Data Science and Artificial Intelligence

Machine Learning

Bayesian learning

Lecture No. 2

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — MLE
- Topic — MAP
- Topic — (decision)
- Topic
- Topic

Topics to be Covered

- **Topic** — Bayes classifier ✓
- **Topic** — Why it is not used ✓
- **Topic** — Naive Bayes ✓
- **Topic**
- **Topic**

STOP DOUBTING YOURSELF.
WORK HARD AND MAKE IT HAPPEN.

Summary of the last class

2 approach for decision

Approach ⇒ $\underbrace{P(X/C_1) \quad P(X|C_2)}_{\text{Class Conditioned PDF}}$

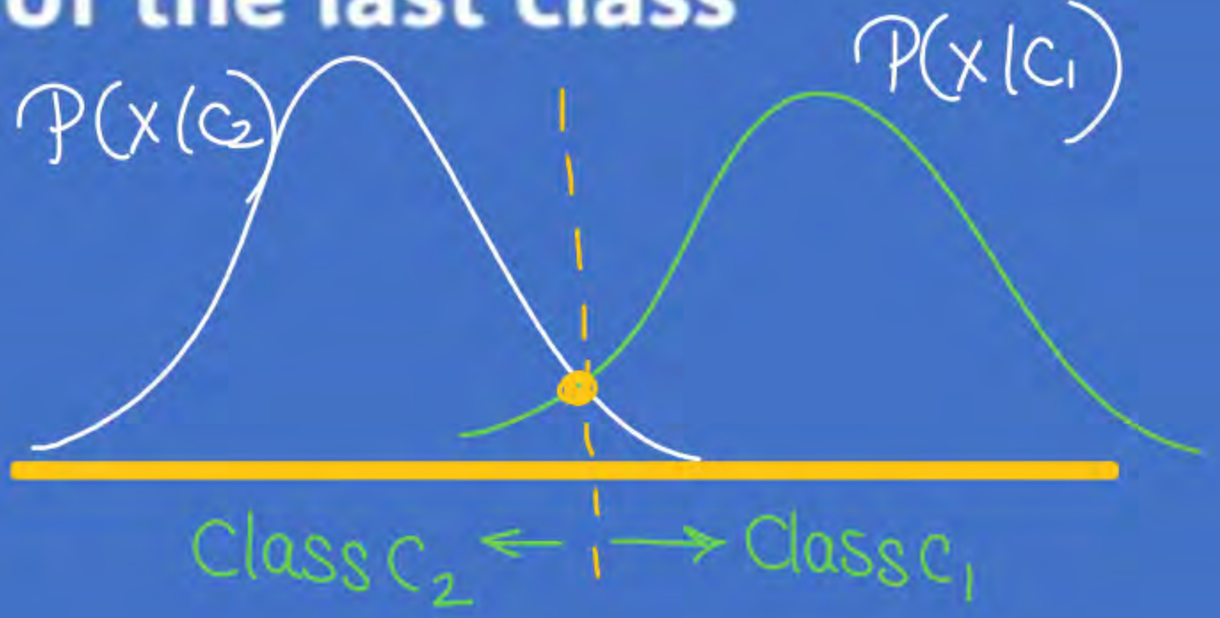
$\underbrace{P(X|C_1) \overset{C_1}{>} \overset{C_2}{<} P(X|C_2)}$

