# Artificial Intelligence/Machine Learning/Deep Learning: 'Bridging the Skills Gap'

**Optional: Math Refresher - Statistics**

A machine-learning system is trained rather than explicitly programmed. It is presented with many examples (features, labels) relevant to a task, and it **finds statistical structure** in these examples that eventually allows the system to come up with a model for automating the task.

Topics of this session include:

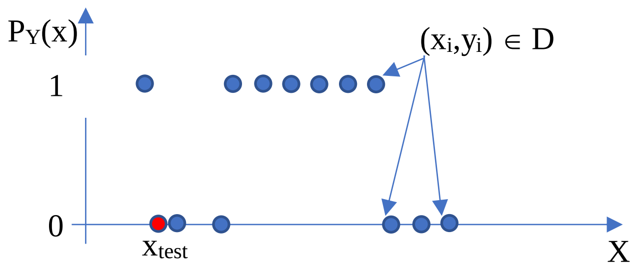
1. Sample Space, Random Variable, Probability Distribution of a Random Variable
2. Conditional Probability & Bayes Rule
    - Maximum Likelihood Estimate (MLE)
    - Maximize a Posterior (MAP)
3. Expected Value and Variance of a Random Variable
4. Discreet Probability Distributions: Bernoulli, Binomial
5. Continuous Probability Distributions: Gaussian/Normal Distribution
6. Weak Law of Large Numbers (WLLN)
7. Central Limit Theorem
8. Covariance/Variance Matrix of a Random Variable

A ML learning classification problem can be defined as follows:
We observe a dataset D= $\{(\boldsymbol{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \rightarrow$ drawn from a distribution $P_Y(Y|X)$ that we do now know!

$\mathbf{x}_i$: feature vector
y: label (class)

*A binary classification problem is about learning the distribution of the label $P_Y$ so that we can predict the label given a test point: $P(y|x_{test})$ – where $P(y|x_{test})$ is the conditional probability*

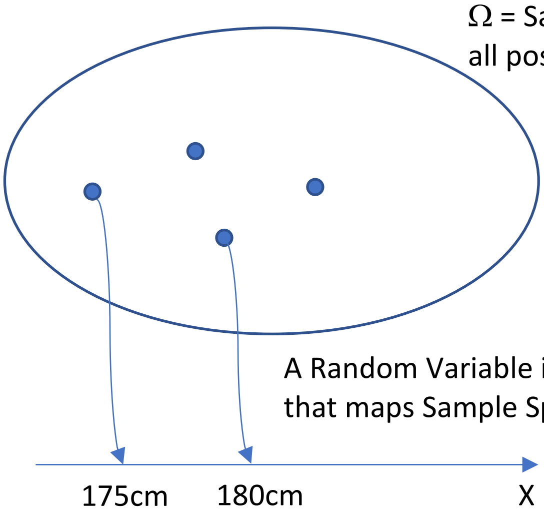
Example: given the brain scans of a patient $x_{test}$
We want to be able to predict if a patient has a tumor (y=1) or not (y=0)



**<u>Sample Space, Random Variable, Probability Distribution</u>**

Assume we have an experiment

$\Omega$ = Sample Space:
all possible outcomes of the experiment



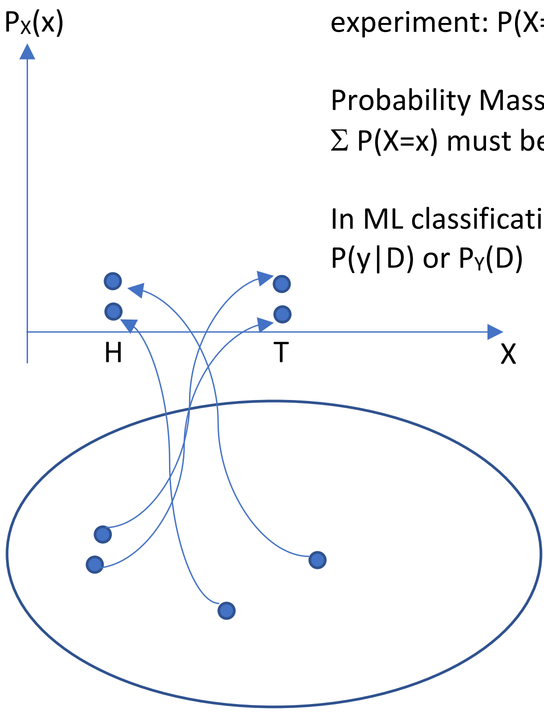

A Random Variable is a function that maps Sample Space to $\mathbb{R}$

Height of Students is a (continuous) RV if $\Omega$ would represents all Grade 4 students

