Data Science and Artificial Intelligence

Machine Learning

Bayesian learning

Lecture No. 4

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- **Topic** Bayes classifier
- **Topic** Bayes optimum cl.
- **Topic** Naive Bayes
- **Topic**
- **Topic**

Topics to be Covered

- **Topic** Naive Bayes
- **Topic** Discriptive/Discriminative & Generative learning
- **Topic**
- **Topic**
- **Topic**

Patience

REVISION

Bayes Classifier

- ⇒ work on MAP Rule
- ⇒ for any point x_i it find $\underbrace{P(x_i/C_j) \cdot P_{C_j}}$

  C_j = any class
- ⇒ for all classes we find $P(x_i/C_j)P_{C_j}$
- ⇒ So whichever class max $\left(P(x_i/C_j)P_{C_j}\right)$ that class will be assigned to x_i

- $P_{C_j} = \dfrac{\text{Number of points in class } C_j}{\text{Total No of Points in training data}}$ ⇒ M Probab

REVISION

Bayes Classifier

- $P(x_i^{\circ}/C_j)$ = Calculated from training data

M: No of classes, D: No of dimension, each dimension can take a values,

$$P(x_i^1, x_i^2, \ldots x_i^D/C_j) \Rightarrow (a^D \times M)$$

- Total parameters $= (a^D M + M) \Rightarrow$ These parameters are used by the hypothesis to find class of the point

Bayes Classifier

@ Time of testing we get a new point

$(x_j) \Rightarrow \underbrace{(x_j^1, x_j^2, x_j^3 - - - x_j^D)}$

For this we need to find $P(x_j/C_k)P_{C_k}$

Since x_j will be a combination of D dimensions

thus $P(x_j/C_k)$ will be available

So find $\max\limits_{C_k} P(x_j/C_k)P_{C_k} \Rightarrow$ the class with $\max P(x_j/C_k)P_{C_k}$ is assigned

REVISION

Bayes Optimum Classifier

- Similar to bayes classifier but here we have more than one hypothesis
- For each hypothesis we need to find $\underbrace{P(x_i|C_j)}P_{C_j}$

$\Downarrow$

$(a^D M + M)$

- For each hypothesis $\Rightarrow (a^D M + M)$ Parameter

Basics of Machine Learning

REVISION

Bayes Optimum Classifier

O(No of Points in dataset × No of hypothesis)

at time of testing ⇒ $\sum_{i=1}^{H} P(h_i/D)\, P(C_j/h_i)$

⇒ H : No of hypothesis

⇒ all these parameters are used by each hypothesis to find class for given point i.e $P(C_j/h_i)$

⇒ For any given point this is calculated for each class whichever class has max value that is assigned

REVISION

Naïve Bayes Classifier

Bayes class.

$\max \left[P(x_i^o / C_j) P_{C_j} \right]$

- dimension are Ind
- all dimension have equal Contribution

$\max P(x_i^1 / C_j) P(x_i^2 / C_j) \dots P(x_i^D / C_j) P_{C_j}$

a a a

(aDM) + M

No of Parameters.

REVISION

Naïve Bayes Classifier

Bayes classo

$P(x_i^{o}/c_j)Pc_j$

Probab for all
Combi available

Naive Bayes

$P(x_i^{1}/c_j)P(x_i^{2}/c_j)$ - -
- - $P(x_i^{p}/c_j)Pc_j$

Bayesian Decision Theory

Class for
(Sunny, Hot, High, Strong)

Outlook	Temp.	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Weak	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Strong	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Outlook	P(O/Yes)	P(O/No)
Sunny	2/9 $P(S|Y)$	3/5
Overcast	4/9 $P(O|Y)$	0
Rain	3/9 $P(R/Y)$	2/5

