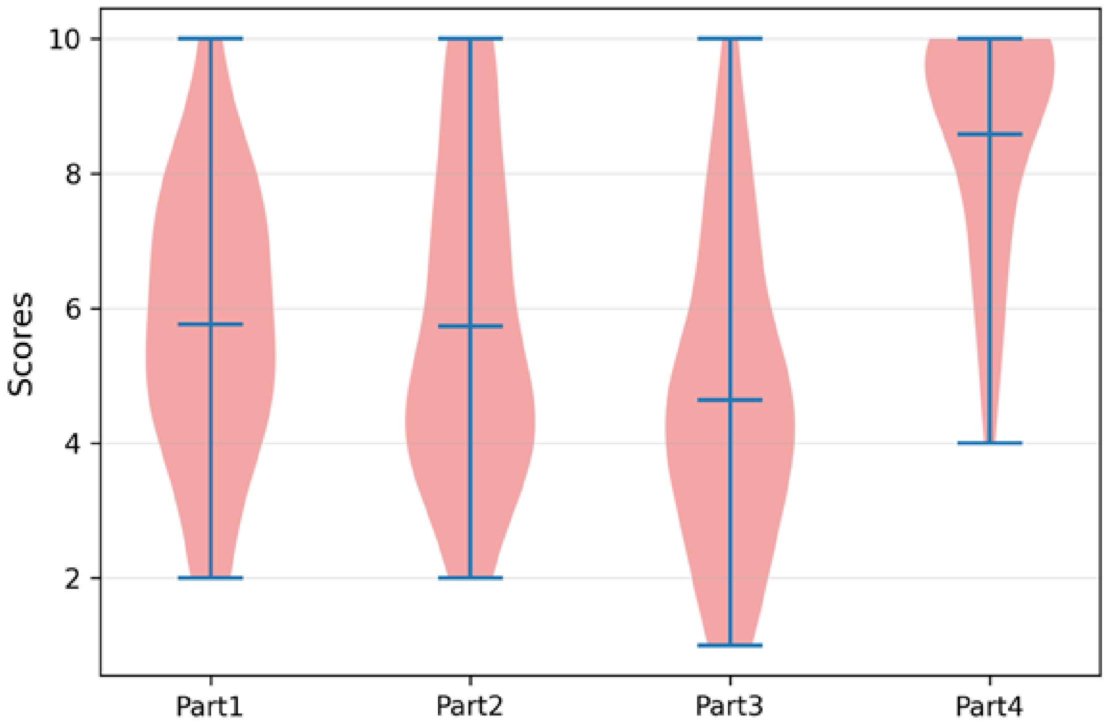
Monday: 2:00-3:00pm, *Coda 2nd floor*: Changhao Li

Thursday: 3:00-4:00pm, *Coda 2nd floor*: Chenxiao Gao

Basic / Prerequisites

- Probability
  - Distributions, densities, marginalization, conditioning
- Statistics
  - Mean, variance, maximum likelihood estimation
- Linear Algebra and Optimization
  - Vector, matrix, multiplication, inversion, eigen-value decomposition
- Coding Skills
  - Pytorch and/or JAX

Statistics of Background Test



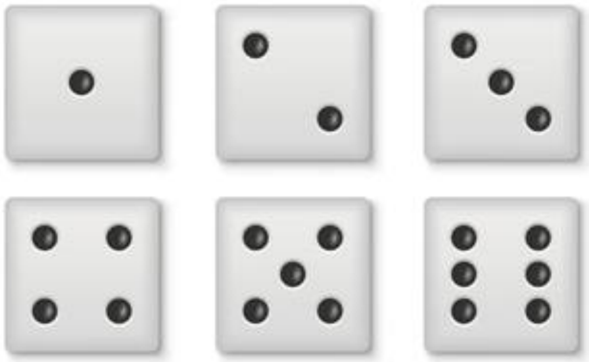
Probability and Statistics Revist

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.)

