Machine Learning Principles

Class2 : Sept. 8

Probability for ML

Instructor: Diana Kim

Today's Lecture

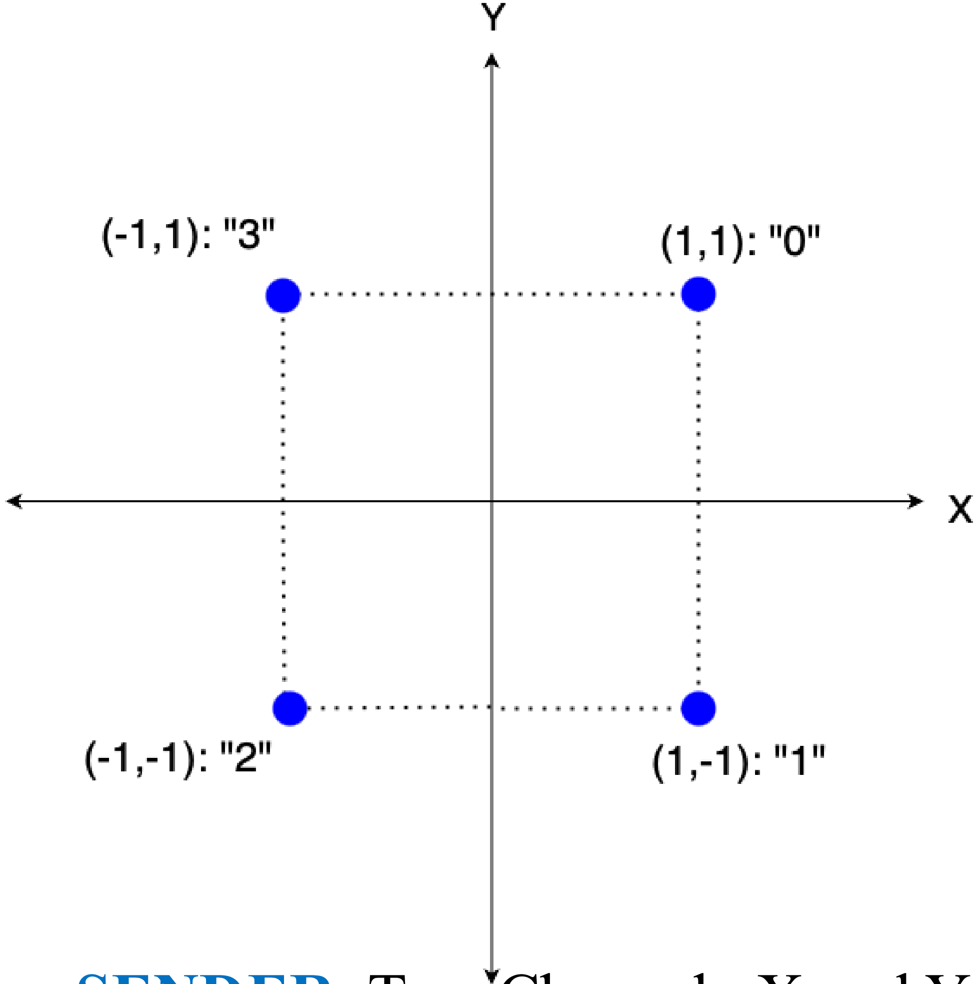
0. Why Probability?
1. Two Interpretation of Probability: frequentist vs Bayesian
2. Probability Space / Axioms
3. Random Variables/ Random Vectors
4. Computing Probability: Joint & Conditional Prob/ Marginalization
5. Bayes Rules
6. Important Statistics : mean & variance (random scalar & vectors)
7. Gaussian Density (defined by mean and variance)
8. Maximum Likelihood Estimation (MLE)
9. Sample mean and Sample Variance

[1] Why Probability?
to measure uncertainty

DALL-E Image

We could predict dice # if we know every factor that
affects the outcome: knowing initial grip, angle to throw, air condition, etc.
however, in practice, we have partial information.

[2] Why Probability?
In the real world, we have limited knowledge



SENDER: Two Channels-X and Y

RECEIVER: One Channel-X

- If we receive (1): "0", "1"
- If we receive (-1): "2", "3"

[3] Why Probability?

Probability Theory is to provide mathematical machinery to measure uncertainty associated events.

[4] Why Probability in ML?

1. Probabilistic Modeling: (target to learn joint/ conditional density)
2. Modeling Errors : even for non-probabilistic/deterministic modeling, we need to consider random error (ε).

- causes of random error ($y = f_w(x) + \varepsilon$: as we predict "y" based on "x")
  (1) limited knowledge/ accessibility to the features
  (2) limited hypothesis space/ limited by # data points
  (3) measurement errors