Temperature	P(T/Yes)	P(T/No)
Hot	2/9	2/5
Mild	4/9	2/5
Cold	3/9	1/5

Humidity	P(H/Yes)	P(H/No)
High	3/9	4/5
Normal	6/9	1/5

Wind	P(W/Yes)	P(W/No)
Weak	6/9	2/5
Strong	3/9	3/5

$P(\text{Sunny, Hot, high, strong} | \text{Yes})\, P_{yes}$

$\Rightarrow P(S|Y)\, P(H/Y)\, P(H/Y)\, P(S|Y)\, P_Y$

$\Rightarrow \left(2/9 \times 2/9 \times 3/9 \times 3/9 \times \frac{9}{14} \right)$

$P(\text{Sunny, Hot, high, stron} / \text{No})$

$P(S/N)\, P(H/N)\, P(H/N)\, P(S/N)\, P_{No}$

$\frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} \times \frac{5}{14}$

$\Rightarrow$ No ✓

Bayesian Decision Theory

Test point ⇒ (Overcast, mild, Normal, weak)

Outlook	Temp.	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Weak	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Strong	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Outlook	P(O/Yes)	P(O/No)
Sunny	2/9 P(S/Y)	3/5
Overcast	4/9 P(O/Y)	0
Rain	3/9 P(R/Y)	2/5

Temperature	P(T/Yes)	P(T/No)
Hot	2/9	2/5
Mild	4/9	2/5
Cold	3/9	1/5

Humidity	P(H/Yes)	P(H/No)
High	3/9	4/5
Normal	6/9	1/5

Wind	P(W/Yes)	P(W/No)
Weak	6/9	2/5
Strong	3/9	3/5

Sol $P(O, M, N, W / Yes)\, P_{Yes} \Rightarrow P(O|Y)\, P(M|Y)\, P(N|Y)\, P(W|Y)\, P_Y$

$\Rightarrow \frac{4}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{6}{9} \times \frac{9}{14}$

$P(O, M, N, W / No)\, P_{No} \Rightarrow P(O|N)\, P(M|N)\, P(N|N)\, P(W|N)\, P_N$

$0 \times \ldots \Rightarrow 0$

- Since $P(Overcast / No) = 0$

  thus in naive bayes if any test point has 1st dimension = overcast then $P(x_i / No)\, P_{No} = 0$ and test point $\Rightarrow$ Yes class always

Naïve Bayes Algorithm

Zero probability problem…

* This is called Zero probab Problem

* Reason ⇒ bcoz in whole training data no data point with class No has dimension 1 = Overcast

(lack of data)

Naïve Bayes Algorithm

Zero probability problem...

Solution to Zero probab problem ⇒

2 solutions

1. inc the data

2. add the arbitrary points

↳ Smoothening process ✓

Naïve Bayes Algorithm

- Smoothing techniques, adding virtual entries…
- Additional data…

Solving the zero-probability problem…

• Smoothing Process ⇒

1st dimension → 3 values

Values for Yes class

S
S
O
O
O
O
R
R
R

α Values: S O R, S O R, S O R, ...

$$P(S/Y) = \frac{2+\alpha}{9+3\alpha}$$

$$P(O/Y) = \frac{4+\alpha}{9+3\alpha}$$

$$P(R/Y) = \frac{3+\alpha}{9+3\alpha}$$

1st dimension → 3 values

Values for No class.

S
S
S
R
R

α Values: S R O, S R O, S R O, ...

