Machine Learning

Angshuman Paul

Assistant Professor

Department of Computer Science & Engineering

Bayes Decision Theory

Joint Probability Distribution

- Consider two random variables
- X corresponding to weather $\{sunny, rainy, cloudy\}$
  - $P(X) = \{0.6, 0.1, 0.3\}$
- Y corresponding to power cut $\{power\ cut, no\ power\ cut\}$
  - $P(Y) = \{0.15, 0.85\}$
- A joint probability distribution of $X\ and\ Y$
  - Probability distribution on all possible pairs of outputs

Joint Probability Distribution

- X corresponding to weather $\{sunny, rainy, cloudy\}$
  - $P(X) = \{0.6, 0.1, 0.3\}$
- Y corresponding to power cut $\{power\ cut, no\ power\ cut\}$
  - $P(Y) = \{0.15, 0.85\}$
- A joint probability distribution of $X\ and\ Y$
  - Probability distribution on all possible pairs of outputs
- A 3×2 matrix of values

Joint Probability Distribution

- X corresponding to weather $\{sunny, rainy, cloudy\}$
  - $P(X) = \{0.6, 0.1, 0.3\}$
- Y corresponding to power cut $\{power\ cut, no\ power\ cut\}$
  - $P(Y) = \{0.15, 0.85\}$
- A joint probability distribution of $X\ and\ Y$
  - Probability distribution on all possible pairs of outputs
- A 3×2 matrix of values

	Power cut	**No power cut**
Sunny	0.01	0.4
Rainy	0.2	0.1
Cloudy	0.09	0.2

Joint Probability Distribution

- Sample space corresponding to X is $S_X = \{s, r, c\}$
- Sample space corresponding to Y is $S_Y = \{pc, npc\}$
- Sample space corresponding to the joint distribution is

	Power cut (pc)	**No power cut (npc)**
Sunny (s)	0.01 $P(s \cap pc)$	0.4 $P(s \cap npc)$
Rainy (r)	0.2 $P(r \cap pc)$	0.1 $P(r \cap npc)$
Cloudy (c)	0.09 $P(c \cap pc)$	0.2 $P(r \cap npc)$

Joint Probability Distribution

- Sample space corresponding to X is $S_X = \{s, r, c\}$
- Sample space corresponding to Y is $S_Y = \{pc, npc\}$
- Sample space corresponding to the joint distribution is

$$S_J = \{(s, pc), (s, npc), (r, pc), (r, npc), (c, pc), (c, npc)\}$$

	Power cut (pc)	No power cut (npc)
Sunny (s)	0.01 $P(s \cap pc)$	0.4 $P(s \cap npc)$
Rainy (r)	0.2 $P(r \cap pc)$	0.1 $P(r \cap npc)$
Cloudy (c)	0.09 $P(c \cap pc)$	0.2 $P(r \cap npc)$

Chain Rule

- If $A_1, A_2, \ldots, A_n$ are n events, then
  - $P(A_n \cap A_{n-1} \cap \cdots \cap A_1) = P(A_n | A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} \cap \cdots \cap A_1)$ (1)
- Similarly,
  - $P(A_{n-1} \cap A_{n-2} \cap \cdots \cap A_1) = P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1)\, P(A_{n-2} \cap \cdots \cap A_1)$ (2)
- Extending this for the subsequent events and putting in (1), we get,
  - $P(A_n \cap A_{n-1} \cap \cdots \cap A_1)$

$$= P\,(A_n | A_{n-1} \cap \cdots \cap A_1)\, P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) P(A_{n-2} | A_{n-3} \cap \cdots \cap A_1) \ldots P(A_1)$$

Chain Rule

- If A_1, A_2, A_3 are 3 events, then we use
  - $P(A_n \cap A_{n-1} \cap \cdots \cap A_1)$

    $= P\ (A_n | A_{n-1} \cap \cdots \cap A_1)\ P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) P(A_{n-2} | A_{n-3} \cap \cdots \cap A_1) \ldots P(A_1)$

- We get
  - $P(A_4 \cap A_3 \cap A_2 \cap A_1)$

    $= P\ (A_4 | A_3 \cap A_2 \cap A_1)\ P(A_3 | A_2 \cap A_1) P(A_2 | A_1) P(A_1)$

Inference by Enumeration

	fever		¬ fever	
	cough	¬ cough	cough	¬ cough
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- $P(fever) =$

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- For any proposition, add all the boxes where the proposition is true
- $P(fever) = 0.21 + 0.10 + 0.11 + 0.07 = 0.49$

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- For any proposition, add all the boxes where the proposition is true
- $P(fever \lor covid) =$

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- For any proposition, add all the boxes where the proposition is true
- $P(fever \vee covid) = 0.21 + 0.10 + 0.11 + 0.07 + 0.11 + 0.08 = 0.68$

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- For any proposition, add all the boxes where the proposition is true
- $P(\neg covid | \neg fever) =$

Inference by Enumeration

	fever		¬ fever	
	cough	¬ cough	cough	¬ cough
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- For any proposition, add all the boxes where the proposition is true

- $P(\neg covid|\neg fever) = \frac{P(\neg covid \cap \neg fever)}{P(\neg fever)} = \frac{0.09+0.23}{P(\neg fever)}$

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- For any proposition, add all the boxes where the proposition is true

- $P(\neg covid|\neg fever) = \frac{P(\neg covid \cap \neg fever)}{P(\neg fever)} = \frac{0.09+0.23}{0.11+0.08+0.09+0.23} \approx 0.627$

Inference by Enumeration

	fever		¬ fever	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- If we have the complete joint distribution, I can answer any related queries
- But, what is the problem with this approach?

Inference by Enumeration

	fever		**¬ fever**	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- But, what is the problem with this approach?
  - For a system with many causes and effects, we have to maintain a large set of values and operate on those

Our Old Example in a New Form

- In a box, there are 40 Samsung phones and 20 MI phones
- Out of these, 10 Samsung phones and 2 MI phones are not working.
- You pick up a phone and find that the phone is not working.

- Can you tell me whether the phone is a Samsung Phone or MI Phone?

How Will You Solve This Problem?

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).

How Will You Solve This Problem?

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).

- Find $P(SP|NW)$ and $P(MIP|NW)$

How Will You Solve This Problem?

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).

- Find $P(SP|NW)$ and $P(MIP|NW)$
- If $P(SP|NW) > P(MIP|NW)$
  - The phone that I picked up is Samsung
- Else if $P(SP|NW) < P(MIP|NW)$
  - The phone that I picked up is MI
- Else if $P(SP|NW) = P(MIP|NW)$
  - We can't take any decision

How Will You Solve This Problem?

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).

- Find $P(SP|NW)$ and $P(MIP|NW)$
- If $P(SP|NW) > P(MIP|NW)$
  - The phone that I picked up is Samsung
- Else the phone is MI

Conditional Probability

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).
- You pick up a phone and find that the phone is not working.
- What is the probability that the phone you picked up is a Samsung phone?
- Find $P(SP|NW)$

- $$P(SP|\mathrm{NW}) = \frac{\#\ Samsung\ phones\ that\ are\ not\ working}{\#\ Phones\ that\ are\ not\ working}$$
  $$= \frac{10}{12}$$
  $$= \frac{\frac{10}{40} \times \frac{40}{60}}{\frac{12}{60}}$$
- $\frac{10}{40}$: Given an SP, probability that it is NW ($P(NW|SP)$)
- $\frac{40}{60}$: Probability of SP in the box $P(SP)$
- $\frac{12}{60}$: Probability of finding a NW phone in the box $P(NW)$

Conditional Probability

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).
- You pick up a phone and find that the phone is not working.
- What is the probability that the phone you picked up is a Samsung phone?
- Find $P(SP|NW)$

- $P(SP|\mathrm{NW}) = \dfrac{\#\ \textit{Samsung phones that are not working}}{\#\ \textit{Phones that are not working}}$

  $= \dfrac{10}{12}$

  $= \dfrac{\frac{10}{40} \times \frac{40}{60}}{\frac{12}{60}}$

- $\dfrac{10}{40}$: Given an SP, probability that it is NW ($P(NW|SP)$)
- $\dfrac{40}{60}$: Probability of SP in the box $P(SP)$
- $\dfrac{12}{60}$: Probability of finding a NW phone in the box $P(NW)$

- $P(SP|\mathrm{NW}) = \dfrac{P(NW|SP)\,P(SP)}{P(NW)}$

Similarly

- In a box, there are 40 Samsung phones (SP) and 20 MI phones (MIP)
- Out of these, 10 Samsung phones and 2 MI phones are not working (NW).
- You pick up a phone and find that the phone is not working.
- What is the probability that the phone you picked up is a Samsung phone?
- Find $P(SP|NW)$

- $P(MI|\text{NW}) = \dfrac{\#\ MI\ phones\ that\ are\ not\ working}{\#\ Phones\ that\ are\ not\ working}$
- $P(MI|\text{NW}) = \dfrac{2}{12}$

The Decision

- $P(SP|\mathrm{NW}) = \frac{10}{12}$
- $P(MI|\mathrm{NW}) = \frac{2}{12}$

Since $P(SP|\mathrm{NW}) > P(MI|\mathrm{NW})$, I conclude that I picked up Samsung phone

A Closer Look

- Consider a two-class classification problem with classes cl_1 and cl_2
  - For example, let
    - $cl_1: cat$
    - $cl_2: tiger$
- Suppose, I have a data point with a feature x that I need to classify to one of these two classes
  - Let x be the weight of the animal
- To use Bayes decision rule, I will find out $P(cl_1|x)\ and\ P(cl_2|x)$

A Closer Look

- $P(cl_1|x) > P(cl_2|x) \Rightarrow cl_1$

- Now let's use Bayes theorem

- $P(cl_1|x) = \frac{P(x|cl_1)P(cl_1)}{P(x)}$ $\qquad$ $P(cl_2|x) = \frac{P(x|cl_2)P(cl_2)}{P(x)}$

A Closer Look

- $P(cl_1|x) > P(cl_2|x) \Rightarrow cl_1$

- Now let's use Bayes theorem

Class conditional probability

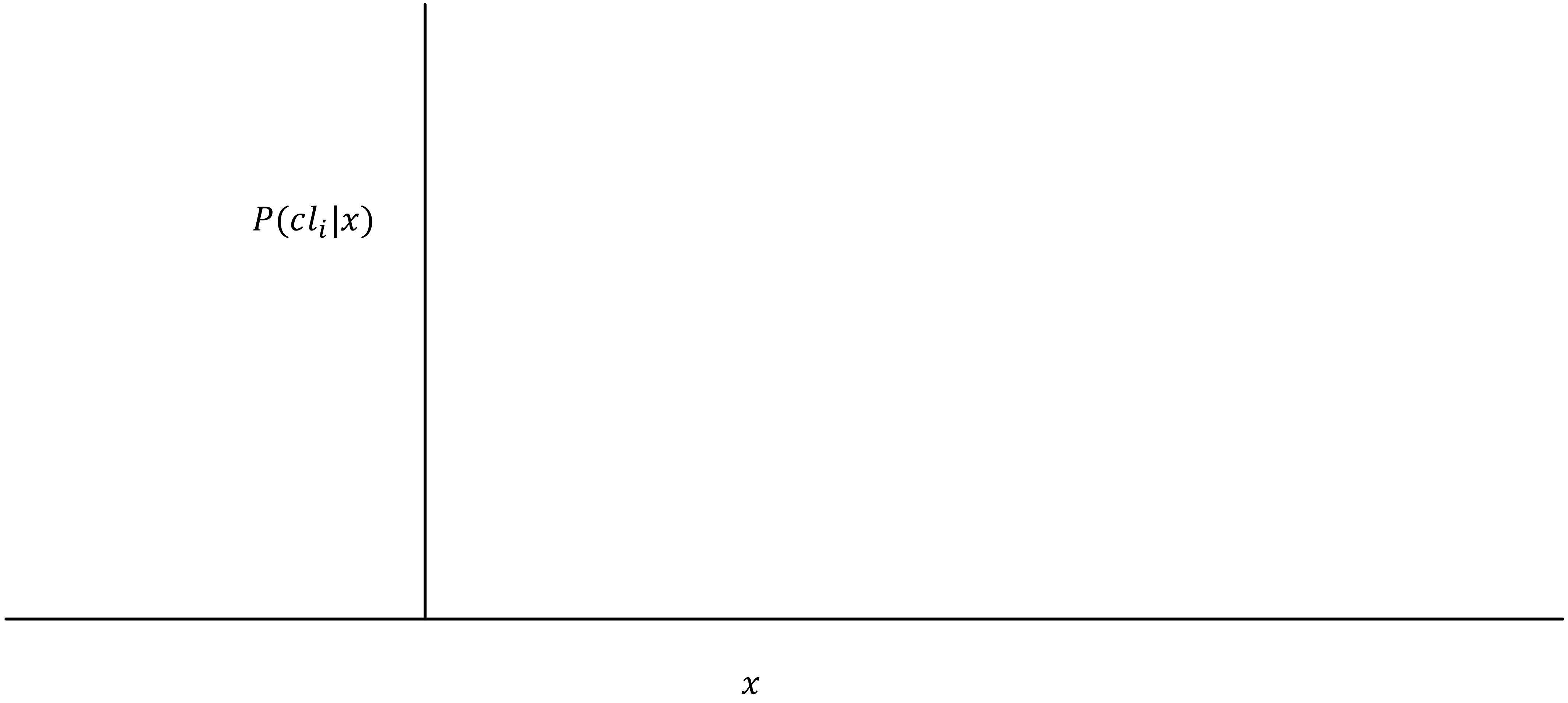
- $P(cl_1|x) = \dfrac{P(x|cl_1)P(cl_1)}{P(x)}$ $\qquad$ $P(cl_2|x) = \dfrac{P(x|cl_2)P(cl_2)}{P(x)}$

Prior probability
(my belief about the
existence of the
particular class)

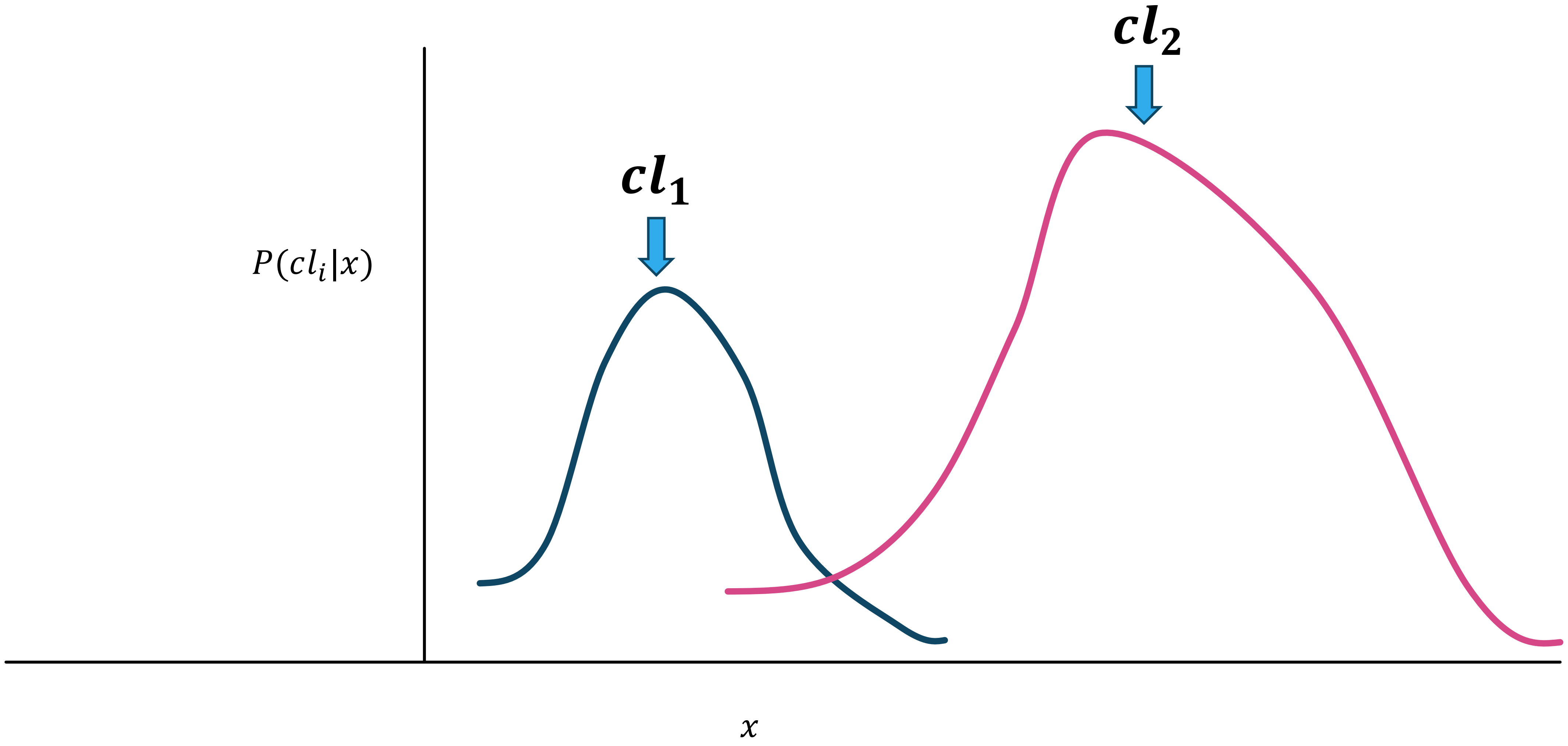
A Closer Look

- $P(cl_1|x) > P(cl_2|x) \Rightarrow cl_1$
- Now let's use Bayes theorem
  - $P(cl_1|x) = \frac{P(x|cl_1)P(cl_1)}{P(x)}$ $\qquad P(cl_2|x) = \frac{P(x|cl_2)P(cl_2)}{P(x)}$
- $P(x|cl_1)P(cl_1) > P(x|cl_2)P(cl_2) \Rightarrow cl_1$