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 $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$



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 $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$
  - E.g., S may be the set of all possible nucleotides of a DNA site: S $(\mathrm{A} \quad \mathrm{C} \quad \mathrm{G} \quad \mathrm{T})$

- An Event A is any subset of S
  - Seeing "1" or "6" in a dice roll; observing a "G" at a site

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 $\sum_i p_X(x_i) = 1$

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:

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

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:

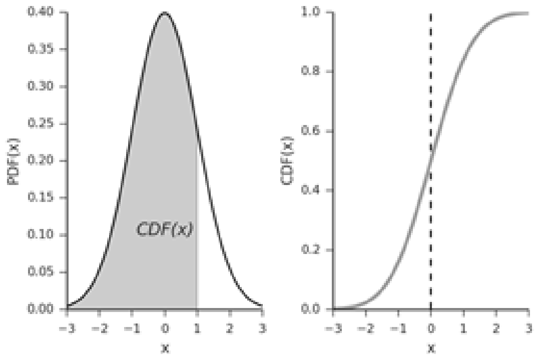
$$\begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

  - 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.

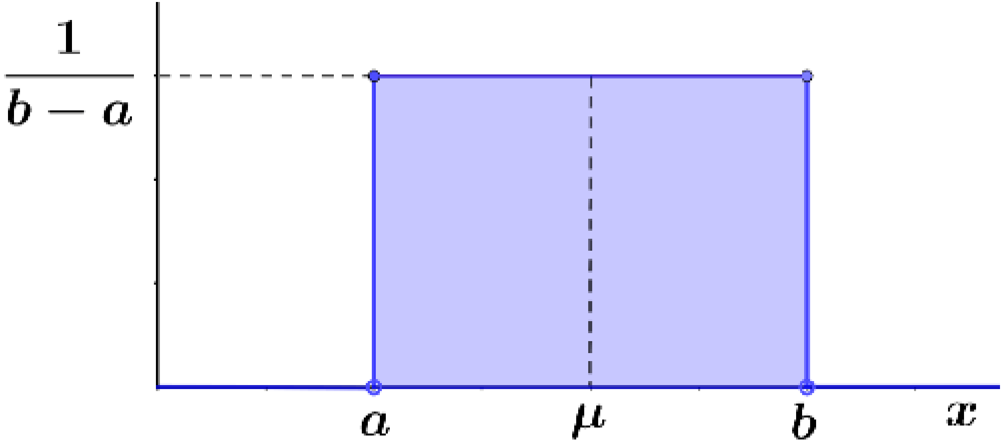


  - Cumulative Distribution Function (CDF): $F_x(x) = P[X \leq x]$
  - Probability Density Function (PDF): $F_x(x) = \int_{-\infty}^{x} f_x(x)dx$ or $f_x(x) = \frac{dF_x(x)}{dx}$
  - Properties: $f_x(x) \geq 0$ and $\int_{-\infty}^{\infty} f_x(x)dx = 1$
  - Interpretation: $f_x(x) = P\left[X \in \frac{x,x+\Delta}{\Delta}\right]$

Continuous Probability Distribution

- Examples:
  - Uniform Density Function:

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

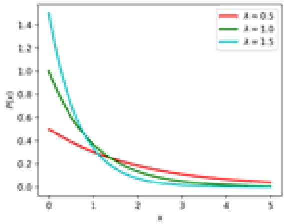


Continuous Probability Distribution

- Examples:
  - Uniform Density Function:

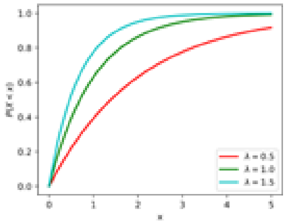
$$f_x(x) = \begin{cases} \dfrac{1}{b-a} & for\, a \leq x \leq b \\ 0 & otherwise \end{cases}$$

  - Exponential Density Function:



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

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



Continuous Probability Distribution

- Examples:
  - Uniform Density Function:

$$f_x(x) = \begin{cases} \dfrac{1}{b-a} & for\, a \le x \le b \\ 0 & otherwise \end{cases}$$

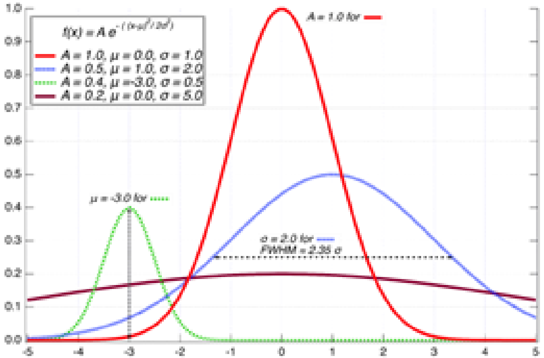
  - Exponential Density Function:

$$f_x(x) = \lambda e^{-\lambda x} \qquad for\, x \ge 0$$

$$F_x(x) = 1 - e^{-\lambda x} \qquad for\, x \ge 0$$

  - Gaussian(Normal) Density Function

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

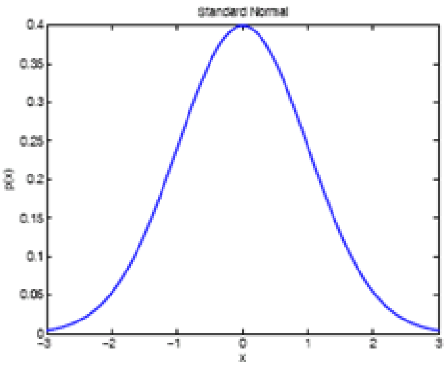


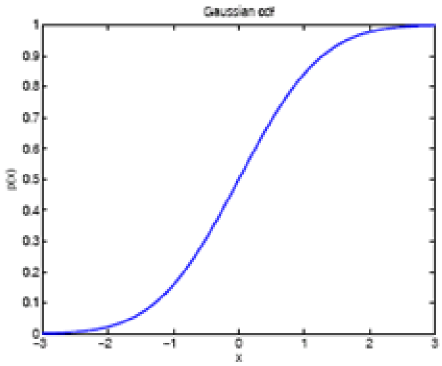
Continuous Probability Distribution

- 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:

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

Statistics

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

$$E_X[g(X)] = \int_{-\infty}^{\infty} g(x) p_X(x) dx \quad \approx \quad \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$

Statistics

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

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

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n \quad \mu = E_X[x]$

Statistics

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

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

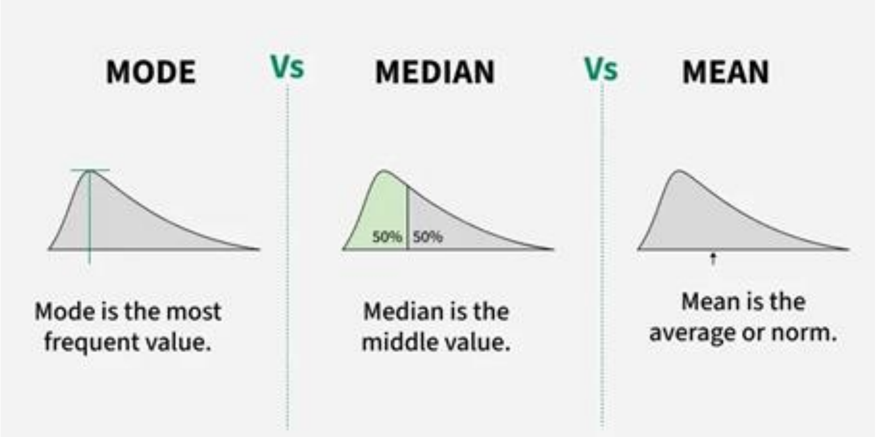
- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n \quad \mu = E_X[x]$
- Mean: $E_X[X] = \int_{-\infty}^{\infty} xp_X(x)dx$
  - $E[\alpha X] = \alpha E[X]$
  - $E[\alpha + X] = \alpha + E[X]$
- Variance(Second central moment): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
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

Statistics

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

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

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n$
- 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 $Var[X_i]$

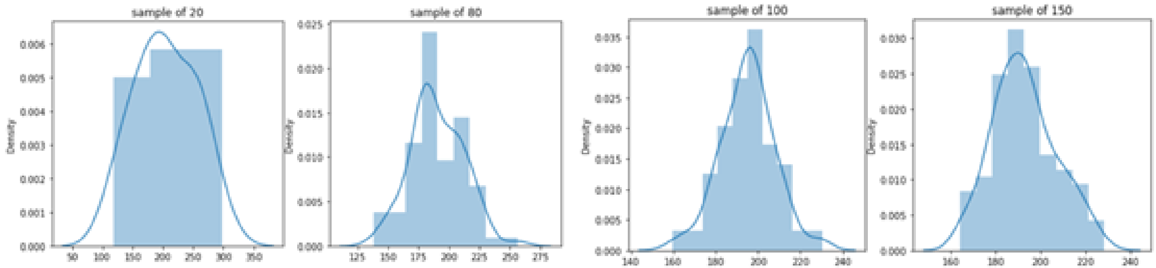
$$\bar{X} = f(X_1, X_2, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{as } n \rightarrow \infty$$