$$P(S/N) = \frac{3+\alpha}{5+3\alpha}$$

$$P(R/N) = \frac{2+\alpha}{5+3\alpha}$$

$$P(O/N) = \frac{0+\alpha}{5+3\alpha}$$

Temp

2nd dimension

$$P(H/Y) = \frac{2+\alpha}{9+3\alpha}$$

$$P(m/Y) = \frac{4+\alpha}{9+3\alpha}$$

$$P(C/Y) = \frac{3+\alpha}{9+3\alpha}$$

H
H
m
m
m
m
C
C
C
H m C
H m C
| | |

H
H
m
m
C
H m C
H m C
H m C
| | |

$$P(H/N) = \frac{2+\alpha}{5+3\alpha}$$

$$P(m/N) = \frac{2+\alpha}{5+3\alpha}$$

$$P(C/N) = \frac{1+\alpha}{5+3\alpha}$$

3^{rd} dimension

$$P(H/Y) = \frac{3+\alpha}{9+2\alpha}$$

$$P(N/Y) = \frac{6+\alpha}{9+2\alpha}$$

H
H
H
N
N
N
N
N
N

H N
H N
H |
| |
|

After smoothing

new Probab ⇒ $\dfrac{\text{Old Value} + \alpha}{\text{Old Value} + K\alpha}$

K = No of values a dimension Can take

Bayesian Decision Theory

Test point ⇒ (overcast, mild, Normal, weak)

Outlook	Temp.	Humidity	Wind	Play Tennis
Sunny ✗	Hot	High	Weak	No
Sunny ✗	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain ✗	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Weak	Yes
Sunny ✗	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Strong	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain ✗	Mild	High	Strong	No

Outlook	P(O/Yes)	P(O/No)
Sunny	2/9 ⇒ $2+\alpha/9+3\alpha$	3/5 ⇒ $3+\alpha/5+3\alpha$
Overcast	4/9 ⇒ $4+\alpha/9+3\alpha$	0 ⇒ $0+\alpha/5+3\alpha$
Rain	3/9 ⇒ $3+\alpha/9+3\alpha$	2/5 ⇒ $2+\alpha/5+3\alpha$

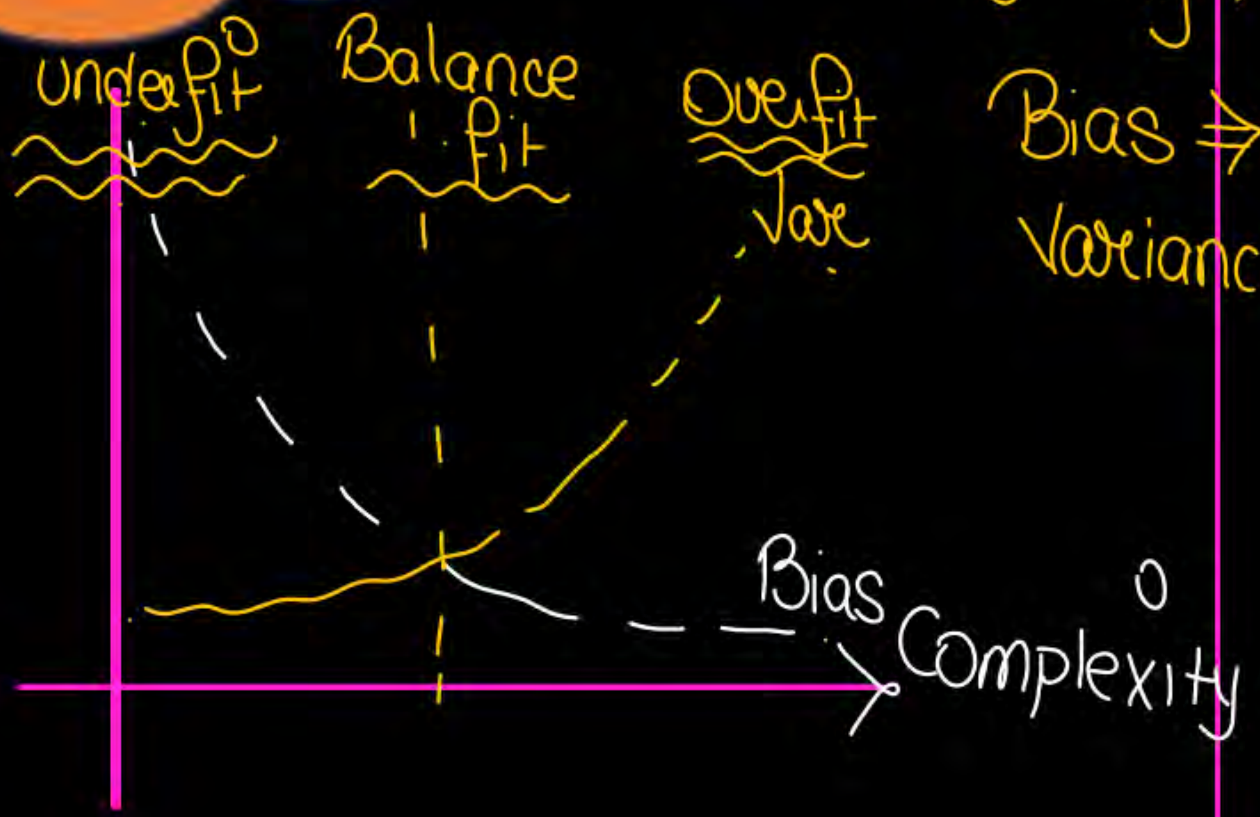
Temperature	P(T/Yes)	P(T/No)
Hot	2/9	2/5
Mild	4/9	2/5
Cold	3/9	1/5

