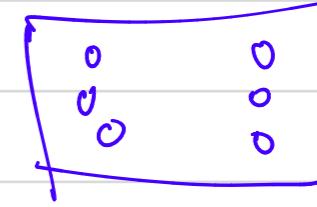
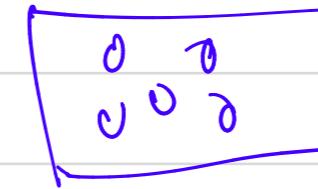
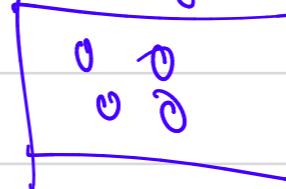
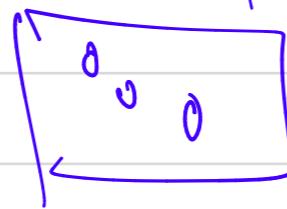
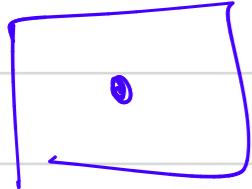


Day-73, Feb-12, 2025 (Mogh 29, 2081)

F Conditional Probability.



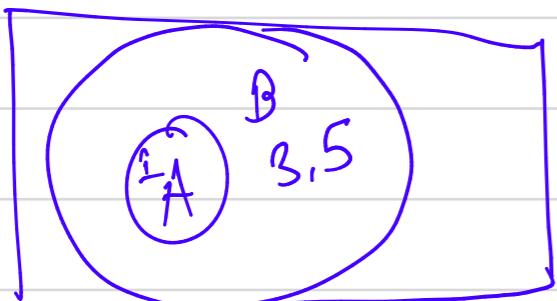
$$A = \{1, 3, 5\}$$

$$B = \{1, 3, 5\}$$

$P(\text{one given that roll is odd}) = P(A|B)$

$$\Rightarrow \frac{P(A \cap B)}{P(B)}$$

$P(A \cap B) = P(A)$ because



$$\Rightarrow \frac{P(A)}{P(B)}$$

$$\Rightarrow \frac{\frac{1}{6}}{\frac{3}{6}} \Rightarrow \frac{1}{3}$$

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

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

→ Choose set B be an event
so that $P(B) > 0$.

$$\neq P(A)$$

If Bayes' Rule (Bayesian Theorem):

Compute $P(B|A)$ When we have $P(A|B)$ so,

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

mostly used in diagnostic test.

Diagnostic tests

let f and $-$ be the events that the result of a diagnostic test is positive or negative respectively -
 let D and D^c be the event that the subject of the test has or doesn't have the disease respectively -

$$\text{Sensitivity} = P(+|D)$$

$$\text{Specificity} = P(-|D^c)$$

for example, in an HIV blood test, we could take people who we know to have disease and apply the diagnostic test to that blood. Similarly the people not having disease.

More Definitions

$$(1) \text{ positive predictive value} = P(D|+)$$

Probability of Having Disease given that the disease is positive

(2)

$$\text{Negative Predictive value} = P(D^C | -)$$

(3)

$$\text{Prevalence of disease} = P(D)$$

Example:

Sensitivity of 99.7%.

Specificity of 98.5%.

Population with a 1%. Prevalence of HIV

$$P(D|+)$$

Source: John Hopkins University Inferential Statistics Coursera Course

Using Bayes' formula:

$$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+|D) \cdot P(D) + P(+|D^C) \cdot P(D^C)}$$

$$\Rightarrow \frac{P(+|D) \cdot P(D)}{P(+|D) \cdot P(D) + \{1 - P(-|D^C)\} \{1 - P(D)\}}$$

$$\Rightarrow \frac{.997 \times .001}{.997 \times 0.001 + .015 \times .999} \Rightarrow 0.062$$

Likelihood Ratios:

$$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+|D) \cdot P(D) + P(+|D^c) \cdot P(D^c)}$$

population with a 1% prevalence of HIV,

$$P(D^c|+) = \frac{P(+|D^c) \cdot P(D^c)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

$$P(D^c|+) \Rightarrow \frac{P(+|D^c) \cdot P(D^c)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

Now dividing the eqns

$$\frac{P(D|+)}{P(+|D)} = \frac{P(+|D) \cdot P(D)}{P(+|D) \cdot P(D) + P(+|D^c) \cdot P(D^c)} \quad \text{--- eqn 0.}$$

$$\frac{P(D^c|+)}{P(+|D^c)} \Rightarrow \frac{P(+|D^c) \cdot P(D^c)}{P(+|D) \cdot P(D) + P(+|D^c) \cdot P(D^c)} \quad \text{--- eqn 11}$$

Now eqn 0 and eqn 11 dividing we get -

dikelihuan ratio's

$$\frac{P(D|+)}{P(D^c|+)} \Rightarrow \frac{P(+|D) \cdot P(D)}{P(+|D^c) \cdot P(D^c)}$$

When we take a probability and divide it by $1 - p =$ odds

→ Diagnostic test result
likelihood ratio.

$$\frac{P(D|+)}{P(D^c|+)} = \frac{P(+|D) \cdot P(D)}{P(+|D^c) \cdot P(D^c)} \rightarrow \text{odds of disease absence in the event of the result.}$$

