

Data Science and Artificial Intelligence

Machine Learning



Bayesian learning

Lecture No. 6

By- SIDDHARTH SABHARWAL SIR



Recap of Previous Lecture



Topic

Naive Bayes.

Topic

Topic

Topic

Topic

Topics to be Covered



Topic

Naive Bayes-theory

Topic

Question

Topic

Decisiontree

Topic

Topic

STOP DOUBTING
YOURSELF.
WORK HARD AND
MAKE IT HAPPEN.

1. What type of algorithm is Naive Bayes used for in machine learning?
- a. Classification
 - b. Regression ~~x~~
 - c. Clustering
 - d. Reinforcement learning

3. What is the "naive" assumption in Naive Bayes?

a. It assumes that all features are equally important.

b. It assumes that features are independent of each other, given the Class.

c. It assumes that the dataset is small.

d. It assumes that features are dependent on each other.

9. In the context of Naive Bayes, what is Laplace smoothing (additive smoothing) used for?

- ☒ a. Reducing the impact of rare features ✓
- b. Increasing the model's complexity
- c. Decreasing the training time
- d. Ignoring missing data

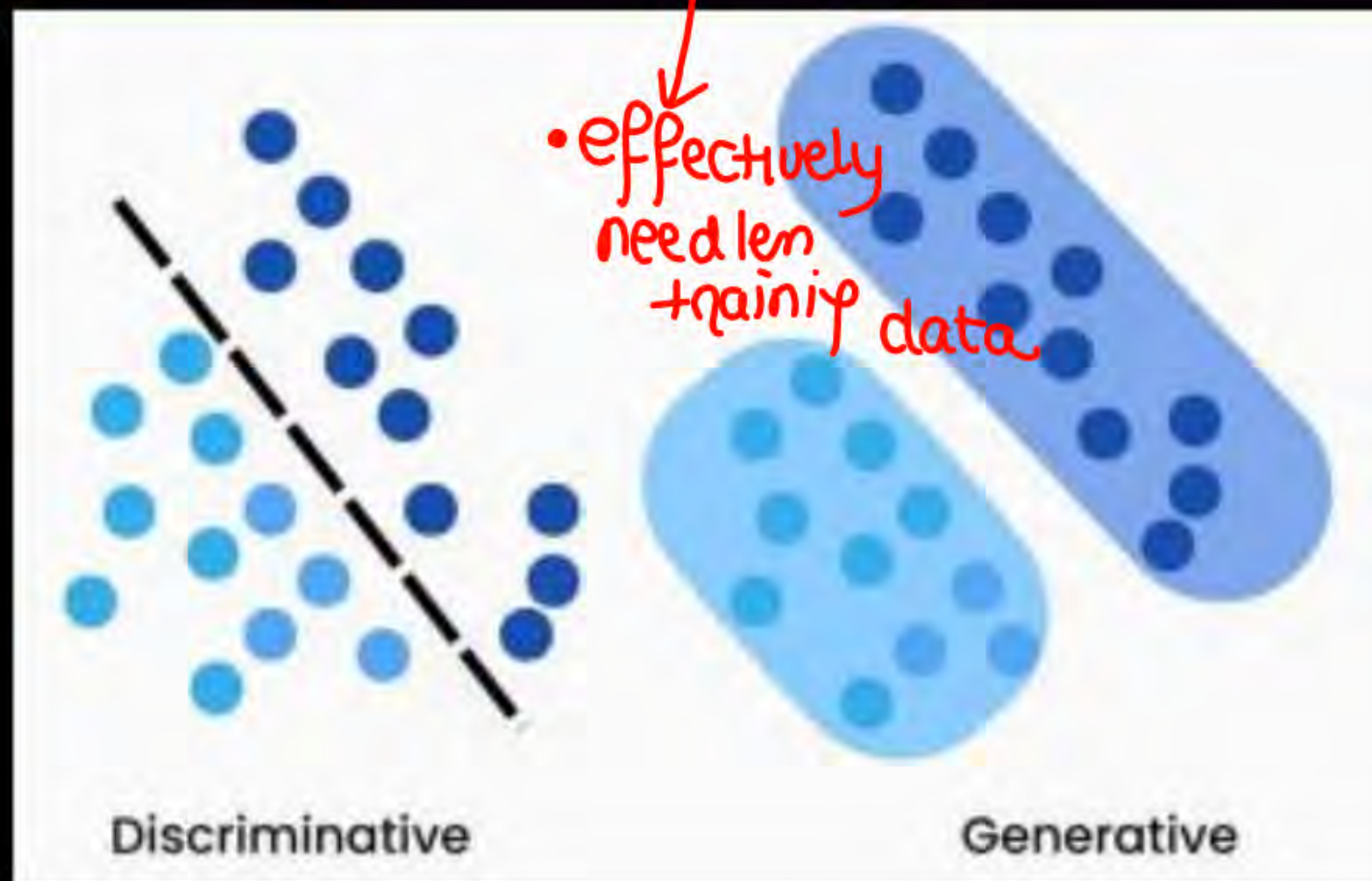


Bayesian Decision Theory

learn classifier



Discriminative vs. Generative Learning



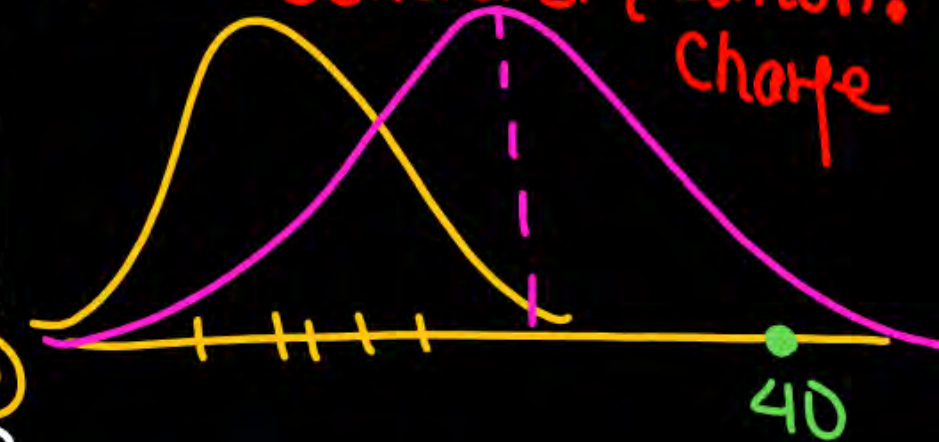
effectively need less training data

learn data distribution

highly effected by outlier

need more training data
dimension • Poisson distribution. Change

40
50



Naive Bayes disadvantages.