Humidity	P(H/Yes)	P(H/No)
High	3/9 ⇒ $\frac{3+\alpha}{9+2\alpha}$	4/5 ⇒ $\frac{4+\alpha}{5+2\alpha}$
Normal	6/9 ⇒ $\frac{6+\alpha}{9+2\alpha}$	1/5 ⇒ $\frac{1+\alpha}{5+2\alpha}$

Wind	P(W/Yes)	P(W/No)
Weak	6/9	2/5
Strong	3/9	3/5

Naïve Bayes Algorithm

Laplace smoothing ⇒ means $\alpha = 1$



	$\alpha = 0$ ⟶		$\alpha \Rightarrow$ v. large
Fitting ⇒	Overfit	Balance fit	underfitting
Bias ⇒	Low	low	High
Variance ⇒	High	low	low

- How to solve for continuous dimension in naive bayes.

In this case $P(x_i/C_j)P_{C_j}$

$$\left(\underbrace{P(x_i^{1}/C_j)}_{\Downarrow}\underbrace{P(x_i^{2}/C_j)}_{\Downarrow} \dots P_{C_j}\right)$$

Categorical done

* we create PDF of dimention for given class
* Gaussian PDF

Bayesian Decision Theory

Naïve Bayes Algorithm

What if the dimension are continuous in nature

$P(S/Y) = 2/9$

$P(O/Y) = 4/9$

$P(R/Y) = 3/9$

$P(S/N) = 3/5$

$P(O/N) = 0$

$P(R/N) = 2/5$

Categorical

The numeric weather data with summary statistics

outlook			temperature		humidity		windy			play	
	yes	no	yes	no	yes	no		yes	no	yes	no
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
	9	5	64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						

$P(F/Y) = 6/9$

$P(F/N) = 2/5$

$P(T/Y) = 3/9$

$P(T/N) = 3/5$

$P_Y = 9/14$

$P_N = 5/14$

Bayesian Decision Theory

The numeric weather data with summary statistics											
outlook			temperature		humidity		windy			play	
	yes	no	yes	no	yes	no		yes	no	yes	no
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
			64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						

mean $= 86.2$

$\sigma^2 = 75.76$

mean $= 74.6$

$\sigma^2 = 49.84$

mean $= 79.11$

$\sigma^2 = 92.76$

mean $= \frac{1}{9}\sum$ all values $= 73$

$\sigma^2 = \frac{1}{9}\sum(x_i - 73)^2 = 34.66$

- PDF of Temp and Humidity $\Rightarrow$

$$P(T/Y) \Rightarrow \frac{1}{\sqrt{2\pi \times 34.66}} e^{-\frac{(x-73)^2}{2\times 34.66}}$$

$$P(T/N) = \frac{1}{\sqrt{2\pi \times 49.84}} e^{-\frac{(x-74.6)^2}{2\times 49.84}}$$

$$P(h/Y) = \frac{1}{\sqrt{2\pi \times 92.76}} e^{-\frac{(x-79.11)^2}{2\times 92.76}}$$

$$P(h/N) = \frac{1}{\sqrt{2\pi \times 75.76}} e^{-\frac{(x-86.2)^2}{2\times 75.76}}$$