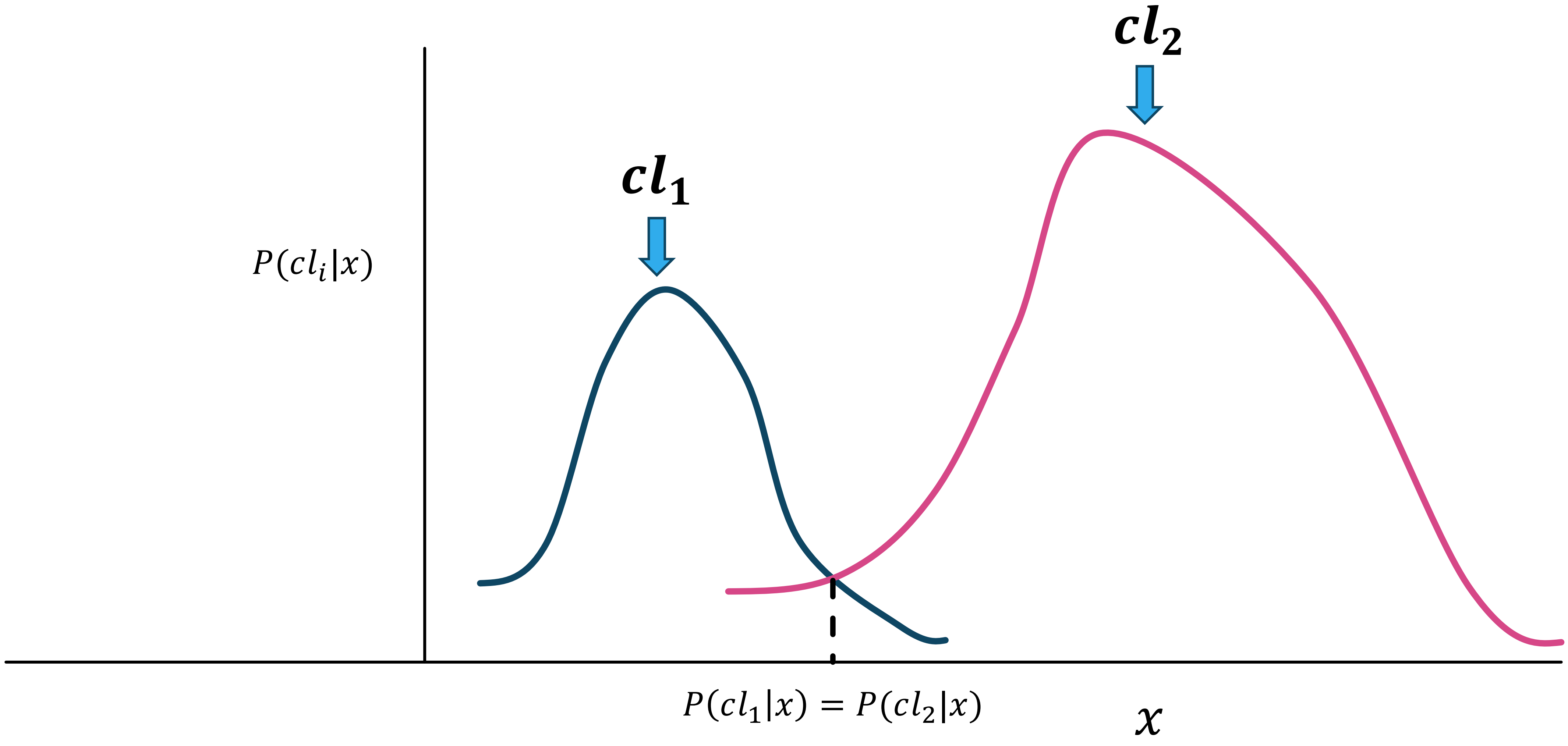
A Closer Look



A Closer Look



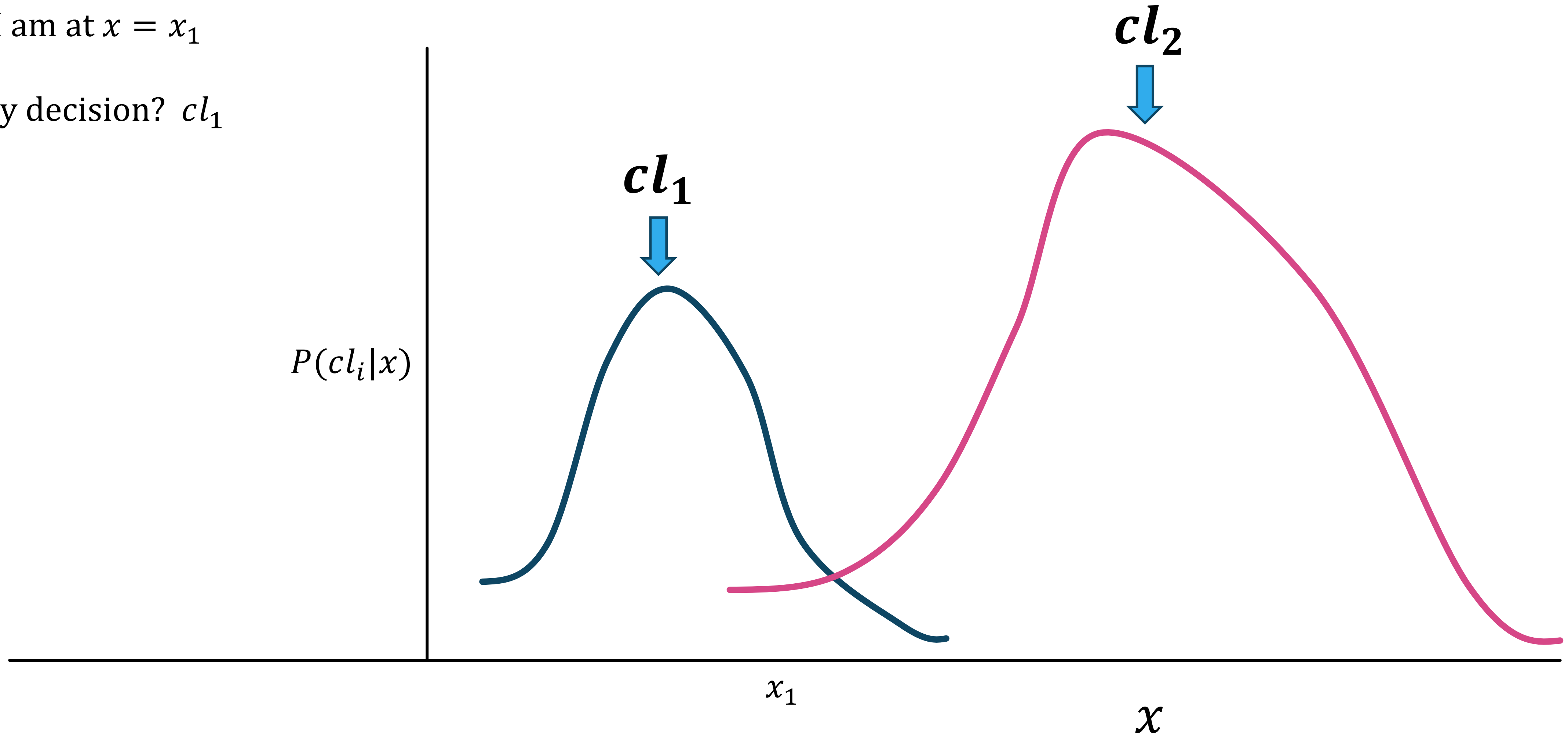
A Closer Look



Error

Suppose, I am at $x = x_1$

What is my decision? cl_1

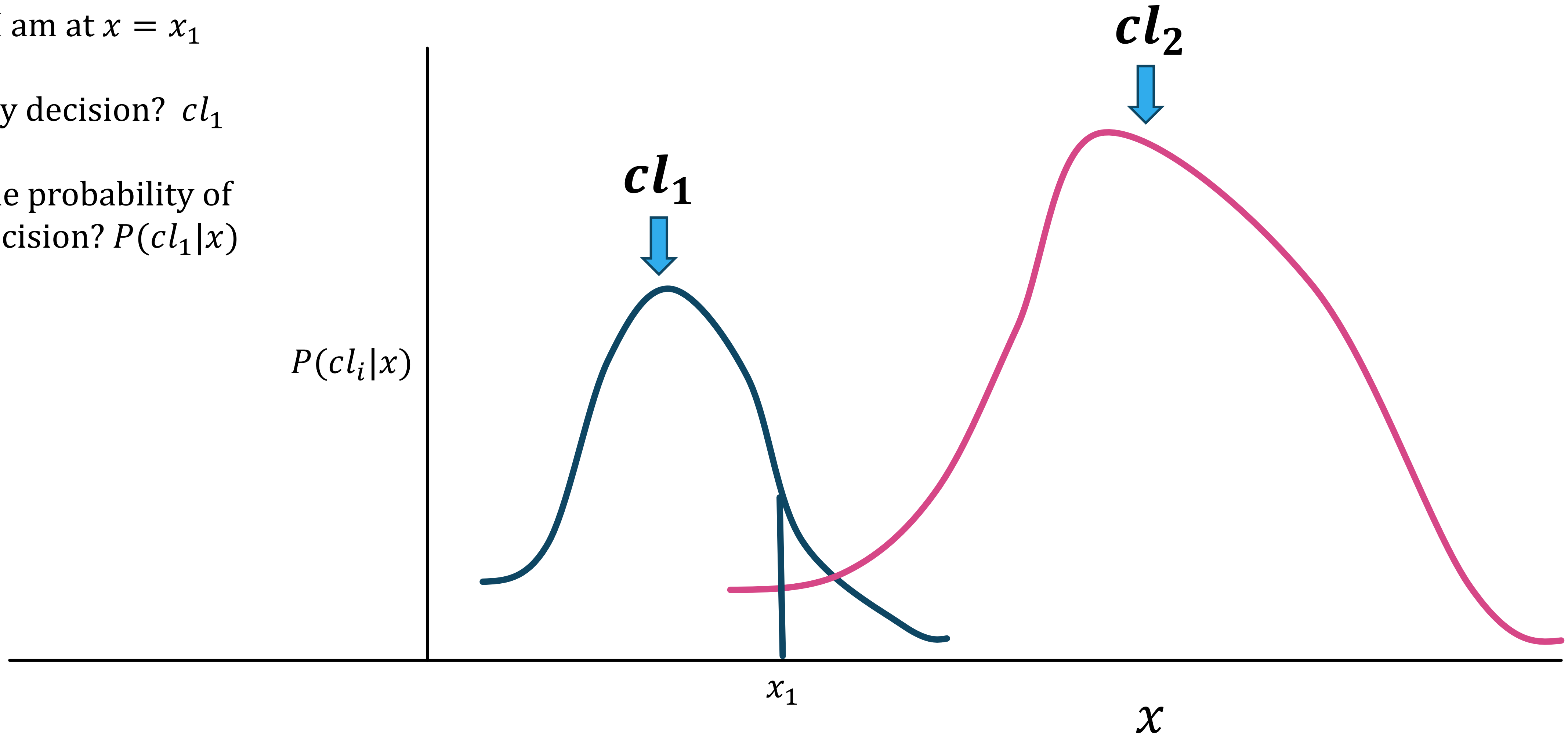


Error

Suppose, I am at $x = x_1$

What is my decision? cl_1

What is the probability of correct decision? $P(cl_1|x)$



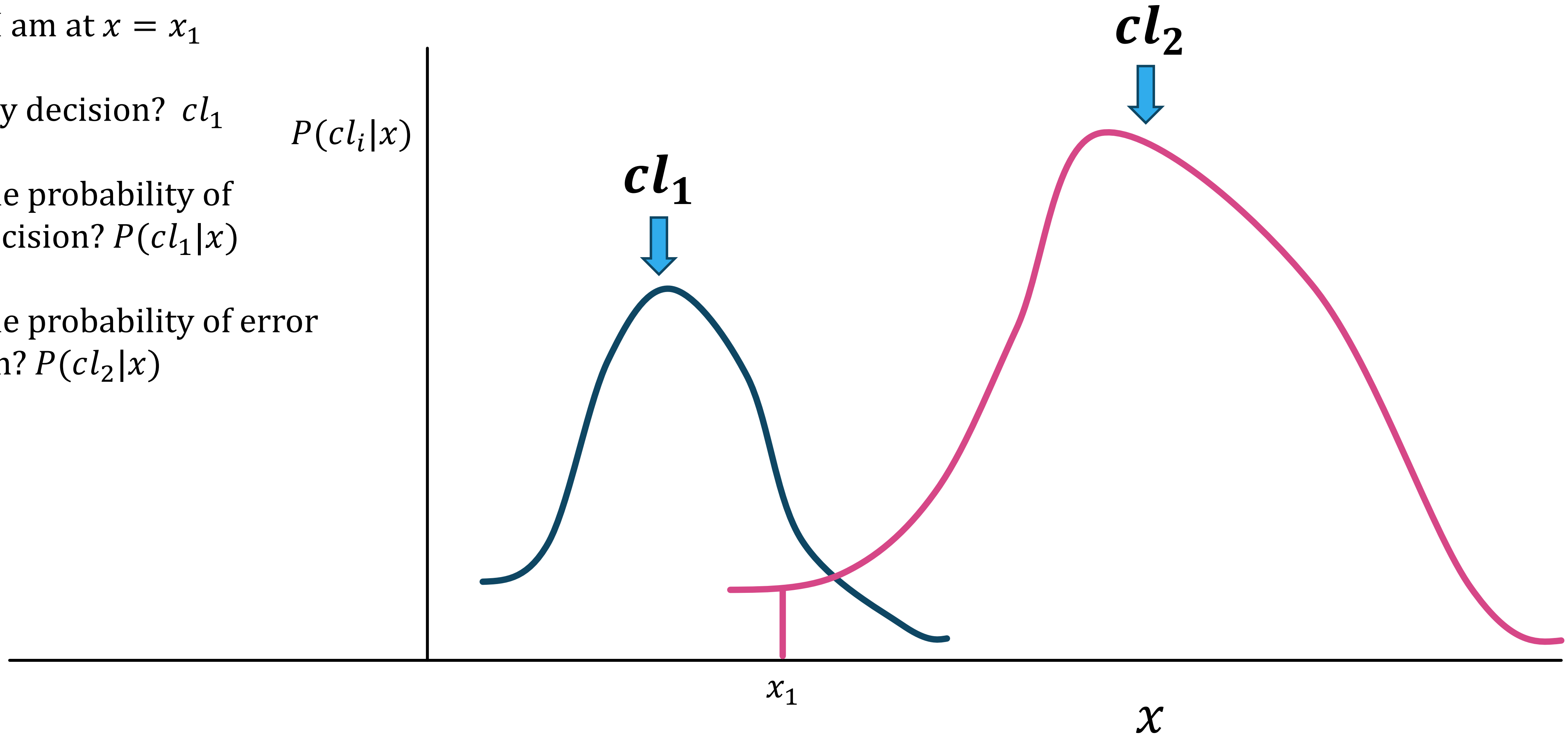
Error

Suppose, I am at $x = x_1$

What is my decision? cl_1

What is the probability of correct decision? $P(cl_1|x)$

What is the probability of error in decision? $P(cl_2|x)$



Error

Suppose, I am at $x = x_1$

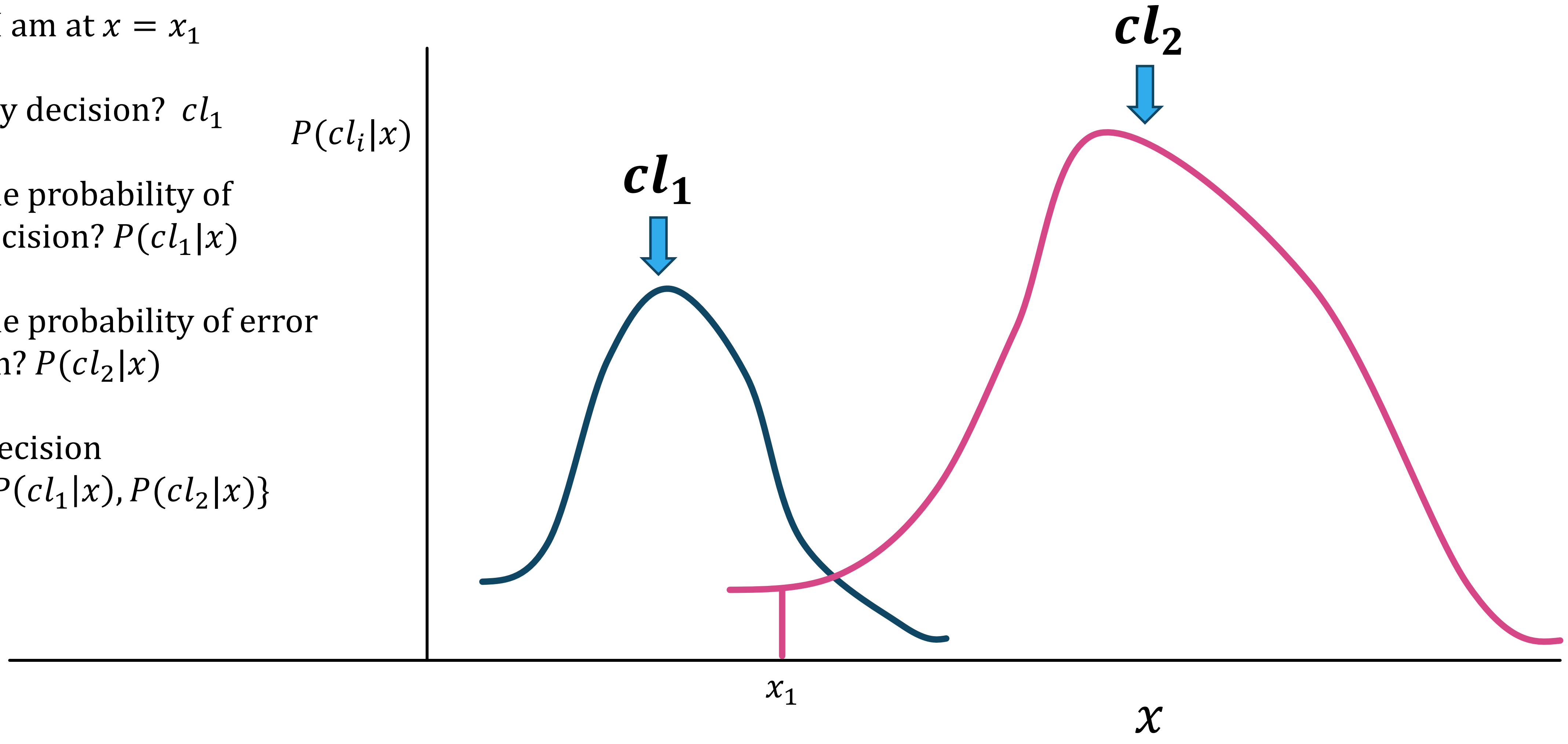
What is my decision? cl_1

What is the probability of correct decision? $P(cl_1|x)$

What is the probability of error in decision? $P(cl_2|x)$

Error in decision

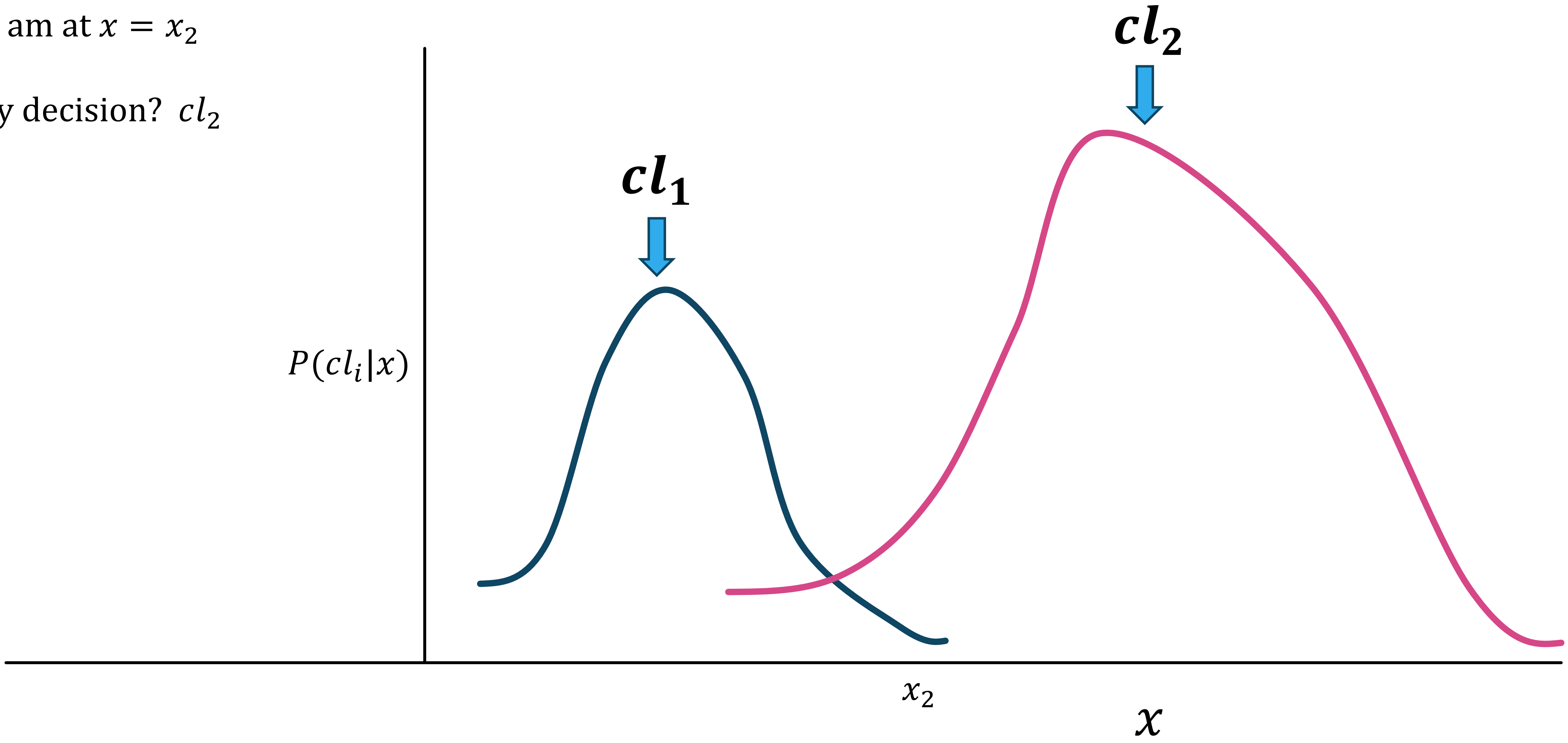
$\min\{P(cl_1|x), P(cl_2|x)\}$



Error

Suppose, I am at $x = x_2$

What is my decision? cl_2

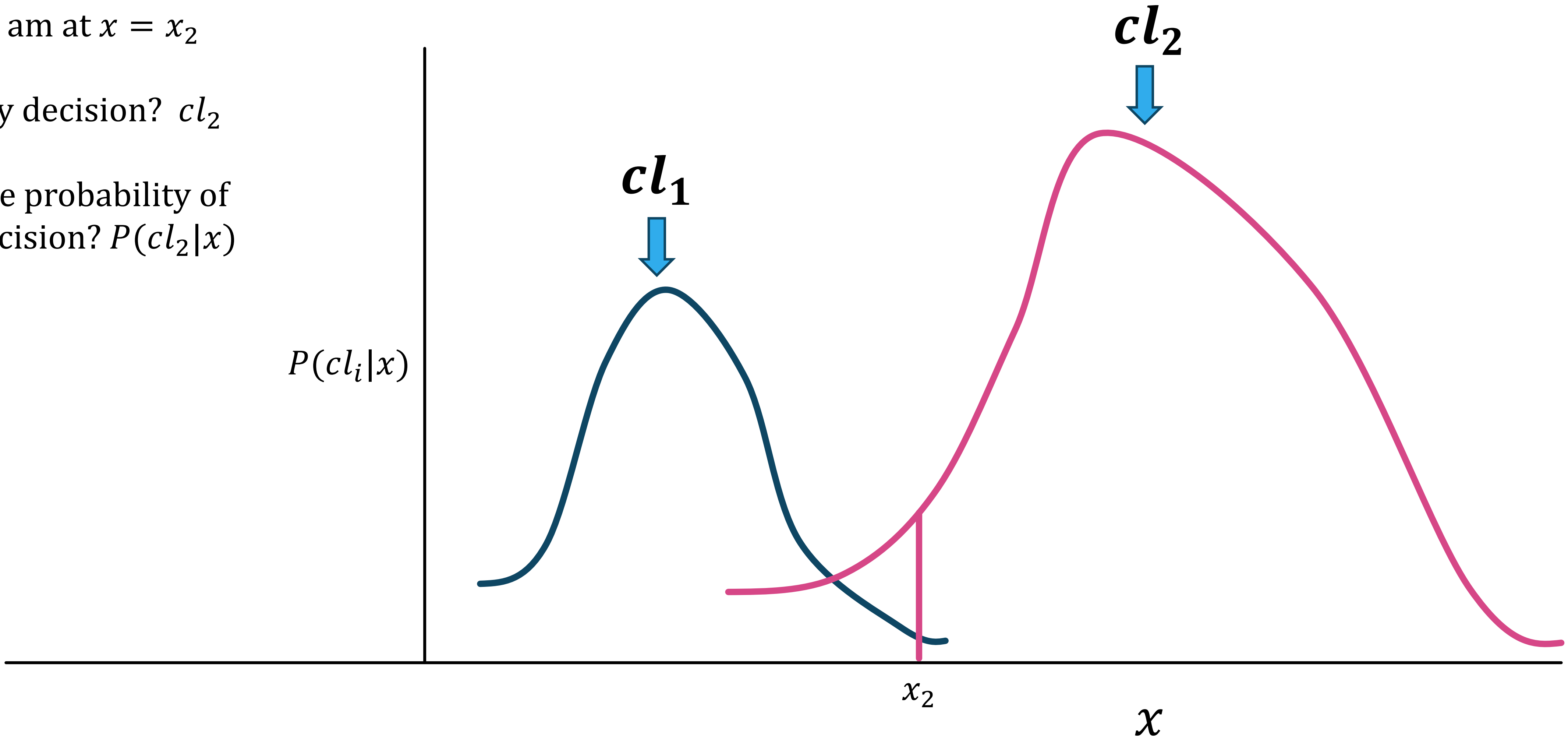


Error

Suppose, I am at $x = x_2$

What is my decision? cl_2

What is the probability of correct decision? $P(cl_2|x)$



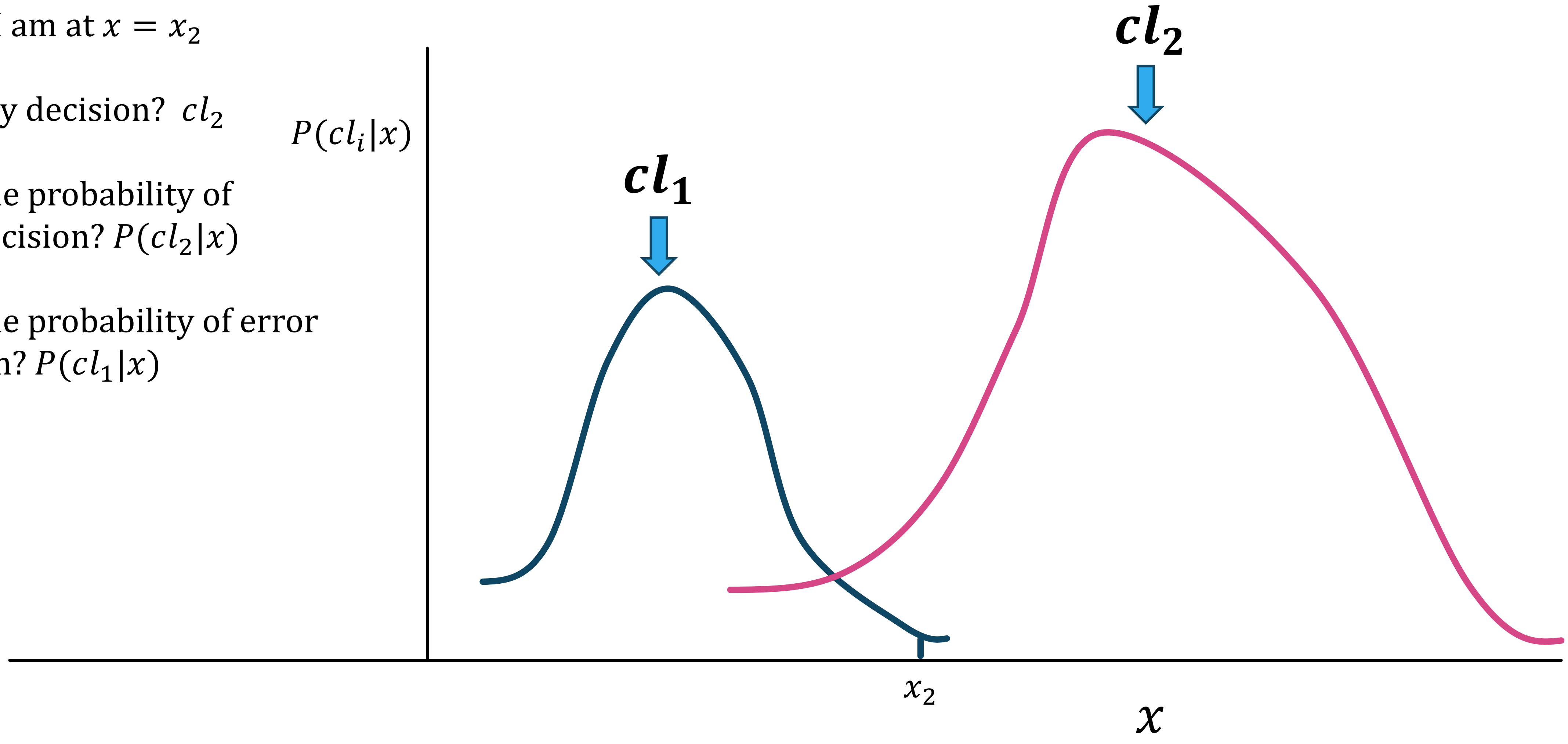
Error

Suppose, I am at $x = x_2$

What is my decision? cl_2

What is the probability of correct decision? $P(cl_2|x)$

What is the probability of error in decision? $P(cl_1|x)$



Error

Suppose, I am at $x = x_2$

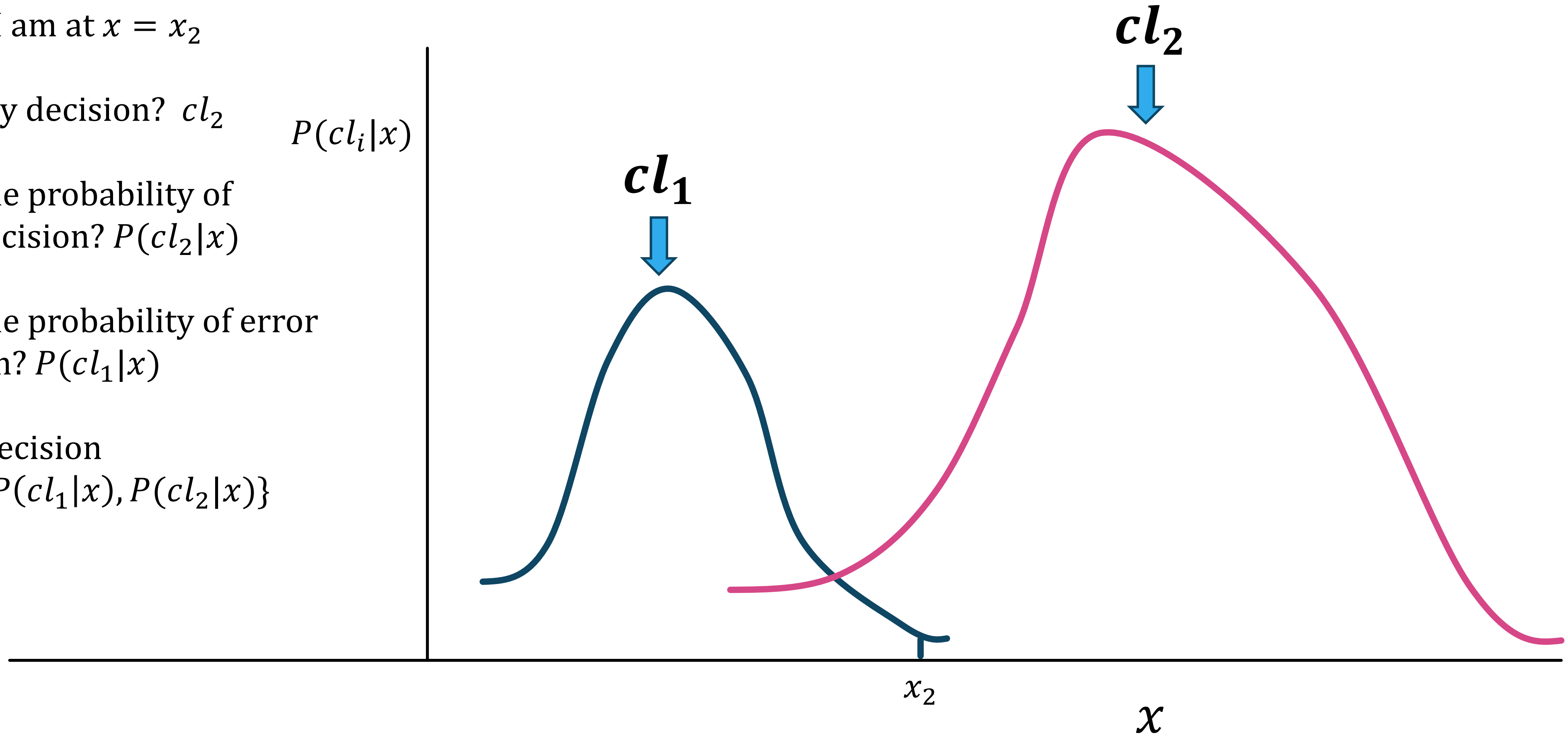
What is my decision? cl_2

What is the probability of correct decision? $P(cl_2|x)$

What is the probability of error in decision? $P(cl_1|x)$

Error in decision

$\min\{P(cl_1|x), P(cl_2|x)\}$

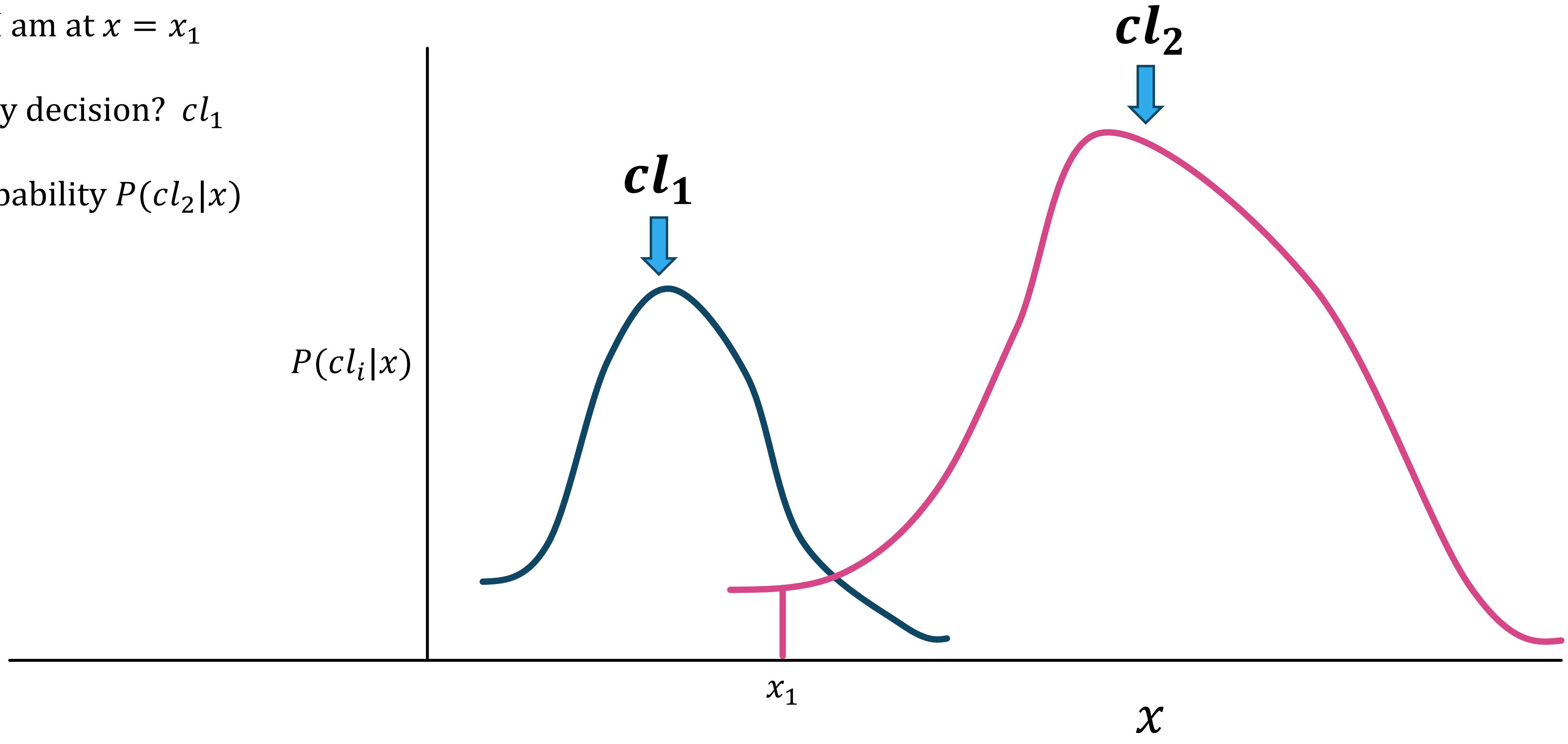


Error

Suppose, I am at $x = x_1$

What is my decision? cl_1

Error probability $P(cl_2|x)$



Error

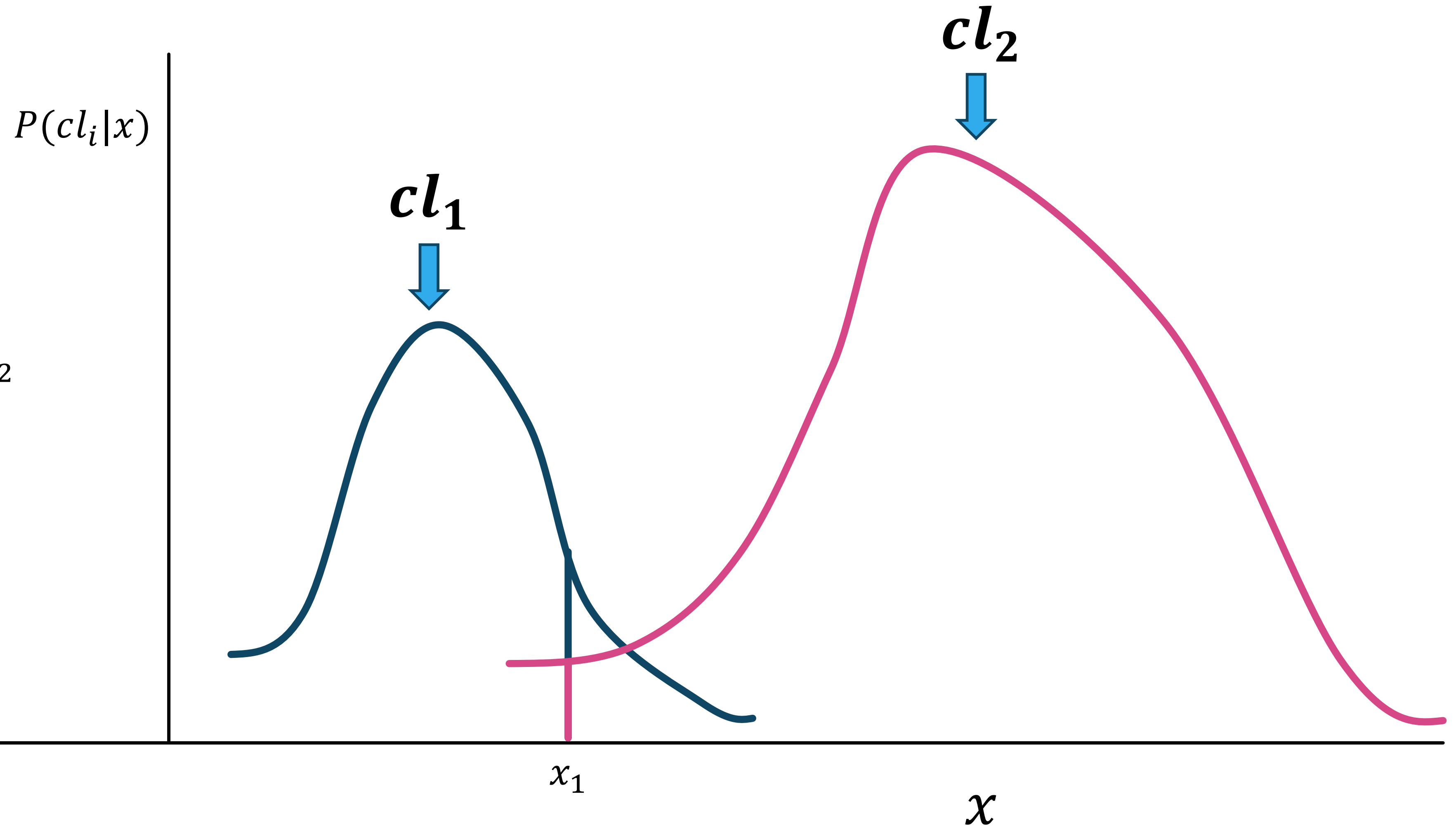
Suppose, I am at $x = x_1$

What is my decision? cl_1

Error probability $P(cl_2|x)$

But if I take decision to be cl_2

Error probability $P(cl_1|x)$



Error

Suppose, I am at $x = x_1$

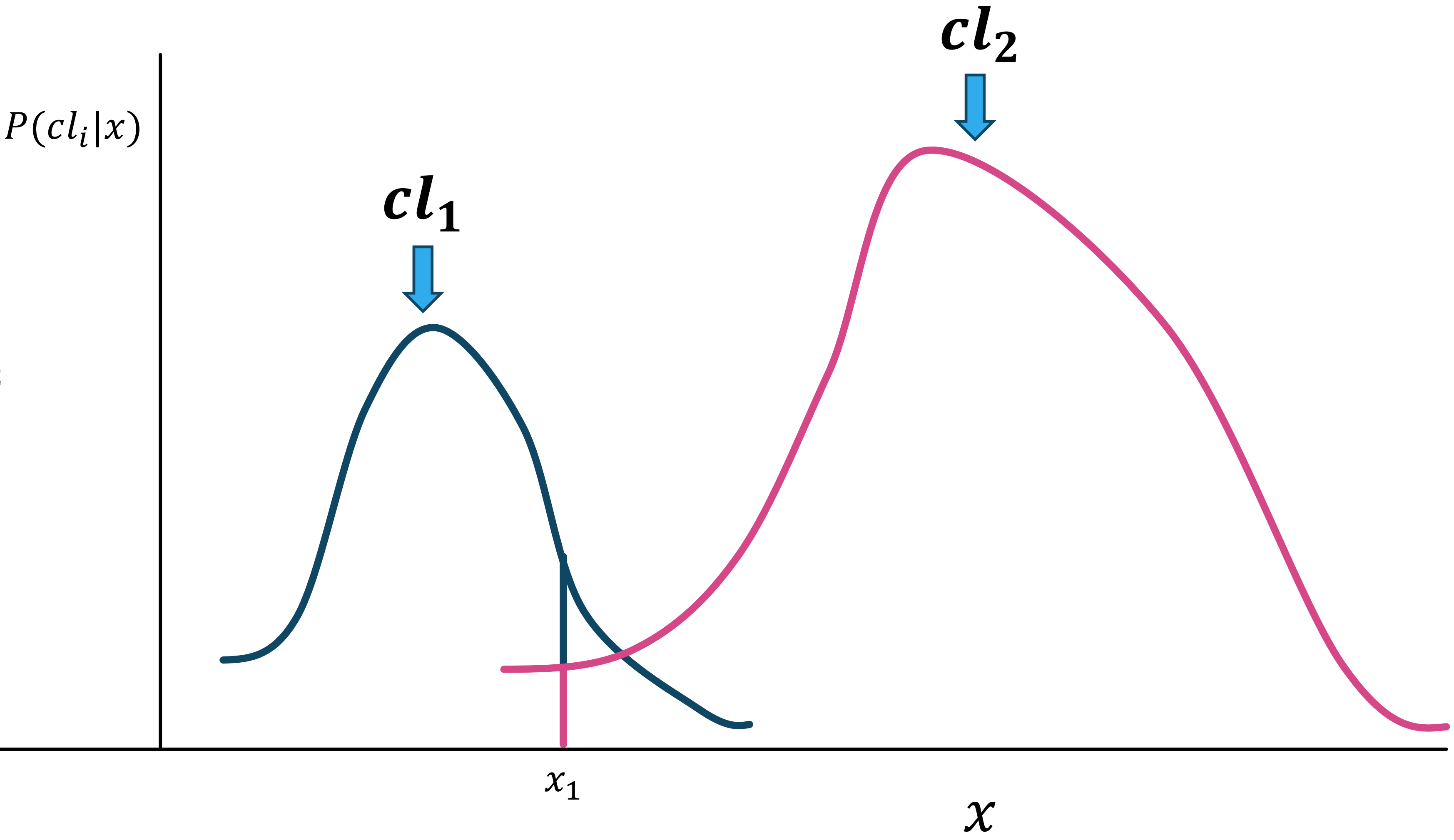
What is my decision? cl_1

Error probability $P(cl_2|x)$

But if I take decision to be cl_2

Error probability $P(cl_1|x)$

Min of the two error probabilities is $P(cl_2|x)$



Correct decision is obtained only when we consider the minimum error probability

Minimum Error Classification

Suppose, I am at $x = x_1$

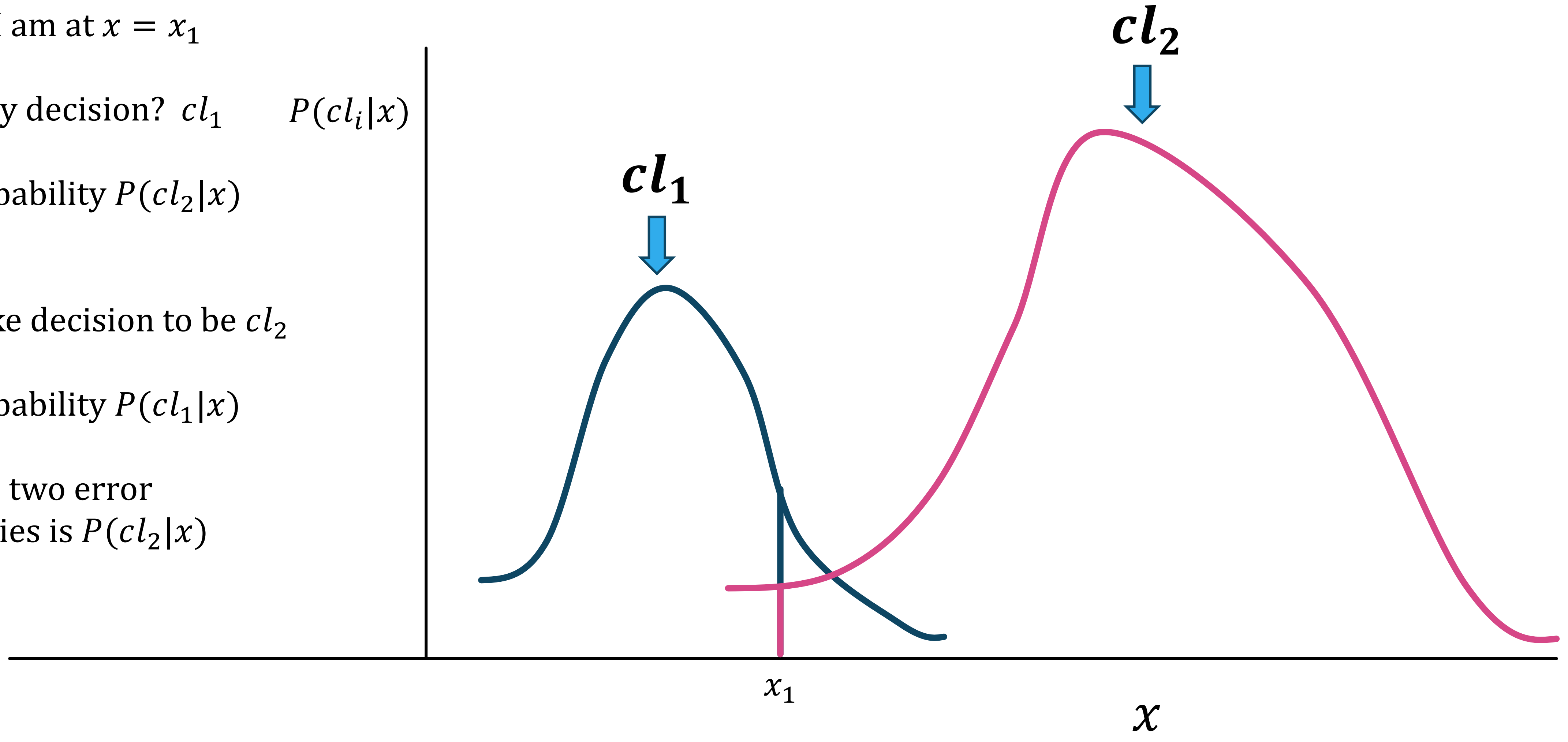
What is my decision? cl_1

Error probability $P(cl_2|x)$

But if I take decision to be cl_2

Error probability $P(cl_1|x)$

Min of the two error probabilities is $P(cl_2|x)$



Correct decision is obtained only when we consider the minimum error probability

Loss Function: A Generic Approach

- Consider that there are k number of classes
  - $cl_1, cl_2, \dots, cl_k$
  - Also called states of nature
- Consider that there are a number of actions
  - $\alpha_1, \alpha_2, \dots, \alpha_a$
  - Action can be assigning one class to the data
  - Action can also be assigning no class when there is a tie
- Loss function
  - $\lambda(\alpha_i|cl_j)$: Loss incurred for taking action α_i when the true state of nature is cl_j
- We consider a data point x to be a d-dimensional feature vector

Loss Function: A Generic Approach

- Loss function
  - $\lambda(\alpha_i|cl_j)$: Loss incurred for taking action α_i when state of nature is cl_j
- We consider a data point x to be a d-dimensional feature vector