- ① large data to be collected, for finding correct distribution
 - ② naive assumption: for a given class dimensions are independent.
 - ③ effected by outlier
- © If dataset is small then \Rightarrow then the probabilities will try to overfit data

- Linear regression \Rightarrow Noise data analysis.

y
 Actual value

x
 Actual data

$$Y = X\beta + \underbrace{\varepsilon}_{\text{noise}}$$

ε is Gaussian, zero mean RV.

actual
 label
 value

actual data

$$\begin{aligned}
 y_1 &= x_1\beta + \varepsilon_1 \\
 y_2 &= x_2\beta + \varepsilon_2 \\
 &\vdots
 \end{aligned}$$

$$\begin{aligned}
 P(y_1 | x_1) &= N(x_1\beta, \sigma^2) \\
 &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_1 - x_1\beta)^2 / 2\sigma^2}
 \end{aligned}$$

Z is a $N(0, \sigma^2)$
 $(Z+5)$ is also RV $N(5, \sigma^2)$

all data points independent

$$P(Y/x) = \prod P(y_1|x_1)P(y_2|x_2) \dots$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \left[e^{-(y_1-x_1\beta)^2/2\sigma^2} \cdot e^{-(y_2-x_2\beta)^2/2\sigma^2} \dots \right]$$

MLE

$$P(y/x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \left[e^{-\sum_{i=1}^N (y_i - x_i\beta)^2 / 2\sigma^2} \right]$$

→ maximize ⇒

LR

$$\max P(y/x) \Rightarrow \min \sum_{i=1}^N (y_i - x_i\beta)^2$$
$$\max P(y/x) \Rightarrow \min \text{RSS}$$

In logistic Reg

$$\max P(y/x) \Rightarrow \prod (p_i)^{y_i} (1-p_i)^{1-y_i}$$

max likelihood

max log likelihood

→ min $-\log \text{likelihood}$
CE

Discriminative models

- In linear regression and logistic regression \Rightarrow we directly find $P(y|x)$

But in naive Bayes, however the rule is $P(y_1|x) > P(y_2|x)$

Generative model

- but we do not find $P(y|x)$ directly.

- we use $\underline{P(y|x)} = \frac{\underline{P(x|y)} P_y}{\underline{P_x}}$



A father has two kids, Kid A and Kid B. Kid A has a special character whereas he can learn everything in depth. Kid B have a special character whereas he can only learn the differences between what he saw.

One fine day, The father takes two of his kids (Kid A and Kid B) to a zoo. This zoo is a very small one and has only two kinds of animals say a lion and an elephant. After they came out of the zoo, the father showed them an animal and asked both of them **"is this animal a lion or an elephant?"**

The Kid A, the kid suddenly draw the image of lion and elephant in a piece of paper based on what he saw inside the zoo. He compared both the images with the animal standing before and answered based on the **closest match** of image & animal, he answered: "The animal is Lion".

The Kid B knows only the differences, based on **different properties learned**, he answered: "The animal is a Lion".

Here, we can see both of them is finding the kind of animal, but the way of learning and the way of finding answer is entirely different. In Machine Learning, We generally call Kid A as a Generative Model & Kid B as a Discriminative Model.



Discriminative vs. Generative Learning

Let's consider an example.

Imagine yourself as a language classification system.



There are two ways you can classify languages.

- ☐ Learn every language and then classify a new language based on acquired knowledge.
- ☐ Understand some distinctive patterns in each language without truly learning the language. Once done, classify a new language.

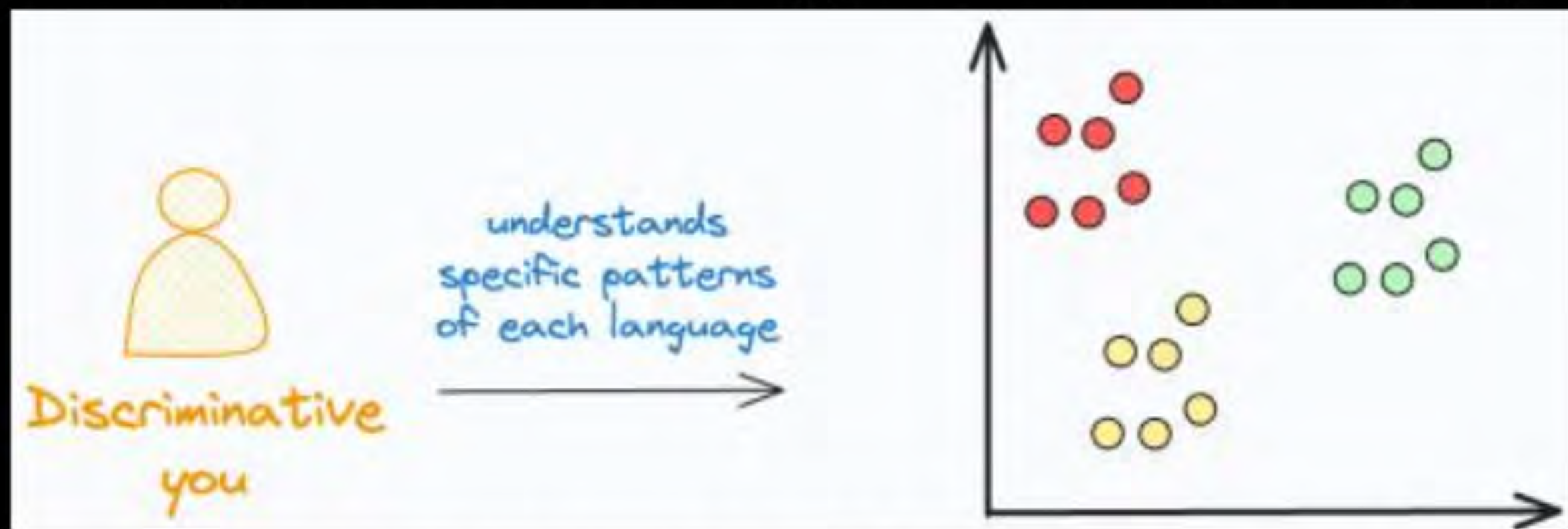
Can you figure out which of the above is generative and which one is discriminative?



Discriminative vs. Generative Learning

The second approach is a **discriminative approach**. This is because you only learned specific distinctive patterns of each language. It is like:

- If so and so words appear, it is likely "Language A."
- If this specific set of words appear, it is likely "Language B." and so on.



In other words, you learned the conditional distribution $P(\text{Language}|\text{Words})$.



Discriminative vs. Generative Learning

- ❑ Also, the above description might persuade you that generative models are more generally useful, but it is not true.
- ❑ This is because generative models have their own modeling complications.
- ❑ For instance, typically, generative models require more data than discriminative models.
- ❑ Relate it to the language classification example again.
- ❑ Imagine the amount of data you would need to learn all languages (generative approach) vs. the amount of data you would need to understand some distinctive patterns (discriminative approach).
- ❑ Typically, discriminative models outperform generative models in classification tasks.



Discriminative vs. Generative Learning

- ❑ In General, A Discriminative model models the **decision boundary between the classes.**
- ❑ A Generative Model explicitly models the **actual distribution of each class.**
- ❑ In final both of them is predicting the conditional probability $P(\text{Animal} | \text{Features})$. But Both models learn different probabilities.
- ❑ A Generative Model learns the **joint probability distribution $p(x,y)$** . It predicts the conditional probability with the help of **Bayes Theorem.**
- ❑ A Discriminative model learns the **conditional probability distribution $p(y|x)$** . Both of these models were generally used in supervised learning problems.



