



Xi'an Jiaotong-Liverpool University

西交利物浦大学

**INT305**

# **Probability Review for Machine Learning**

**Jimin Xiao**

**Department Intelligence Science**

**[Jimin.xiao@xjtlu.edu.cn](mailto:Jimin.xiao@xjtlu.edu.cn)**

# Motivation

---

Uncertainty arises through:

- Noisy measurements
- Variability between samples
- Finite size of data sets

Probability provides a consistent framework for the quantification and manipulation of uncertainty.

# Sample Space

---

**Sample space**  $\Omega$  is the set of all possible outcomes of an experiment.

**Observations**  $\omega \in \Omega$  are points in the space also called sample outcomes, realizations, or elements.

**Events**  $E \subset \Omega$  are subsets of the sample space.

In this experiment we flip a coin twice:

**Sample space** All outcomes  $\Omega = \{HH, HT, TH, TT\}$

**Observation**  $\omega = HT$  valid sample since  $\omega \in \Omega$

**Event** Both flips same  $E = \{HH, TT\}$  valid event since  $E \subset \Omega$

# Sample Space

---

The probability of an event  $E$ ,  $P(E)$ , satisfies three axioms:

1:  $P(E) \geq 0$  for every  $E$

2:  $P(\Omega) = 1$

3: If  $E_1, E_2, \dots$  are disjoint then

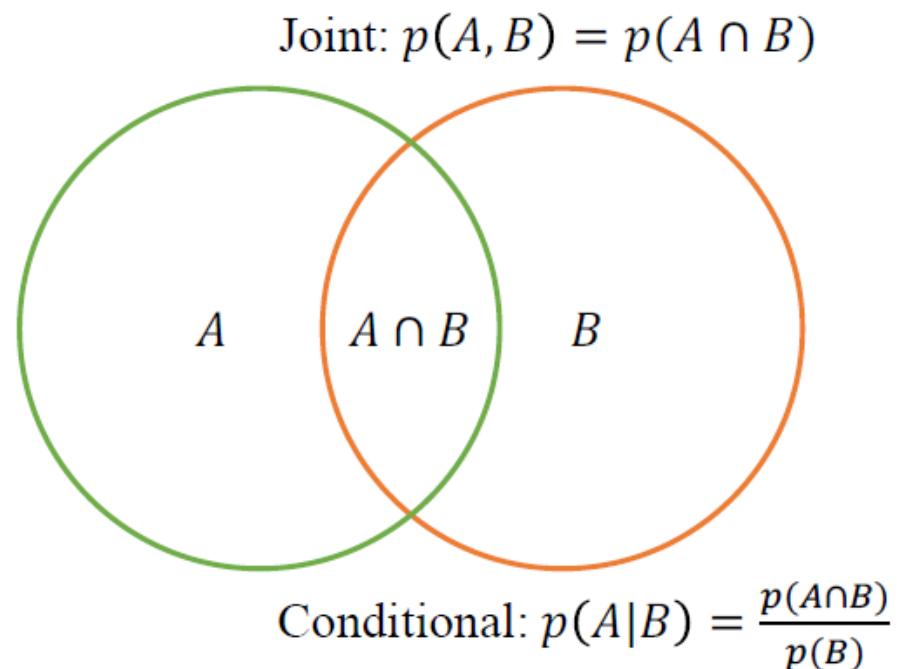
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

# Joint and Conditional Probabilities

---

**Joint Probability** of  $A$  and  $B$  is denoted  $P(A, B)$ .

**Conditional Probability** of  $A$  given  $B$  is denoted  $P(A|B)$ .



$$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$$

## Conditional Example

---

Probability of passing the midterm is 60% and probability of passing both the final and the midterm is 45%.

What is the probability of passing the final given the student passed the midterm?

$$\begin{aligned}P(F|M) &= P(M, F)/P(M) \\&= 0.45/0.60 \\&= 0.75\end{aligned}$$

# Independence

---

Events  $A$  and  $B$  are **independent** if  $P(A, B) = P(A)P(B)$ .