- Expected loss for taking action α_i when we observe data x

$$R(\alpha_i|x) = \sum_{j=1}^{k} \lambda(\alpha_i|cl_j) P(cl_j|x)$$

Risk function/ conditional risk/ expected loss

Minimum Risk Classifier

- Expected loss for taking action α_i when we observe data x

$$R(\alpha_i|x) = \sum_{j=1}^{k} \lambda(\alpha_i|cl_j)P(cl_j|x)$$

Risk function/ conditional risk/ expected loss

- We want to take an action which minimizes the risk
  - Minimum risk classifier

Minimum Risk Classifier: For Two-class Problem

- Expected loss for taking action α_i when we observe data x

$$R(\alpha_i|x) = \sum_{j=1}^{k} \lambda(\alpha_i|cl_j) P(cl_j|x)$$

- Let $\lambda(\alpha_i|cl_j) = \lambda_{ij}$

- For two-class problem

$$R(\alpha_i|x) = \sum_{j=1}^{2} \lambda_{ij} P(cl_j|x) =$$

Minimum Risk Classifier: For Two-class Problem

- For two-class problem

$$R(\alpha_i|x) = \sum_{j=1}^{2} \lambda_{ij} P(cl_j|x)$$

- So,

$$R(\alpha_1|x) = \sum_{j=1}^{2} \lambda_{1j} P\left(cl_j \middle| x\right) = \lambda_{11} P(cl_1|x) + \lambda_{12} P(cl_2|x)$$

$$R(\alpha_2|x) = \sum_{j=1}^{2} \lambda_{2j} P\left(cl_j \middle| x\right) = \lambda_{21} P(cl_1|x) + \lambda_{22} P(cl_2|x)$$

Minimum Risk Classifier: For Two-class Problem

- Let
  - α_1 be the action of assigning class 1 to the input data x
  - α_2 be the action of assigning class 2 to the input data x
- We have

$$R(\alpha_1|x) = \sum_{j=1}^{2} \lambda_{1j} P\left(cl_j \middle| x\right) = \lambda_{11} P(cl_1|x) + \lambda_{12} P(cl_2|x)$$

$$R(\alpha_2|x) = \sum_{j=1}^{2} \lambda_{2j} P\left(cl_j \middle| x\right) = \lambda_{21} P(cl_1|x) + \lambda_{22} P(cl_2|x)$$

Minimum Risk Classifier: For Two-class Problem

- We have

$$R(\alpha_1|x) = \sum_{j=1}^{2} \lambda_{1j} P\left(cl_j \middle| x\right) = \lambda_{11} P(cl_1|x) + \lambda_{12} P(cl_2|x)$$

$$R(\alpha_2|x) = \sum_{j=1}^{2} \lambda_{2j} P\left(cl_j \middle| x\right) = \lambda_{21} P(cl_1|x) + \lambda_{22} P(cl_2|x)$$

- If we want to assign class 1 to the input data x, we want
  - $R(\alpha_1|x) < R(\alpha_2|x)$
  - $\left(\lambda_{11} P(cl_1|x) + \lambda_{12} P(cl_2|x)\right) < \left(\lambda_{21} P(cl_1|x) + \lambda_{22} P(cl_2|x)\right)$
  - $(\lambda_{11} - \lambda_{21}) P(cl_1|x) < (\lambda_{22} - \lambda_{12}) P(cl_2|x)$

Minimum Risk Classifier: For Two-class Problem

- If we want to assign class 1 to the input data x, we want
  - $(\lambda_{11} - \lambda_{21})P(cl_1|x) < (\lambda_{22} - \lambda_{12})P(cl_2|x)$
  - $(\lambda_{21} - \lambda_{11})P(cl_1|x) > (\lambda_{12} - \lambda_{22})P(cl_2|x)$

- Recall from Bayes decision rule, we need to have the following to assign class 1 to the input data x
  - $P(cl_1|x) > P(cl_2|x)$

Minimum Risk Classifier: For Two-class Problem

- If we want to assign class 1 to the input data x, we want
  - $(\lambda_{21} - \lambda_{11})P(cl_1|x) > (\lambda_{12} - \lambda_{22})P(cl_2|x)$

Weight

- Recall from Bayes decision rule, we need to have the following to assign class 1 to the input data x
  - $P(cl_1|x) > P(cl_2|x)$

Minimum Risk Classifier: How to Know λ_{ij}

- If we want to assign class 1 to the input data x, we want
  - $(\boldsymbol{\lambda_{21}} - \boldsymbol{\lambda_{11}})P(cl_1|x) > (\boldsymbol{\lambda_{12}} - \boldsymbol{\lambda_{22}})P(cl_2|x)$
- All the losses $(\boldsymbol{\lambda_{ij}s})$ are predefined depending on the problem
- In many occasions, we can consider $\boldsymbol{\lambda_{ii} = 0}$
  - Why?