Find class of

(Sunny, 78, 86, true)

No class

$$P(S|Y)\,\overbrace{P(78|Y)}\,\overbrace{P(86|Y)}\,P(T|Y)\,P_Y$$

$$2/9 \times \frac{1}{\sqrt{2\pi \times 34.66}} e^{-(78-73)^2/2\times 34.66} \cdot \frac{1}{\sqrt{2\pi \times 92.76}} e^{-\frac{(86-79.11)^2}{2\times 92.76}} \times \frac{3}{9} \times \frac{9}{14}$$

$$\frac{2}{9} \times .0472 \times .032 \times \frac{3}{9} \times \frac{9}{14} \Rightarrow 7.19 \times 10^{-5}$$

$$P(S|N)\,P(78|N)\,P(86|N)\,P(T|N)\,P_N$$

$$\frac{3}{5} \times .05032 \times .0458 \times \frac{3}{5} \times \frac{5}{14}$$

$$= 2.96 \times 10^{-4}$$

Bayesian Decision Theory

The numeric weather data with summary statistics											
outlook			temperature		humidity		windy			play	
	yes	no	yes	no	yes	no		yes	no	yes	no
sunny	2	3	83	85	86	85	false	6	2	9	5
overcast	4	0	70	80	96	90	true	3	3		
rainy	3	2	68	65	80	70					
			64	72	65	95					
			69	71	70	91					
			75		80						
			75		70						
			72		90						
			81		75						

Q: Consider a classification problem with 10 classes $y \in \{1,2,\ldots,10\}$, and two binary features $x1, x2 \in \{0,1\}$.

Suppose:

$p(Y=y)=1/10$,
$p(x1=1|Y=y)=y/10$,
$p(x2=1|Y=y)=y/540$

Which class will naïve Bayes classifier produce on a test item with $(x1=0,x2=1)$?

A. 1
B. 3
C. 5
D. 8
E. 10

How additive smoothing effect bias and variance

1. What type of algorithm is Naive Bayes used for in machine learning?

a. Classification

b. Regression

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.
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

13. In a binary classification problem, a Naive Bayes classifier correctly classifies 85% of Class A instances and 90% of Class B instances. If the prior probabilities are P(Class A) = 0.4 and P(Class B) = 0.6, what is the overall accuracy of the classifier?