- Two Different Interpretation of Probability

  frequentist vs. Bayesian

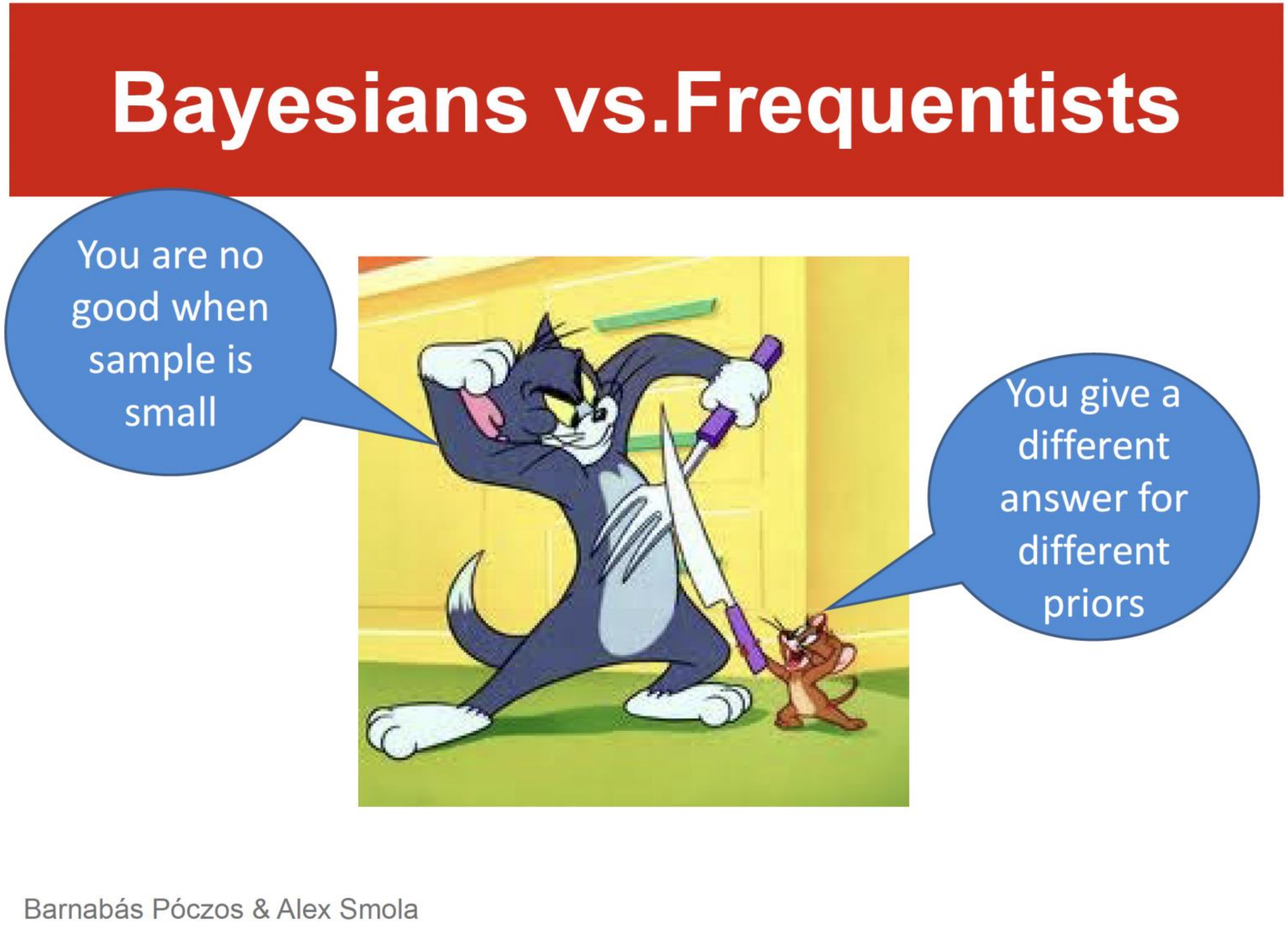
[1] Two Different Interpretation of Probability

- Bayesian Probability: P[event | evidence] $\propto$ P[evidence | event]P[event]
  measuring uncertainty or belief
  Q: The chances of detecting life on Mars?
  Q: The chances of the Artic ice cap will have disappeared
  by the end of century?

- Frequentist Probability:
  measuring relative frequency

$$P[\text{event A}] = \frac{\#event\ A}{\#\ trials}$$

[2] Two Different Interpretation of Probability



[3] Two Different Interpretation of Probability

Suppose someone claims $P[\{\text{coin head}\}] = 1$ based on two trials.
Can you accept the probability?

- Probability 101

Probability Space [$\boldsymbol{\Omega}$, $\mathbf{2}^{|\boldsymbol{\Omega}|}$, P]

[1] Experiment: any process of obtaining or generating an observation

Ex] Inspection of an instance item is defective or non-defective

[2] **Sample Space ($\boldsymbol{\Omega}$): a set** of all possible outcomes

Ex] $\boldsymbol{\Omega}$ = {non-defective, defective}

[3] **Events Set: ($A \subset \boldsymbol{\Omega}$ or $A \in \mathbf{2}^{|\boldsymbol{\Omega}|}$): a set of** all possible subsets of Ω

Ex] $2^{|\Omega|} = \{\emptyset$, {non-defective}, {defective}, $\boldsymbol{\Omega}\}$

[4] **Probability Measure P[E]: a function** P: $\mathbf{2}^{|\boldsymbol{\Omega}|} \rightarrow [0, 1]$

Ex] P[{defective}] = monitor assembly line for a period of time, compute the relative frequency.

Probability Axioms

Probability Measure follows the three axioms.

- **Non-negativity**: $P[A] \geq 0$
- **Total Proablity**: $P[\Omega] = 1$
- **Countable Additive**: $A_i \cap A_j = \phi \; if \; i \neq j \implies \mathrm{P}[\cup_k A_k] = \sum_k P[A_k]$

Probability Axioms and Corollaries

- Non-negativity: $P[A] \geq 0$

- Total Proablity: $P[\Omega] = 1$

- Countable Additive: $A_i \cap A_j = \phi \; if \; i \neq j \;\; \Longrightarrow \mathrm{P}[\cup_k A_k] = \sum_k P[A_k]$

- $P[A^c] = 1 - P[A]$

  by countable additivity and total probability, $P[A^c \cup A] = \mathrm{P}[A] + \mathrm{P}[A^c] = 1$
- $P[\phi] = 1 - P[\Omega] = 0$

- Random Variables:

  to handle numerical outcomes / events

  (data samples & internal/output representations of ML systems)

[1] Random Variables

A random Variable **X** is a function
that <u>assigns a real number</u>
to each of outcome $\boldsymbol{\omega}$ in the sample space $\boldsymbol{\Omega}$ of a random experiment.

[2] Random Variables (Bernoulli R.V)