Minimum Risk Classifier: How to Know λ_{ij}

- If we want to assign class 1 to the input data x, we want
  - $(\lambda_{21} - \lambda_{11})P(cl_1|x) > (\lambda_{12} - \lambda_{22})P(cl_2|x)$

- All the losses $\left(\lambda_{ij}s\right)$ are predefined depending on the problem

- In many occasions, we can consider $\lambda_{ii} = 0$
  - Why?
  - Because λ_{ii} indicates correct decision

One Strategy of Defining λ_{ij}

- Let's define

$$\lambda_{ij} = \lambda(\alpha_i | cl_j) = \begin{cases} 0 \ \ when\ i = j \\ 1\ when\ i \neq j \end{cases}$$

Calculation of Risk

- We have

$$\lambda_{ij} = \lambda(\alpha_i | cl_j) = \begin{cases} 0 \;\; when \; i = j \\ 1 \; when \; i \neq j \end{cases}$$

- We also have

$$R(\alpha_i|x) = \sum_{j=1}^{k} \lambda(\alpha_i|cl_j) P(cl_j|x)$$

- Combining the two, we get

$$R(\alpha_i|x) = \sum_{i \neq j} P(cl_j|x)$$

Calculation of Risk

- Combining the two, we get

$$R(\alpha_i|x) = \sum_{i \neq j} P(cl_j|x)$$

- Let $i = 2$ and there are three classes cl_1, cl_2, cl_3
- So,

$$\sum_{j=1}^{3} P\left(cl_j \middle| x\right) = ?$$

Calculation of Risk

- Combining the two, we get

$$R(\alpha_i|x) = \sum_{i \neq j} P(cl_j|x)$$

- Let $i = 2$ and there are three classes cl_1, cl_2, cl_3
- So,

$$\sum_{j=1}^{3} P\left(cl_j \middle| x\right) = 1$$

Calculation of Risk

- Combining the two, we get

$$R(\alpha_i|x) = \sum_{i \neq j} P(cl_j|x)$$

- Let $i = 2$ and there are three classes cl_1, cl_2, cl_3
- So,

$$\sum_{j=1}^{3} P(cl_j|x) = 1 = P(cl_1|x) + P(cl_2|x) + P(cl_3|x)$$

- So,

$$\sum_{i \neq j} P(cl_j|x) = P(cl_1|x) + P(cl_3|x) = 1 - P(cl_2|x)$$

Calculation of Risk

- So, we get

$$R(\alpha_i|x) = \sum\nolimits_{i \neq j} P(cl_j|x) = 1 - P(cl_i|x)$$

- So, if I want to minimize risk for action α_i, I have to maximize $P(cl_i|x)$

- That means, given the observation x
  - If the probability of class label cl_i is maximum, i.e. if $P(cl_i|x)$
  - The corresponding risk of error is minimum

Minimum Error Rate Classifier

- So, if I want to minimize risk for action α_i, I have to maximize $P(cl_i|x)$

- That means, given the observation x
  - If the probability of class label cl_i is maximum, i.e. if $P(cl_i|x)$
  - The corresponding risk of error is minimum

- This is called minimum error rate classifier

- This is similar to the Bayes decision rule $P(cl_1|x) > \ P(cl_2|x) \ \Rightarrow cl_1$

Inference by Enumeration

	fever		¬ fever	
	cough	¬ cough	cough	¬ cough
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- But, what is the problem with this approach?
  - For a system with many causes and effects, we have to maintain a large set of values and operate on those

Independent Events

- Two events A and B are said to be independent if
  - $\boldsymbol{P(A|B) = P(A)}$ (1)

- We already have
  - $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}}$ (2)

- From (1) and (2), we get
  - $\boldsymbol{P(B|A) = P(B)}$
  - $\boldsymbol{P(A \cap B) = P(A)P(B)}$

Independent Events

- Suppose I want to deal with Fever, Cough, Covid, and Internet Speed (I_Sp)

- We intuitively know that internet speed does not depend on the other three

- So, if we want to find out the joint distribution
  - $P(Fever, Cough, Covid, I\_Sp)$
- We can write

$$P(Fever, Cough, Covid, I\_Sp) = P(Fever, Cough, Covid)\ P(I\_Sp)$$

Independent Events

- Suppose the sample space for Internet Speed (I_Sp) is

$$\{slow, medium, fast, very\ fast\}$$

- So, if we want to find out the joint distribution
  - $P(Fever, Cough, Covid, I\_Sp)$

		fever		¬ fever	
		cough	**¬ cough**	**cough**	**¬ cough**
covid	**slow**	a1	a2	a3	a4
	medium	a5	a6	a7	a8
	fast	a9	a10	a11	a12
	very fast	a13	a14	a15	a16
¬ covid	**slow**	a17	a18	a19	a20
	medium	a21	a22	a23	a24
	fast	a25	a26	a27	a28
	very fast	a29	a30	a31	a32

Independent Events

		fever		¬ fever	
		cough	**¬ cough**	**cough**	**¬ cough**
covid	**slow**	a1	a2	a3	a4
	medium	a5	a6	a7	a8
	fast	a9	a10	a11	a12
	very fast	a13	a14	a15	a16
¬ covid	**slow**	a17	a18	a19	a20
	medium	a21	a22	a23	a24
	fast	a25	a26	a27	a28
	very fast	a29	a30	a31	a32

- How many variables do I need to store this table?

Independent Events

		fever		**¬ fever**	
		cough	**¬ cough**	**cough**	**¬ cough**
covid	**slow**	a1	a2	a3	a4
	medium	a5	a6	a7	a8
	fast	a9	a10	a11	a12
	very fast	a13	a14	a15	a16
¬ covid	**slow**	a17	a18	a19	a20
	medium	a21	a22	a23	a24
	fast	a25	a26	a27	a28
	very fast	a29	a30	a31	a32

- How many entries do I need to store this table? **32 (31 parameters)**

Independent Events

- Total entries: **32 (31 parameters)**

- Now I know that

$$P(Fever, Cough, Covid, I\_Sp) = P(Fever, Cough, Covid)\ P(I\_Sp)$$

- Now I know that
  - To store the above table, I need to store $P(Fever, Cough, Covid)\ and\ P(I\_Sp)$

- Total entries: **32 (31 parameters)**

- Now I know that

$$P(Fever, Cough, Covid, I\_Sp) = P(Fever, Cough, Covid)\ P(I\_Sp)$$

- Now I know that
  - To store the above table, I need to store $P(Fever, Cough, Covid)\ and\ P(I\_Sp)$

Independent Events

	fever		**¬ fever**	
	cough	**¬ cough**	**cough**	**¬ cough**
covid	0.21	0.10	0.11	0.08
¬ covid	0.11	0.07	0.09	0.23

- Table for $P(Fever, Cough, Covid)$
- How many entries: **8 (7 parameters)**

Independent Events

slow	medium	fast	very fast
0.2	0.4	0.25	0.15

- Table for $P(I\_Sp)$
- How many entries? **4 (3 parameters)**

Independent Events

		fever		**¬ fever**	
		cough	**¬ cough**	**cough**	**¬ cough**
covid	**slow**	a1	a2	a3	a4
	medium	a5	a6	a7	a8
	fast	a9	a10	a11	a12
	very fast	a13	a14	a15	a16
¬ covid	**slow**	a17	a18	a19	a20
	medium	a21	a22	a23	a24
	fast	a25	a26	a27	a28
	very fast	a29	a30	a31	a32

- Total entries required was: **32 (31 parameters)**
- After performing factorization $P(Fever, Cough, Covid, I\_Sp) = P(Fever, Cough, Covid)\ P(I\_Sp)$
  - Total entries required: 8+4= **12 (7+3=10 parameters)**

Independent Events

- Total entries required was: **32**

- After performing factorization

$$P(Fever, Cough, Covid, I\_Sp) = P(Fever, Cough, Covid)\ P(I\_Sp)$$

  - Total entries required: 8+4= **12**

- Complete independence is extremely powerful

- But in our model, why should we include something that does not have any relation to the problem?
  - We will not include such factors

Independent Events

- But in our model, why should we include something that does not have any relation to the problem?
  - We will not include such factors
- **Although complete independence is extremely powerful, it is useless for our modeling purpose**
- So, what do we do?

Conditional Independence

- Suppose, I want to calculate the probability of fever given that a person has covid and cough

- So, we want to model
  - $P(fever|\,covid, cough\,)$

- I know that the person has covid
  - So, does the knowledge about cough give us any additional information to predict the probability of fever?

Conditional Independence

- Suppose, I want to calculate the probability of fever given that a person has covid and cough

- So, we want to model
  - $P(fever|\ covid, cough\ )$

- I know that the person has covid
  - So, does the knowledge about cough give us any additional information to predict the probability of fever?
    - **No**

Conditional Independence

- So, we can say
  - $P(fever|\, covid, cough\,) = P(fever|\, covid)$
- And also
  - $P(fever|\, \neg covid, cough\,) = P(fever|\neg covid)$
- **We say that fever is conditionally independent of cough given covid**
  - $P(Fever|\, Covid, Cough\,) = P(Fever|\, Covid)$

Conditional Independence

- Suppose we want to model
  - $P(Fever, Covid, Cough)$
  - Typically, I need **8** entries and **7** parameters

- $P(Fever, Covid, Cough) = P(Fever|Covid, Cough)\, P(Covid, Cough)$

$$= P(Fever|Covid)\, P(Covid, Cough)$$

$$= P(Fever|Covid)\, P(Covid|Cough)\, P(Cough)$$

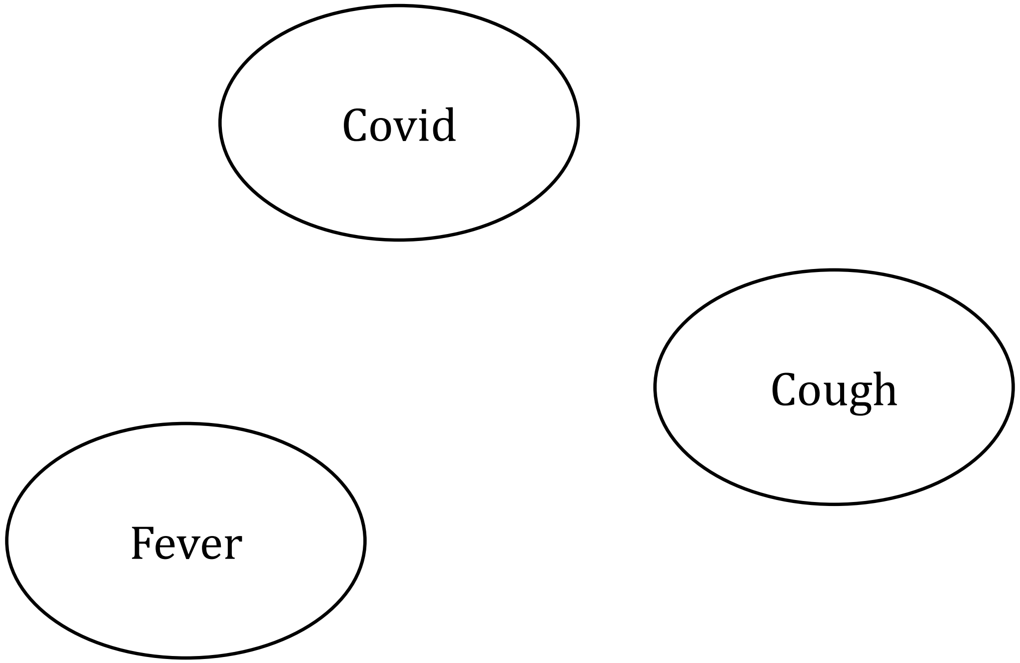
Conditional Independence

- $P(Fever, Covid, Cough) = \underbrace{P(Fever|Covid)}_{\text{2 parameters}} \, \underbrace{P(Covid|Cough)}_{\text{2 parameters}} \, \underbrace{P(Cough)}_{\text{1 parameter}}$

Total: **5** parameters

Graphical Representation

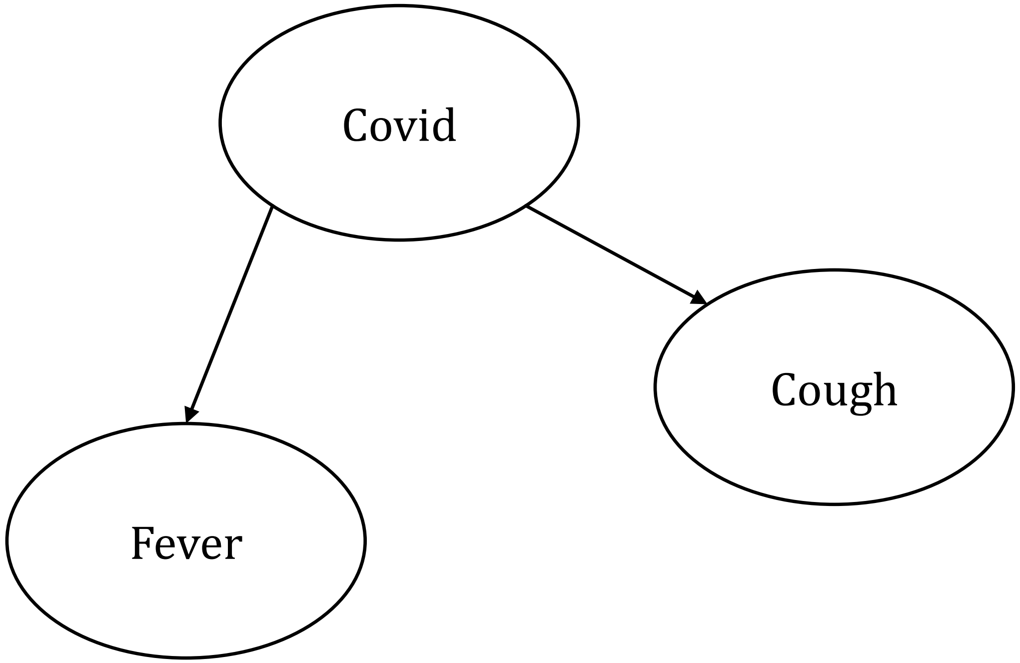
- A graph representing influences



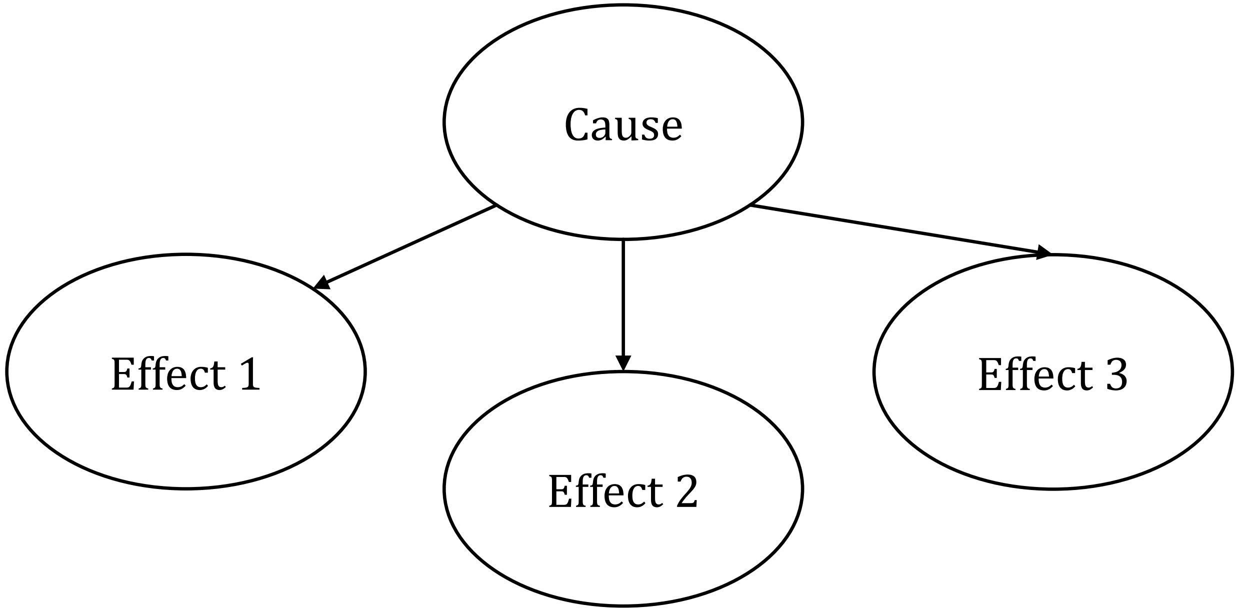
Graphical Representation

- A graph representing influences

One cause resulting in
multiple independent effects



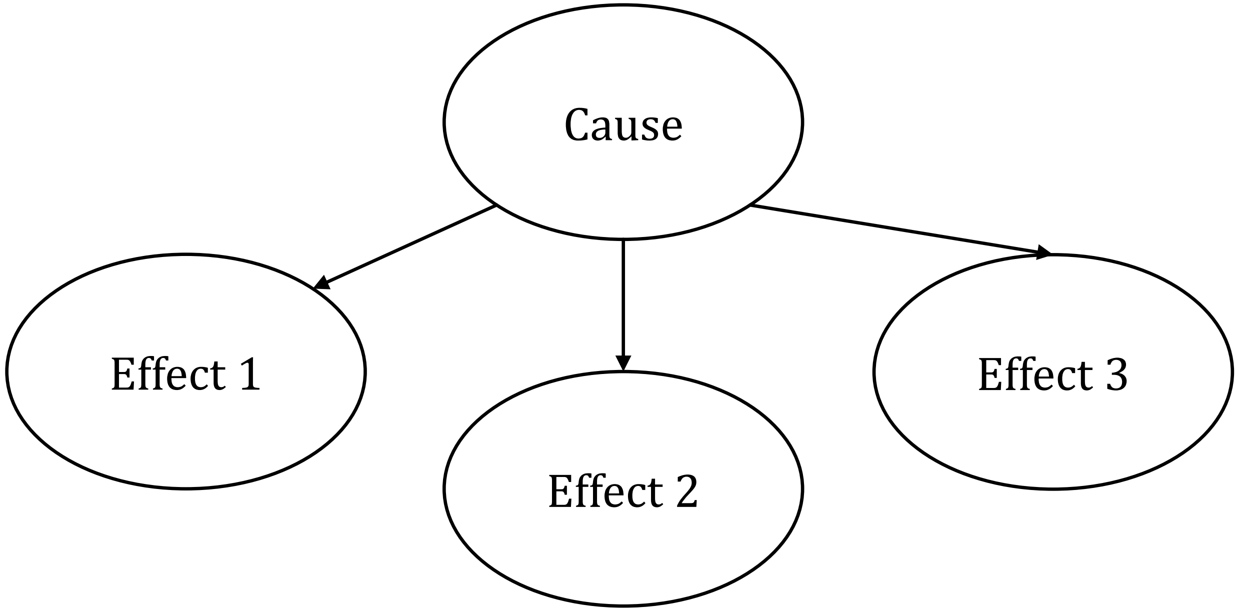
Graphical Representation



One cause resulting in multiple independent effects

A naïve assumption

Graphical Representation



One cause resulting in multiple independent effects

Naïve Bayes' model

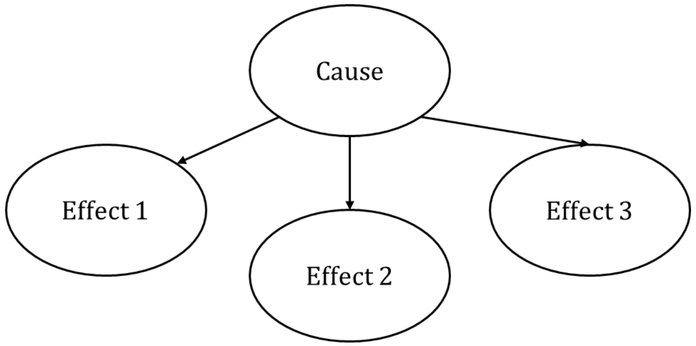
Naïve Bayes' Model

$$P(\text{Effect 1}, \text{Effect 2}, \text{Effect 3}, \text{Cause})$$

$$= P(\text{Effect 1}|\text{Effect 2}, \text{Effect 3}, \text{Cause}) \quad P(\text{Effect 2}|\text{Effect 3}, \text{Cause}) \quad P(\text{Effect 3}|Cause)P(Cause)$$

$$= P(\text{Effect 1}|Cause)P(\text{Effect 2}|\ Cause)\ P(\text{Effect 3}|Cause)P(Cause)$$

$$= P(Cause)\prod_{k=1}^{3} P(\text{Effect } k|Cause)$$



- For n number of effects

$$P(Effect_1, Effect_2, Effect_3, \dots, Effect_n, Cause) = P(Cause)\prod_{k=1}^{n} P(Effect_k|Cause)$$

Naïve Bayes' Model

$$P(\text{Effect 1}, \text{Effect 2}, \text{Effect 3}, \text{Cause})$$

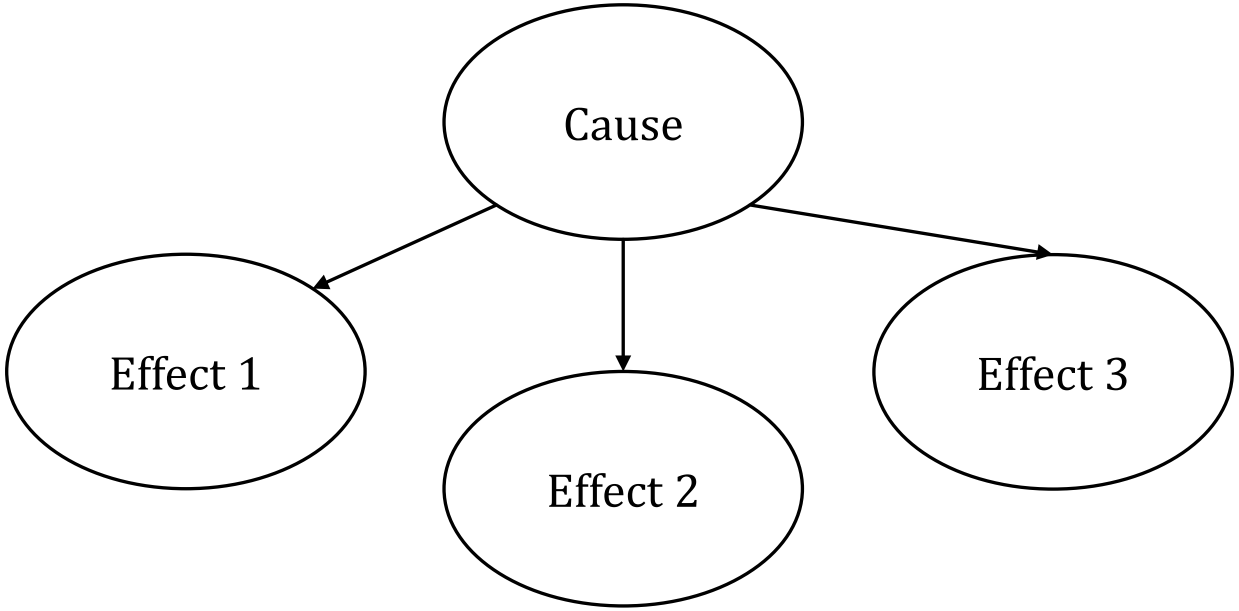
$$= P(Cause) \prod_{k=1}^{3} P(\text{Effect } k | Cause)$$

- For d number of effects

$$P(\text{Effect 1}, \text{Effect 2}, \text{Effect 3}, \dots, \text{Effect } d, Cause) = P(Cause) \prod_{k=1}^{d} P(\text{Effect } k | Cause)$$

Naïve Bayes' Model

One cause resulting in multiple independent effects



Naïve Bayes' model

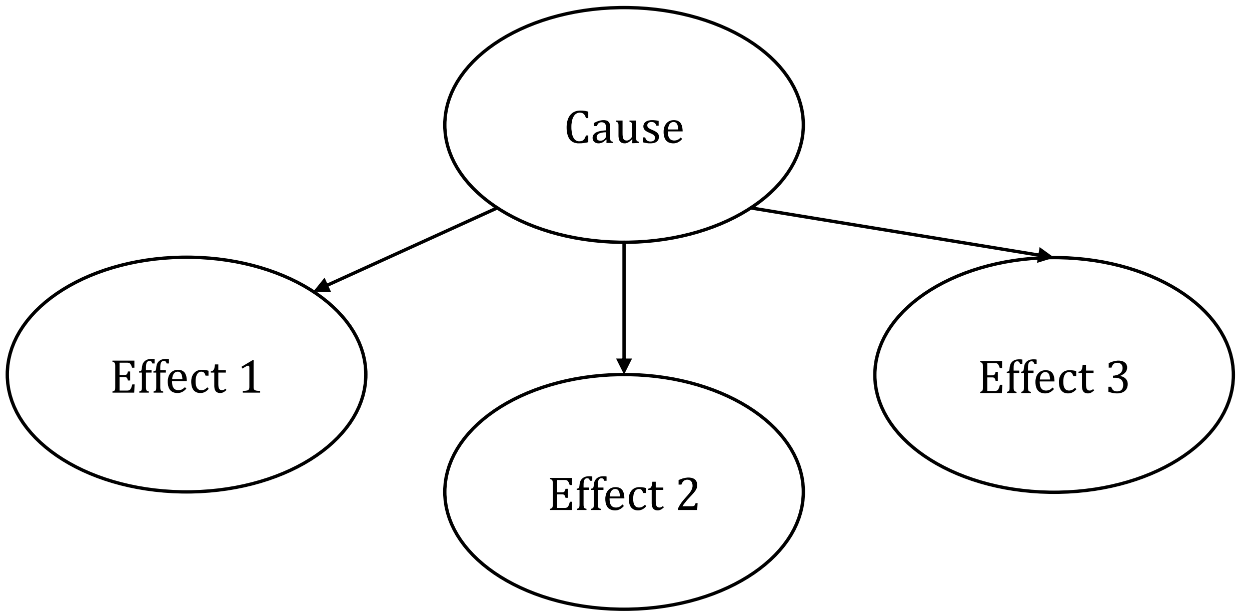
Naïve Bayes' Model

One cause resulting in multiple independent effects

Classification Problem

Cause: **Class Label (class 1, class2, etc.)**

Effect 1: Feature 1; Effect 2: Feature 2, etc.