- Independent:  $A$ : first toss is HEAD;  $B$ : second toss is HEAD;

$$P(A, B) = 0.5 * 0.5 = P(A)P(B)$$

- Not Independent:  $A$ : first toss is HEAD;  $B$ : first toss is HEAD;

$$P(A, B) = 0.5 \neq P(A)P(B)$$

# Independence

---

Events  $A$  and  $B$  are **conditionally independent** given  $C$  if

$$P(A, B|C) = P(B|C)P(A|C)$$

Consider two coins <sup>2</sup>: A regular coin and a coin which always outputs HEAD or always outputs TAIL.

$A$ =The first toss is HEAD;  $B$ =The second toss is HEAD;  $C$ =The regular coin is used.  $D$ =The other coin is used.

Then  $A$  and  $B$  are conditionally independent given  $C$ , but  $A$  and  $B$  are NOT conditionally independent given  $D$ .

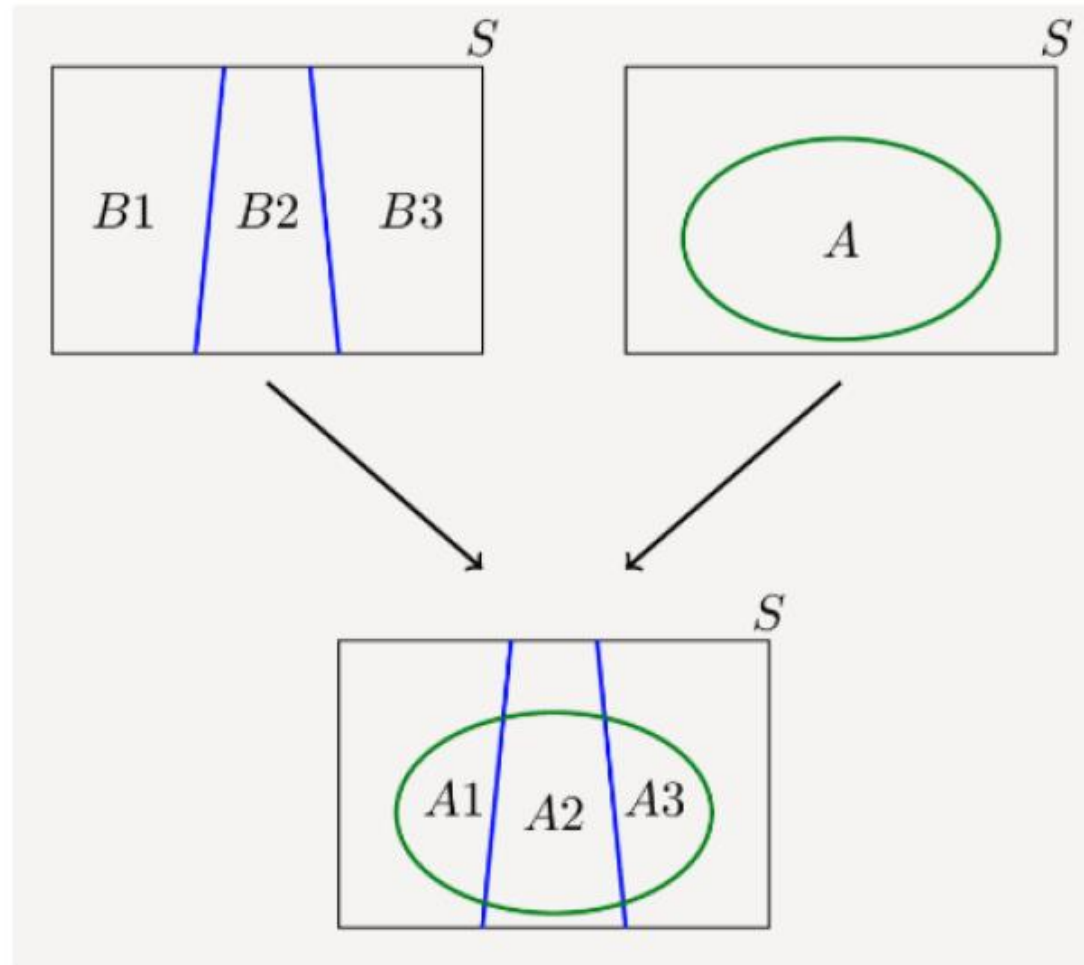


# Marginalization and Law of Total Probability

---

Law of Total Probability <sup>3</sup>

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



# Bayes' Rule

---

Bayes' Rule:

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

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$\textit{Posterior} \propto \textit{Likelihood} \times \textit{Prior}$$

# Bayes' Example

---

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

This depends on the prior probability of the disease:

- $P(T = 1|D = 1) = 0.95$  (likelihood)
- $P(T = 1|D = 0) = 0.10$  (likelihood)
- $P(D = 1) = 0.1$  (prior)

So  $P(D = 1|T = 1) = ?$

# Bayes' Example

---

Suppose you have tested positive for a disease. What is the probability you actually have the disease?

$$P(T = 1|D = 1) = 0.95 \text{ (true positive)}$$

$$P(T = 1|D = 0) = 0.10 \text{ (false positive)}$$

$$P(D = 1) = 0.1 \text{ (prior)}$$

So  $P(D = 1|T = 1) = ?$

Use Bayes' Rule:

$$P(D = 1|T = 1) = \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1)} = \frac{0.95 * 0.1}{P(T = 1)} = 0.51$$

$$\begin{aligned} P(T = 1) &= P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0) \\ &= 0.95 * 0.1 + 0.1 * 0.90 = 0.185 \end{aligned}$$

# Random Variable

---

How do we connect sample spaces and events to data?

A **random variable** is a mapping which assigns a real number  $X(\omega)$  to each observed outcome  $\omega \in \Omega$

For example, let's flip a coin 10 times.  $X(\omega)$  counts the number of Heads we observe in our sequence. If  $\omega = HHTHTHHTHT$  then  $X(\omega) = 6$ .

# Discrete and Continuous Random Variables

---

## Discrete Random Variables

- Takes countably many values, e.g., number of heads
- Distribution defined by probability mass function (PMF)
- Marginalization:  $p(x) = \sum_y p(x, y)$

## Continuous Random Variables

- Takes uncountably many values, e.g., time to complete task
- Distribution defined by probability density function (PDF)
- Marginalization:  $p(x) = \int_y p(x, y) dy$

## I.I.D.

---

Random variables are said to be **independent and identically distributed** (i.i.d.) if they are sampled from the same probability distribution and are mutually independent.

This is a common assumption for observations. For example, coin flips are assumed to be iid.

# Probability Distribution Statistics

---

**Mean:** First Moment,  $\mu$

$$E[x] = \sum_{i=1}^{\infty} x_i p(x_i) \quad (\text{univariate discrete r.v.})$$

$$E[x] = \int_{-\infty}^{\infty} x p(x) dx \quad (\text{univariate continuous r.v.})$$

**Variance:** Second (central) Moment,  $\sigma^2$

$$\begin{aligned} Var[x] &= \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx \\ &= E[(x - \mu)^2] \\ &= E[x^2] - E[x]^2 \end{aligned}$$