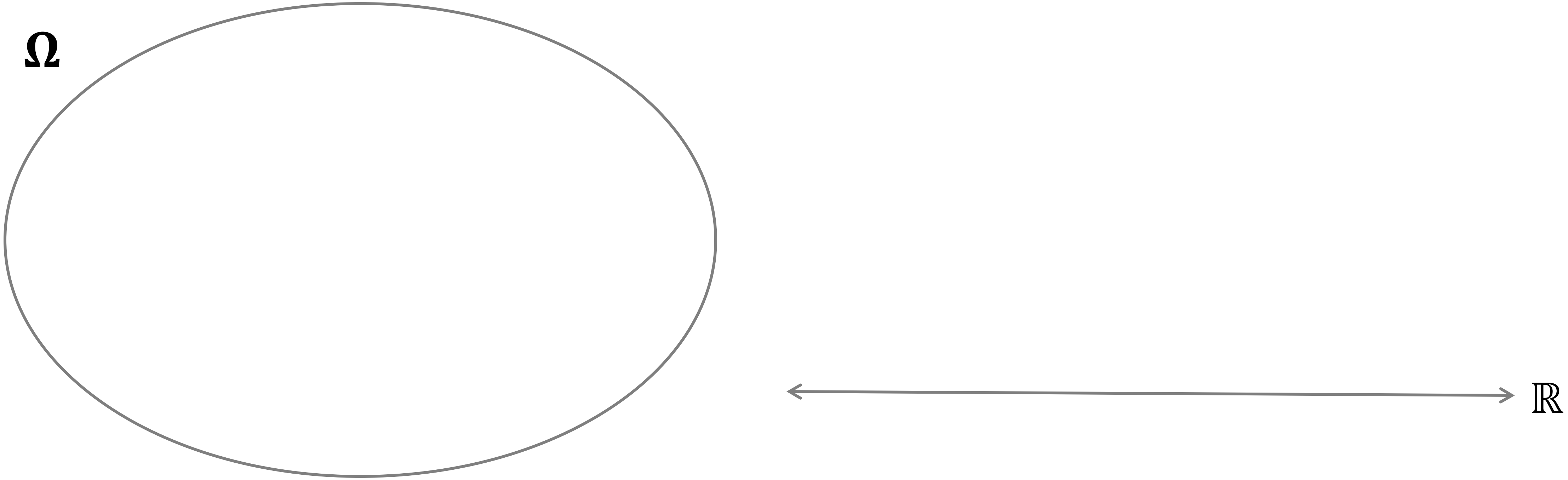
ex] Suppose a coin tossed one time.
Let R.V X be the indicator function for **tail event.**

[3] Random Variables (Binomial R.V)

ex] Let Y be the number of heads in the three-coin tosses.

Define an event $\{Y = 3\}$ *or* $\{Y \geq 2\}$

what is the equivalent event of $\{Y \geq 2\}$?



[4] Random Variables (The **C**umulative **D**istribution **F**unction: CDF)

The cumulative distribution function (CDF) of a R.V **X** is defined as the probability of the event $\{\mathbf{X} \leq \boldsymbol{x}\}$

$$F_X(x) = P[X \leq x] \qquad \text{for } -\infty \leq \boldsymbol{x} \leq +\infty$$

[5] Random Variables (The **C**umulative **D**istribution **F**unction: CDF)

ex] Suppose a coin tossed one time.
Let R.V X be the indicator function for **tail event.**
CDF of X: $F_X(x)$:

- $F_x(-\infty)$: $\mathrm{P}[X \le -\infty]$
- $F_x(0)$: $\mathrm{P}[X \le 0]$
- $F_x(1)$: $\mathrm{P}[X \le 1] = \mathrm{P}[X = 0] + \mathrm{P}[X = 1]$
- $F_x(2)$: ?
- $F_x(3)$: ?

[6] Random Variables: PDF is the derivative of CDF

The probability density function (PDF) of a R.V **X** is defined as the derivative of $F_x(x)$.

$$f_X(x) = \frac{dF_X(x)}{dx}, \; f_{XY}(x,y) = \frac{\partial^2}{\partial xy} F_{XY}(x,y)$$

proof)

$$\begin{aligned} P[x < X \leq x+h] &= F_X(x+h) - F_X(x) \\ &= \frac{\{F_X(x+h) - F_X(x)\} \cdot h}{h} \end{aligned}$$

[7] Random Variables: PDF is the derivative of CDF

ex] Suppose a coin tossed one time.
Let R.V X be the indicator function for **tail event.**
CDF of X: $F_x(x)$ and PDF/ (PMF) of X: $f_X(x)$

[8] Random Variables: Computing Probability by using PDF

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

ex] the pdf of the amplitude of speech waveform
is proposed $as\ f_X(x) = 1/2 \cdot e^{-|x|}$
compute $P[|X| \leq 1]$?

Sol)

$$\begin{aligned} P[-1 \leq X \leq 1] &= 2\int_0^1 1/2 \exp^{-x} dx \\ &= \int_0^1 \exp^{-x} dx \\ &= 1 - e^{-1} \end{aligned}$$

- Computing Probability:

  Conditional & Joint Probability
  Independence & Conditional Independence
  Marginalization & Partition

[1] Computing Probability (Equally Likely Outcomes)

As equally likely outcomes, P[A] becomes counting problem.

$$P[A] = \frac{|A|}{|\Omega|}$$

Ex] When tossing a fair coin N times, compute P [k times H]

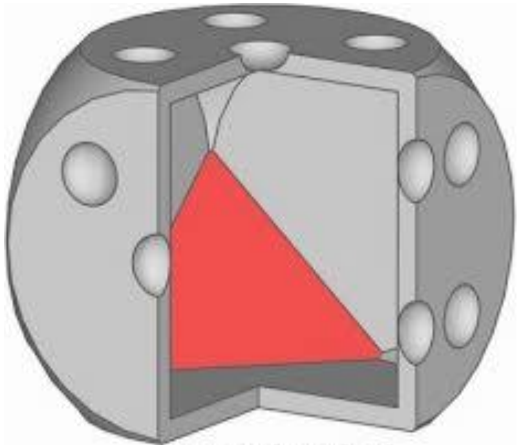
$$P[\ k \text{ times H }] = \binom{N}{k}\frac{1}{2^N}$$

- $|\Omega|$ = choose H or T N times : 2^N
- $|A|$ = choose k among different N without orders:

2000, Denver Native American One Dollar coin

[2] Computing Probability (Biased Outcomes)

However,
not always the outcomes are equally and likely.



Corner is loaded with lead!

[3] Computing Probability (Biased Outcomes)

However, if outcomes are not equally likely?

$$P[A] = \sum_{\omega_k \in A} P[\{\omega_k\}]$$
$$= \sum_{A_k \subset A} P[A_k], A_k \cap A_j = \emptyset \text{ if } k \neq j$$