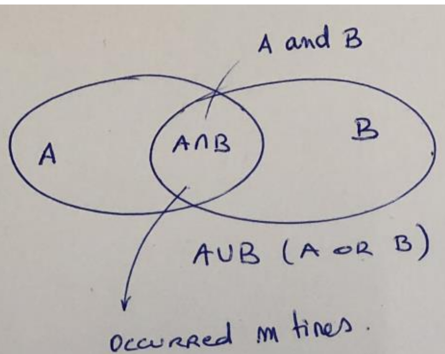
Probability Distribution $P_X(x)$ a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment: P(X=x). For a coin toss this would be P(H) and P(T)

Probability Mass Function (pmf) if X is discreet
$\Sigma$ P(X=x) must be 1

In ML classification problem we are estimating P(label y|data D) or P(y|D) or $P_Y(D)$



**Conditional Probability & Bayes Rule**




2 EVENTS

A occurred $m_1$ times

B occurred $m_2$ times

$n = m_1 + m_2$ total occurances!

$P(A \wedge B) = P(A) \cdot P(B)$

$$P(A \cup B) = P(A-B) + P(B-A) + P(AB)$$
$$= P(A) - P(AB) + P(B) - \cancel{P(AB)} + \cancel{P(AB)}$$
$$= P(A) + P(B) - P(AB).$$

let $P(A|B)$ be the probability of A given that B has already occurred

↑ conditional probability.

B happened in $m_2$ ways

A restricted to the event $A \cap B$ that can occur in $m$ ways.

$$\rightarrow P(A|B) = \frac{m}{n_2}$$

$$\rightarrow P(A|B) = \frac{m/n}{n_2/n} = \frac{P(AB)}{P(B)}$$

divide both by $n$

$$\rightarrow P(A|B) = \frac{P(AB)}{P(B)} \rightarrow P(AB) = P(B)\,P(A|B) \quad (1)$$
$$P(AB) = P(A)\,P(B|A) \quad (2)$$

$$(1)\,(2) \rightarrow P(B)\,P(A|B) = P(A)\,P(B|A)$$

$$\rightarrow \boxed{P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}} \rightarrow \text{Bayes Rule}$$

---

Why is Bayes Rule so important for ML?

Dataset D = {($\boldsymbol{x}_1$, $y_1$,), …, ($\mathbf{x}_n$, $y_n$)} → observed from some probability distribution **P** that we do now know → but maybe we can approximate the distribution from the data!

Assume a simple 1-dimensional experiment of several coin tosses:
D = {H, T, T, H, H, H, T, T, T, T}

**Maximum Likelihood Estimation (MLE)**: given that I observe the data D, which parameters θ would make it most likely that I observe the data D → MLE = argmax (D|θ) where argmax refers to the parameters θ at which the function outputs are as large as possible.

For our coin toss experiment: looking at the data D we can estimate P(H) and P(T) as follows:

$P(H) \approx \frac{n_H}{n_H+n_T}$ = 4/10 → not very accurate especially if sample size is small

Could be problematic when for example $n_H$=0 (small number of coin tosses)

In this case we can for example add 1 to numerator and add 2 to denominator → this is called smoothing. Alternatively, you can add m tosses of Hs and m tosses of Ts because you have a prior belief over the distribution P(θ)

**Maximize a Posterior (MAP)**
Bayesians consider θ as a RV with a known distribution P(θ). P(θ) is called the prior and encodes your belief of what θ should be. MLE supporters claim that there is no given sample space where you can draw θ from.