- ❑ The discriminative model learn the boundaries between classes or labels in a dataset.
- ❑ Discriminative models focus on modelling the decision boundary between classes in a classification problem. The goal is to learn a function that maps inputs to binary outputs, indicating the class label of the input.
- ❑ Maximum likelihood estimation is often used to estimate the parameters of the discriminative model, such as the coefficients of a logistic regression model or the weights of a neural network.
- ❑ Discriminative models (just as in the literal meaning) separate classes. But these models are not capable of generating new data points. Therefore, the ultimate objective of discriminative models is to separate one class from another.
- ❑ If we have some outliers present in the dataset, discriminative models work better compared to generative models i.e., discriminative models are more robust to outliers.
- ❑ But overall the accuracy of discriminative model is less than the generative models.



Generative and Descriptive Learning

☐ Examples of Discriminative Models

- ☒ Logistic regression
- ☒ Support vector machines (SVMs)
- ☒ Traditional neural networks
- ☒ Nearest neighbor
- ☒ Conditional Random Fields (CRFs)
- ☒ Decision Trees and Random Forest

☐ Outliers have little to no effect on these models. They are a better choice than generative models, but this leads to misclassification problems which can be a major drawback.

They need less data

to train

- Generative model affected by outliers
- more data to train



- ❑ Generative models are machine learning models that learn to generate new data samples similar to the training data they were trained on. They capture the underlying distribution of the data and can produce novel instances.
- ❑ So, the Generative approach focuses on the distribution of individual classes in a dataset, and the learning algorithms tend to model the underlying patterns or distribution of the data points (e.g., gaussian). These models use the concept of joint probability and create instances where a given feature (x) or input and the desired output or label (y) exist simultaneously.
- ❑ These models use probability estimates and likelihood to model data points and differentiate between different class labels present in a dataset. Unlike discriminative models, these models can also generate new data points.
- ❑ However, they also have a major drawback – If there is a presence of outliers in the dataset, then it affects these types of models to a significant extent.

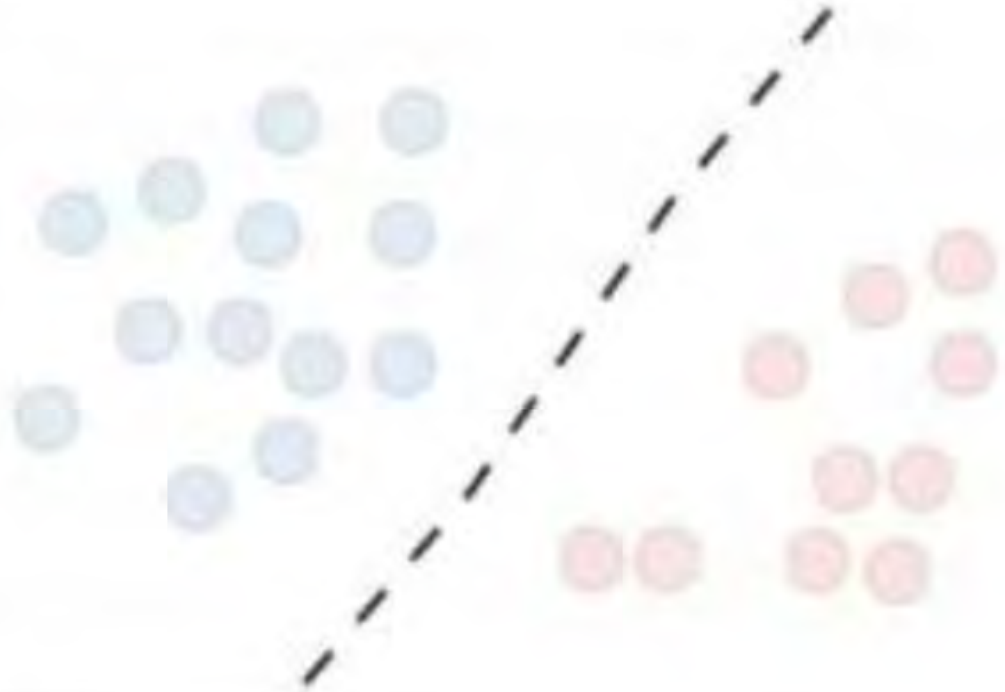
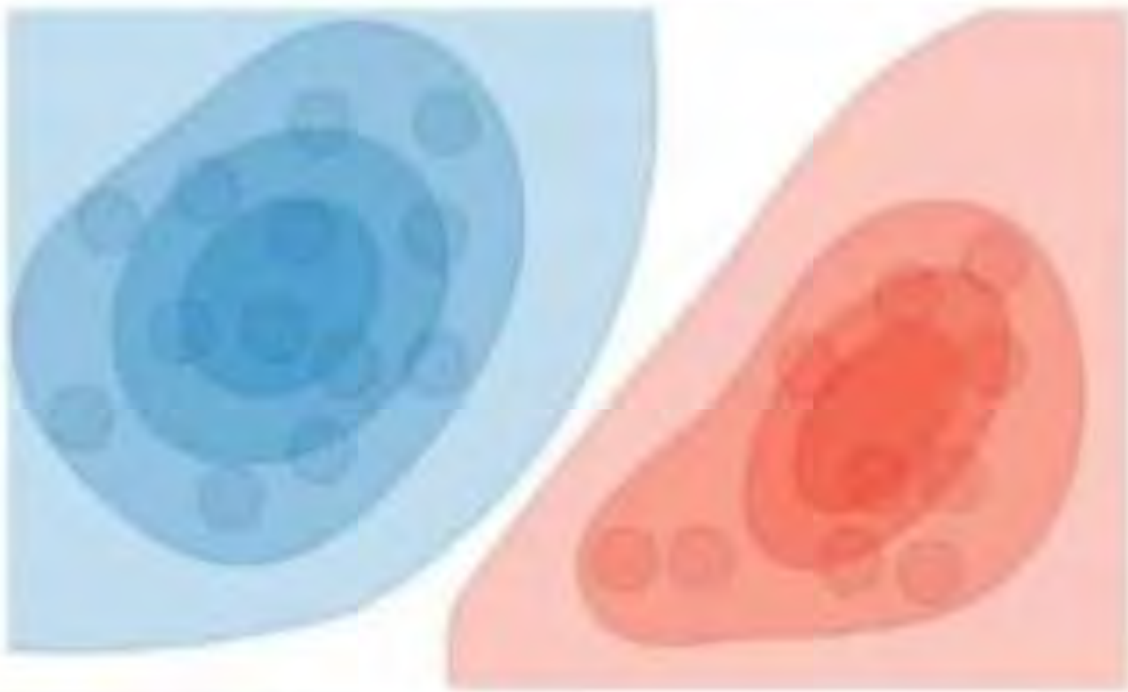


Generative and Descriptive Learning

- **Generative model**
- As the name suggests, generative models can be used to generate new data points. These models are usually used in unsupervised machine learning problems.
- Generative models go in-depth to model the actual data distribution and learn the different data points, rather than model just the decision boundary between classes.
- These models are prone to outliers, which is their only drawback when compared to discriminative models. The mathematics behind generative models is quite intuitive too. The method is not direct like in the case of discriminative models. To calculate $P(Y|X)$, they first estimate the prior probability $P(Y)$ and the likelihood probability $P(X|Y)$ from the data provided.