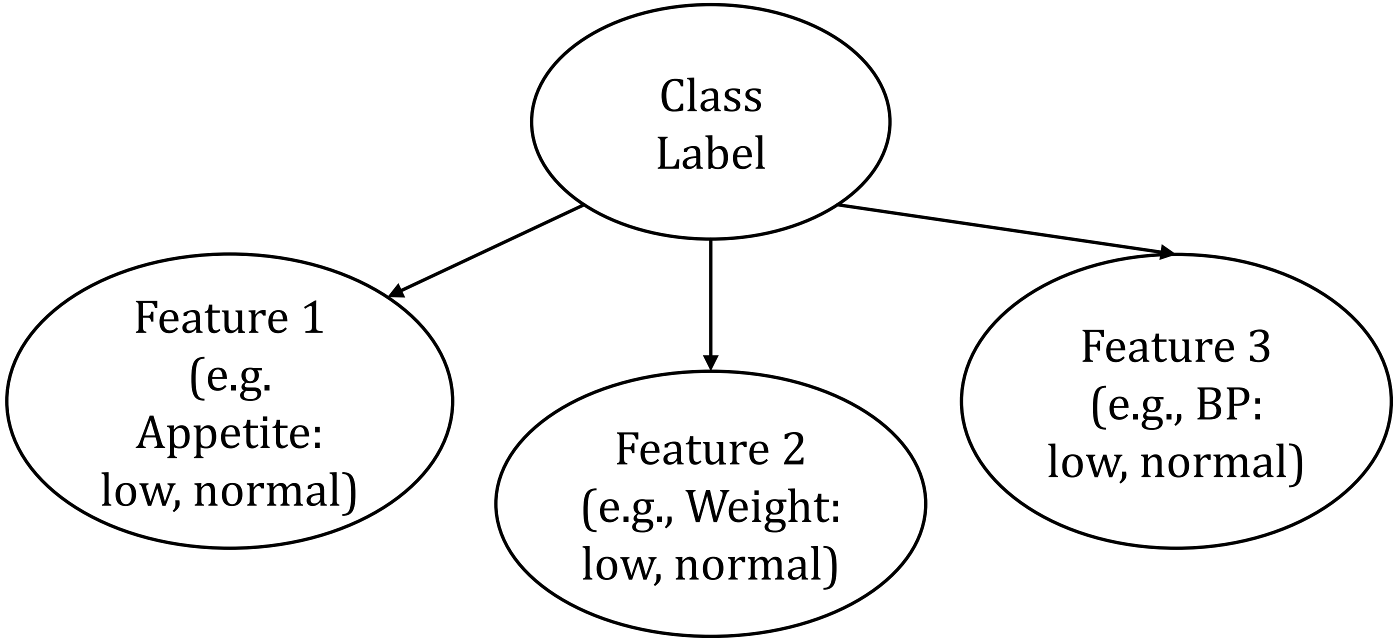
Naïve Bayes' model

Naïve Bayes' Model

Classification Problem: Classification of persons with anemia

Cause: **Class Label (anemia, no anemia)**

Effect 1: Feature 1; Effect 2: Feature 2, etc.



Naïve Bayes' model

Naïve Bayes' Model

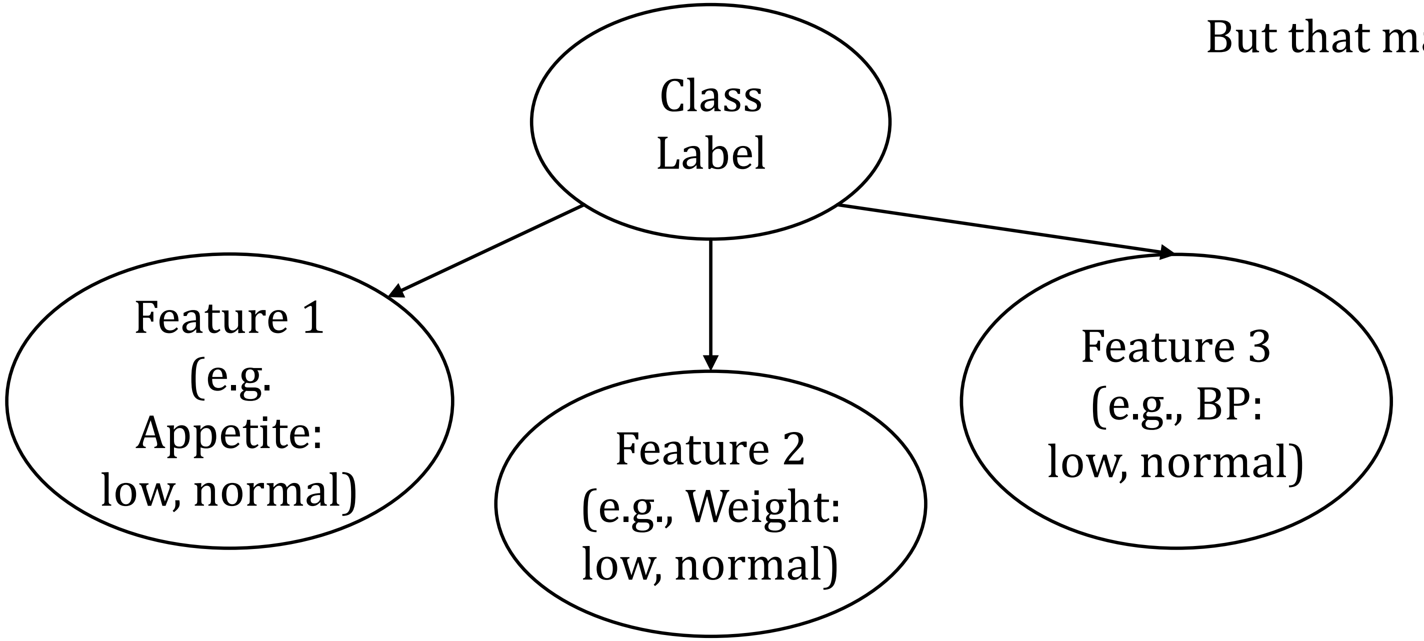
Classification Problem: Classification of persons with anemia

Cause: **Class Label (anemia, no anemia)**

Effect 1: Feature 1; Effect 2: Feature 2, etc.

We consider the features to be independent

But that may not be the case



Naïve Bayes' model

Naïve Bayes' Model

- Bayes' decision rule

$$P(cl_1|x) > P(cl_2|x) \Rightarrow cl_1$$

- Now let's use Bayes theorem
  - $P(cl_1|x) = \frac{P(x|cl_1)P(cl_1)}{P(x)}$

- Consider x to be a $d-$ dimensional feature vector $\{x_1, x_2, \ldots, x_d\}$
- All features are conditionally independent of each other given the class

Naïve Bayes' Model

- Consider x to be a $d-$dimensional feature vector $\{x_1, x_2, \ldots, x_d\}$
- All features are conditionally independent of each other given the class
- Now let's use Bayes theorem

$$\boldsymbol{P(\textbf{Effect 1}, \textbf{Effect 2}, \textbf{Effect 3}, \ldots, \textbf{Effect } d, Cause)}$$
$$\boldsymbol{= P(Cause) \prod_{k=1}^{d} P(\textbf{Effect } k | Cause)}$$

$$P(cl_1|x) = \frac{P(x, cl_1)}{P(x)}$$
$$= \frac{P(x_1 x_2 \ldots x_d, cl_1)}{P(x)}$$
$$= \frac{\boldsymbol{P(x_1|cl_1) P(x_2|cl_1) \ldots P(x_d|cl_1) P(cl_1)}}{P(x)}$$

Naïve Bayes' Model

- Consider x to be a $d-$dimensional feature vector $\{x_1, x_2, \dots, x_d\}$
- All features are conditionally independent of each other given the class
- Now let's use Bayes theorem

$$P(cl_1|x) = \frac{\boldsymbol{P(x_1|cl_1)P(x_2|cl_1)\dots P(x_d|cl_1)P(cl_1)}}{P(x)}$$

Similarly

$$P(cl_2|x) = \frac{\boldsymbol{P(x_1|cl_2)P(x_2|cl_2)\dots P(x_d|cl_2)P(cl_2)}}{P(x)}$$

Naïve Bayes' Model

- Bayes' decision rule

$$P(cl_1|x) > P(cl_2|x) \Rightarrow cl_1$$

$$\frac{\boldsymbol{P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1)}}{P(x)} > \frac{\boldsymbol{P(x_1|cl_2)P(x_2|cl_2)\ldots P(x_d|cl_2)P(cl_2)}}{P(x)} \Rightarrow cl_1$$

$$\boldsymbol{P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1) > P(x_1|cl_2)P(x_2|cl_2)\ldots P(x_d|cl_2)P(cl_2)} \Rightarrow cl_1$$

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Training data

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Training data

Suppose, we see a new sample who has normal appetite, normal weight, and low BP. Find if the person has anemia

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 1: Calculate $P(anemia), P(no\ anemia)$ from the training data

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 1: Calculate $P(anemia), P(no\ anemia)$ from the training data

$$P(anemia) = \frac{4}{7}$$

$$P(no\ anemia) = \frac{3}{7}$$

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

		Anemia	No Anemia
Appetite	Low		
	Normal		

$P(Appetite = Low|Anemia)$

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

		Anemia	No Anemia
Appetite	Low		
	Normal		

$P(Appetite = Low|Anemia)$

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

Similarly, we can calculate for other features

		Anemia	No Anemia
Weight	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

Similarly, we can calculate for other features

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

Naïve Bayes': An Example

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$P(anemia) = \frac{4}{7}$ $\quad P(no\ anemia) = \frac{3}{7}$

Step 3: Testing:

Suppose, I observe a test data with normal appetite, normal weight, and low BP

Predict if the person has anemia

Naïve Bayes': An Example

$P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1)$

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$P(anemia) = \frac{4}{7}$ $\quad$ $P(no\ anemia) = \frac{3}{7}$

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **normal** weight, and **low** BP

Predict if the person has anemia

Let's first evaluate the chance of anemia

$$\pi_{anemia} = P(appetite = normal|anemia)P(weight = normal|anemia)P(BP = low|anemia)P(anemia)$$

$$= \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{7} = 0.009$$

Naïve Bayes': An Example

$$P(x_1|cl_2)P(x_2|cl_2)\dots P(x_d|cl_2)P(cl_2)$$

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$P(anemia) = \frac{4}{7}$ $\quad$ $P(no\ anemia) = \frac{3}{7}$

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **normal** weight, and **low** BP

Predict if the person has anemia

Let's first evaluate the chance of no anemia

$$\pi_{no\ anemia} =$$

$$P(appetite = normal|no\ anemia)P(weight = normal|no\ anemia)P(BP = low|no\ anemia)P(no\ anemia)$$

$$= \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{3}{7} = 0.06$$

Naïve Bayes': An Example

$$\pi_{anemia} = P(appetite = normal|anemia)P(weight = normal|anemia)P(BP = low|anemia)P(anemia) = \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{7} = 0.009$$

$$\pi_{no\ anemia} = P(appetite = normal|no\ anemia)P(weight = normal|no\ anemia)P(BP = low|no\ anemia)P(no\ anemia) = \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{3}{7} = 0.06$$

$$P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1) > P(x_1|cl_2)P(x_2|cl_2)\ldots P(x_d|cl_2)P(cl_2) \Rightarrow cl_1$$

So, our conclusion is that the person does not have anemia as per the Naïve Bayes' classifier

Naïve Bayes' Classifier: Algorithm

- Training Data
  - Assume N training samples and class label pairs $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(N)}, y^{(N)}\right)$
  - Each training sample $x^{(i)}$ is a $D -$ dimensional feature vector $\{x_1^{(i)}, x_2^{(i)}, \dots, x_D^{(i)}\}$
  - Assume that attribute x_k can take values $x_{k_1}, x_{k_2}, \dots, x_{k_v}$
  - The values of the attributes are discrete
  - We have a total of C number of classes $cl_1, cl_2, \dots, cl_C$

Naïve Bayes' Classifier: The Algorithm

- Training Data
  - Assume N training samples and class label pairs $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(N)}, y^{(N)}\right)$
  - Each training sample $x^{(i)}$ is a $D-$ dimensional feature vector $\{x_1^{(i)}, x_2^{(i)}, \ldots, x_D^{(i)}\}$
  - Assume that attribute x_k can take values $x_{k_1}, x_{k_2}, \ldots, x_{k_v}$
  - We have a total of C number of classes $cl_1, cl_2, \ldots, cl_C$

Training method

for $j = 1: C$
    Calculate $P(cl_j)$ from training data
for $d = 1: D$
    for $q = 1: v$
        Calculate $P(x_{d_q} | cl_j)$

Naïve Bayes' Classifier: The Algorithm

Training method

for $j = 1{:}\,C$

Calculate $P(cl_j)$ from training data

for $d = 1{:}\,D$

for $q = 1{:}\,v$

Calculate $P(x_{d_q}|cl_j)$

Inference

for a new test sample $x^{(new)}$, suppose the feature vector is $\left\{x_1^{(new)}, x_2^{(new)}, \dots, x_d^{(new)}\right\}$

for $j = 1{:}\,C$

- get $P(cl_j)$ computed during training
- get $P\left(x_1^{(new)}\middle|cl_j\right), P\left(x_2^{(new)}\middle|cl_j\right), \dots, P\left(x_d^{(new)}\middle|cl_j\right)$ computed during training
- Calculate $\pi_j = P\left(x_1^{(new)}\middle|cl_j\right) P\left(x_2^{(new)}\middle|cl_j\right) \dots P\left(x_d^{(new)}\middle|cl_j\right) P(cl_j)$

Find out j for which π_j is maximum

$$j^* = \underset{j}{\operatorname{argmax}}\, \boldsymbol{\pi_j}$$

Assign the class label j^* to the test data

Life without Naïve Bayes'

- Consider x to be a $d-$dimensional feature vector $\{x_1, x_2, \dots, x_d\}$
- If all features **are not** conditionally independent of each other given the class
- The Bayes theorem would be

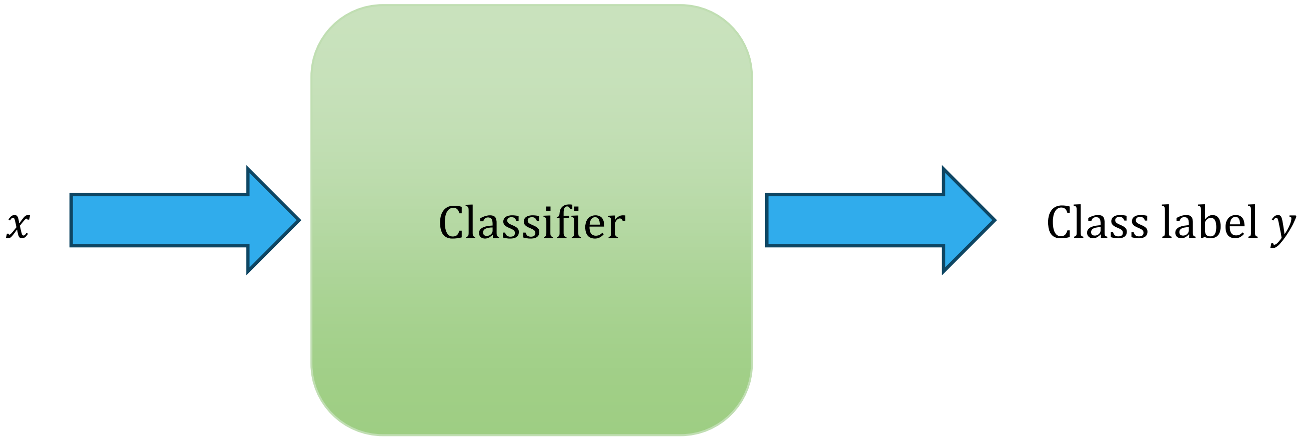
Computing these conditional probabilities would be extremely difficult

$$P(cl_1|x) = \frac{P(x, cl_1)}{P(x)}$$

$$= \frac{P(x_1 x_2 \dots x_d, cl_1)}{P(x)}$$

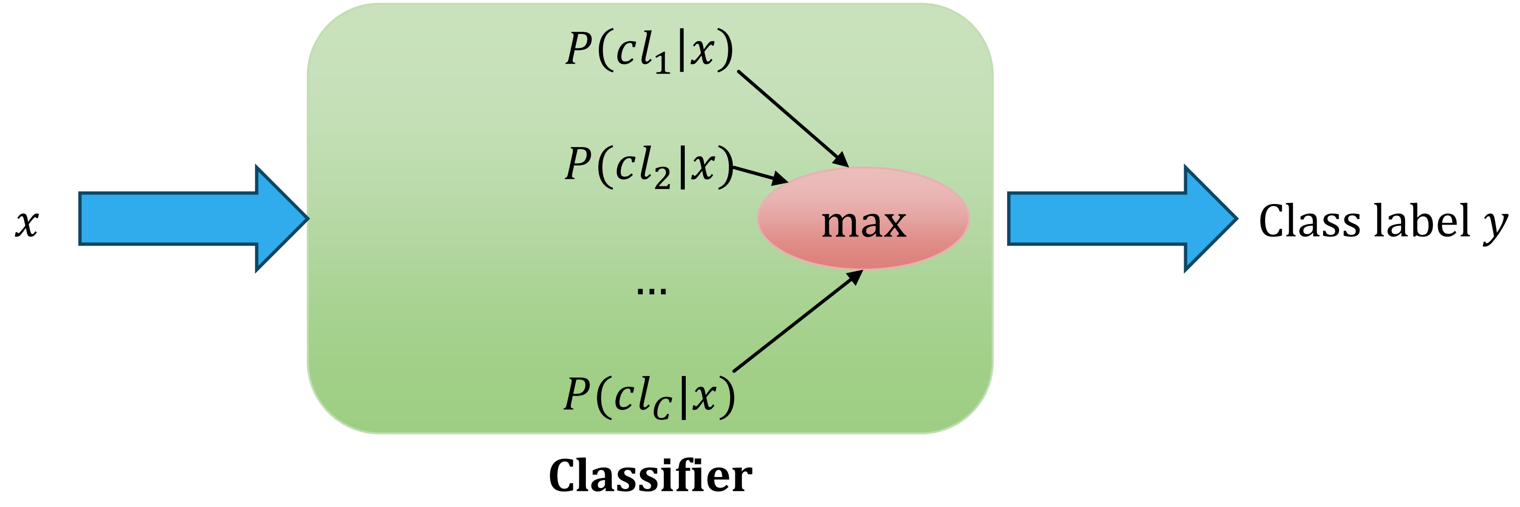
$$= \frac{\boldsymbol{P(x_1|x_2 \dots x_d, cl_1) P(x_2|x_3, \dots, cl_1) \dots P(x_d|cl_1) P(cl_1)}}{P(x)}$$

The Classifier

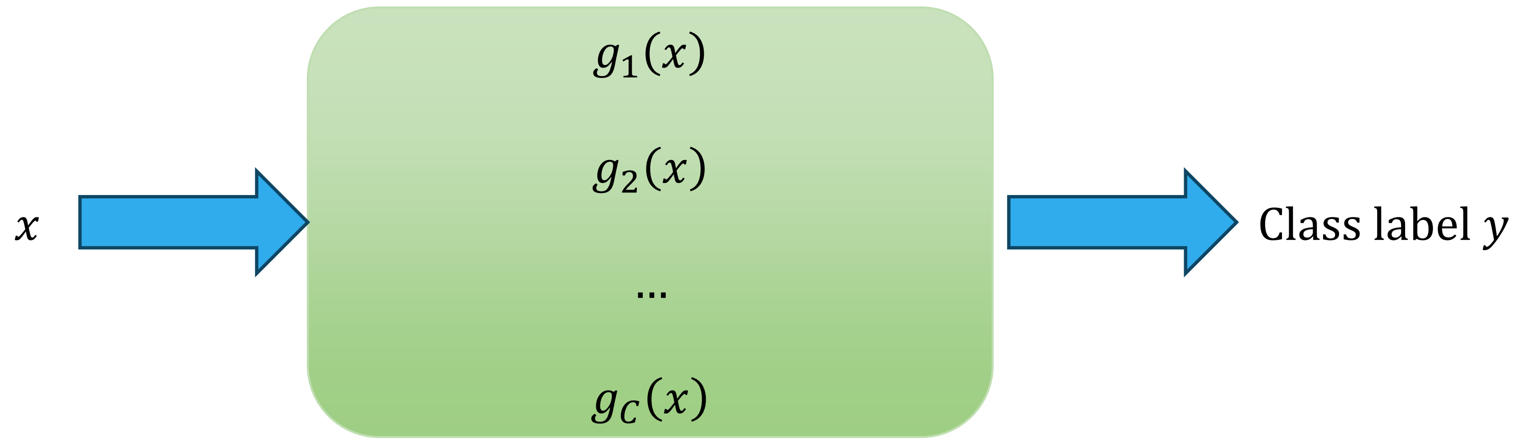


How Does the Classifier Do it?