a. 0.48

b. 0.87

c. 0.90

d. 0.84

Naïve Bayes Classifier

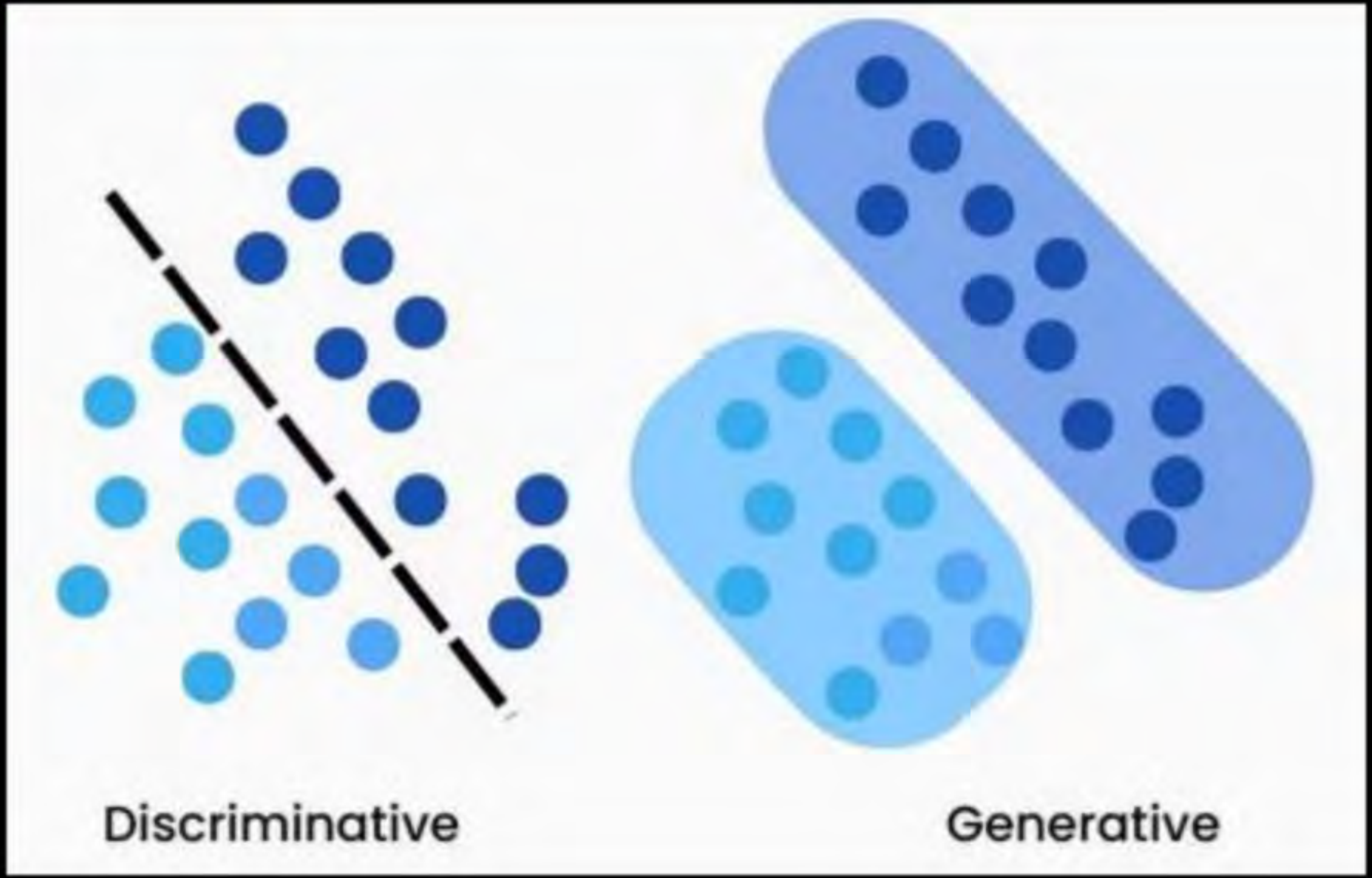
Advantages of Naïve Bayes Classifier:

- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for **text classification problems.**

Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.

Discriminative vs. Generative Learning



Bayesian Decision Theory

Story...

- 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.