Odds of disease given a positive test result.

$$\frac{P(D|+)}{P(D^c|+)} = \frac{P(+|D)}{P(+|D^c)} \times \frac{P(D)}{P(D^c)}$$

post-test odds of $D = DLR_+ \times$ pre-test odds of D

HIV example revisited:

[Specificity of 98.5%] [Sensitivity of 99.7%]

$$DLR_+ = \frac{0.997}{(1 - 0.985)}$$

≈ 66 times

or pre-test odds are

66 times more supported by the data than the hypothesis of no disease.

$$DLR_- = (1 - .997) | .985 \approx 0.03$$

post-test odds of disease in the light of negative test result is now $0.3 | .985$ times of the pre-test odds of disease. In other words, the hypothesis of disease is supported 0.003 times that of the hypothesis of the absence of disease given the negative test result.

Brief Idea of Independence in Probability:

→ Event A is Independent of event B if

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

Example:

$$A = \{ \text{Head in flip 1} \} \cup P(A) = 0.5$$

$$B = \{ \text{Head in flip 2} \} \cup P(B) \Rightarrow 0.5$$

$$A \cap B = \{ \text{Head in flips 1 and 2} \}$$

$$P(A \cap B) = P(A) \cdot P(B) \Rightarrow 0.25$$

Example: prevalence of Sudden Infant Death Syndrome of 1 out of 8,543,

$$\left(\frac{1}{8543}\right)^2$$

Multiplying Probabilities without Independence

i.e. $p(A_1 \cap A_2)$ is not necessarily equal to $p(A_1) \cdot p(A_2)$

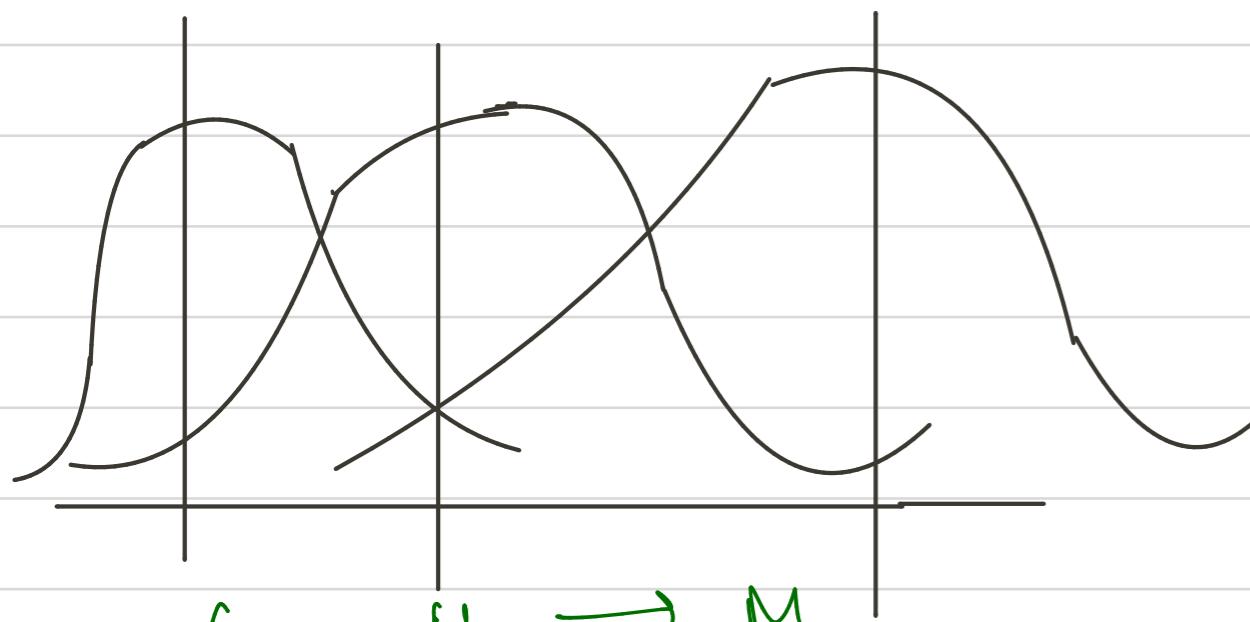
Random Variables are Independent (Default model for Random Samples)

Source: The power of Statistics, Coursera Course -

Expected Values:

$$E[X] = \sum_x x p(x).$$

- The process of making conclusions about populations from noisy data that was drawn from it.
- The most useful characterization are so called expected value.
- Mean is a characterization of its center.



Central mass

→ The variance and standard deviation are characterizations of how spread out it is

→ Our sample expected values (the sample mean and variance) will estimate the population versions

→ Expected value or mean of a random variable is the center of its distribution.