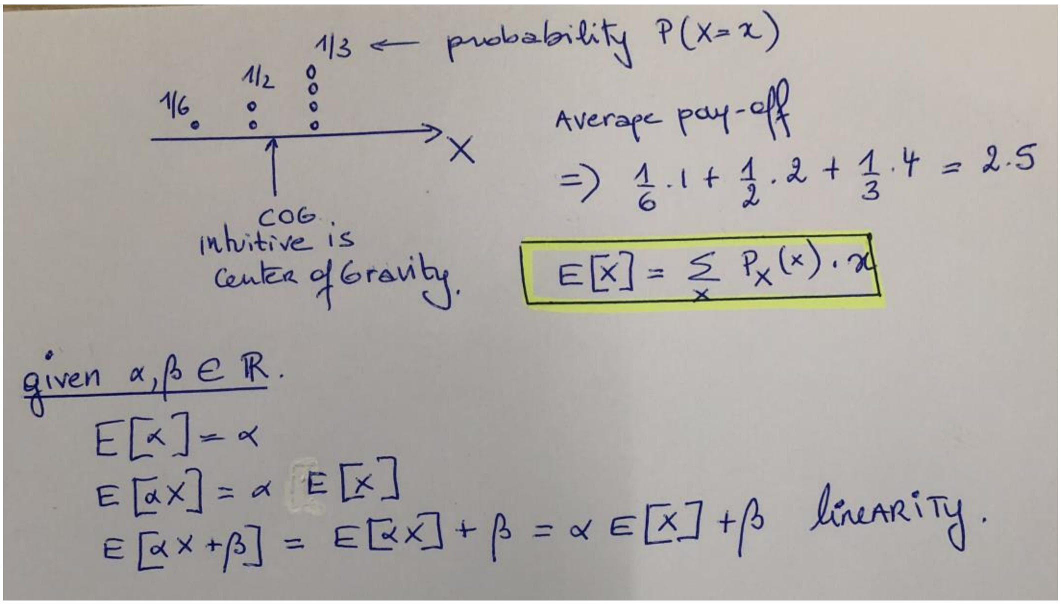
P(D|θ) is the MLE

Using Bayes we can estimate the distribution of the parameters θ

$P(\theta|D) = \frac{P(D|\theta).P(\theta)}{P(D)}$

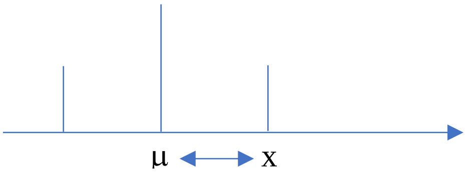
**<u>Expectation and Variance of a Random variable</u>**

The expected value is the average value of a RV over a large number of experiments.



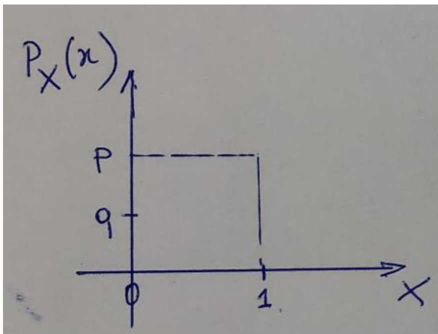
Variance of a RV: Var(X)

- Variance of a RV X with $E[X]=\mu$ is defined as $Var[X]=E[(X-\mu)^2]$ or $Var[X]=E[(X-\mu)(X-\mu)^T]$ or $Var[X]=E[X^2]-(E[X])^2$
- Var[X] is a RV
- Var(X) is a measure of the spread of the distribution with $\mu$ as reference point. Far away points are more penalized through squaring
- $Var[X] \geq 0$
- $Var[\alpha X+\beta]=a^2\ Var[X]$ with $\alpha,\beta$ scalars
- $\sqrt{Var}=\sigma$ standard deviation




**<u>Discreet Probability Distributions: Bernoulli, Binomial</u>**

*Bernoulli Distribution:* discrete distribution with parameter **p**
Random variable X takes the value 1 with probability p and value 0 with probability q=1-p. Example: a coin toss




$$p(x=1)=p.$$
$$P(x=0)=q=1-p.$$
$$E[X]=p \quad (\text{coin toss } p=50\%)$$
$$Var(X)=E[X^2]-(E[X])^2$$
$$=\underbrace{P(X=1)\cdot 1^2+P(X=0)\cdot 0^2}_{p}-(E[X])^2$$
$$\Rightarrow Var(X)=p-p^2=p(1-p)=pq.$$

Independence of 2 events A and B & Counting:

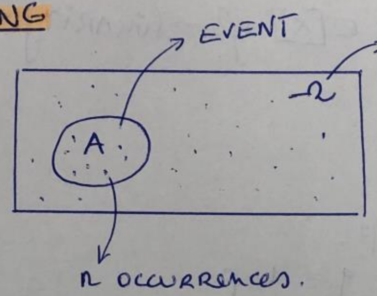
$$P(B|A) = P(B)$$

independent: occurrence of event A does not give you any information about B's occurrence

$$P(A \cap B) = P(A)\, P(B|A)$$

$$\Rightarrow \boxed{P(A \cap B) = P(A) \cdot P(B)}$$

**COUNTING**



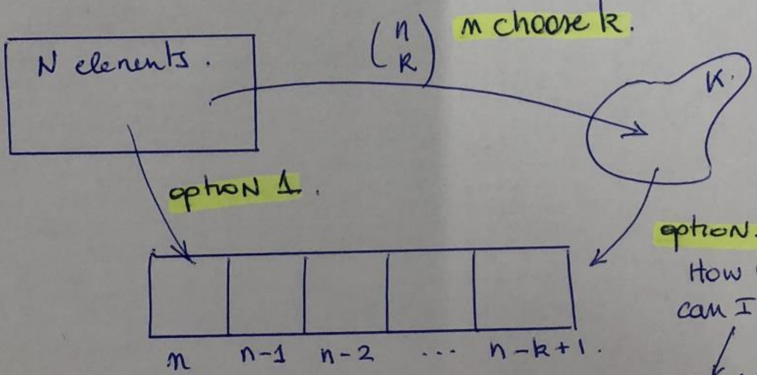

sample space : N occurrences

$$|\Omega| = N$$
$$|A| = n$$
$$P(A) = \frac{n}{N}$$

example: how many licence plates can you make if you have 3 letters and 4 digits.

| | letter 1 | letter 2 | letter 3 | digit 1 | digit 2 | digit 3 | digit 4. |
|---|---|---|---|---|---|---|---|
| put back in → | 26 | 26 | 26 | 10 | 10 | 10 | 10. |
| don't put back in → | 26 | 25 | 24 | 10 | 9 | 8 | 7. |

$\binom{n}{k}$ n choose k.




PERMUTATIONS

How many ordered list can I make with k elements → $k!$

↳ use example if not clear.!

$$\frac{n!}{(n-k)!} \overset{\text{must}}{=} \binom{n}{k} k!$$

$$\longrightarrow \boxed{\binom{n}{k} = \frac{n!}{k!\,(n-k)!}}$$

BINOMIAL Coefficients!

$0! \overset{\text{set}}{=} 1$

*Binomial distribution**:* discreet distribution with parameters **n** and **p**
Reflects the number of successes in a sequence of n independent (Bernoulli) experiments. If n=1 the binomial distribution is a Bernoulli distribution.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. In ML we will discuss this when we talk about Support Vector Machines (SVMs) and Naïve Bayes. For naïve Bayes we will assume that the prior P(θ) is binomial.

$$\boxed{X \sim B(\underbrace{n}_{\#\text{ trials}}, \overbrace{p}^{\#\text{ successes}})} \qquad p \in \{0, 1\}$$

# successes from n trials

assume k successes with probability p.

$\longrightarrow p^k.$

$\Rightarrow$ failure rate $= (n-k) \rightarrow$ will occur with probability $(1-p)$

$\longrightarrow (1-p)^{n-k}.$

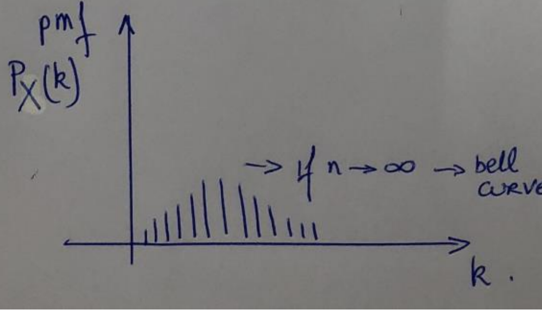
can occur anywhere among n trials.
and there are $\binom{n}{k}$ different ways of distributing k successes in a sequence of n trials!

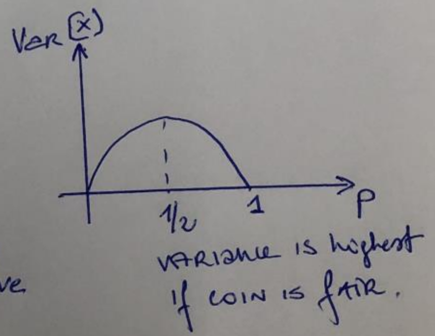
$$\longrightarrow \boxed{f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}.} \quad \text{pmf } P_X(k)$$