Generative and Descriptive Learning

	Discriminative model	Generative model
Goal	Directly estimate $P(y x)$	Estimate $P(x y)$ to then deduce $P(y x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Q1-1: Which of the following about Naive Bayes is incorrect?

- Attributes = Features*
- A ✓ Attributes can be nominal or numeric
 - B ✓ Attributes are equally important
 - C ✗ Attributes are statistically dependent of one another given the class value
 - D ✓ Attributes are statistically independent of one another given the class value
 - E All of above

Q1-2: Consider a classification problem with two binary features, $x_1, x_2 \in \{0, 1\}$. Suppose $P(Y = y) = 1/32$, $P(x_1 = 1 | Y = y) = y/46$, $P(x_2 = 1 | Y = y) = y/62$. Which class will naive Bayes classifier produce on a test item with $x_1 = 1$ and $x_2 = 0$?

• A 16

• B 26

• C 31

• D 32

$$\Downarrow (P(x/y) P_y)^{\max}$$

$$\frac{1}{32} \times P(x_1=1/y) \times P(x_2=0/y)$$

$$\frac{1}{32} \times (y/46) \times (1 - y/62)$$

$$\frac{1}{32} (y/46 - y^2/46 \times 62) \Rightarrow \frac{d}{dy} = 0$$

$$y = 31$$

Q1-3: Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other. We want to classify a new instance with **Confident=Yes, Studied=Yes, and Sick=No.**

• $P_P = 3/5$ $P_F = 2/5$

label.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

• $P(Y/P)P(Y/P)P(N/P)P_P \Rightarrow$
 $\frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \times \frac{3}{5} \Rightarrow$

• $P(Y/F)P(Y/F)P(N/F)P_F \Rightarrow$
 $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{2}{5} \Rightarrow$

Pass



data ke distribution
ya data ke pattern ke liye
kuch assumption

12. Identify the parametric machine learning algorithm.

a) CNN (Convolutional neural network)

b) KNN (K-Nearest Neighbours)

c) Naïve Bayes

d) SVM (Support vector machines)

⇒ Nonparametric

→ KNN: No assumption
on data

→ Parametric

→ Parametric

Yes it make certain assumption
on underlying data
numerical attribute ⇒ we assume
Gaussian dist.
dimension are independent for
a given class.

6. Impact of outliers?

Naive Bayes is **highly impacted by outliers** and completely robust in this case (depending on the USE case we are working on). The reason is the NB classifier assigns the **0 probability** for all the data instances it has not seen in the **training set**, which creates an issue during the **prediction time**, and the same goes with outliers also, as it would have been the same data that the classifier has not seen before.

4) Consider the following statements-

- ☒ a. Naive Bayes assumes independence among predictors.
- ☒ b. Naive Bayes can perform multi-class prediction.

Which of the following statements are correct-

- ☒ Both a and b
- ☐ Only a
- ☐ Only b

determine the how this - -

4) The table given below contains some of the bigram frequencies of $(determine, w_i)$ where w_i represents every word in the column

	the	how	this	a	his
determine	0.115	0	0.0125	0.006	0.0013

What is the conditional probability of $P(his|determine)$ if the probability of *determine* as the starting word is 0.6?

- ☐ 0.0031
- ☒ 0.0022
- ☐ 0.0122
- ☐ 0.0128

$P(determine, w_i)$ • w_i represent word in Column

$$P(his|determine) = \frac{P(his, determine)}{P(determine)} = \frac{0.0013}{0.6} = 0.0022$$

5) Given that

$$P(Y|X) = \frac{P(X|Y) \times P(Y)}{P(X)}$$

Match the following.

Handwritten annotations:
 $P(Y|X)$ is circled and labeled "Posterior".
 $P(X|Y)$ is circled and labeled "likelihood".
 $P(Y)$ is circled and labeled "Prior".
 $P(X)$ is circled and labeled "evidence".

- $P(Y|X)$
- $P(X|Y)$
- $P(X)$
- $P(Y)$
- i. Evidence
- ii. Prior
- iii. Posterior
- iv. Likelihood



(a \rightarrow iii, b \rightarrow iv, c \rightarrow i, d \rightarrow ii)



(a \rightarrow i, b \rightarrow iv, c \rightarrow iii, d \rightarrow ii)



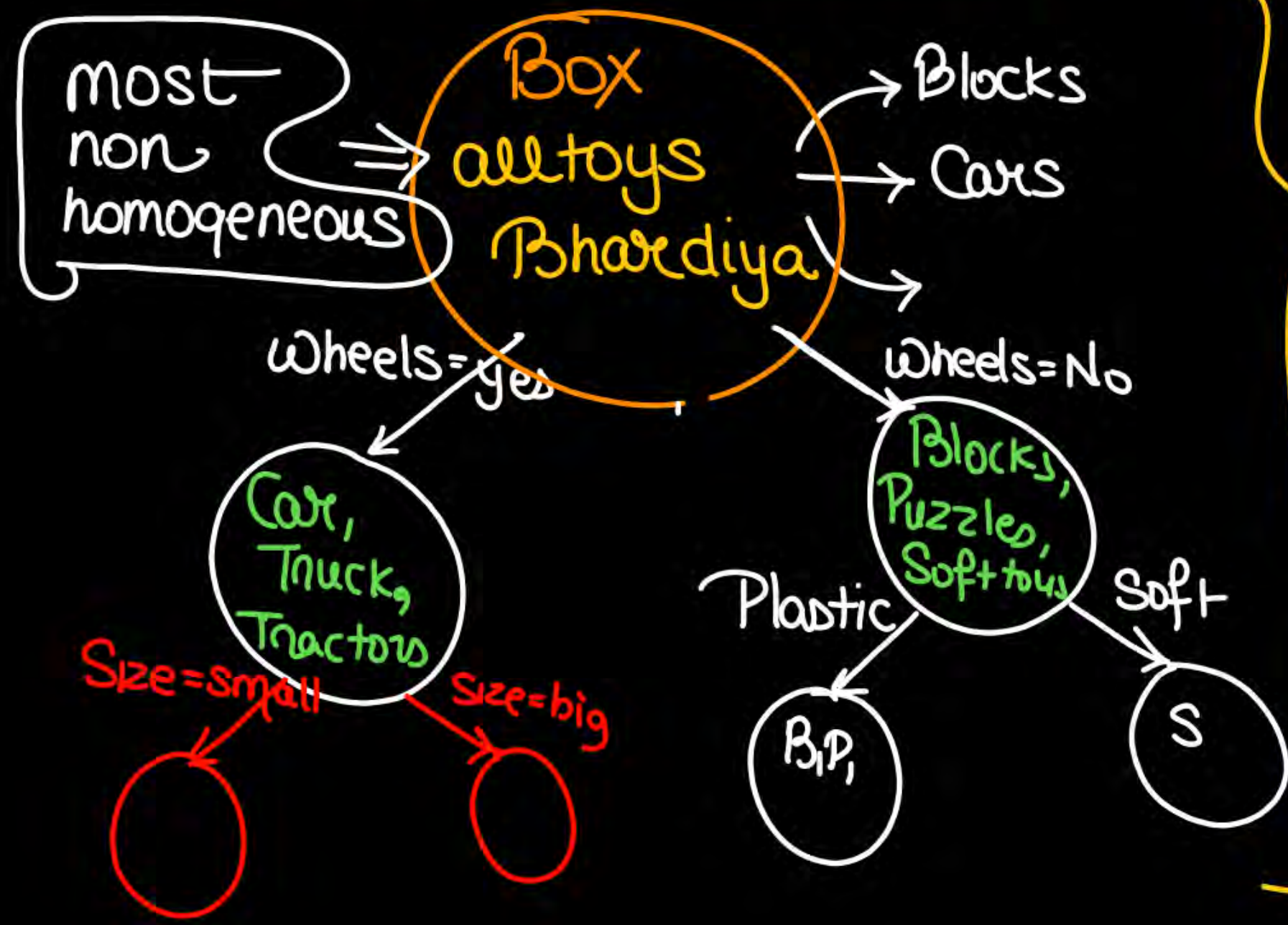
(a \rightarrow iii, b \rightarrow ii, c \rightarrow i, d \rightarrow iv)