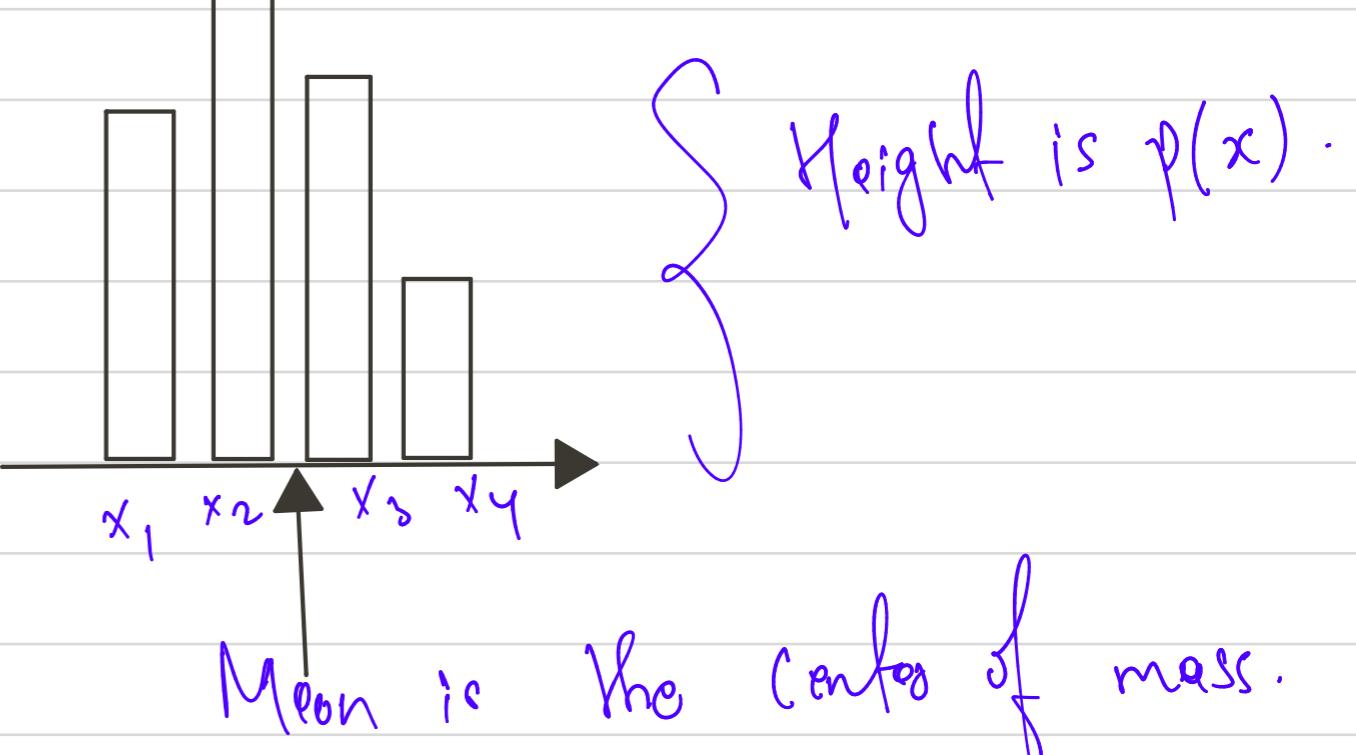
$E[X]$ represents the center of mass of a collection of locations and weights, $\{x_i p(x)\}$.

$$E[X] = \sum_x x p(x).$$

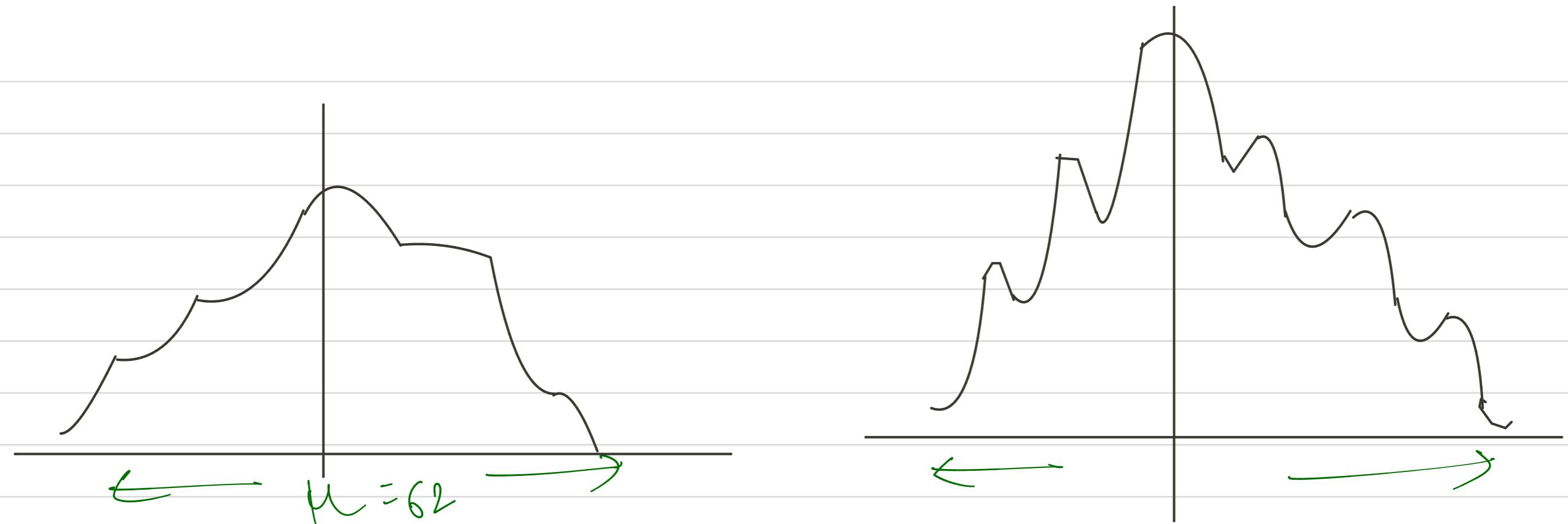
Sample Mean estimates the population mean.