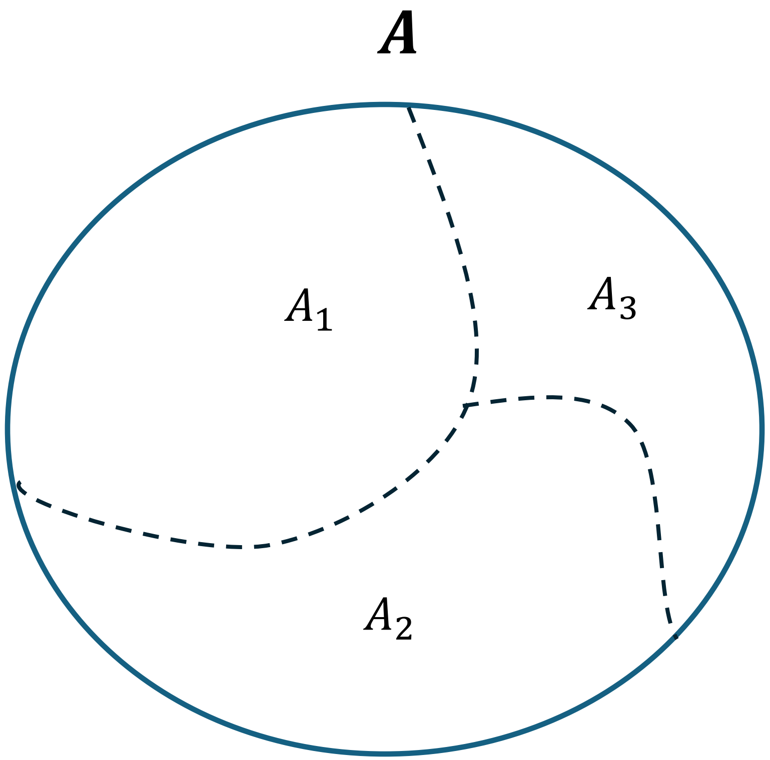


Fig: Event A can be divided into three disjoint sets.

We can divide a complex event into disjoint events, which are tracible.

[4] Computing Probability (Conditional Probability)

Computing Diagnostic Probability (Posterior Prob)

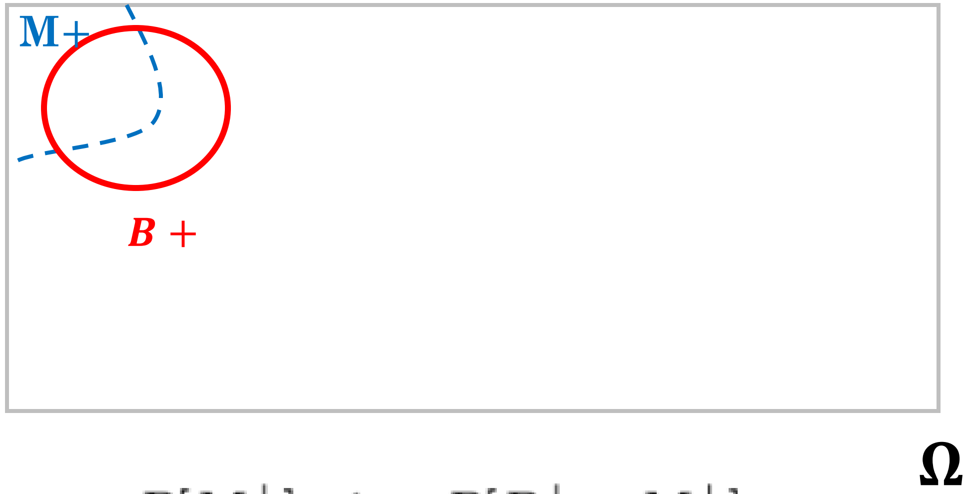
ex] **Breast Cancer** is a deadly disease that claims thousands of lives every year.

If you are a doctor who had a patient having a positive mammogram.

You know that mammogram accuracy is between 90% - 95%.

Which probability would you tell the patient?

- P[B+]=0.008?
- P[M accurate]=?
- P[B+|M+]? = 9%



$$P[M^+] : 1 = P[B^+ \cap M^+] : x$$

$$P[B+|M+] = \frac{P[B^+ \cap M^+]}{P[M^+]}$$

[5] Computing Probability (Conditional Probability and Joint Density)

- Chain Rule

$$P[A \cap B] = P[A] \cdot P[B|A] = P[B] \cdot P[A|B]$$

$$P[A \cap B \cap C] = P[A] \cdot P[B|A] \cdot P[C|A \cap B]$$

[6] Computing Probability (Intendent Events)