(a \rightarrow i, b \rightarrow iv, c \rightarrow ii, d \rightarrow iii)

- Naive Bayes ⇒ we need more accuracy.
⇒ we have large data available
⇒ language model
⇒ To create new data points.
Generative model.

Decision tree



So we distribute/
Segregate to inc

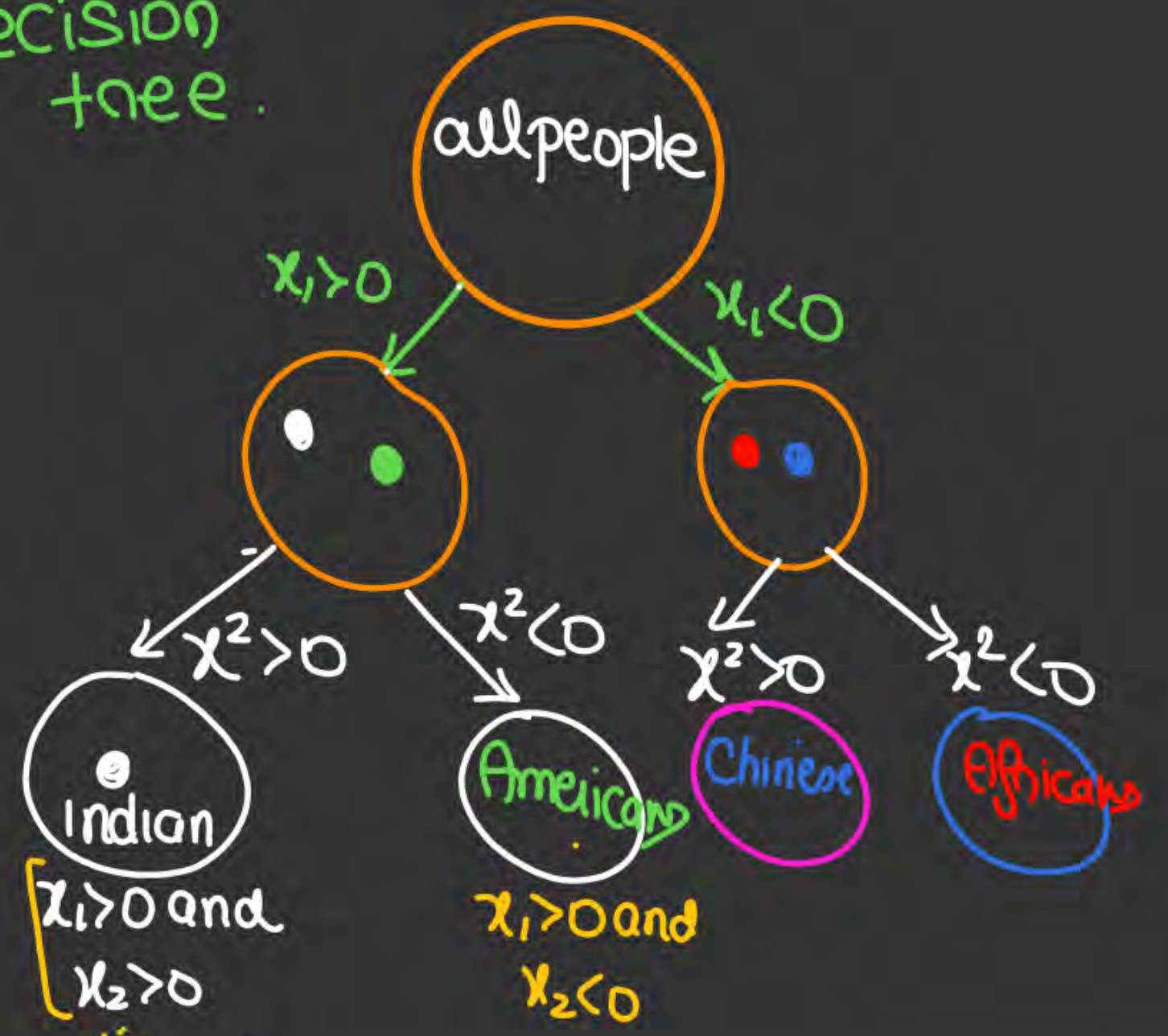
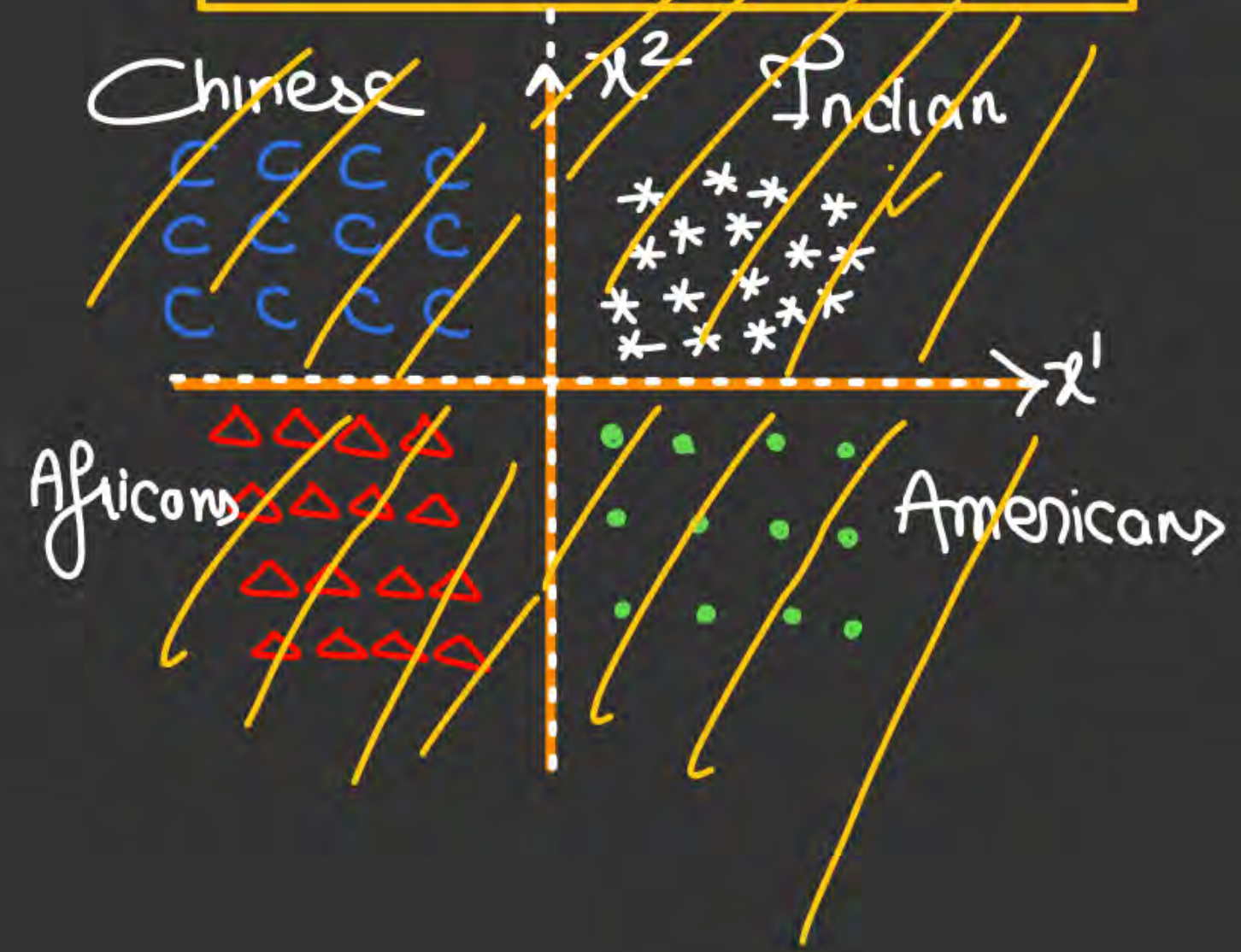
Homogeneity