$$\bar{X} = \sum_{i=1}^n x_i p(x_i)$$

which is the center of mass of the data is the empirical mean.



Mean is the center of mass.



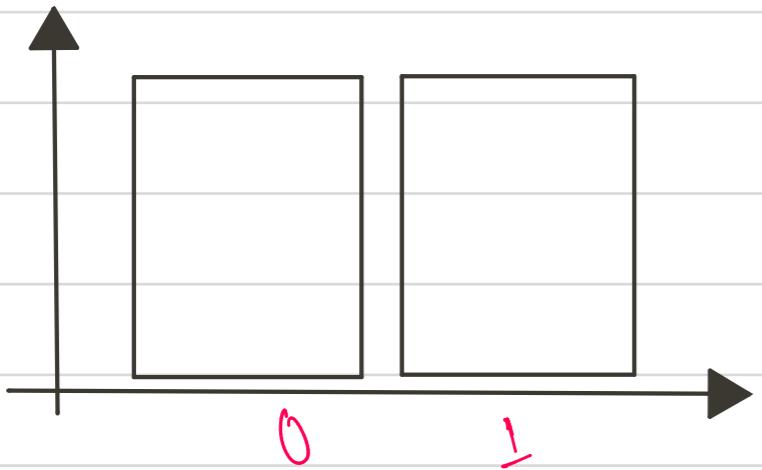
Suppose. $MSE = 43.403$

\rightarrow MSE helps to balance the distribution as $\uparrow \mu$ and $\downarrow MSE$.

Example of a Population Mean.

Suppose a coin is flipped and X declared 0 or 1 corresponding to a head or a tail. What is the expected value of X ? So, again the expected value is the Property of the Population.

$$E[X] = .5 \times 0 + .5 \times 1 \Rightarrow 5$$



Q. What about a biased coin?

\Rightarrow Suppose that a random variable X_1 is so that $P(X=1) = p$ and $P(X=0) = (1-p)$

$$E[X] = 0 \cdot (1-p) + 1 \cdot p \\ = p$$

Q. What about a dice? Suppose that a dice is rolled, and X is the number that is face up. What is the expected value of X ?

$$E[X] \Rightarrow 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} \\ \Rightarrow 3.5$$

Expanded Version of Bayes theorem

$$\therefore p(A|B) = \frac{p(B|A) * p(A)}{p(B|A) * p(A) + p(B|A^c) * p(A^c)}$$

- This version of Bayes Theorem are used for quality control and testing diagnostic medical tests.
- False positives and false Negatives are the Errors Account for the probability test.

False Positive: Test Result that indicates something is present when it really is not. Pm email as email, Software testing & medical testing.

False Negative: Test result that indicates Something is not present when it really is. Spam email as ham email, applied to quality control tests and other kinds of tests.

Evaluate the accuracy of a Diagnostic test:

- % of the population has the medical condition
- If a person has the condition, there's a gr. chance that the test is positive.
- If a person does NOT have the condition, there is still a 2% chance that the test is positive.

Some have allegy after the test means updating the posterior probability.

• Prior probability = the probability that a person has the medical condition.

• Posterior probability = the probability that the condition is present GIVEN that the test is positive.

• Event A = actually having the medical condition

• Event B = testing positive

$$\bullet P(A) = 1\% \quad (\text{allergy})$$

$$\bullet P(B|A) = 95\% \quad (\text{person having allergy tested positive})$$

$$\bullet P(B|\text{not } A) = 2\% \quad (\text{person not having allergy tested negative})$$

Having Not Allergy

$$P(A') = 1 - p(A)$$

$$= 1 - 0.01$$

$$= 0.99$$

$$= 99\%$$

Bayes' theorem (Basic Version)

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

Alternatively we can use expanded version if not calculating $P(B)$.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}$$

$$\Rightarrow \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.02 \times 0.99} \Rightarrow 0.324.$$

$P(A|B) \Rightarrow 32.4\%$ means probability of having Allergy given that the test is positive is 32.4% .

Day-74, Feb 12, 2025 (Magh 30, 2082 BS)

- Application of Multi-variate Gaussian Distribution and
- Conditional Probability in establishing the Perception and Bayes Classifier Under Gaussian Distribution or Environment.

Relation between the Perception and Bayes Classifier.