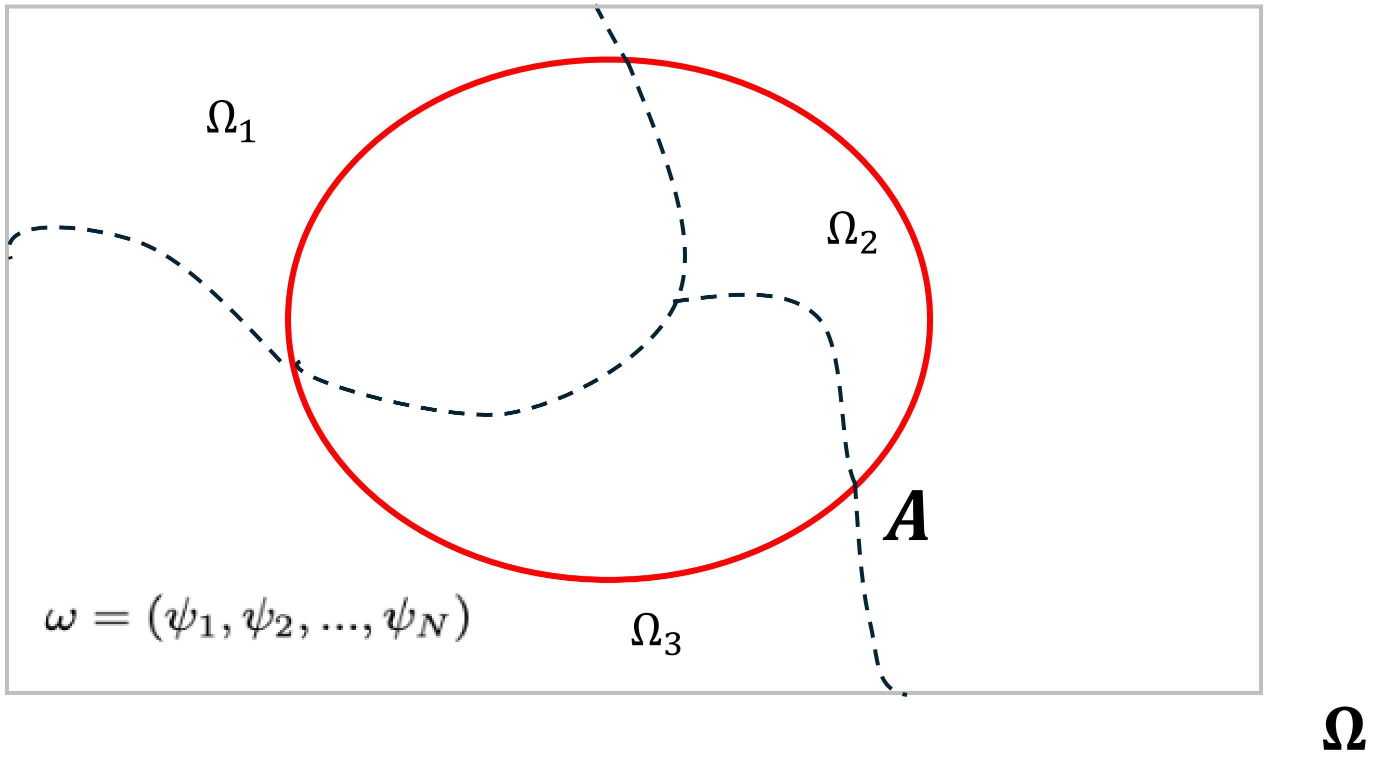
Independent Events $\leftrightarrow$

$$P[A \cap B] = P[A] \cdot P[B]$$

$$P[A \cap B \cap C] = P[A] \cdot P[B] \cdot P[C]$$

[7] Computing Probability (using Partition)

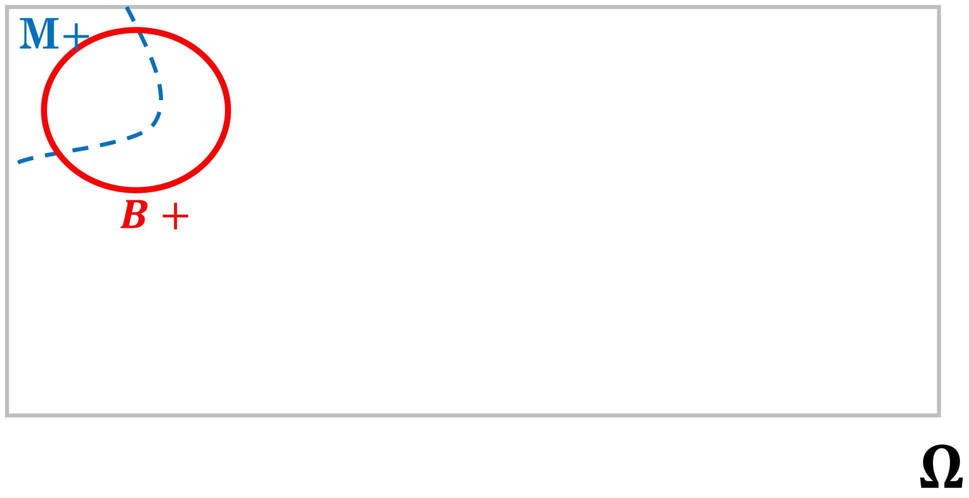
Partition the sample space Ω and measure the probability A



$$P[A] = P[A \cap \Omega_1] + P[A \cap \Omega_2] + P[A \cap \Omega_3]$$
$$P[A] = P[A|\Omega_1]P[\Omega_1] + P[A|\Omega_2]P[\Omega_2] + P[A|\Omega_3]P[\Omega_3]$$

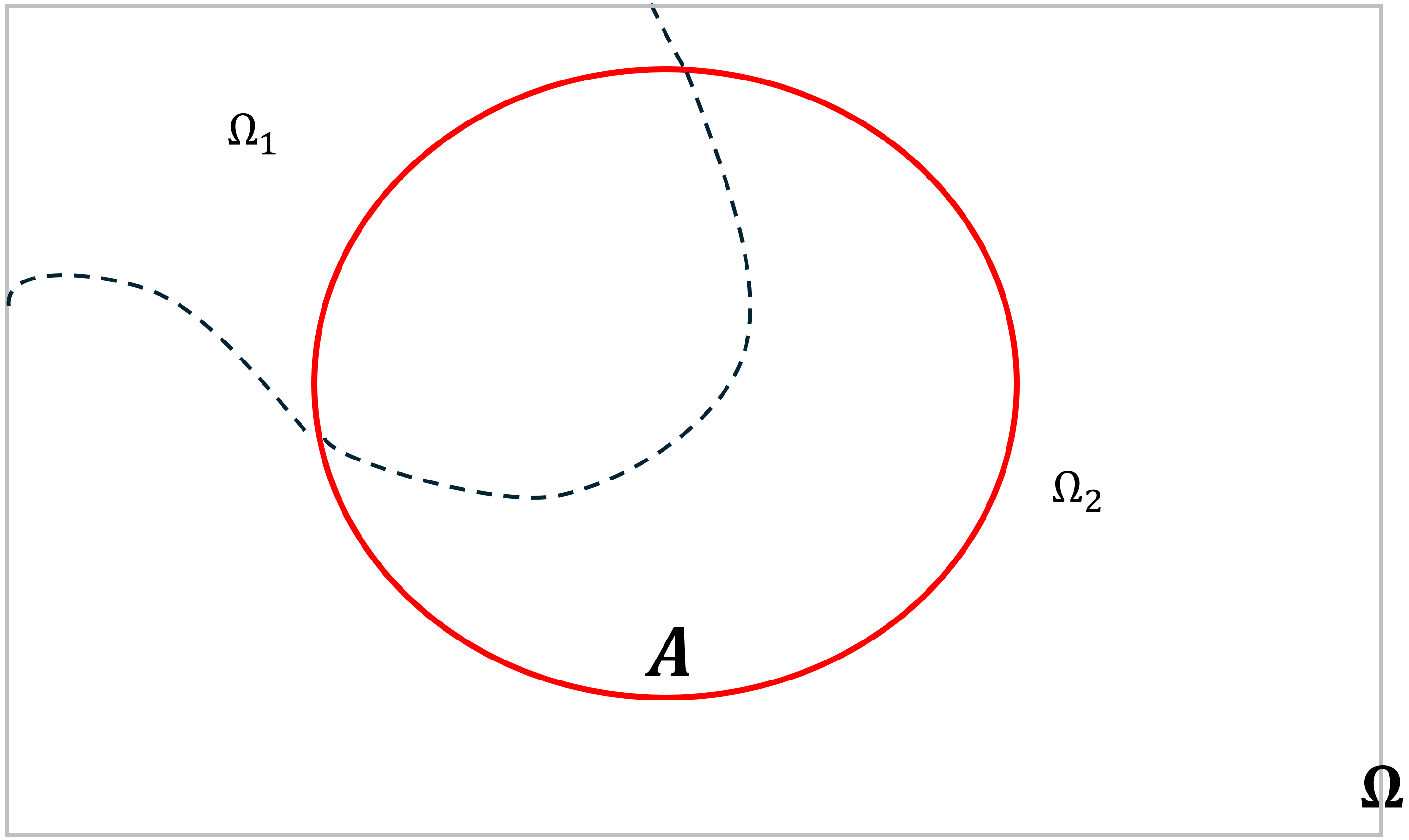
ex) revisit the breast cancer example and compute the diagnostic probability?