Kya Kar rahi hai

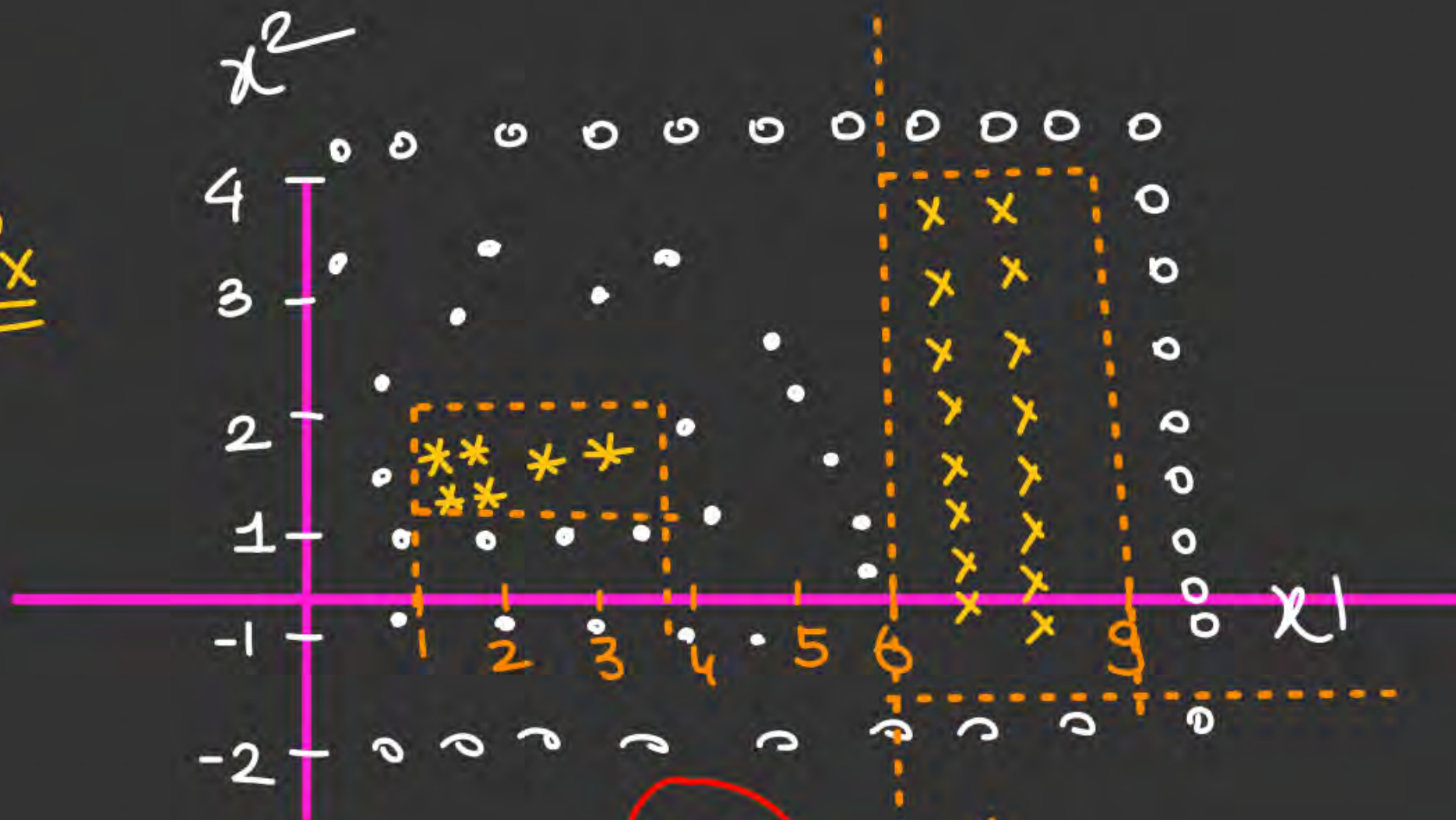
- from most non homogeneous Starting node, Starting point we segregate, such that we create nodes which has homogeneous content