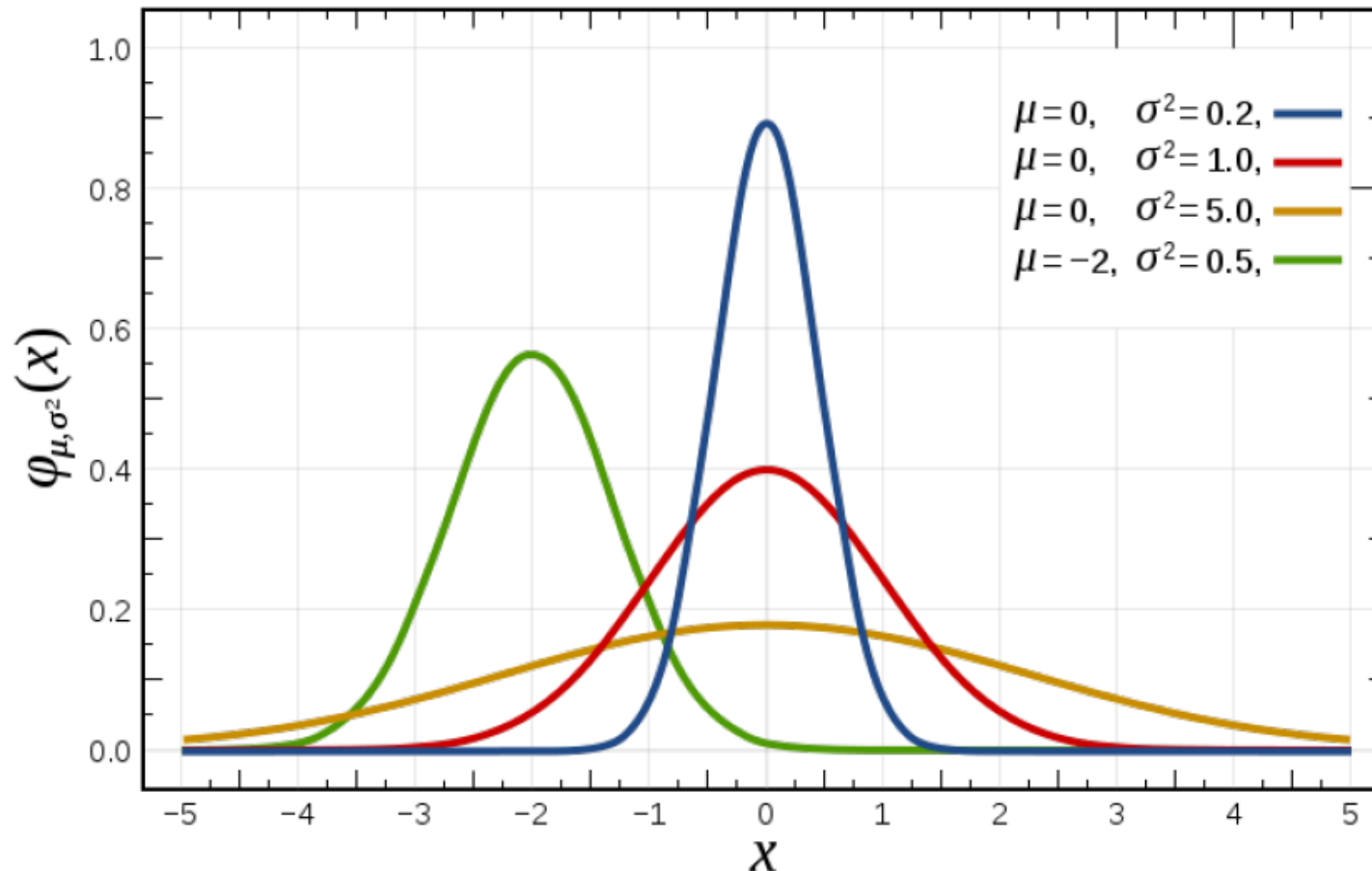


# Univariate Gaussian Distribution

---

Also known as the **Normal Distribution**,  $\mathcal{N}(\mu, \sigma^2)$

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



# Multivariate Gaussian Distribution

---

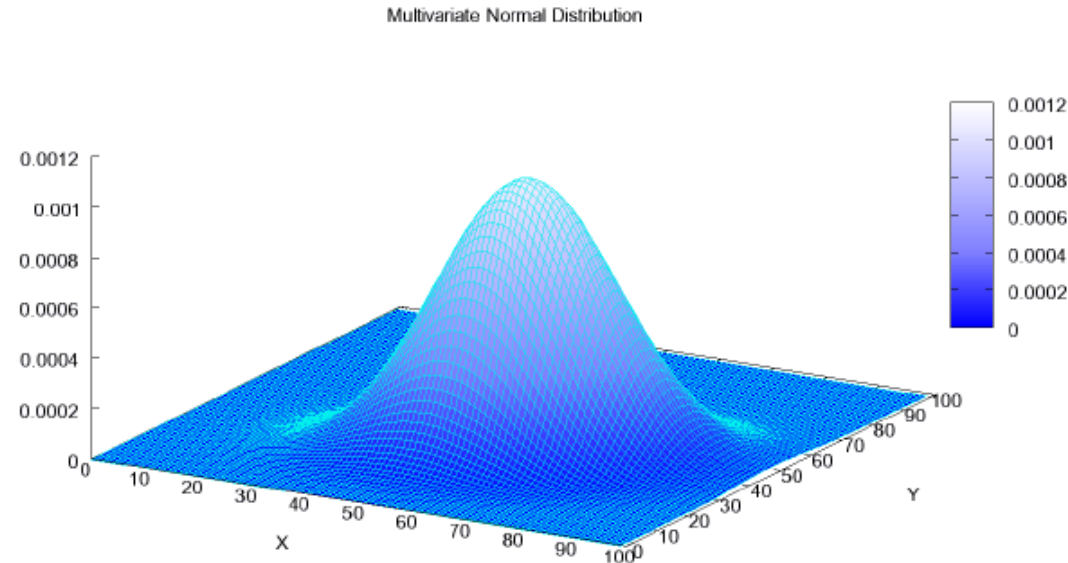
Multidimensional generalization of the Gaussian.