Before we move,

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

σ = Standard deviation and μ = mean $f(x) = \text{PDF.}$

Multivariate Gaussian Distribution formula,

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

Where $|\Sigma|^{1/2}$ is a real number

$(x - \mu)^T$ is a $[n \times 1]$ or $[m \times 1]$ dimensional.

$$X \sim N(\mu, \Sigma)$$

μ (mean vector) = A column vector representing the mean of each variable

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Σ (Covariance Matrix):

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots \\ \sigma_{21} & \sigma_{22} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \cdots & \sigma_{nn} \end{bmatrix}$$

So, $|\Sigma| \Rightarrow$ determinant of the Covariance matrix

Σ^{-1} = inverse of the Covariance matrix.

Bayes classifier:

Here we minimize the average risk, denoted by R .

Suppose we have φ_1 and φ_2 then using Conditional probability

Concept first -

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \rightarrow \text{eqn } \textcircled{I}$$

where $P(A)$ and $P(B)$ are known or Prior Probability.

So, let $P(A) \in \mathcal{Q}_1$ [sample \mathcal{H}_1]

and $P(B) \in \mathcal{Q}_2$ [sample \mathcal{H}_2]

Then, for \mathcal{Q}_1 and \mathcal{Q}_2 using Conditional Probability,

$$P(\mathcal{E}_1|x) \Rightarrow \frac{P(x|\mathcal{E}_1) \times P(\mathcal{E}_1)}{P(x)} \rightarrow \text{eqn } \textcircled{II}$$

$$P(\mathcal{E}_2|x) = \frac{P(x|\mathcal{E}_2) \times P(\mathcal{E}_2)}{P(x)} \rightarrow \text{eqn } \textcircled{II}$$

Combining eqn ① and eqn ② we get -

$$P(\ell_1 | x) = \frac{P(x | \ell_1) \times P(\ell_1)}{P(x)}$$

$$P(\ell_2 | x) = \frac{P(x | \ell_2) \times P(\ell_2)}{P(x)}$$

Now for Ideal Use $P(\ell_1) = P(\ell_2)$ means the classes ℓ_1 and ℓ_2 equal in size we get -

$$\frac{P(\ell_1 | x)}{P(\ell_2 | x)} = \frac{P(x | \ell_1)}{P(x | \ell_2)} \rightarrow \text{eqn III}$$

$$P(\ell_1 | x) \\ P(\ell_2 | x)$$

$$= \frac{P(x | \ell_1)}{P(x | \ell_2)}$$

Now using these two terms of the distribution can be derived since -

$$f = f_1 + f_2$$

$$f = 1$$

Because the area under the curve or the bell is always 1. (Rejection Error -)



For two-class problem, represented by classes ℓ_1 and ℓ_2 , the average risk is defined by

$$R = C_{11}P_1 \int_{\mathcal{H}_1} P_x(x|\ell_1) \cdot dx + C_{22}P_2 \int_{\mathcal{H}_2} P_x(x|\ell_2) \cdot dx$$

$$+ C_{21}P_1 \int_{\mathcal{H}_2} P_x(x|\ell_1) \cdot dx + C_{12}P_2 \int_{\mathcal{H}_1} P_x(x|\ell_2) \cdot dx$$

Explaining

$C_{11} \rightarrow$ True Class ℓ_j where j is the actual class and i is the predicted/classified class

$C_{21} \rightarrow$ Incorrect because class ℓ_1 is classified as class ℓ_2

$P_1 | P_2$ or $P_i \Rightarrow$ Prior Probability so here P_1 and P_2 denoted
can be taken as P_L belongs to class ℓ_L .

Our task is to minimize R so, $H - H_1 = H_2$

$$R = C_{11} P_L \int_{H_1} p_x(x | \ell_1) \cdot dx + C_{22} P_2 \int_{H-H_1} p_x(x | \ell_2) \cdot dx \\ + C_{21} P_1 \int_{H-H_1} p_x(x | \ell_1) \cdot x + C_{12} P_2 \int_{H_1} p_x(x | \ell_2) \cdot dx$$

$$R = C_{11} P_1 \int_{H_1} p_x(x | \ell_1) \cdot dx + C_{22} P_2 \int_{H_1} p_x(x | \ell_2) \cdot dx - C_{22} P_2 \int_{H_1} p_x(x | \ell_2) \cdot dx \\ + C_{12} P_2 \int_{H_1} p_x(x | \ell_2) \cdot dx$$

+ ~~$C_{22} P_2 \int_{H_1} p_x(x | \ell_2) \cdot dx$~~ becomes δ .

Now we know that -

$$\int_{\mathcal{H}} p_x(x | \ell_1) \cdot dx = \int_{\mathcal{H}} p_x(x | \ell_2) \cdot dx = 1$$

Because $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$

$$[\mathcal{H} = 1]$$

whose $C_{11} < C_{21}$ and $C_{22} < C_{12}$ we observed

$$\int_{\mathcal{H}} p_x(x | \ell_1) \cdot dx = \int_{\mathcal{H}} p_x(x | \ell_2) \cdot dx$$

Because Correct classification C_{11} risk factor < Misclassified C_{21} risk factor