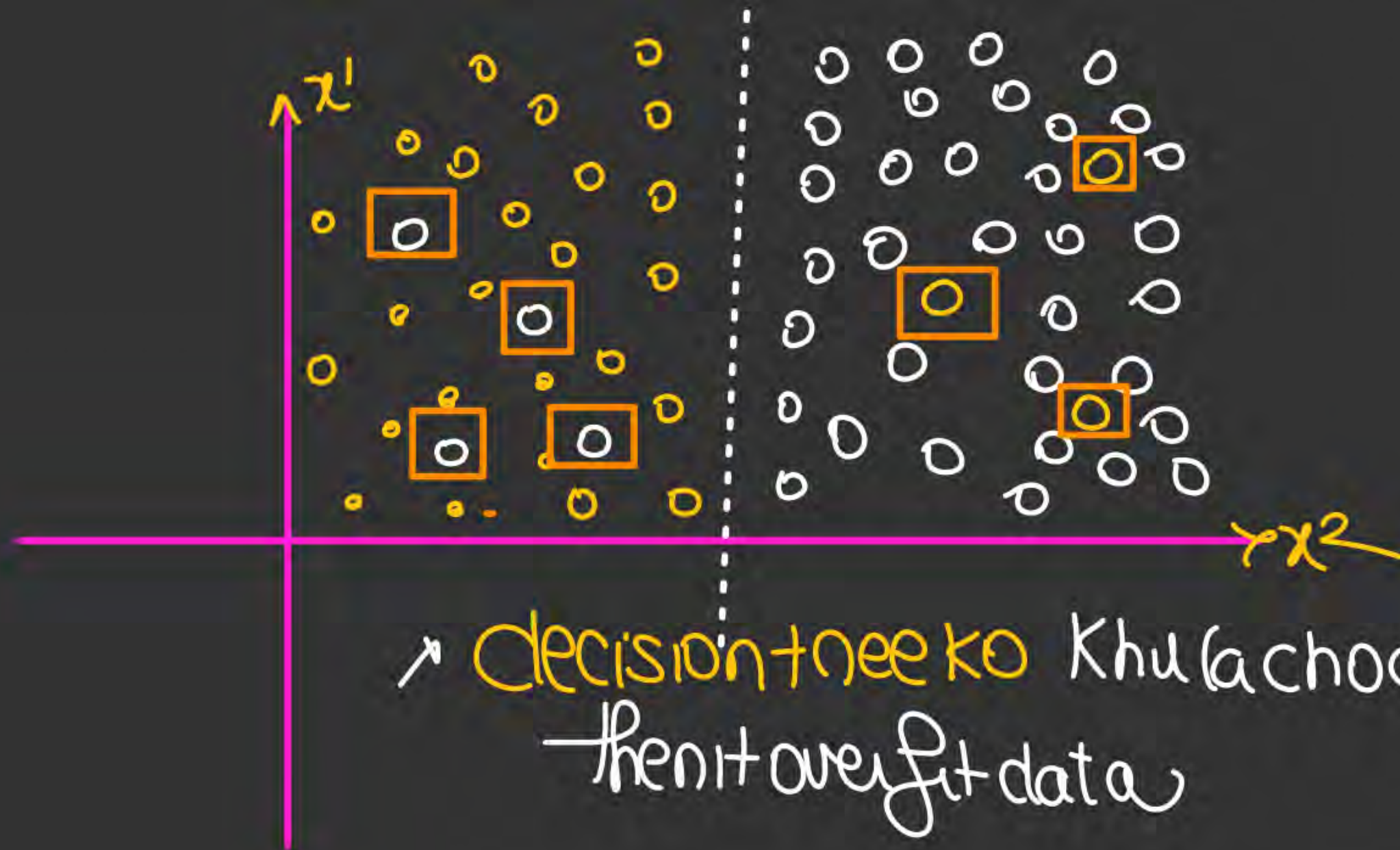
Decision tree

decision tree.

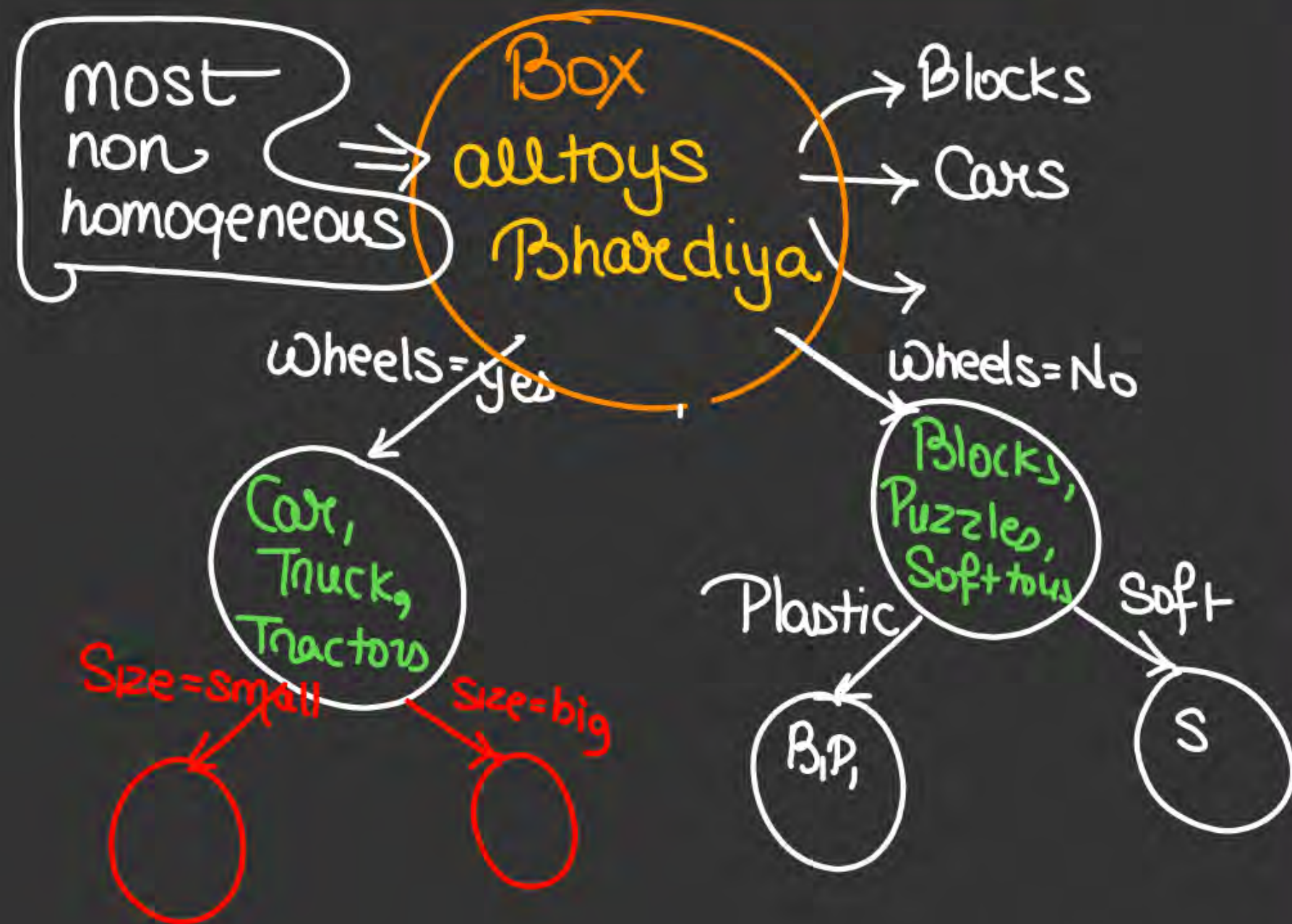


\mathcal{E}_x





Decision tree



why we need Segregation that Create homogeneous nodes ??

• decision tree finally Homogeneous node de deg a

* Regions in D dimension Space which have similar Points.