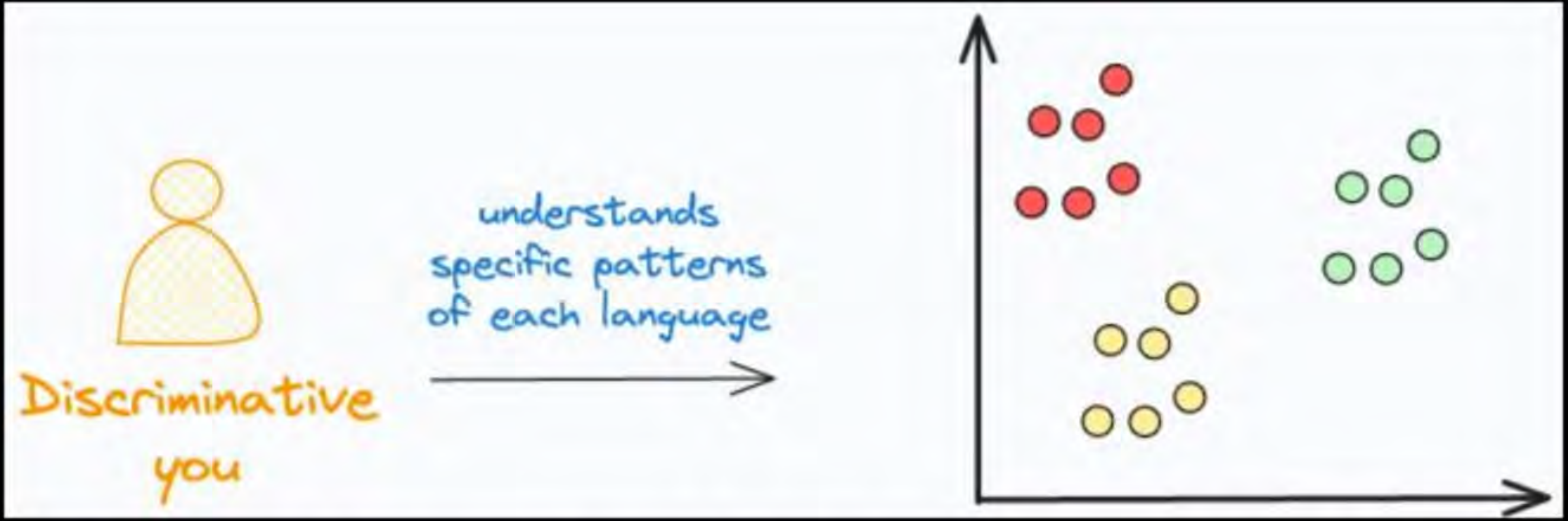
- ❑ Learn every language and then classify a new language based on acquired knowledge.
- ❑ Understand some distinctive patterns in each language without truly learning the language. Once done, classify a new language.

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(Language|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(Animal | 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.

Discriminative Learning

- 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.