Given Information
- P[B+]=0.008
- P[M+|B+]=0.9 and P[M+|B-]=0.07
- P[B+|M+] =?



- Bayes Theorem

[1] Bayes Theorem (Inference)



$$P[\Omega_1|A] = \frac{P[A|\Omega_1] \cdot P[\Omega_1]}{P[A|\Omega_1] \cdot P[\Omega_1] + P[A|\Omega_2] \cdot P[\Omega_2]}$$

[2] Bayes Theorem (Example)

3. [Bayes Rule] Before going on vacation, you ask your friend to water your ailing plant. Without water, the plant has an 80 percent chance of dying. Even with proper watering, it has a 20 percent chance of dying. And the probability that your friend will forget to water it is 30 percent.

3.1 What's the chance that your plant will survive the week?

3.2 If your friend forgot to water it, what's the chance it'll be dead when you return?

3.3 If it's dead when you return, what's the chance your friend forgot to water it?

- Computing Statistics (Summarized Information about R.V)

  Mean & Variance & Covariance

[1] First Order Statistic (Mean)

$$E[X] = \sum_x x P(x)$$ (if X is a discrete R.V)

$$E[X] = \int_x x f(x) dx$$ (if X is a continuous R.V)

- Linearity of Expectation

$$\begin{aligned} E[aX^2 + bX + c] &= \sum_x (ax^2 + bx + c) P(x) \\ &= a \sum_x x^2 P(x) + b \sum_x x P(x) + c \sum_x P(x) \\ &= aE[X^2] + bE[X] + c \end{aligned}$$

[2] First Order Statistic (computing mean)

$$E[X] = \sum_x xP(x)$$

$$E[X] = \int_x xf(x)dx$$

ex] Compute $E[X]$ if $P[X = 1] = 1/3$ and $P[X = 0] = 2/3$

ex] Suppose X is Bernoulli ($P[X = 1] = 1/3$) and Y is binomial 3 choose y then $E[X]$ and $E[Y]$

[3] Second Order Statistic (computing variance)

$$VAR[X] = E[(X - E[X])^2]$$

$$\begin{aligned} VAR[X] &= E[(X^2 - 2XE[X] + E[X]^2)] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

ex] A random variable X has mean 2 and variance 7. Find $E[X^2]$.

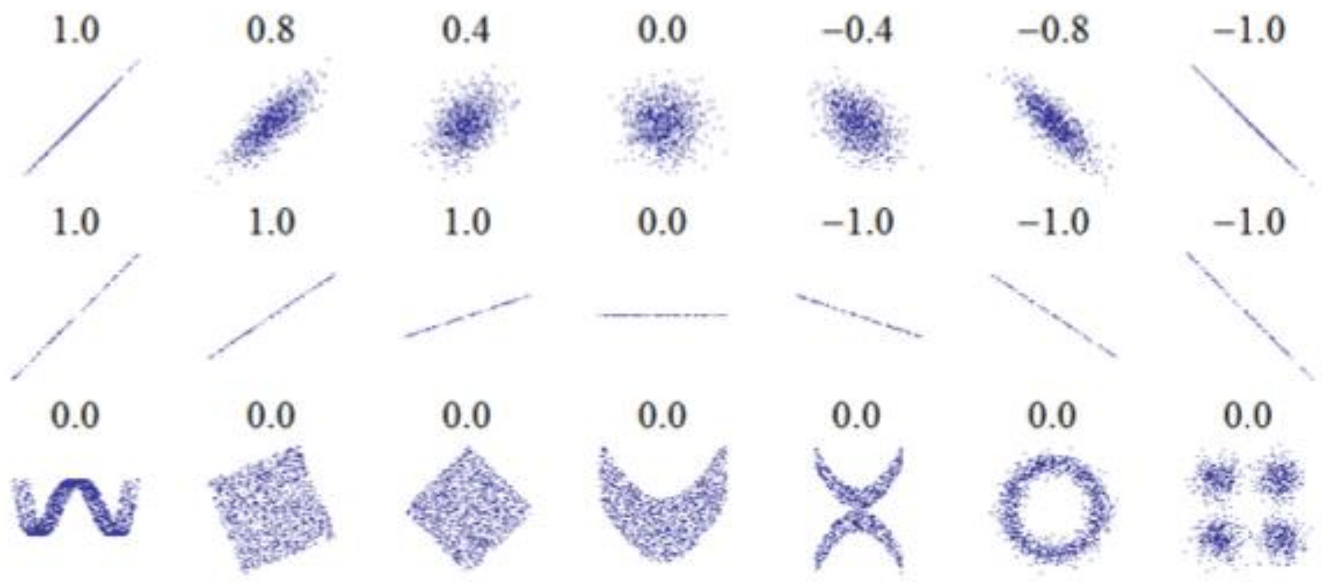
[4] Second Order Statistic (computing covariance)

$$
\begin{aligned}
VAR[aX+bY] &= E[(aX+bY-aE[X]-bE[Y])^2] \\
&= E[a^2(X-E[X])^2 + b^2(Y-E[Y])^2 + 2ab\cdot(X-E[X])\cdot(Y-E[Y])] \\
&= a^2E[(X-E[X])^2] + b^2E[(Y-E[Y])^2] + 2abE[(X-E[X])\cdot(Y-E[Y])] \\
&= a^2VAR[X] + b^2VAR[Y] + \boxed{2abE[(X-E[X])\cdot(Y-E[Y])]}
\end{aligned}
$$

COVARIANCE

- $COV(X,Y)+$: how the two R.Vs <u>linearly</u> covary?
- $COV(X,Y)-$
- $COV(X,Y)\ 0$

[5] Second Order Statistic (Covariance & Data Scatter plots)