Suppose I have C number of classes. I consider Bayes' decision rule



In a More Generic Form



Classifier

$g_i(\cdot)$ is called discriminant function

If $\boldsymbol{g_i(x) > g_j(x), \forall j \neq i}$, we assign class label $\boldsymbol{i}$ to the input data $\boldsymbol{x}$

Nature of the Discriminant Function

For the minimum risk classifier

We assign the class label corresponding to the action of minimum risk

If the risk of action α_i is $R(\alpha_i|x)$

$g_i(x) = -R(\alpha_i|x)$

If $\boldsymbol{g_i(x) > g_j(x), \forall j \neq i}$, we assign class label $\boldsymbol{i}$ to the input data $\boldsymbol{x}$

Nature of the Discriminant Function

For the minimum error rate classifier

We assign the class label corresponding to the maximum posterior probability

If the posterior probability for class cl_i is $P(cl_i|x)$

$g_i(x) \propto P(cl_i|x)$

If $\boldsymbol{g_i(x) > g_j(x), \forall j \neq i}$, we assign class label $\boldsymbol{i}$ to the input data $\boldsymbol{x}$

The Choice of the Discriminant Function

The choice of the discriminator function is not unique

If $g_i(x)$ is a discriminant function and $f(\cdot)$ is a monotonically increasing function, $f(g_i(x))$ is also a valid discriminant function

Why?

If $\boldsymbol{g_i(x) > g_j(x), \forall j \neq i}$, we assign class label $\boldsymbol{i}$ to the input data $\boldsymbol{x}$

The Discriminant Function for Minimum Error Rate Classification

$g_i(x) \propto P(cl_i|x)$

$$\propto \frac{P(x|cl_i)P(cl_i)}{P(x)}$$

Since the denominator is common for every class i

We take
$\boldsymbol{g_i(x) = P(x|cl_i)P(cl_i)}$

If $\boldsymbol{g_i(x) > g_j(x), \forall j \neq i}$, we assign class label $\boldsymbol{i}$ to the input data $\boldsymbol{x}$

The Discriminant Function for Minimum Error Rate Classification

$g_i(x) \propto P(cl_i|x)$

$$\propto \frac{P(x|cl_i)P(cl_i)}{P(x)}$$

Since the denominator is common for every class i

We take
$\boldsymbol{g_i(x) = P(x|cl_i)P(cl_i)}$

We can also take

$$\begin{aligned} \boldsymbol{g_i(x)} &= \mathbf{ln}(\boldsymbol{P(x|cl_i)P(cl_i)}) \\ &= \mathbf{ln}\,\boldsymbol{P(x|cl_i)} + \mathbf{ln}\,\boldsymbol{P(cl_i)} \end{aligned}$$

Why can we take this?

The Discriminant Function for Minimum Error Rate Classification

For a two-class problem, I have two discriminant functions $g_1(x)$ and $g_2(x)$

If $g_1(x) > g_2(x)$, we conclude that x belongs to class 1

If $g_1(x) < g_2(x)$, we conclude that x belongs to class 2

We can also take

$$\begin{aligned} \boldsymbol{g_i(x)} &= \mathbf{ln}(\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{cl_i})\boldsymbol{P}(\boldsymbol{cl_i})) \\ &= \mathbf{ln}\,\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{cl_i}) + \mathbf{ln}\,\boldsymbol{P}(\boldsymbol{cl_i}) \end{aligned}$$

The decision boundary is $g_1(x) = g_2(x)$

The Discriminant Function for Minimum Error Rate Classification

For a two-class problem, I have two discriminant functions $g_1(x)$ and $g_2(x)$

If $g_1(x) > g_2(x)$, we conclude that x belongs to class 1

If $g_1(x) < g_2(x)$, we conclude that x belongs to class 2

So, instead of two discriminant functions $\boldsymbol{g_1(x)}$ and $\boldsymbol{g_2(x)}$, I can take one discriminant function

$$\boldsymbol{g(x) = g_1(x) - g_2(x)}$$

$$g(x) > 0 \;\Rightarrow \text{class } 1$$
$$g(x) < 0 \;\Rightarrow \text{class } 2$$

The Discriminant Function for Minimum Error Rate Classification

So, instead of two discriminant functions $\boldsymbol{g_1(x)}$ and $\boldsymbol{g_2(x)}$, I can take one discriminant function

$$\boldsymbol{g(x) = g_1(x) - g_2(x)}$$

$$g(x) > 0 \ \Rightarrow \text{class } 1$$
$$g(x) < 0 \ \Rightarrow \text{class } 2$$

$$\begin{aligned} g_i(x) &= \ln P(x|cl_i)P(cl_i) \\ &= \ln P(x|cl_i) + \ln P(cl_i) \end{aligned}$$

So,

$$\boldsymbol{g(x) = \ln P(x|cl_1) + \ln P(cl_1) - \ln P(x|cl_2) - \ln P(cl_2)}$$

$$\boldsymbol{= \ln \frac{P(x|cl_1)}{P(x|cl_2)} + \ln \frac{P(cl_1)}{P(cl_2)}}$$

Probability Density and the Discriminant Function

$$g_i(x) = \ln P(x|cl_i)P(cl_i)$$
$$= \ln P(x|cl_i) + \ln P(cl_i)$$

$P(x|cl_i)$ is a probability distribution which can take many forms

For our discussion, assume that $P(x|cl_i)$ follows Gaussian or Normal distribution in $\boldsymbol{d}$ − dimensions

So,

$$\boldsymbol{P(x|cl_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}{2}\right]}$$

Probability Density and the Discriminant Function

$$g_i(x) = \ln P(x|cl_i)P(cl_i)$$
$$= \ln P(x|cl_i) + \ln P(cl_i)$$

$P(x|cl_i)$ is a probability distribution which can take many forms

For our discussion, assume that $P(x|cl_i)$ follows Gaussian or Normal distribution in $\boldsymbol{d}$ − dimensions

So,

$$\boldsymbol{P(x|cl_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}{2}\right]}$$

$\boldsymbol{\mu_i}$: Expected value of $\boldsymbol{x}$ (samples) given that $\boldsymbol{x}$ belongs to class $\boldsymbol{cl_i}$

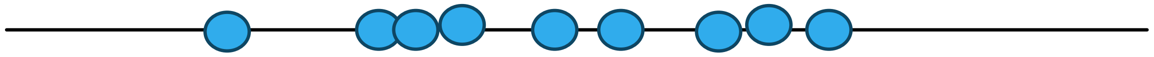
$\boldsymbol{\Sigma_i}$: Covariance matrix computed from $\boldsymbol{x}$ (samples) given that $\boldsymbol{x}$ belongs to class $\boldsymbol{cl_i}$

Variance

- Variance

- $Var(X) = \mathbb{E}[(X - \mu)^2]$

Suppose, I want to measure the mean weights



Weights of Persons

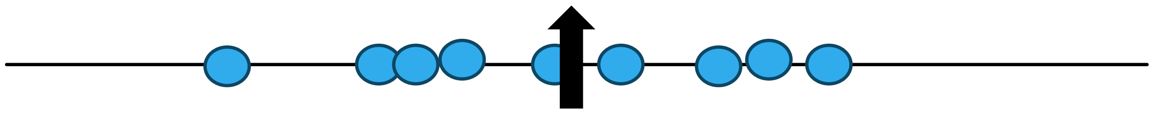
Variance

- Variance

- $Var(X) = \mathbb{E}[(X - \mu)^2]$

Suppose, I want to measure the mean weights

Mean $\mu = \mathbb{E}(X)$



Weights of Persons

Variance

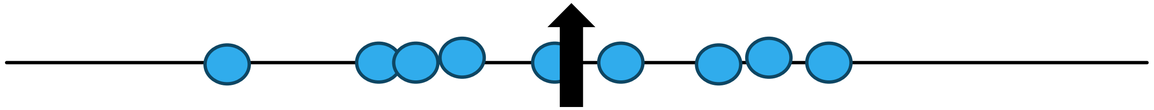
- Variance

- $Var(X) = \mathbb{E}[(X-\mu)^2]$

Mean $\mu = \mathbb{E}(X) = \int xp(x)dx$

For N distinct (equally likely) data points $x_1, \dots, x_N$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$



Weights of Persons

Suppose, I want to measure the mean weights

Variance

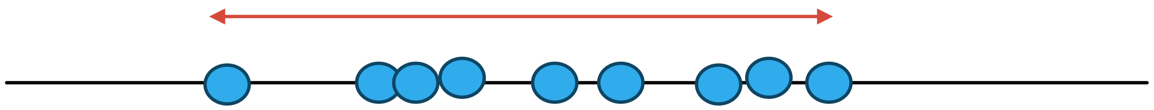
- Variance

- $Var(X) = \mathbb{E}[(X - \mu)^2]$

Variance indicates the spread

For N distinct (equally likely) data points $x_1, \dots, x_N$

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

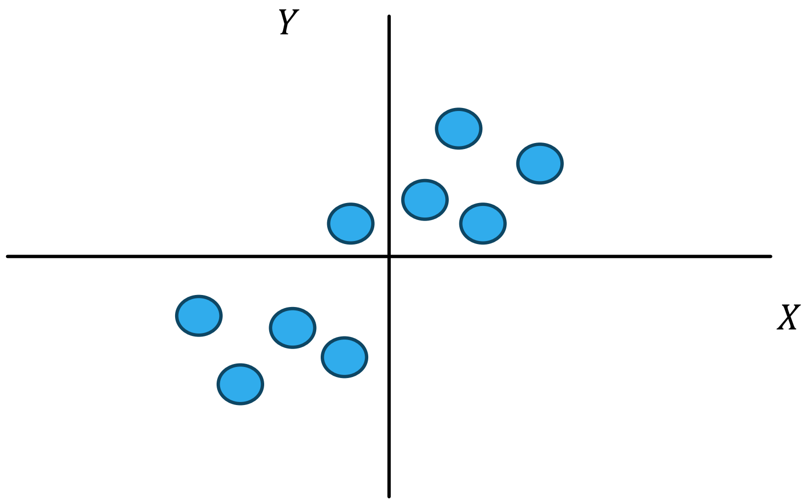


Weights of Persons

Covariance

- Covariance

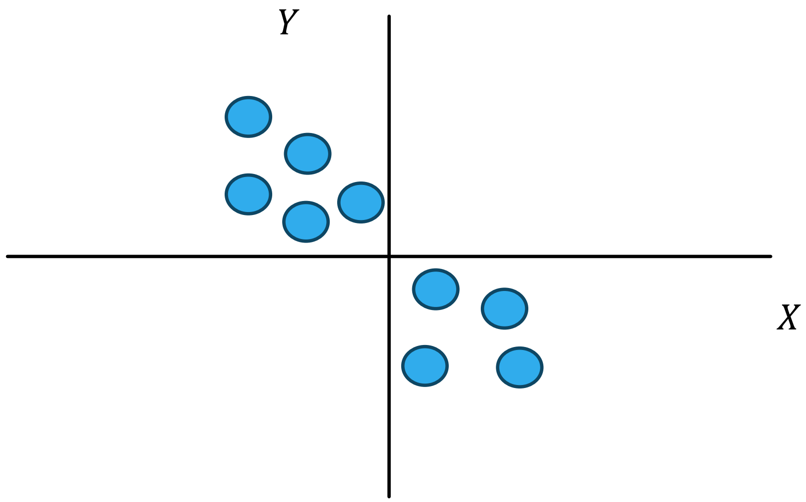
We observe that as X increases, Y also increases



Covariance

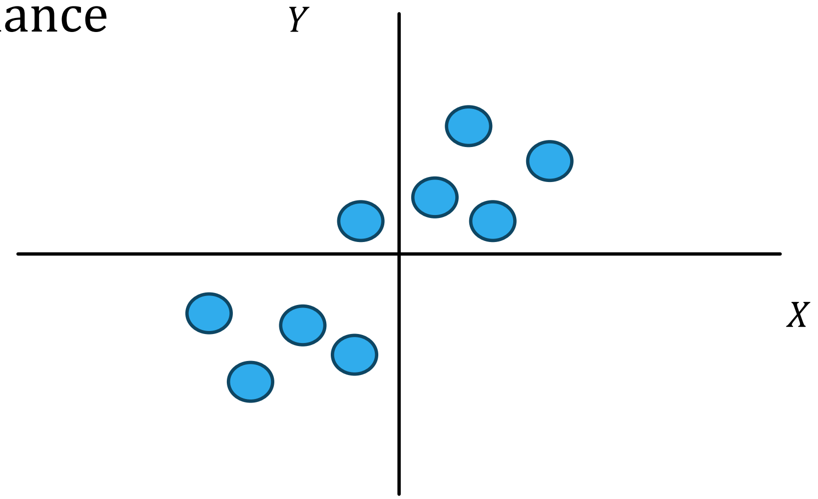
- Covariance

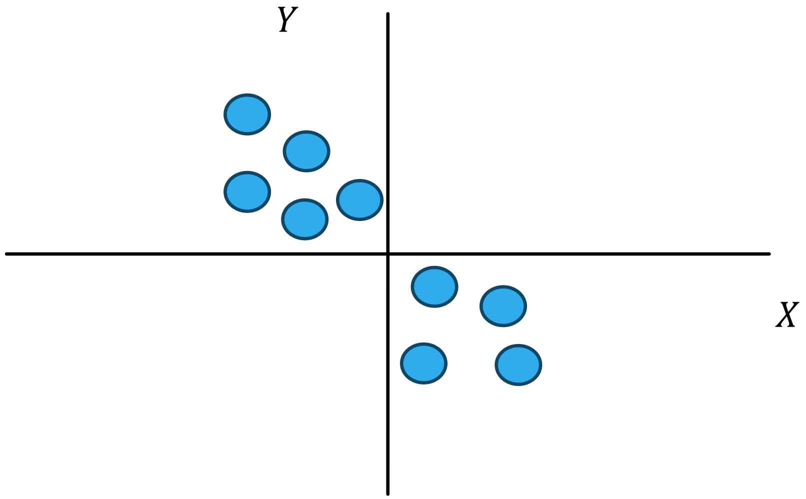
We observe that as X increases, Y decreases



Covariance

- Covariance



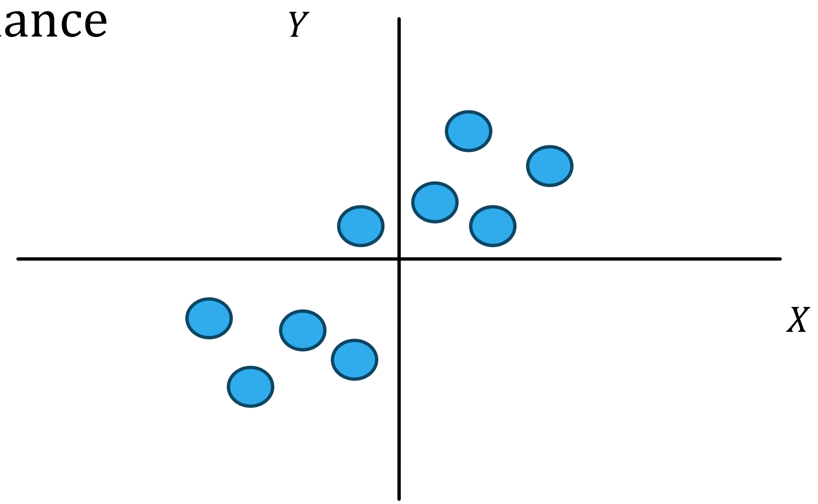


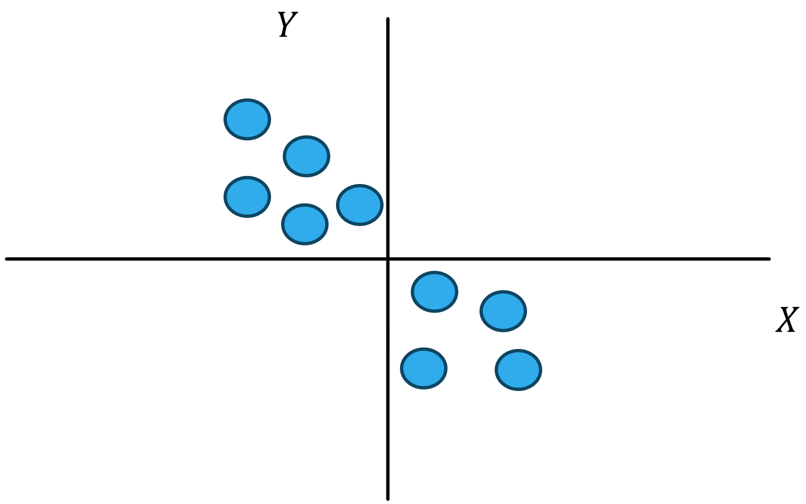
But, in both cases, X variances are almost same

Similar situation happens for Y variance

Covariance

- Covariance





But, in both cases, X variances are almost same

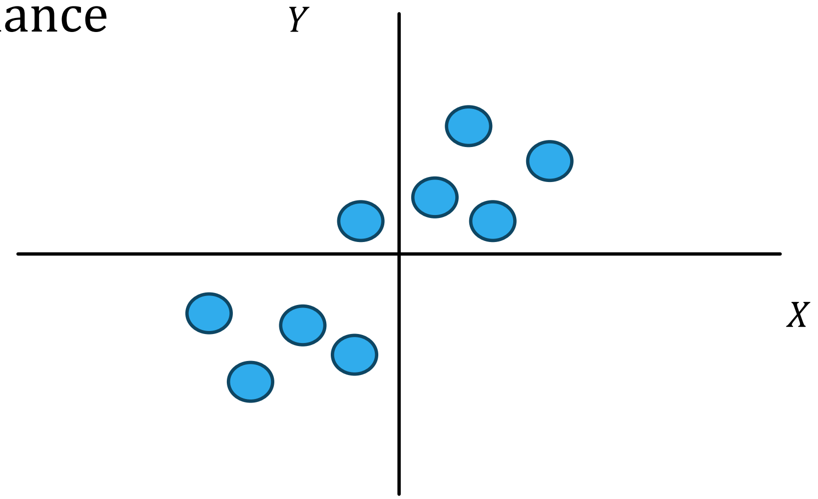
Similar situation happens for Y variance

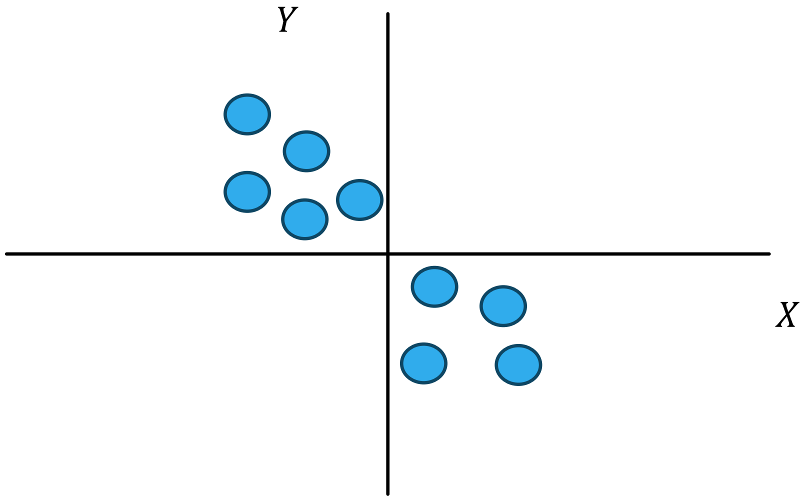
Individual variances do not capture the trend

To differentiate between these distributions, we introduce covariance

Covariance

- Covariance

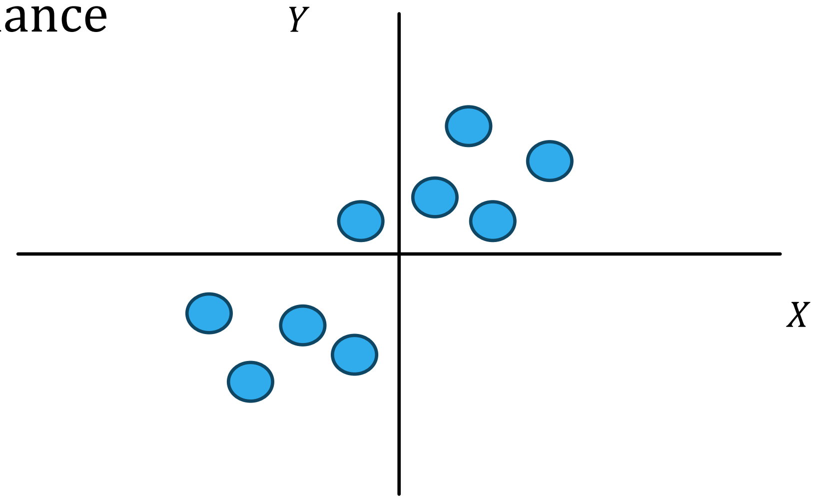


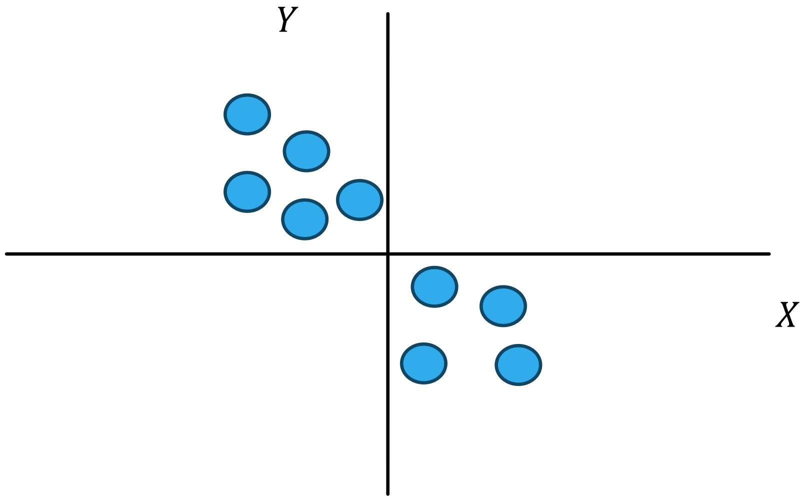


$$cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T]$$

Covariance

- Covariance



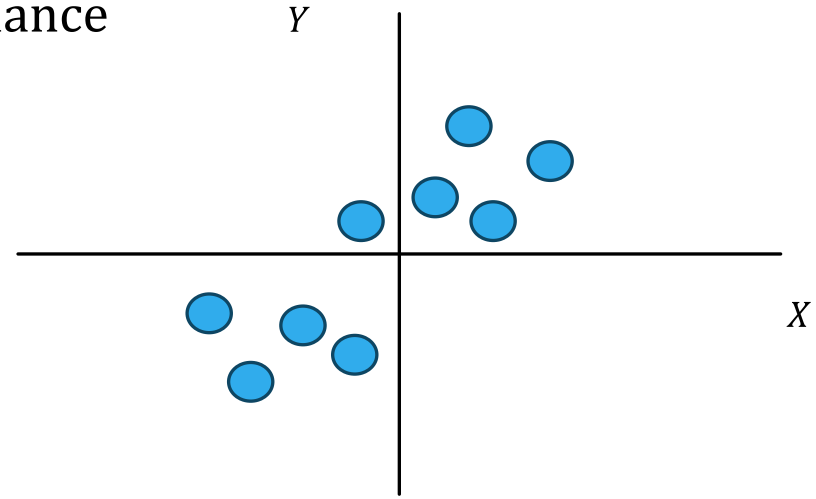


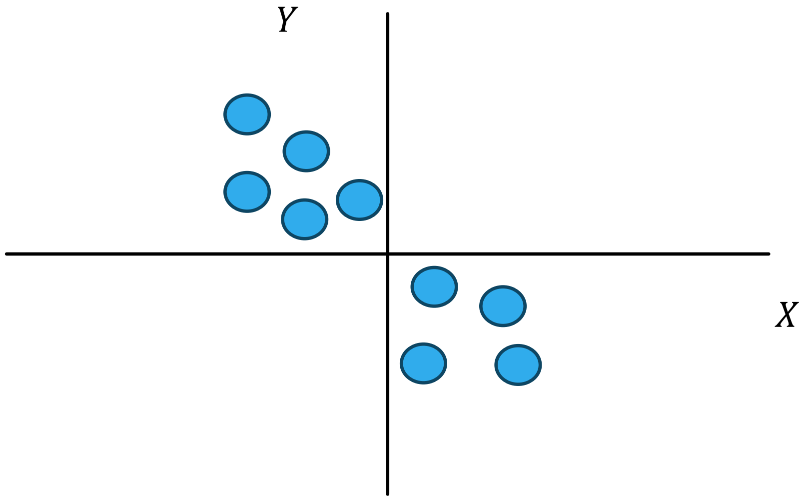
For N distinct (equally likely) data points $(x_1, y_1), \ldots, (x_N, y_N)$

$$cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

Covariance

- Covariance





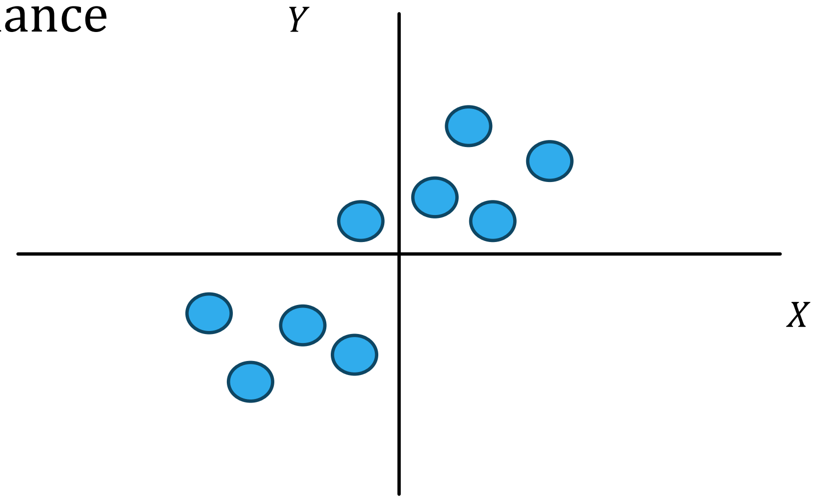
$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