<u>example</u>: H, H, H $\rightarrow p = 3$    $n = 5$
T, T $\rightarrow (1-p) = 2$

$$\longrightarrow \binom{5}{3} = \frac{5!}{3!\,2!} = 10$$

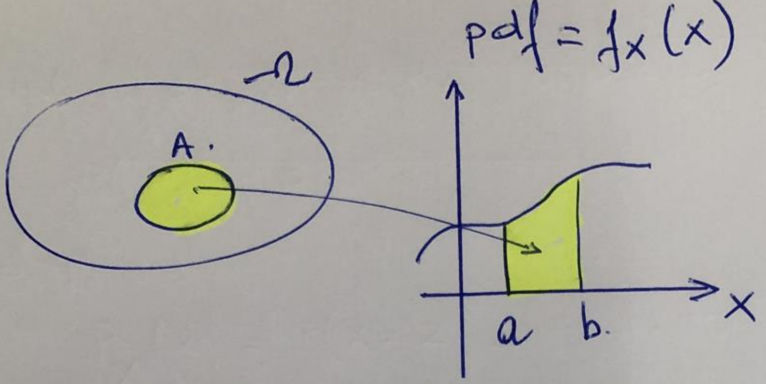
$E[X] = np \longrightarrow n$ Bernoulli trials with $E[X] = p$
$Var[X] = np(1-p)$
$= n\,Var[X]$



## Continuous Probability Distributions: Gaussian/Normal Distribution

A continuous RV is described by its probability density function (pdf)




$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx \qquad \int_{-\infty}^{+\infty} f_X(x)\,dx = 1.$$

↳ $\Sigma \rightarrow \int$

$P_X \rightarrow f_X$

$P(X=a)=0$ !

$f_X(x) \geq 0$

$$E[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\,dx.$$

$$VAR[X] = \sigma_X^2 = \int_{-\infty}^{+\infty} (x - E[X])^2 f_X(x)\,dx.$$

Gaussian or Normal Distribution

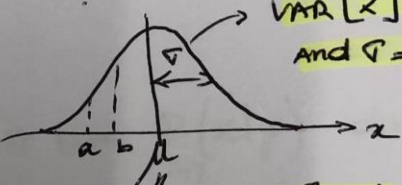
Example: a continuous random variable X is used to denote the height of all adult males in Singapore. In this specific case the distribution is a **Normal (Gaussian)**: $\mathcal{N}(\mu, \sigma^2)$
Lots of experiments tend to be Gaussian mainly because of the **central limit theorem**

## GAUSSIAN (normal) PDF

(10

↳ if you measure a quantity that is made off of lots of random contributions

independent of distributions of little random contributors ⟶ Normal distribution $\mathcal{N}(\mu, \sigma^2)$




$VAR[X] = \sigma^2$ and $\sigma$ = std deviation

mean $E[X] = \mu$

$\mu \pm \sigma \rightarrow 68\%$ of data
$\mu \pm 2\sigma \rightarrow 95\%$
$\mu \pm 3\sigma \rightarrow 99.99\%$

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

$$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1.$$

to make sur $\int = 1$.

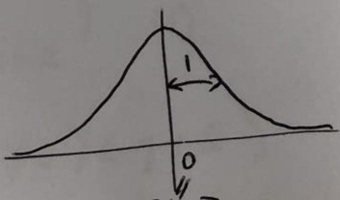
$$P(a \leq x \leq b) = \int_a^b f_X(x)\, dx$$

$\mathcal{N}(0,1) \rightarrow$ std normal distribution

$\mu = 0$ $\quad$ $\sigma^2 = 1$

from $\mathcal{N}(\mu, \sigma^2)$ substract $\mu$ and devide by. $\sigma$

$$\left(\frac{x-\mu}{\sigma}\right)$$




$E[X] = 0$
$VAR[X] = 1$

A probability distribution whose sample space is the set of real numbers is called **univariate**, while a distribution whose sample space is a vector space ($X_1$, $X_2$, $X_3$…) is called **multivariate**.

## Weak Law of large numbers (WLLN)

*Sample mean $E[\overline{X}]$ will converge to the population mean E[X] if n $\to\propto$*

For Independent and Identically Distributed (i.i.d) RVs: $X_1$, $X_2$,…,$X_n$, the sample mean $\overline{X}$ is denoted by:

$\overline{X}=\frac{X_1+X_2+\cdots+X_n}{n}$

Since the $X_i$'s are RVs, the sample mean $\overline{X}$ is also a RV

Assume a repetitive experiment (n times) of 100 coin tosses $\to$ count the number of Hs.