So

$$R = C_{11}P_1 \int_{\mathcal{H}_1} P_X(x | \ell_1) \cdot dx + C_{22}P_2 \int_{\mathcal{H}-\mathcal{H}_1} P_X(x | \ell_2) \cdot dx + C_{21}P_1 \cdot \int_{\mathcal{H}-\mathcal{H}_1} P_X(x | \ell_1) x \\ + C_{12}P_2 \int_{\mathcal{H}_1} P_X(x | \ell_2) \cdot dx \quad \text{--- eqn A}$$

Eqn A Reduces to

$$R = (C_{21}P_1 + C_{22}P_2) + \int_{\mathcal{H}_1} P_2 C_{12} - P_2 C_{22} \quad P_X(x | \ell_2) - \\ P_1 C_{21} - P_1 C_{11} \quad P_X(x | \ell_1) \cdot dx$$

Fixed cost.

$$R = C_{21}P_1 + C_{22}P_2 + \int_{\mathcal{H}_1} [P_2(C_{12} - C_{22})P_X(\bar{x})\ell_2) - P_1(C_{21} - C_{11}) \cdot P_X(x|\ell_1)]dx$$

↗ ↗
 fixed. dynamic.

→ -ve assigned to class ℓ_1
 → +ve vector excluded from
 $\mathcal{H}_1(\ell_2)$ and it contribute to R .

So, More and less if values X for which the Integrand = 0 have no effect on the average risk R .

if Condition $P_1(C_{21} - C_{11})P_X(x|\ell_1) > P_2(C_{12} - C_{22})P_X(x|\ell_2)$
 holds, assign the observation vector X to Subspace \mathcal{H}_1 (class ℓ_1).

else:

assign X to ℓ_2 (i.e. ℓ_2)

Defining, $N(x) = \frac{p_x(x|\ell_1)}{p_x(x|\ell_2)}$

and $\ell_1 = \frac{p_2(c_{12} - c_{22})}{p_1(c_{21} - c_{11})}$

Given Above Equation

$$p_1(c_{21} - c_{11})p_x(x|\ell_1) > p_2(c_{12} - c_{22})p_x(x|\ell_2)$$

$$\frac{P_1(C_{21} - C_{11})}{P_2(C_{12} - C_{22})} \leq \frac{P_X(x | \ell_2)}{P_X(x | \ell_1)}$$

that means

$$\left(\frac{P_1(C_{21} - C_{11})}{P_2(C_{12} - C_{22})} \right) = \left(\frac{P_X(x | \ell_2)}{P_X(x | \ell_1)} \right)$$

Above quantity $\Lambda(x)$ Ratio of two Conditional Probability Density Functions we call it likelihood Ratios -

Quantity ϵ is the threshold of the test . Both $\Lambda(x)$ and ℓ_i are always positive -

If for an observation vector x , the likelihood ratio $\lambda(x) \geq \epsilon$
belongs to ℓ_1