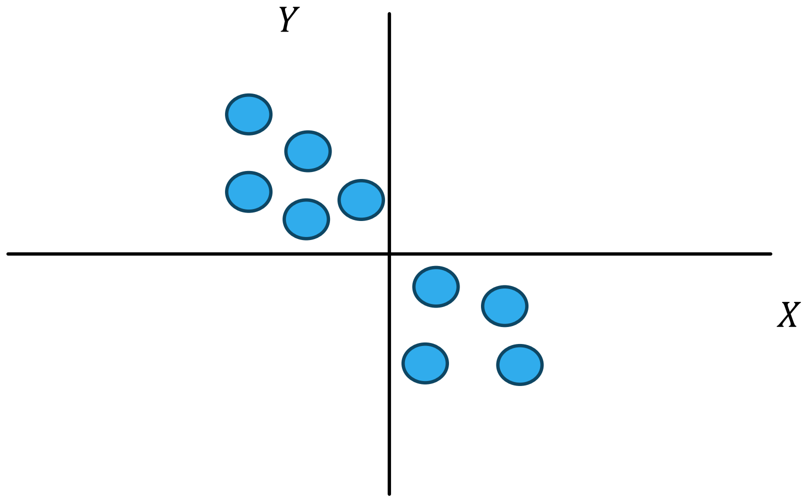
Point-wise product of the differences from the mean

Shows a joint trend of the different variables

Covariance

- Covariance





$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

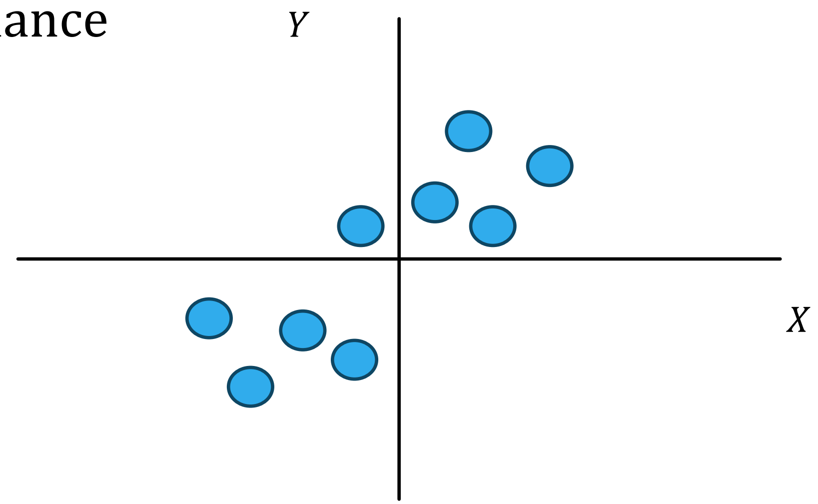
Point-wise product of the differences from the mean

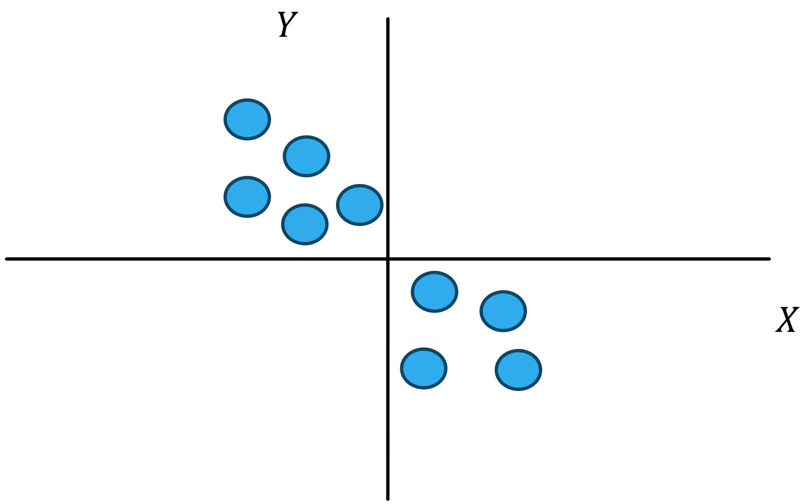
Shows a joint trend of the different variables

In our case, the covariance will be positive for the left and negative for the right example

Covariance Matrix

- Covariance





$$\Sigma = \begin{pmatrix} var(X) & cov(X,Y) \\ cov(X,Y) & var(Y) \end{pmatrix}$$

Covariance Matrix

- Covariance for three random variables X_1, X_2, X_3

$$\Sigma = \begin{pmatrix} Var\ (X_1) & Cov\ (X_1, X_2) & Cov\ (X_1, X_3) \\ Cov\ (X_2, X_1) & Var\ (X_2) & Cov\ (X_2, X_3) \\ Cov\ (X_3, X_1) & Cov\ (X_3, X_2) & Var\ (X_3) \end{pmatrix}$$

Covariance Matrix

- Covariance for three random variables X_1, X_2, X_3

$$\Sigma = \begin{pmatrix} Var\,(X_1) & Cov\,(X_1, X_2) & Cov\,(X_1, X_3) \\ Cov\,(X_2, X_1) & Var\,(X_2) & Cov\,(X_2, X_3) \\ Cov\,(X_3, X_1) & Cov\,(X_3, X_2) & Var\,(X_3) \end{pmatrix}$$

Similar covariance matrices can be constructed for more number of random variables (features in our context)

Probability Density and the Discriminant Function

$$\begin{aligned} g_i(x) &= \ln P(x|cl_i)P(cl_i) \\ &= \ln P(x|cl_i) + \ln P(cl_i) \end{aligned}$$

$$\boldsymbol{P(x|cl_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[-\frac{(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}{2}\right]}$$

It can be shown that

$$\boldsymbol{g_i(x) = -\frac{(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}{2} - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\Sigma_i) + \ln P(cl_i)}$$

Normal Density and the Discriminant Function

$$\boldsymbol{g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\Sigma_i) + \ln P(cl_i)}$$

Parameters: $\boldsymbol{\mu_i}$, $\boldsymbol{\Sigma_i}$

If $\boldsymbol{\mu_i}$, $\boldsymbol{\Sigma_i}$ are given, the Gaussian pdf can be uniquely identified

Normal Density and the Discriminant Function

$$\boldsymbol{g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(\Sigma_i) + \ln P(cl_i)}$$

Independent of class label

So we may ignore this term in
constructing the discriminant function

Normal Density and the Discriminant Function

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln(\Sigma_i) + \ln P(cl_i)$$

Assume

$$\Sigma_i = \sigma^2 I \ \ \forall i$$

In 3D

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Only the diagonal elements (variance terms) are non zero

Off-diagonal elements (covariance terms) are zero

It means that the features of the data are statistically independent of each other, i.e., no pair of features show any trend

Normal Density and the Discriminant Function

$$g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) - \frac{1}{2}\ln(\Sigma_i) + \ln P(cl_i)$$

$$\Sigma_i = \sigma^2 I \ \ \forall i$$

$|\Sigma_i| = \sigma^{2d}$ **(assuming I to be a $d \times d$ identity matrix)**

$$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$$

Normal Density and the Discriminant Function

$$\boldsymbol{g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln\left(\sigma^2 I\right) + \ln P(cl_i)}$$

Assume

$$\boldsymbol{\Sigma_i = \sigma^2 I \;\; \forall i}$$

Independent of class label

So we may ignore this term in
constructing the discriminant function

Normal Density and the Discriminant Function

$$\boldsymbol{g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln P(cl_i)}$$

$$\boldsymbol{\Sigma_i^{-1} = \frac{1}{\sigma^2} I}$$

Normal Density and the Discriminant Function

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln P(cl_i)$$

$$\Sigma_i^{-1} = \frac{1}{\sigma^2} I$$

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(cl_i)$$

Normal Density and the Discriminant Function

$$
\begin{aligned}
g_i(x) &= -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(cl_i) \\
&= -\frac{1}{2\sigma^2}\left[x^T x - 2\mu_i^T x + \mu_i^T \mu_i\right] + \ln P(cl_i)
\end{aligned}
$$

Normal Density and the Discriminant Function

$$\boldsymbol{g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(cl_i)}$$

$$\boldsymbol{= -\frac{1}{2\sigma^2}\left[x^T x - 2\mu_i^T x + \mu_i^T \mu_i\right] + \ln P(cl_i)}$$

Independent of class label

So we may ignore this term in constructing the discriminant function

Normal Density and the Discriminant Function

$$g_i(x) = -\frac{1}{2\sigma^2}\left[-2\mu_i^T x + \mu_i^T \mu_i\right] + \ln P(cl_i)$$

$$= \frac{\mu_i^T}{\sigma^2} x + \ln P(cl_i) - \frac{1}{2\sigma^2}\mu_i^T \mu_i$$

$$= w_i^T x + w_{i0}$$

$$w_i^T = \frac{\mu_i^T}{\sigma^2}$$

$$w_{i0} = \ln P(cl_i) - \frac{1}{2\sigma^2}\mu_i^T \mu_i$$

Normal Density and the Discriminant Function

$$\boldsymbol{g_i(x) = w_i^T x + w_{i0}}$$

The discriminant function is a linear function of the input

This is called linear machine

Naïve Bayes': An Example

Sample Number	Appetite	Weight	BP	Class
1	Low	Normal	Low	No Anemia
2	Low	Low	Low	Anemia
3	Normal	Low	Low	Anemia
4	Low	Low	Normal	No Anemia
5	Normal	Low	Normal	Anemia
6	Normal	Normal	Low	Anemia
7	Normal	Normal	Normal	No Anemia

Training data

Naïve Bayes': An Example

$$P(x_1|cl_1)P(x_2|cl_1)\dots P(x_d|cl_1)P(cl_1)$$

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$P(anemia) = \frac{4}{7}$ $\quad P(no\ anemia) = \frac{3}{7}$

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **normal** weight, and **low** BP

Predict if the person has anemia

Let's first evaluate the chance of anemia

$$\pi_{anemia} = P(appetite = normal|anemia)P(weight = normal|anemia)P(BP = low|anemia)P(anemia)$$

$$= \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{7} = 0.009$$

Naïve Bayes': An Example

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$P(anemia) = \frac{4}{7}$ $\quad P(no\ anemia) = \frac{3}{7}$

$$P(x_1|cl_2)P(x_2|cl_2)\ldots P(x_d|cl_2)P(cl_2)$$

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **normal** weight, and **low** BP

Predict if the person has anemia

Let's first evaluate the chance of no anemia

$$\pi_{no\ anemia} = P(appetite = normal|no\ anemia)P(weight = normal|no\ anemia)P(BP = low|no\ anemia)P(no\ anemia)$$

$$= \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{3}{7} = 0.06$$

Naïve Bayes': An Example

$$\pi_{anemia} = P(appetite = normal|anemia)P(weight = normal|anemia)P(BP = low|anemia)P(anemia)$$
$$= \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{7} = 0.009$$

$$\pi_{no\ anemia} = P(appetite = normal|no\ anemia)P(weight = normal|no\ anemia)P(BP = low|no\ anemia)P(no\ anemia)$$
$$= \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{3}{7} = 0.06$$

$$P(x_1|cl_1)P(x_2|cl_1) \ldots P(x_d|cl_1)P(cl_1) > P(x_1|cl_2)P(x_2|cl_2) \ldots P(x_d|cl_2)P(cl_2) \Rightarrow cl_1$$

So, our conclusion is that the person does not have anemia as per the Naïve Bayes' classifier

Naïve Bayes': A Slightly Different Example

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Training data

Naïve Bayes': An Example

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Step 1: Calculate $P(anemia), P(no\ anemia)$ from the training data

Naïve Bayes': An Example

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Step 1: Calculate $P(anemia), P(no\ anemia)$ from the training data

$$P(anemia) = \frac{4}{7}$$

$$P(no\ anemia) = \frac{3}{7}$$

Naïve Bayes': An Example

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$$P(Appetite = low|Anemia), \qquad P(Appetite = normal|Anemia)$$

$$P(Appetite = low|No\ Anemia), \qquad P(Appetite = normal|No\ Anemia)$$

All these can be calculated

Naïve Bayes': An Example

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

$$P(BP = low|Anemia), \qquad P(BP = normal|Anemia)$$

$$P(BP = low|No\ Anemia), \qquad P(BP = normal|No\ Anemia)$$

All these can be calculated

Now What?

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

What should we put in the blank boxes?

		Anemia	No Anemia
Weight			

Now What?

Sample Number	Appetite	Weight (Kg)	BP	Class
1	Low	61	Low	No Anemia
2	Low	49	Low	Anemia
3	Normal	47	Low	Anemia
4	Low	51	Normal	No Anemia
5	Normal	50	Normal	Anemia
6	Normal	63	Low	Anemia
7	Normal	58	Normal	No Anemia

Step 2: Calculate $P(feature|Anemia), P(feature|No\ Anemia)$ from the training data

What should we put in the blank boxes? Do we want to calculate this

		Anemia	No Anemia
Weight	61	0	1
	49	1	0
	47	1	0
	51	0	1
	50	1	0
	63	1	0
	58	0	1

Naïve Bayes': An Example

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	61	0	1
	49	1	0
	47	1	0
	51	0	1
	50	1	0
	63	1	0
	58	0	1

$P(anemia) = \frac{4}{7}$ $\quad P(no\ anemia) = \frac{3}{7}$

Step 3: Testing:

Suppose, I observe a test data with normal appetite, 56 Kg weight, and low BP

Predict if the person has anemia

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	61	0	1
	49	1	0
	47	1	0
	51	0	1
	50	1	0
	63	1	0
	58	0	1

$P(anemia) = \frac{4}{7}$ $\quad P(no\ anemia) = \frac{3}{7}$

$$P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1)$$

Naïve Bayes': An Example

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **56 Kg** weight, and **low** BP.
Predict if the person has anemia

Let's first evaluate the chance of anemia

$$\pi_{anemia} = P(appetite = normal|anemia)P(weight = 56\ Kg|anemia)P(BP = low|anemia)P(anemia)$$

Naïve Bayes': An Example

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	61	0	1
	49	1	0
	47	1	0
	51	0	1
	50	1	0
	63	1	0
	58	0	1

$P(anemia) = \frac{4}{7}$ $\quad P(no\ anemia) = \frac{3}{7}$

$$P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1)$$

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **56 Kg** weight, and **low** BP.
Predict if the person has anemia

Let's first evaluate the chance of anemia

$$\pi_{anemia} = P(appetite = normal|anemia)P(weight = 56\ Kg|anemia)P(BP = low|anemia)P(anemia)$$

$$= \frac{1}{4} \times ? \times \frac{1}{4} \times \frac{4}{7}$$

From my training data, I can't find $P(weight = 56\ Kg|anemia)$

If I just look at the table, $P(weight = 56\ Kg|anemia)$ does not have an entry

Naïve Bayes': An Example

		Anemia	No Anemia
BP	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Appetite	Low	$\frac{3}{4}$	$\frac{1}{3}$
	Normal	$\frac{1}{4}$	$\frac{2}{3}$

		Anemia	No Anemia
Weight	61	0	1
	49	1	0
	47	1	0
	51	0	1
	50	1	0
	63	1	0
	58	0	1

$P(anemia) = \frac{4}{7}$ $\quad$ $P(no\ anemia) = \frac{3}{7}$

$$P(x_1|cl_1)P(x_2|cl_1)\ldots P(x_d|cl_1)P(cl_1)$$

Step 3: Testing:

Suppose, I observe a test data with **normal** appetite, **56 Kg** weight, and **low** BP.

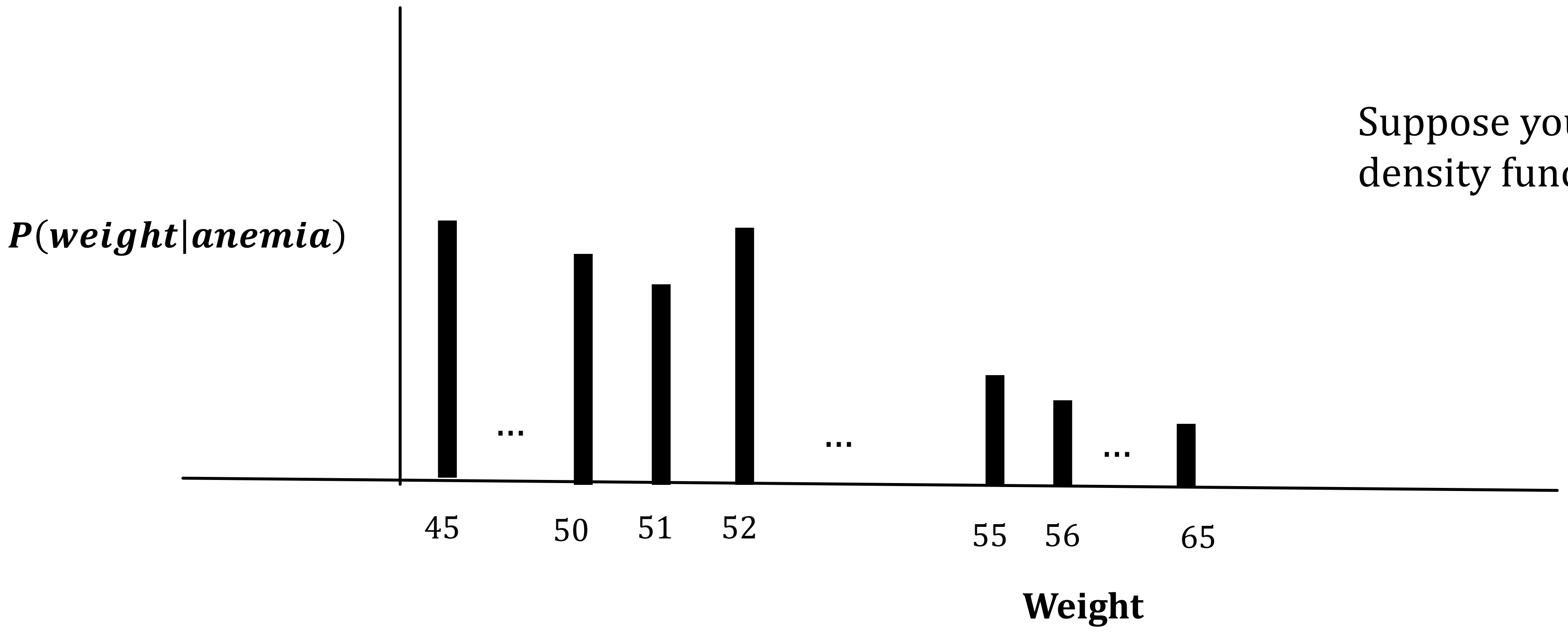
$$\pi_{anemia} =$$
$$P(appetite = normal|anemia)P(weight = 56\ Kg|anemia)P(BP = low|anemia)P(anemia)$$
$$= \frac{1}{4} \times ? \times \frac{1}{4} \times \frac{4}{7}$$

From my training data, I can't find $P(weight = 56\ Kg|anemia)$

This is a problem when we have real numbers as features

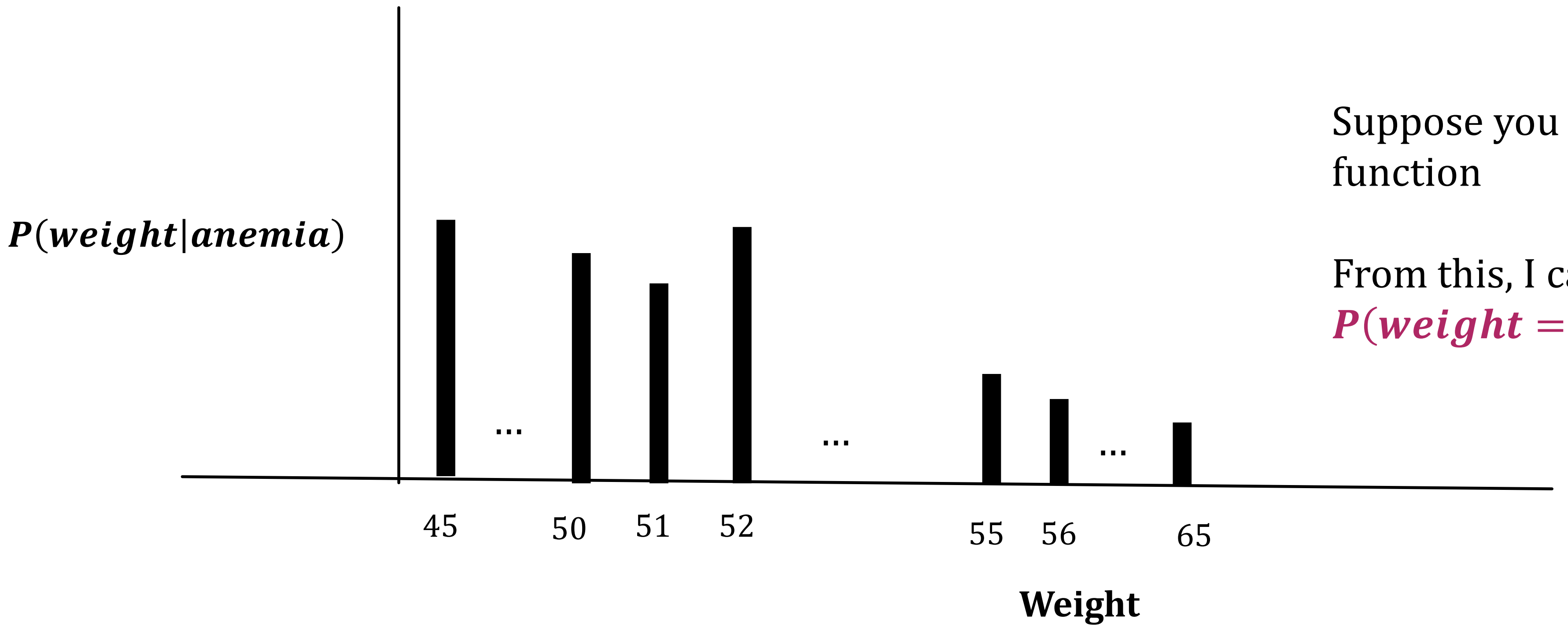
What is the way out? Let's see

What If You Can Find This?