$\mathbf{x}$  is a  $D$ -dimensional vector

$\mu$  is a  $D$ -dimensional mean vector

$\Sigma$  is a  $D \times D$  covariance matrix with determinant  $|\Sigma|$

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right\}$$



# Covariance Matrix

---

Recall that  $\mathbf{x}$  and  $\mu$  are D-dimensional vectors

**Covariance matrix**  $\Sigma$  is a matrix whose  $(i, j)$  entry is the covariance

$$\begin{aligned}\Sigma_{ij} &= Cov(\mathbf{x}_i, \mathbf{x}_j) \\ &= E[(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)] \\ &= E[(\mathbf{x}_i \mathbf{x}_j)] - \mu_i \mu_j\end{aligned}$$

so notice that the diagonal entries are the variance of each elements.

The covariant matrix has the property that it is symmetric and positive-semidefinite (this is useful for whitening).

# Inferring Parameters

---

We have data  $X$  and we assume it is sampled from some distribution.  
How do we figure out the parameters that ‘best’ fit that distribution?  
Maximum Likelihood Estimation (MLE)

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} P(X|\theta)$$

Maximum A posteriori Probability (MAP)

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} P(\theta|X)$$

# MLE for Univariate Gaussian Distribution

---

We are trying to infer the parameters for a Univariate Gaussian Distribution, mean ( $\mu$ ) and variance ( $\sigma^2$ ).

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

The **likelihood** that our observations  $x_1, \dots, x_N$  were generated by a univariate Gaussian with parameters  $\mu$  and  $\sigma^2$  is

$$\text{Likelihood} = p(x_1 \dots x_N | \mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

# MLE for Univariate Gaussian Distribution

---

For MLE we want to maximize this likelihood, which is difficult because it is represented by a product of terms

$$\text{Likelihood} = p(x_1 \dots x_N | \mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}$$

So we take the log of the likelihood so the product becomes a sum

$$\begin{aligned} \text{Log Likelihood} &= \log p(x_1 \dots x_N | \mu, \sigma^2) \\ &= \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \end{aligned}$$

Since log is monotonically increasing  $\max L(\theta) = \max \log L(\theta)$

# MLE for Univariate Gaussian Distribution

---

The log Likelihood simplifies to

$$\begin{aligned}\mathcal{L}(\mu, \sigma) &= \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} \\ &= -\frac{1}{2}N \log(2\pi\sigma^2) - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}\end{aligned}$$

Which we want to maximize. How?

# MLE for Univariate Gaussian Distribution

---

To maximize we take the derivatives, set equal to 0, and solve:

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2}N \log(2\pi\sigma^2) - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

Derivative w.r.t.  $\mu$ , set equal to 0, and solve for  $\hat{\mu}$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \mu} = 0 \implies \hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

Therefore the  $\hat{\mu}$  that maximizes the likelihood is the average of the data points.

Derivative w.r.t.  $\sigma^2$ , set equal to 0, and solve for  $\hat{\sigma}^2$

$$\frac{\partial \mathcal{L}(\mu, \sigma)}{\partial \sigma^2} = 0 \implies \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu})^2$$