From Figure 3.1 Murphy, Introduction

This figures presents the correlation coefficient $\rho = \frac{COV(X,Y)}{\sqrt{VAR(X)}\sqrt{VAR(Y)}}$

The correlation reflects (1) the noisiness & direction of a linear relationship

(2) not the slope of a linear relationship

- Computing Statistics for Random Vectors
  (Summarized Information for Multiple Dimensional Data)

  Mean Vector & Covariance Matrix

[1] Random Vectors (Mutli-dimensional / Multiple Realizations)

- In ML system development, *
  data is not one-dimensional; they are **multi-dimensional** (feature)

- In ML system development,
  Collective data is the realization of the **repetitive** process of the R.V (#points)

In this ML class,
we are going to with Random Vectors rather than single variables. $\vec{X} = (X_1, X_2, X_3, \ldots, X_N)$

[2] Computing Statistics for Random Vectors (mean vector and Covariance Matrix)

(1) Mean vector: $E[\overrightarrow{X}] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_D] \end{bmatrix}$

(2) Covariance Matrix

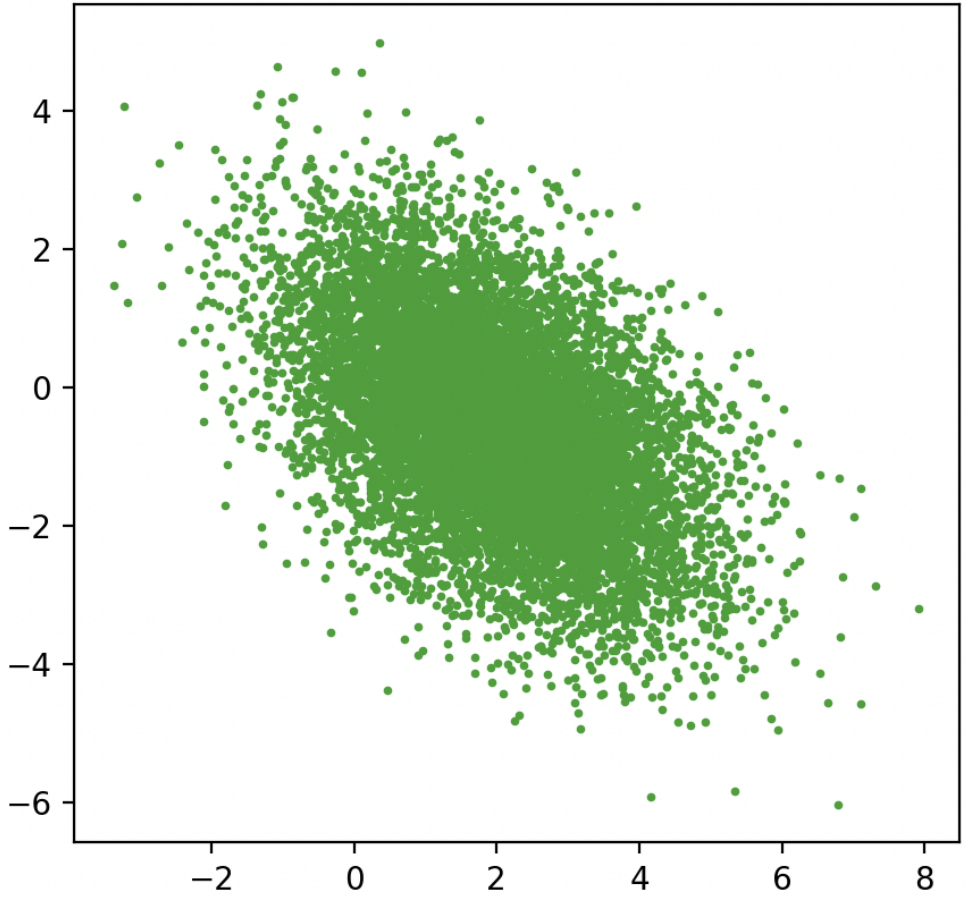
$$\mathrm{Cov}[\boldsymbol{x}] \triangleq \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^{\mathsf{T}}\right] \triangleq \boldsymbol{\Sigma}$$

$$= \begin{pmatrix} \mathbb{V}[X_1] & \mathrm{Cov}[X_1, X_2] & \cdots & \mathrm{Cov}[X_1, X_D] \\ \mathrm{Cov}[X_2, X_1] & \mathbb{V}[X_2] & \cdots & \mathrm{Cov}[X_2, X_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_D, X_1] & \mathrm{Cov}[X_D, X_2] & \cdots & \mathbb{V}[X_D] \end{pmatrix}$$

Mean vector shows the centric location for data points

Covariance Matrix shows how much data points spread and in what direction.

[3] Computing Statistics for Random Vectors (Covariance Matrix and Data Scatter Plots)



[2-d dimensional data scatter plots]

$$\Sigma = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

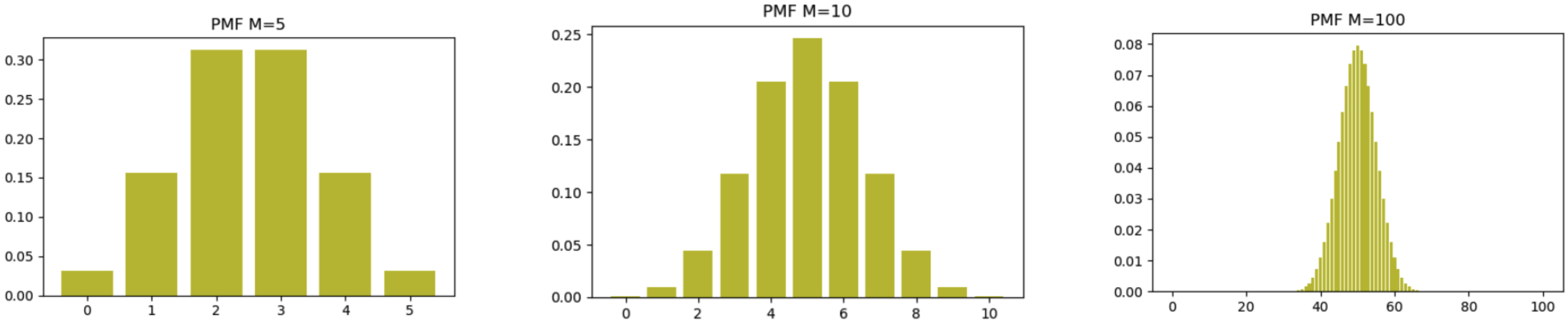
- Gaussian Density

  Central Limit Theorem

[1] Central Limit Theorem

The CDF of the sum of any i.i.d random variables with finite mean and variance approaches CDF of a Gaussian R.V.