Suppose you find the probability density function

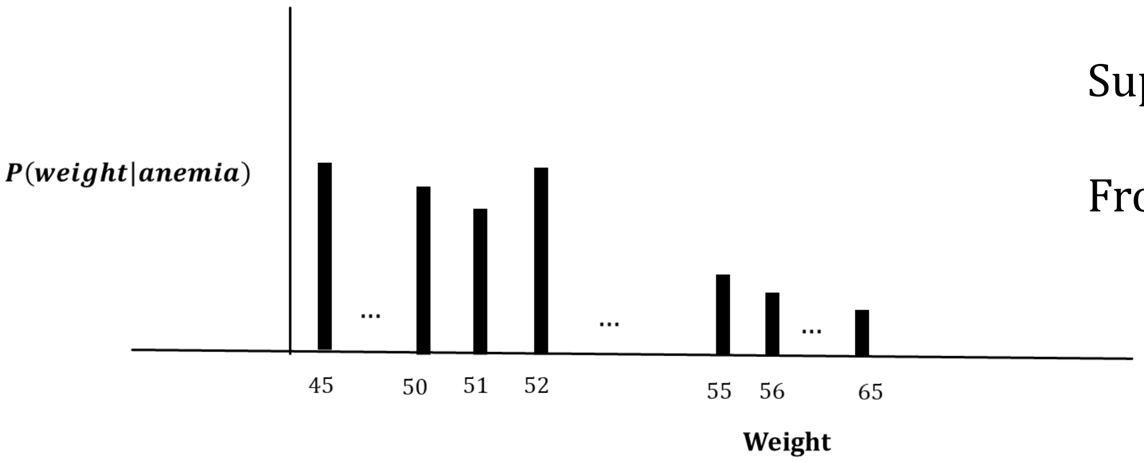
What If You Can Find This?



Suppose you find the probability density function

From this, I can find
$P(weight = 56\ Kg|anemia)$

What If You Can Find This?



Suppose you find the probability density function

From this, I can find $\boldsymbol{P(weight = 56\ Kg|anemia)}$

The question is, how to find the probability density from training data so that we can estimate the likelihood $\boldsymbol{P(weight = 56\ Kg|anemia)}$?

$$P(cl_1|x) = \frac{P(x|cl_1)P(cl_1)}{P(x)}$$

Maximum Likelihood Estimation

Suppose, I have C number of classes $cl_1, cl_2, \ldots, cl_C$ in my dataset

S_1: Training samples of class cl_1
S_2: Training samples of class cl_2
…
S_C: Training samples of class cl_C

I have to find $\boldsymbol{P(x|cl_i)}$ such that $\boldsymbol{P(x|cl_i)}$ is maximized when I use the training data of class cl_i

We assume that $\boldsymbol{P(x|cl_i)}$ has a known parametric form

If $P(x|cl_i)$ Gaussian, it is completely specified by mean μ_i and covariance matrix Σ_i (call them parameters θ_i of the distribution)

So, if I can find out parameters θ_i, I can find out $P(x|cl_i)$

Maximum Likelihood Estimation

Suppose, I have C number of classes $cl_1, cl_2, \dots, cl_C$ in my dataset

S_1: Training samples of class cl_1
S_2: Training samples of class cl_2
…
S_C: Training samples of class cl_C

I have to find $\boldsymbol{P(x|cl_i)}$ such that $\boldsymbol{P(x|cl_i)}$ is maximized when I use the training data of class cl_i

We assume that $\boldsymbol{P(x|cl_i)}$ has a known parametric form

If $P(x|cl_i)$ Gaussian, it is completely specified by mean μ_i and covariance matrix Σ_i (call them parameters θ_i of the distribution)

So, if I can find out parameters θ_i, I can find out $P(x|cl_i)$

So our goal is to use information from training sample in S_i to find parameters θ_i such that $\boldsymbol{P(x|cl_i)}$ is maximized when I use the training data of class cl_i

In this, we assume that information in S_j does not affect θ_i if $i \neq j$

Maximum Likelihood Estimation

Our goal is to use information from training sample in S_i to find parameters θ_i such that $\boldsymbol{P(x|cl_i)}$ is maximized when I use the training data of class cl_i

So, we have to maximize $\boldsymbol{P(x|cl_i, \theta_i)\ \forall i}$

Since $\boldsymbol{\theta_i}$ are the parameters corresponding to $\boldsymbol{cl_i}$ only, we can say we have to maximize $\boldsymbol{P(x|\theta_i)\ \forall i}$

In this, we assume that information in S_j does not affect θ_i if $i \neq j$

So, even if I consider the entire training dataset S instead of just S_i, the parameters θ_i will only be influenced by S_i

Maximum Likelihood Estimation

We can say we have to find $\boldsymbol{\theta_i}$ that maximizes $\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{\theta_i}\,)\ \forall \boldsymbol{i}$

In this, we assume that information in S_j does not affect θ_i if $i \neq j$

So, even if I consider the entire training dataset S instead of just S_i, the parameters θ_i will only be influenced by S_i

So, we can say that we have to find $\boldsymbol{\theta}$ that maximizes $\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{\theta}\,)$

Maximum Likelihood Estimation

So, we can say that we have to find $\boldsymbol{\theta}$ that maximizes $\boldsymbol{P(x|\theta\ )}$

If I have N training samples $x_1, x_2, \ldots, x_N$, we want to maximize the likelihood of each of the training samples

So, we want to maximize $\boldsymbol{P(x_1|\theta\ ), P(x_2|\theta\ ), \ldots, P(x_N|\theta\ )}$

That means, we have to find $\boldsymbol{\theta}$ that maximizes the product

$$\boldsymbol{P(x_1|\theta\ )P(x_2|\theta\ )\ldots P(x_N|\theta\ ) = \prod_{k=1}^{N} P(x_k|\theta\ )}$$

An Example

Suppose, in the classroom, there are

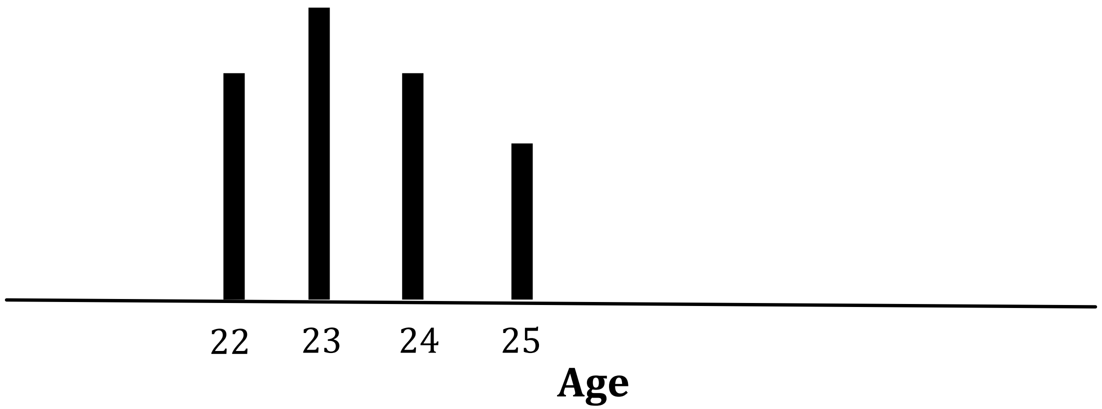
30 students of age 22

35 students of age 23

30 students of age 24

25 students of age 25

Training data



An Example

Suppose, in the classroom, there are

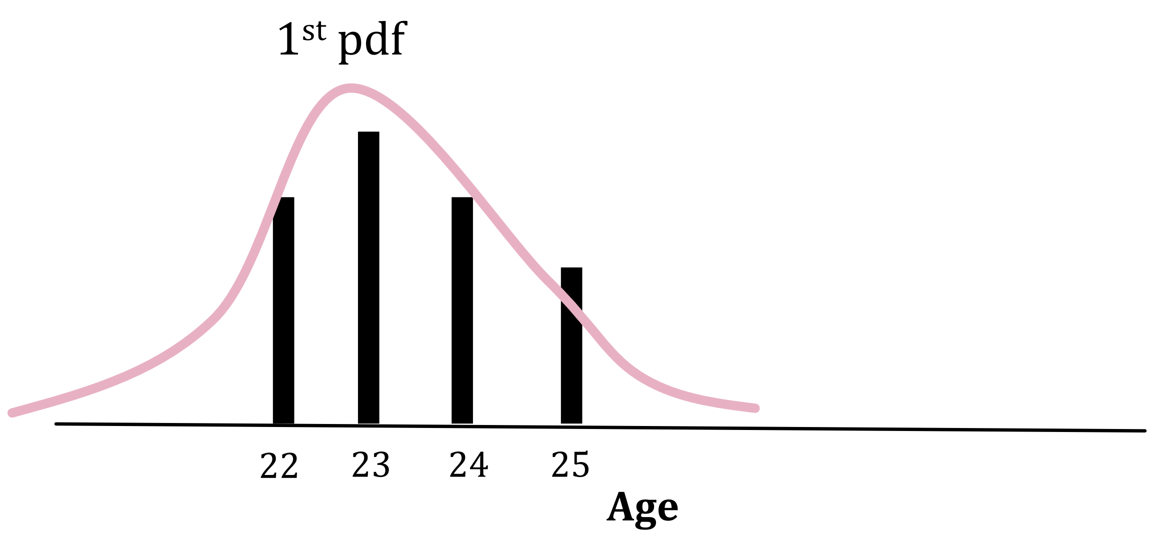
30 students of age 22

35 students of age 23

30 students of age 24

25 students of age 25

Training data



An Example

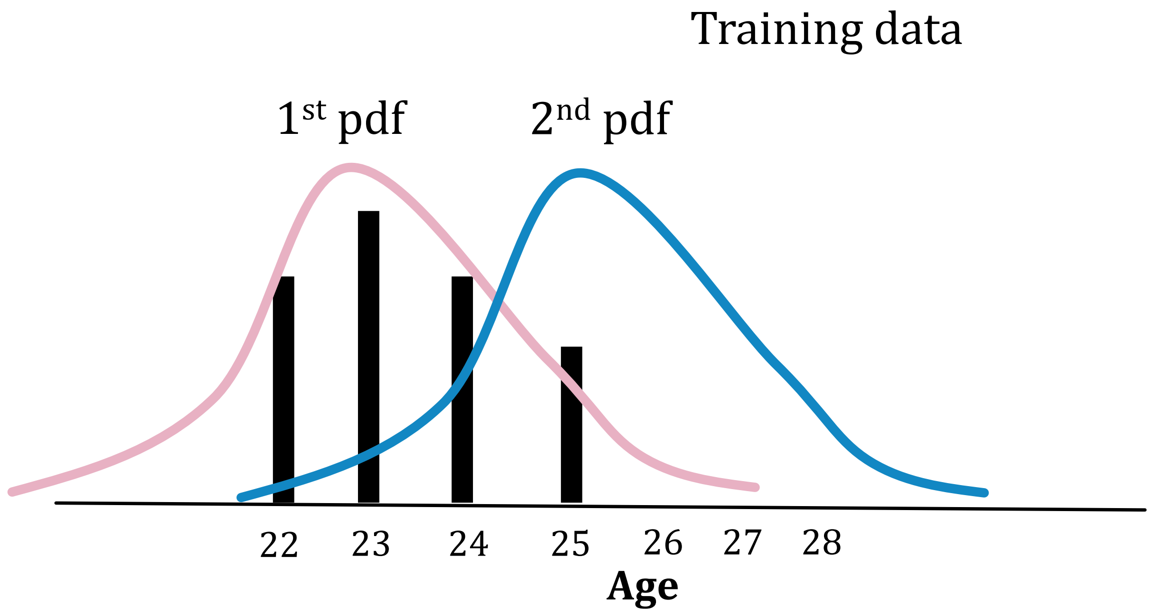
Suppose, in the classroom, there are

30 students of age 22

35 students of age 23

30 students of age 24

25 students of age 25



An Example

Training data ***S***

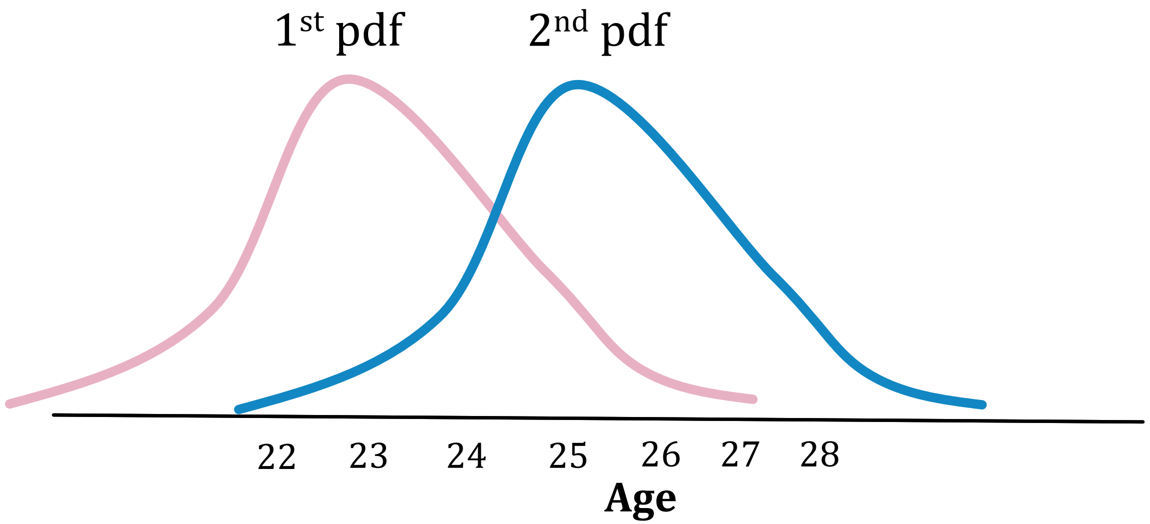
Suppose, in the classroom, there are

30 students of age 22

35 students of age 23

30 students of age 24

25 students of age 25



$P(age = 22|S)P(age = 23|S)P(age = 24|S)P(age = 25|S)$ is maximum for which curve?

An Example

Suppose, in the classroom, there are

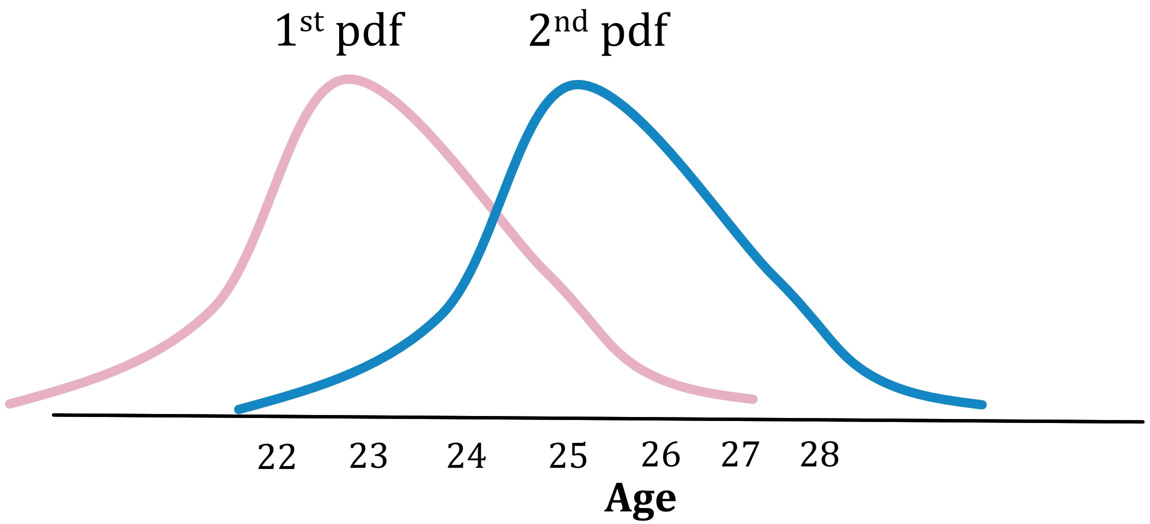
30 students of age 22

35 students of age 23

30 students of age 24

25 students of age 25

Training data ***S***



$P(age = 22|S)P(age = 23|S)P(age = 24|S)P(age = 25|S)$ is maximum for which curve? **Red**

So, the parameters corresponding to the red curve gives me the higher likelihood of the training data among the two curves

Maximum Likelihood Estimation

We have to find $\boldsymbol{\theta}$ that maximizes the product

$$\boldsymbol{P}(\boldsymbol{x}_1|\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{x}_2|\boldsymbol{\theta})\ldots\boldsymbol{P}(\boldsymbol{x}_N|\boldsymbol{\theta}) = \prod_{\boldsymbol{k}=1}^{\boldsymbol{N}}\boldsymbol{P}(\boldsymbol{x}_{\boldsymbol{k}}|\boldsymbol{\theta})$$

It is equivalent to finding $\boldsymbol{\theta}$ that maximizes

$$l(\theta) = \ln\left(\prod_{\boldsymbol{k}=1}^{\boldsymbol{N}}\boldsymbol{P}(\boldsymbol{x}_{\boldsymbol{k}}|\boldsymbol{\theta})\right) = \sum_{\boldsymbol{k}=1}^{\boldsymbol{N}}\ln P(x_k|\theta)$$

Maximum Likelihood Estimation

To maximize, find gradient of $l(\theta)$ and equate to zero

$$\nabla_\theta l(\theta) = 0$$

But this may give me max or min
So, we have to check the second derivative

We may also have multiple maxima
So, we have to find the highest maxima

Maximum Likelihood Estimation

To maximize, find gradient of $l(\theta)$ and equate to zero
$\nabla_\theta l(\theta) = 0$

If I have m number of parameters $\theta_1, \theta_2, \dots, \theta_m$, we have to do

$$\begin{bmatrix} \frac{\partial l(\theta)}{\partial \theta_1} \\ \frac{\partial l(\theta)}{\partial \theta_2} \\ \dots \\ \dots \\ \frac{\partial l(\theta)}{\partial \theta_m} \end{bmatrix} = 0$$

Maximum Likelihood Estimation

To maximize, find gradient of $l(\theta)$ and equate to zero
$\nabla_\theta l(\theta) = 0$

If I have m number of parameters $\theta_1, \theta_2, \ldots, \theta_m$, we have to do

$$\frac{\partial l(\theta)}{\partial \theta_1} = 0$$
$$\frac{\partial l(\theta)}{\partial \theta_2} = 0$$

$$\ldots$$

$$\frac{\partial l(\theta)}{\partial \theta_m} = 0$$

Maximum Likelihood Estimation: Example

Let's assume that the pdf has a Gaussian distribution with mean μ and covariance matrix Σ

Assume that Σ is known, we have to find μ

$$P(x_k|\mu\ ) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)}{2}\right]$$

$Take\ \mu = \theta$ (parameter to be found)

$$\ln P(x_k|\theta\ ) = -\frac{\mathbf{1}}{\mathbf{2}}\ln(2\pi)^d|\Sigma| - \frac{\mathbf{1}}{\mathbf{2}}(x_k-\theta)^T\Sigma^{-1}(x_k-\theta)$$

Maximum Likelihood Estimation: Example

$$P(x_k|\mu\,) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)}{2}\right]$$

Take $\mu = \theta$ (parameter to be found)

$$P(x_k|\theta) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left[-\frac{(x_k-\theta)^T\Sigma^{-1}(x_k-\theta)}{2}\right]$$

$$\ln P(x_k|\theta\,) = -\frac{1}{2}\ln(2\pi)^d|\Sigma| - \frac{1}{2}(x_k-\theta)^T\Sigma^{-1}(x_k-\theta)$$

Maximum Likelihood Estimation: Example

$$l_k(\theta) = \ln P(x_k|\theta\ ) = -\frac{1}{2}\ln(2\pi)^d|\Sigma| - \frac{1}{2}(x_k - \theta)^T\Sigma^{-1}(x_k - \theta)$$

$$l(\theta) = \sum_{k=1}^{N} \ln P(x_k|\theta\ ) = \sum_{k=1}^{N} l_k(\theta)$$

It can be shown
$\nabla_\theta l_k(\theta) = \Sigma^{-1}(x_k - \theta)$

Maximum Likelihood Estimation: Example

$$l(\theta) = \sum_{k=1}^{N} \ln P(x_k|\theta\ ) = \sum_{k=1}^{N} l_k(\theta)$$

It can be shown

$\nabla_\theta l_k(\theta) = \Sigma^{-1}(x_k - \theta)$

$$\nabla_\theta l(\theta) = \sum_{k=1}^{N} \nabla_\theta l_k(\theta) = \sum_{k=1}^{N} \Sigma^{-1}(x_k - \theta)$$

Maximum Likelihood Estimation: Example

For MLE, we make

$$\nabla_\theta l(\theta) = \sum_{k=1}^{N} \nabla_\theta l_k(\theta) = \sum_{k=1}^{N} \Sigma^{-1}(x_k - \theta) = 0$$

Since $\Sigma^{-1} \neq 0$

$$\sum_{k=1}^{N} (x_k - \theta) = 0$$

$$\boldsymbol{\left(\sum_{k=1}^{N} x_k\right) - N\theta = 0}$$

$$\boldsymbol{\theta = \frac{\sum_{k=1}^{N} x_k}{N}}$$

Maximum Likelihood Estimation: Another Example

Consider a univariate Gaussian with unknown μ and σ^2

We have to find the values of μ and σ^2 for MLE

$$P(x_k|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$$

Maximum Likelihood Estimation: Another Example

$$P(x_k|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2}\right]$$

Assuming $\theta_1 = \mu, and\ \theta_2 = \sigma^2$, We write

$$P(x_k|\theta) = \frac{1}{(2\pi\theta_2)^{\frac{1}{2}}}\exp\left[-\frac{(x_k-\theta_1)^2}{2\theta_2}\right]$$

Maximum Likelihood Estimation: Another Example

$$P(x_k|\theta) = \frac{1}{(2\pi\theta_2)^{\frac{1}{2}}} \exp\left[-\frac{(x_k - \theta_1)^2}{2\theta_2}\right]$$

$$\nabla_\theta l_k(\theta) = \begin{bmatrix} \frac{\partial l_k(\theta)}{\partial \theta_1} \\ \frac{\partial l_k(\theta)}{\partial \theta_2} \end{bmatrix}$$

$$l_k(\theta) = \ln P(x_k|\theta) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

Maximum Likelihood Estimation: Another Example

$$l_k(\theta) = \ln P(x_k|\theta) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l_k(\theta) = \begin{bmatrix} \dfrac{\partial l_k(\theta)}{\partial \theta_1} \\ \dfrac{\partial l_k(\theta)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{1}{2\theta_2^2}(x_k - \theta_1)^2 \end{bmatrix}$$

Maximum Likelihood Estimation: Another Example

$$\nabla_\theta l_k(\theta) = \begin{bmatrix} \dfrac{\partial l_k(\theta)}{\partial \theta_1} \\ \dfrac{\partial l_k(\theta)}{\partial \theta_2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\theta_2}(x_k - \theta_1) \\ -\dfrac{1}{2\theta_2} + \dfrac{1}{2\theta_2^2}(x_k - \theta_1)^2 \end{bmatrix}$$

$$\sum_{k=1}^{N} \frac{1}{\theta_2}(x_k - \theta_1) = 0$$

For MLE, we make

$$\sum_{k=1}^{N} \nabla_\theta l_k(\theta) = 0$$

$$\sum_{k=1}^{N} \left(-\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2 \right) = 0$$

Maximum Likelihood Estimation: Another Example

$$\sum_{k=1}^{N} \frac{1}{\theta_2}(x_k - \theta_1) = 0$$

$$\sum_{k=1}^{N} \left(-\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x_k - \theta_1)^2 \right) = 0$$

Assuming $\theta_2 \neq 0$

$$\boldsymbol{\theta_1 = \frac{\sum_{k=1}^{N} x_k}{N}}$$

$$\boldsymbol{\theta_2 = \frac{\sum_{k=1}^{N} (x_k - \theta_1)^2}{N}}$$

This is the estimate of μ

This is the estimate of σ^2

Maximum Likelihood Estimation: Home Assignment

$$P(x_k|\theta) = \begin{cases} \theta \exp(-\theta x_k) & \text{if } \; x_k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$