The E[X]=50% while for example $\overline{X}=\frac{50+44+\cdots+52}{n}$

$E[\overline{X}]=\frac{E[X_1]+E[X_2]+\cdots+E[X_n]}{n}$ $\to$ by linearity of E[X] and each $E[X_i]=E[X]$ because we expect 50 heads for each experiment

$E[\overline{X}]=\frac{nE[X]}{n}=E[X]$

$Var[\overline{X}]=\frac{Var[X_1+X_2+\cdots+X_n]}{n^2}$ and $Var[\alpha\overline{X}]=\alpha^2 Var[\overline{X}]$ …in this case $\alpha$ = 1/n

$Var[\overline{X}]=\frac{Var[X_1]+Var[X_2]+\cdots+Var[X_n]}{n^2}$ since are $X_i$'s independent

$Var[\overline{X}]=\frac{nVar[X]}{n^2}$ because $Var[X_i]=Var[X]$

$Var[\overline{X}]=\frac{Var[X]}{n}$

## Central Limit Theorem:

*The sum of a large number of (i.i.d) RVs $X_1$, $X_2$,…,$X_n$, with $E[X_i]=\mu<\propto$ and $Var[X_i]=\sigma^2$ is approximately normal, no matter what the distribution of the $X_i$'s are.*

## Covariance/Variance matrix

The Covariance of 2 RVs Cov(X,Y) = E[(X-E[X]) (Y-E[Y])] gives information about how the RVs are statistically related and how they move relative to each other.

Note that:

Cov(X,Y) = E[(X-E[X]) (Y-E[Y])]=E[XY-XE[Y]-E[X]Y+E[X]E[Y]] (1)

(1) = E[XY]+E[X]E[Y]-E[X]E[Y]-E[X]E[Y] = E[XY]-E[X]E[Y]

So Cov(X,Y)= E[XY]-E[X]E[Y]

*Properties:*
Cov(X,X)=Var[X]
Cov(X,Y)=Cov(Y,X)
If X and Y are independent Cov(X,Y)=0

Cov(aX,Y)=aCov(X,Y)
Cov(X+c,Y)=Cov(X,Y)

Variance - Covariance Matrix:

Let $X=(X_1,\dots,X_n)^T$ with $X_i$ RVs with finite $Var[X_i]$ and $E[X_i]$

**$X^T.X$** is measure for similarity of the features: (nxd)*(dxn) $\rightarrow$ nxn where n is number of samples and d is the dimension of the feature vector.

$$\mathbf{X^T.X} = \begin{pmatrix} & \boldsymbol{X_1} & \\ \vdots & \ddots & \vdots \\ & \boldsymbol{X_d} & \end{pmatrix} \begin{pmatrix} & \cdots & \\ \boldsymbol{X_1} & \ddots & \boldsymbol{X_d} \\ & \cdots & \end{pmatrix} = \begin{pmatrix} \boldsymbol{X_1.X_1} & & \boldsymbol{X_1.X_d} \\ \vdots & \ddots & \vdots \\ \boldsymbol{X_d.X_1} & & \boldsymbol{X_d.X_d} \end{pmatrix}$$

We assumed that we did mean normalization for all elements in the matrix
Now we take the $E[X^TX]$ and divide by n

$$\rightarrow \begin{pmatrix} \boldsymbol{Var[X_1]} & & \boldsymbol{Cov(X_d.X_1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{Cov(X_d,X_1)} & & \boldsymbol{Var[X_d]} \end{pmatrix} \rightarrow \mathbf{1/n} \begin{pmatrix} \boldsymbol{Var[X_1]} & & \boldsymbol{Cov(X_d.X_1)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{Cov(X_d,X_1)} & & \boldsymbol{Var[X_d]} \end{pmatrix} \quad \mathbf{(2)}$$

matrix is symmetric and square and diagonal shows the variances of the features $X_1, \dots, X_d$

(2) $\rightarrow \sum = \mathrm{Cov}(\mathbf{X}) = \frac{X^T.X}{n}$ and $\sum$ is the Covariance Matrix and $\sum$ is Positive Semi-Definite

A matrix $\sum$ is positive semi-definite if the scalar $X^T \sum X \geq 0$ for $X \in \mathbb{R}^n$

The eigenvalues of a square and symmetric positive semi-definite matrix are all positive.

It means that if we transform a vector X through $\sum$, the new vector $\sum X$ will be pointing in the same general direction ($\theta<90^0$) and will not change sign.
$X^T.(\sum X) = |XT|.|\sum X|.\cos(\theta)$ and cos is positive as long as $\theta<90^0$