[2] Central Limit Theorem (Binomial R.V)



A binomial R.V (B_M) can be represented by sum of Bernoulli R.Vs.
As $M \to \infty$, the shape of PMF approaches to Gaussian like.

[3] Gaussian Density (Central Limit Theorem)

$S_n = (S_1 + S_2 + \ldots + S_n)$
be the sum of n random variables (i.i.d)
with finite mean $E[X] = \mu$ and finite $VAR[X] = \sigma^2$

Then, $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$ (zero mean, unit variance)

s.t $\lim_{n\to\infty} P[Z_n \leq z] = \frac{1}{\sqrt{2\pi}} \boxed{\int_{-\infty}^{z} e^{-\frac{x^2}{2}} \mathrm{dx}}$

+ Gaussian density ~ $N(0,1)$

[4] Gaussian Density

[Scalar Gaussian]

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

[Multivariate Gaussian]

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp^{-1/2(\vec{x}-\vec{\mu})^t \Sigma^{-1} (\vec{x}-\vec{\mu})}$$

Q: Why is the Gaussian density often assumed in ML modeling?

- Maximum Likelihood Estimation

  (estimation of mean and variance)

[1] Bayes Rule (in ML as an inference Method**)

$$\boxed{P(w|D)} = \frac{p(w,D)}{P(D)} = \frac{\boxed{p(D|w)}\boxed{p(w)}}{p(D)}$$

(posterior) **(likelihood)** **(prior)**

Bayesian	Frequentist
• quantification uncertainty • prior density (expert knowledge)	• relative frequency (as # trials goes ∞) • w exists as a fixed point ᥙᦞ

[2] Recall slide **: Bayes Rule (in ML as an inference Method **)

Frequentist vs. Bayes Estimation

- $w* = argmax\ P(D|w)$: **M**aximum **L**liklihood **E**stimation (MLE)
- $w* = argmax\ p(w|D) = \frac{p(D|w)p(w)}{p(D)}$: **M**aximum **A** **P**osteriori Estimation (MAP)

Frequentist assumes w (parameter) as fixed values and perform MLE to estimate the parameters. MLE can be interpreted as a special case of MAP when the prior density $p(w)$ is uniform.

[3] MLE (Computing the Sample Mean)

Problem] suppose we collected i.i.d data points following $\sim N(\mu, \sigma^2)$ estimate the mean value of the samples by using MLE method.

$$f(x_1, x_2, x_3, ..., x_n) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp^{-1/2\sigma^2\{(x_1-\mu)^2+(x_2-\mu)^2+...+(x_n-\mu)^2\}}$$

$$\ln f(x_1, x_2, x_3, ..., x_n) = \ln \frac{1}{(\sqrt{2\pi\sigma^2})^n} + \{(x_1-\mu)^2+(x_2-\mu)^2+...+(x_n-\mu)^2\}$$

$$\frac{\partial}{\partial\mu} f(x_1, x_2, x_3, ..., x_n) = -2\sum_{i=1}^{n}(x_i - \mu) = 0$$

$$n\mu = \sum_{i=1}^{n} x_i$$

$$\boxed{\mu = 1/n \sum_{i=1}^{n} x_i}$$

Sample Mean:

$$M_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

[4] Sample Mean (Unbiased Estimator & Asymptotic Behavior)

Sample Mean:
$M_N = \frac{1}{N}\sum_{i=1}^{N} X_i$

The sample mean (M_N) is a new Random Variable!

[The Two Statistics of M_N]

$$E[1/n \sum_{i=1}^{n} x_i] = \mu$$

$$V[1/n \sum_{i=1}^{n} x_i] = \sigma^2/n$$

- The sample mean is an unbiased estimator
- As N increases M_N gets close to true mean μ

[5] MLE (Computing the Sample Variance)

Sample Variance (biased):

$$V_N = \frac{1}{N}\sum_{i=1}^{N}(X_i - M_n)^2$$

Sample Variance (unbiased):

$$V_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - M_n)^2$$

This problem will be covered in detail during recitation

- Computing Sample Mean Vector & Sample Covariance Matrix (Examples)

[1] Computing Sample Mean (Data Matrix: D)

features

	X_1	X_2	X_3
#1	1.4	2.7	3.2
#2	2.2	3.5	3.3
#3	3.1	5.2	1.2
#4	1.7	1.0	0.2
#5	4.6	1.1	0.9
#6	2.2	4.3	2.7
#7	1.2	2.3	7.6
#8	0.3	0.2	3.2
#9	2.5	0.5	0.9
#10	1.8	1.9	1.1

# of data samples

[2] Computing Sample Mean

features

	X_1	X_2	X_3
#1	1.4	2.7	3.2
#2	2.2	3.5	3.3
#3	3.1	5.2	1.2
#4	1.7	1.0	0.2
#5	4.6	1.1	0.9
#6	2.2	4.3	2.7
#7	1.2	2.3	7.6
#8	0.3	0.2	3.2
#9	2.5	0.5	0.9
#10	1.8	1.9	1.1

# of data samples

$$M_{10} = \begin{bmatrix} 2.1 \\ * \\ * \end{bmatrix}$$

[3] Computing Sample Covariance

features

	X_1	X_2	X_3
#1	1.4	2.7	3.2
#2	2.2	3.5	3.3
#3	3.1	5.2	1.2
#4	1.7	1.0	0.2
#5	4.6	1.1	0.9
#6	2.2	4.3	2.7
#7	1.2	2.3	7.6
#8	0.3	0.2	3.2
#9	2.5	0.5	0.9
#10	1.8	1.9	1.1

# of data samples

$$C_{10} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}$$

$$= \frac{(D - M_{10}\mathbf{1})^t (D - M_{10}\mathbf{1})}{N-1}$$

Summary

0. Why Probability in ML?
1. Two Interpretation of Probability: frequentist vs Bayesian
2. Probability axioms
3. Random Variables/ Random Vectors
3. Computing Probability: Joint & Conditional Prob/ Marginalization
5. Bayes Rules
6. Important statistics : mean & variance & covariance
7. Gaussian Density
8. Maximum Likelihood Estimation (MLE)
9. Sample mean and Sample Variance