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 $Var[X_i]$

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

- 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:
  - $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,\beta) d\beta$
  - $p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j)$

Joint RVs and Marginal Densities

- Joint cumulative distribution:

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

- 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

- $P(X \mid Y)$ = 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"
  - $P(Headche \mid Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?
- Definition:

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

Conditional Probability

- $P(X \mid Y)$ = 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"
  - $P(Headche \mid Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?
- Definition:

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

This is called Bayes Rule

Conditional Probability

- $P(X \mid Y)$ = 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"
  - $P(Headche \mid Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?
- Definition:

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

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

Rules of Independence

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

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

- E and H are (statistically) independent if

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

- Or equivalently

$$P(E) = P(E \mid H)$$

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

Rules of Independence

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

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

- E and H are (statistically) independent if

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

- Or equivalently

$$P(E) = P(E \mid H)$$

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

$$P(E \mid H,F) = P(E \mid H)$$

- Or equivalently

$$P(E,F \mid H) = P(E \mid H)P(F \mid H)$$

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) $E(Y|x)$.

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

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) $E(Y|x)$.

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

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

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

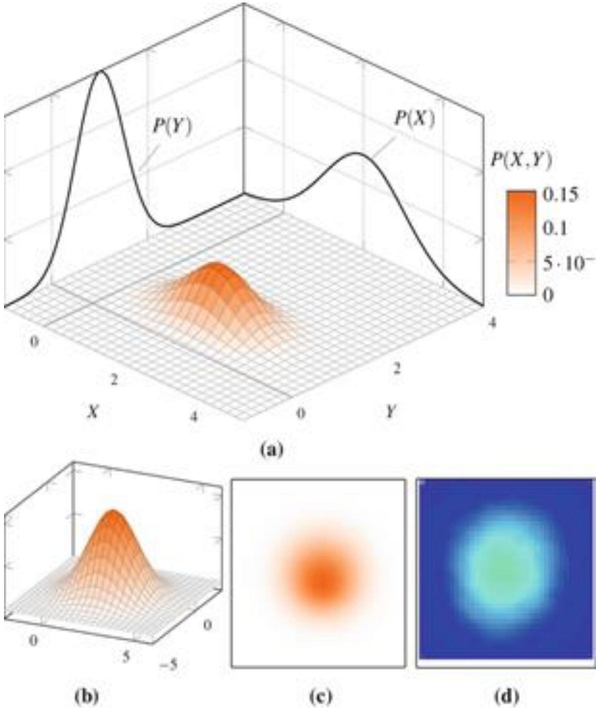
Multivariate Gaussian

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

- 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

- Joint Gaussian $P(X_1, X_2)$

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

- Marginal Gaussian

$$\mu_2^m = \mu_2 \; \Sigma_2^m = \Sigma_2$$

- Conditional Gaussian $P(X_1 \mid X_2 = x_2)$

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

Operations on Gaussian R.V.

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

$$E(AX+b) = AE(X) + b$$
$$Cov(AX+b) = ACov(X)A^\top$$

this means that for Gaussian distributed quantities:

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

- 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)

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

Maximum Log-Likelihood Estimation (MLE)

Given iid samples from Gaussian $\{x_i\}_{i=1}^{n}$

$$p(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Likelihood $$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Maximum Log-Likelihood Estimation (MLE)

Given iid samples from Gaussian $\{x_i\}_{i=1}^n$

$$p(x_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

MLE $$\max_{\mu,\sigma} \ell(\mu,\sigma^2) = \log L(\mu,\sigma^2)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

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:**

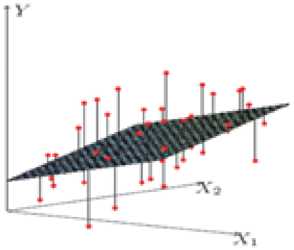
$$\text{monthly rent} = \theta_1(\text{living area}) + \theta_2(\#\ \text{bedroom})$$

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

Gaussian Likelihood

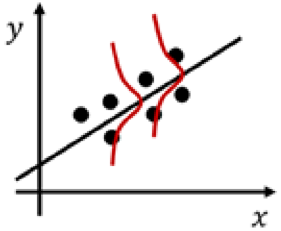
- Assume y is a linear in x plus noise ϵ

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



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

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



- By independence assumption, likelihood is

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

Probability

MLE

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

MLE

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

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

Least Mean Square

Reference

- Chapter 2 in Pattern Recognition and Machine Learning. Springer. 2006

Q&A