else

belongs to ℓ_2

Let's make the $\lambda(x)$ binary or linear - Simply use deg

$$\ln \lambda(x) = \ln \frac{P_x(x|\ell_1)}{P_x(x|\ell_2)}$$

Introduces

Some as

$$\ln(\epsilon) = \ln \frac{P_2(c_{12} - c_{22})}{P_1(c_{21} - c_{11})}$$

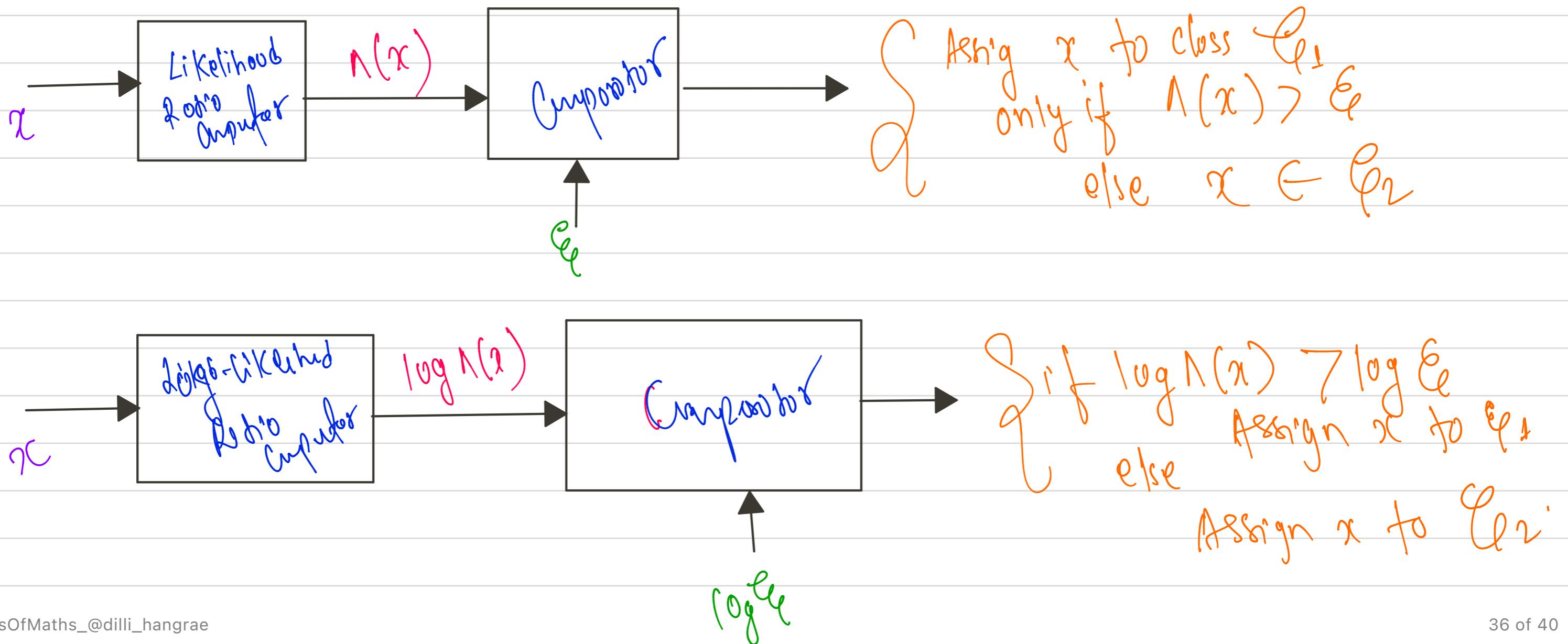
Introduct-

So,

$$\begin{cases} \ln l(x) > \ln (\epsilon) \\ \ln l(x) < \ln (\epsilon) \end{cases} \Rightarrow \begin{cases} x \text{ belongs to } C_1 \\ x \in C_2 \end{cases}$$

→ Now
Decision
Rule

which is equivalent to this visualization -



Bayes classifier for a Gaussian Distribution

$$p_x(x|\theta_i) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2} (x - \mu_i)^T \Sigma^{-1} (x - \mu_i)}, \quad i=1, 2, 3, \dots$$

whose m is the dimensionality of observation vector x .

Now,

$$p_x(x|\theta_1) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1)}$$

$$p_x(x|\theta_2) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2} (x - \mu_2)^T \Sigma^{-1} (x - \mu_2)}$$

The given image contains the probability density function of a **multivariate Gaussian distribution**, specifically:

$$p_X(x|c_i) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)}$$

for classes c_1 and c_2 . The likelihood ratio test involves computing:

$$\Lambda(x) = \frac{p_X(x|c_1)}{p_X(x|c_2)}$$

Step-by-Step Solution:

1. Write the Ratio Explicitly:

$$\Lambda(x) = \frac{\frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}}{\frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)}}$$

2. Cancel Out Common Terms:

- The denominator $(2\pi)^{m/2}|\Sigma|^{1/2}$ cancels out.

$$\Lambda(x) = e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)} \cdot e^{\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)}$$

3. Use Exponent Rules:

- Since $e^a/e^b = e^{a-b}$,

$$\Lambda(x) = e^{\frac{1}{2}[(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) - (x-\mu_1)^T \Sigma^{-1}(x-\mu_1)]}$$

4. Take the Logarithm (Log-Likelihood Ratio Test):

$$\ln \Lambda(x) = \frac{1}{2} [(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) - (x-\mu_1)^T \Sigma^{-1}(x-\mu_1)]$$

Decision Rule:

- If $\ln \Lambda(x) > \ln \eta$, assign x to class c_1 .
- Otherwise, assign x to class c_2 .

This is the quadratic discriminant function used in **Gaussian classification**. Let me know if you

This image explains how the **log-likelihood ratio test** for a Gaussian Bayes classifier results in a **linear decision boundary**. Let's break it down step by step.

1. Log-Likelihood Ratio Expansion

We already derived that:

$$\ln \Lambda(x) = \frac{1}{2} [(x - \mu_2)^T \Sigma^{-1}(x - \mu_2) - (x - \mu_1)^T \Sigma^{-1}(x - \mu_1)]$$

Expanding this further,

$$\ln \Lambda(x) = -\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_2)^T \Sigma^{-1}(x - \mu_2)$$

Rearranging terms,

$$\ln \Lambda(x) = (\mu_1 - \mu_2)^T \Sigma^{-1}x + \frac{1}{2}(\mu_2^T \Sigma^{-1}\mu_2 - \mu_1^T \Sigma^{-1}\mu_1)$$

2. Interpretation as a Linear Classifier

Comparing with a linear function:

$$y = w^T x + b$$

where:

- $y = \ln \Lambda(x)$ (the log-likelihood ratio)
- $w = \Sigma^{-1}(\mu_1 - \mu_2)$ (the weight vector)
- $b = \frac{1}{2}(\mu_2^T \Sigma^{-1}\mu_2 - \mu_1^T \Sigma^{-1}\mu_1)$ (the bias term)

This shows that the **Gaussian Bayes classifier** results in a **linear decision boundary**, meaning it acts like a perceptron but is derived from probabilistic principles.