* Similar to KNN,

Df apply to both
Categorical/numerical data



Decision Tree



What is a Decision Tree

We start with
the full
training data
at the root
node

Now Based
on some
variable the
whole input
space divided

Keep dividing
the input
space till you
reach the final
stopping
criteria

Or you reach
the stopping
condition



Decision Tree



What is a Decision Tree

**How this decision
tree is stored in
memory...**



Decision Tree



What is a Decision Tree

Decision tree is Non parametric, and non linear

Because here we make no assumption on the pattern of the data, we simply take and data and work on it.

Some Important
Terms
Decision Tree is
Non Linear



Decision Tree



Lets see an example

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

**We have to predict
whether we have to play
or not... Classification
Problem**

**At root node we can
choose any attribute for
decision**



Decision Tree



How to select the attribute for splitting ?

We do splitting to reduce the confusion, after splitting we need most homogeneous nodes. Where the concentration of a particular label is very high

We can see that in the starting at the root node we have points with Y/N both... hence there was lot of confusion.
i.e. in whole input space we have all the points we need to divide the input space to get regions of similar points



Decision Tree



Which attribute to choose for decision ?

Decision tree is a greedy approach where we check all the possibilities of split and find the best for us

**Attribute
Selection
Measures ...**



Decision Tree



How to select the attribute for splitting ?

How we select the attribute for splitting ??

Attribute Selection
measure ...



Decision Tree



How to select the attribute for splitting ?

How to measure node impurity/
node purity/ node homogeneity/
degree of randomness...

Attribute Selection
measure ...



Decision Tree



How to select the attribute for splitting ?

**For classification case : Gini
Index and Entropy
For Regression : Variance.**

**Information Gain : After
splitting we measure the
reduction in the
impurity...**



Decision Tree



How to select the attribute for splitting ?

Gini Index (for classification)

Concept is if Probability for misclassify is high then Gini Index is high else it is low...



Decision Tree



How to select the attribute for splitting ?

Gini Index (for classification)

We want gini index as low as possible.



Decision Tree



Lets see an example

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Decision Tree



How to select the attribute for splitting ?

Gini Index (for classification)

Lets find the Gini Index at each node here.

Day	Outlook	Temperature	Humidity	Wind	Play Golf
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Decision Tree



How to select the attribute for splitting ?

Gini Index (for classification)

It is probability of
misclassifying any point
in data...



Decision Tree



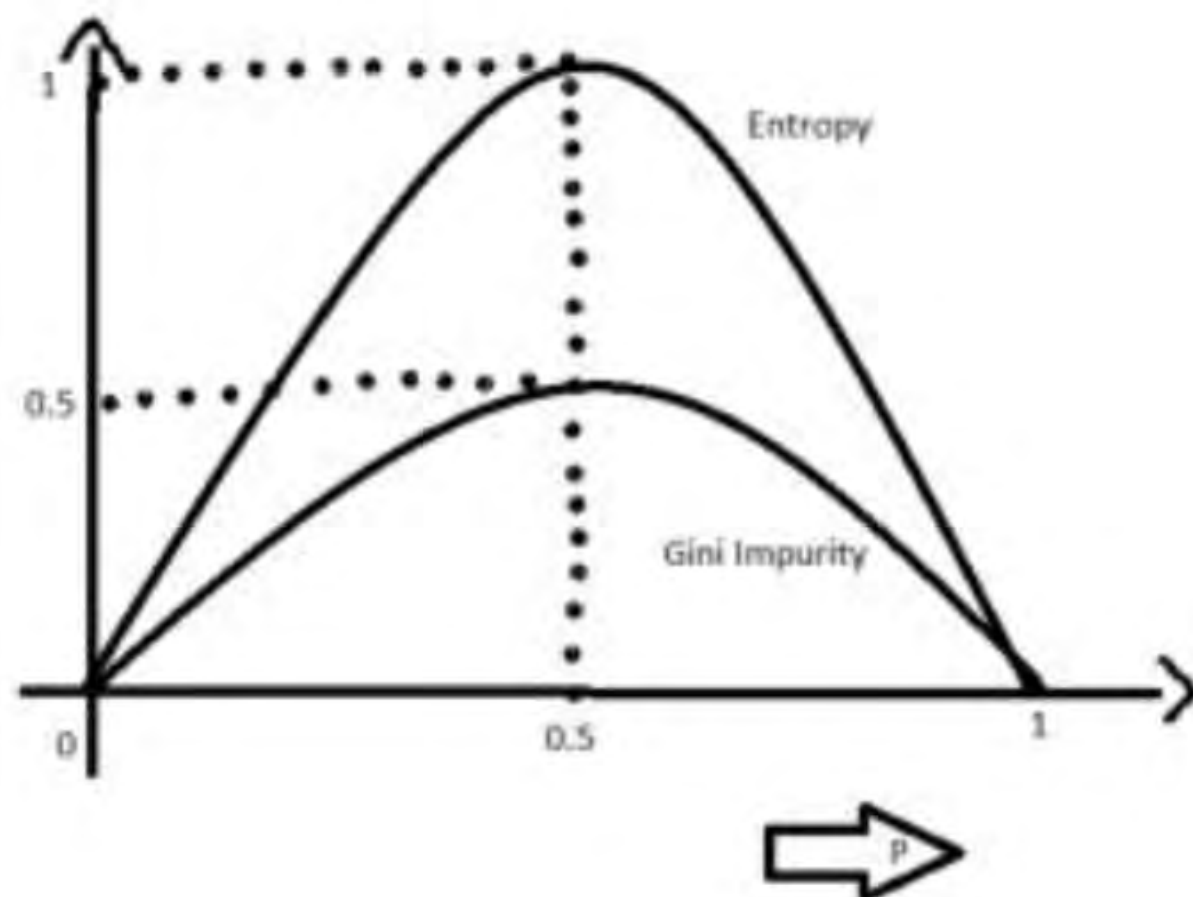
How to select the attribute for splitting ?

Entropy (for classification)

Entropy measure the amount of uncertainty or degree of randomness.



Entropy Vs Gini Index





Entropy Vs Gini Index

It is the probability of misclassifying a randomly chosen element in a set.	While entropy measures the amount of uncertainty or randomness in a set.
The range of the Gini index is $[0, 1]$, where 0 indicates perfect purity and 1 indicates maximum impurity.	The range of entropy is $[0, \log(c)]$, where c is the number of classes.
Gini index is a linear measure.	Entropy is a logarithmic measure.
It can be interpreted as the expected error rate in a classifier.	It can be interpreted as the average amount of information needed to specify the class of an instance.
It is sensitive to the distribution of classes in a set.	It is sensitive to the number of classes.



How to select the attribute for splitting ?

Entropy Vs Gini Index

It is less robust than entropy.

It is more robust than Gini index.

It is sensitive.

It is comparatively less sensitive.

Formula for the Gini index is $Gini(P) = 1 - \sum (P_x)^2$,
where P_i is
the proportion of the instances of class x in a set.

Formula for entropy is $Entropy(P) = -\sum (P_x) \log(P_x)$,
where p_i is the proportion of the instances of class x in
a set.



2 mins Summary



Topic

Bayesian Learning

Topic

Topic

Topic

Topic

THANK - YOU