Generative Learning

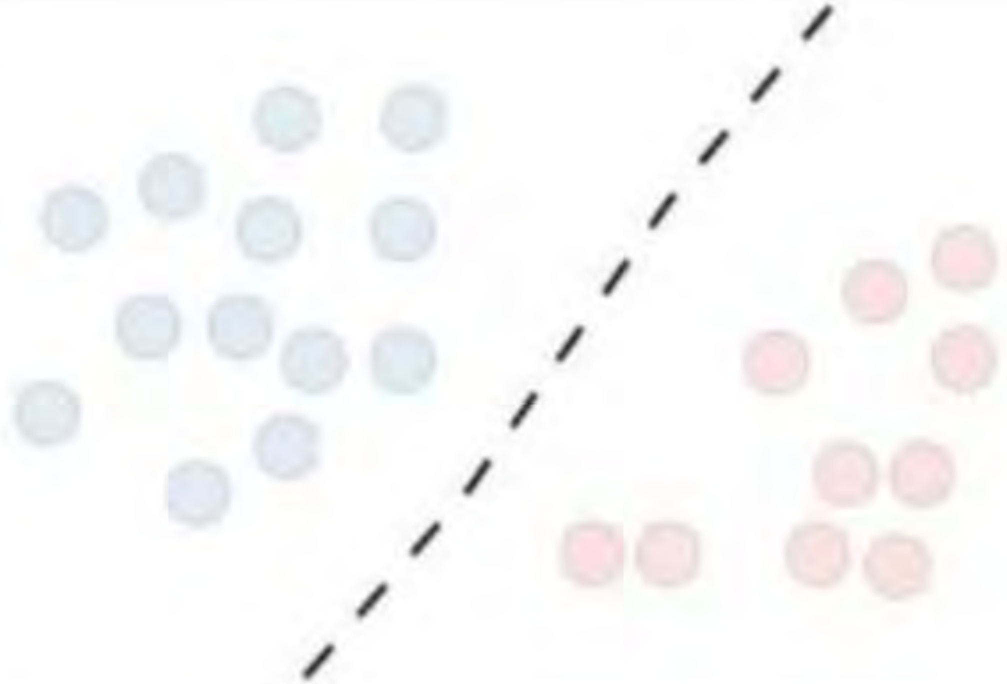
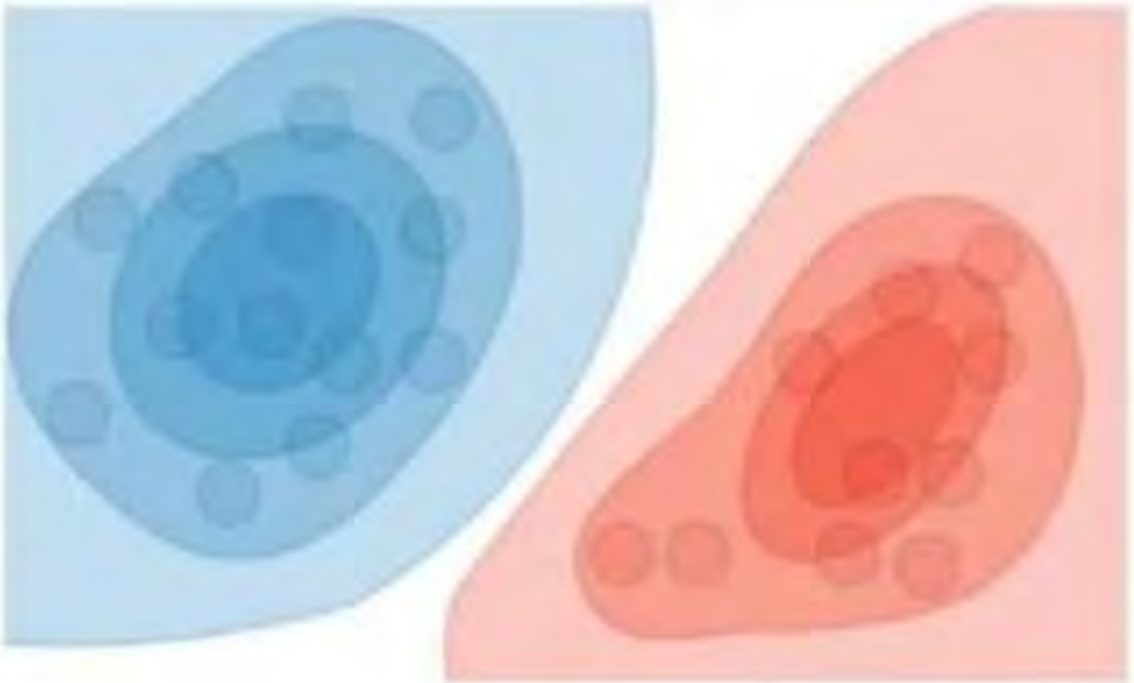
- Generative models are machine learning models that learn to generate new data samples similar to the training data they were trained on. They capture the underlying distribution of the data and can produce novel instances.
- So, the Generative approach focuses on the distribution of individual classes 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.

Bayesian Decision Theory

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

Given a discrete K-class dataset containing N points, where sample points are described using D features with each feature capable of taking V values, how many parameters need to be estimated for Naïve Bayes Classifier?

(A)	$V^D K$	(C)	VDK
(B)	K^{V^D}	(D)	$K(V + D)$

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

- 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

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.

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

- A Pass
- B Fail

THANK - YOU