2. Interpretation as a Linear Classifier

Comparing with a linear function:

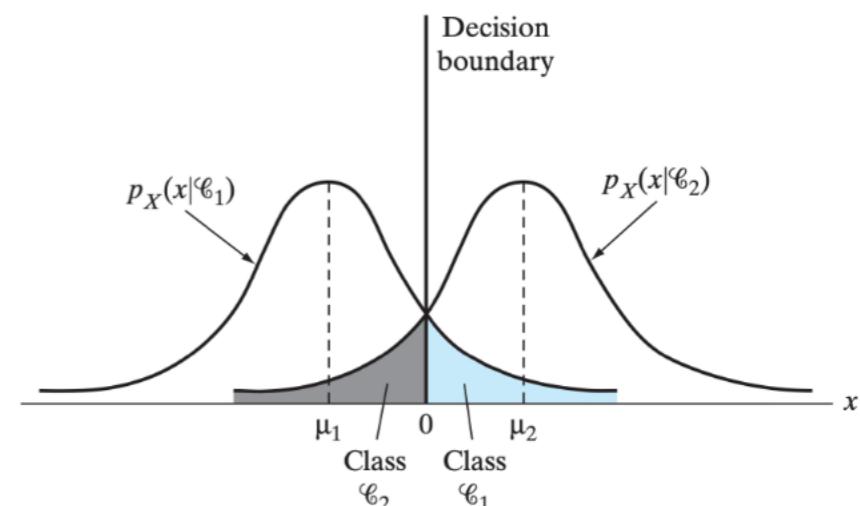
$$y = w^T x + b$$

where:

- $y = \ln \Lambda(x)$ (the log-likelihood ratio)
- $w = \Sigma^{-1}(\mu_1 - \mu_2)$ (the weight vector)
- $b = \frac{1}{2}(\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} \mu_1)$ (the bias term)

This shows that the **Gaussian Bayes classifier** results in a **linear decision boundary**, meaning it acts like a perceptron but is derived from probabilistic principles.

FIGURE 1.7 Two overlapping, one-dimensional Gaussian distributions.



3. Decision Rule

Based on the log-likelihood test:

- If $y = w^T x + b > 0$, assign x to class c_1 .
- Otherwise, assign x to class c_2 .

This means the classifier separates the feature space with a hyperplane defined by $w^T x + b = 0$.

4. Summary

- The Bayes classifier for two Gaussian classes is a **linear classifier**.
- It results in a decision boundary of the form $w^T x + b = 0$.
- The weight vector w depends on the covariance matrix Σ and the mean vectors μ_1, μ_2 .
- The bias term b is derived from the means and covariance as well.

Would you like a worked-out numerical example to clarify this further? 🚀

Source! (SC 372 Neural Networks & Machine Learning)

Neural Networks & Machine Learning Book

Simon Holkin 2009 Textbook

This section describes the **Bayes Classifier** for a Gaussian Distribution, specifically for a two-class problem where:

- The feature vector X follows a **Gaussian (Normal)** distribution.
- The two classes, \mathcal{C}_1 and \mathcal{C}_2 , have different **mean vectors** μ_1 and μ_2 .
- The **covariance matrix** C is **identical** for both classes.

Breaking it Down

1. Class-Specific Means:

- If X belongs to class \mathcal{C}_1 , it has a mean μ_1 .
- If X belongs to class \mathcal{C}_2 , it has a mean μ_2 .

Mathematically:

$$E[X|\mathcal{C}_1] = \mu_1, \quad E[X|\mathcal{C}_2] = \mu_2$$

2. Shared Covariance Matrix:

- The covariance matrix for both classes is **identical**.
- This means that the data distribution for both classes has the **same shape**, but the **means are different**.
- The covariance is defined as:

$$E[(X - \mu_1)(X - \mu_1)^T] = C$$

$$E[(X - \mu_2)(X - \mu_2)^T] = C$$

This assumption simplifies the **Bayes classifier**, leading to a **linear decision boundary**.

3. What Does It Mean?

- Since C is the same for both classes, the likelihood ratio test will depend **only on the means** μ_1 and μ_2 .
- The decision boundary will be **linear** (a straight line in 2D, a plane in 3D, or a hyperplane in higher dimensions).
- If C were different for the two classes, the decision boundary would be **quadratic** instead of linear.

- This means that the data distribution for both classes has the **same shape**, but the **means are different**.

- The covariance is defined as:

$$E[(X - \mu_1)(X - \mu_1)^T] = C$$

$$E[(X - \mu_2)(X - \mu_2)^T] = C$$

This assumption simplifies the **Bayes classifier**, leading to a **linear decision boundary**.

3. What Does It Mean?

- Since C is the same for both classes, the likelihood ratio test will depend **only on the means** μ_1 and μ_2 .
- The decision boundary will be **linear** (a straight line in 2D, a plane in 3D, or a hyperplane in higher dimensions).
- If C were different for the two classes, the decision boundary would be **quadratic** instead of linear.

4. Assumption About Covariance:

- The covariance matrix C is **nondiagonal**, meaning features are **correlated**.
- C is **nonsingular**, meaning it has an inverse C^{-1} that exists, which is crucial for computing the decision rule.

What's Next?

- Using this setup, we can write the **Gaussian likelihood function** for each class.
- Taking the **log-likelihood ratio** gives a **linear decision rule**.
- This leads to the final classification rule:

$$w^T x + b \geq 0$$

where $w = C^{-1}(\mu_1 - \mu_2)$ and b is a bias term.