Summary of the last class

C_1 → mean, σ $\Rightarrow P(x|C_1)$

C_2 → mean, var $\Rightarrow P(x|C_2)$

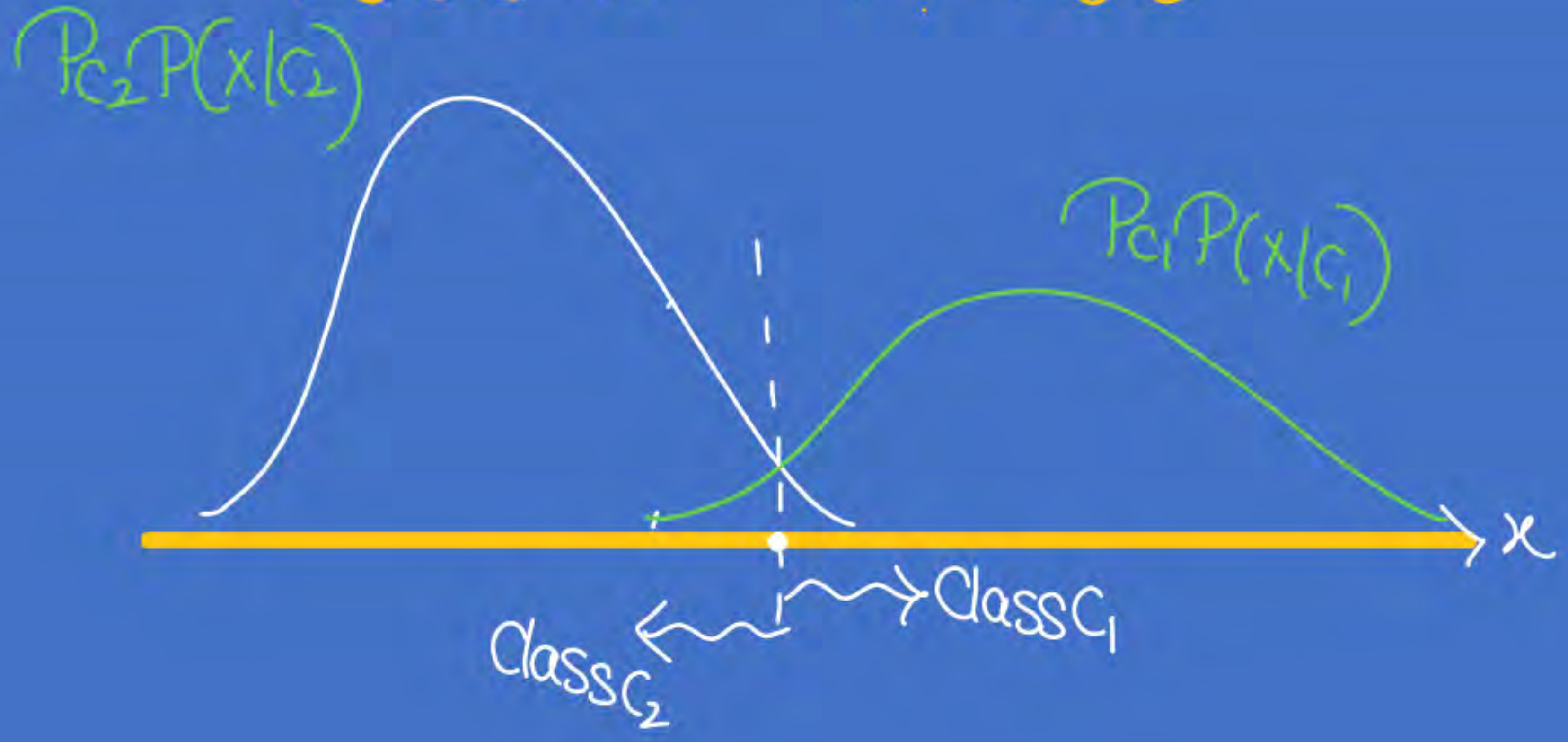
Summary of the last class



The probability of error in MAP case ...

So Aposterior prob ⇒

$$P(x|c_1)P_{c_1} \overset{c_1}{>} \underset{}{\overset{c_2}{<}} P_{c_2}P(x|c_2)$$



The probability of error in MAP case ...

$$P_e = \int_{-\infty}^{\theta} P_{C_1}P(x|C_1)\,dx + \int_{\theta}^{\infty} P_{C_2}P(x|C_2)\,dx$$

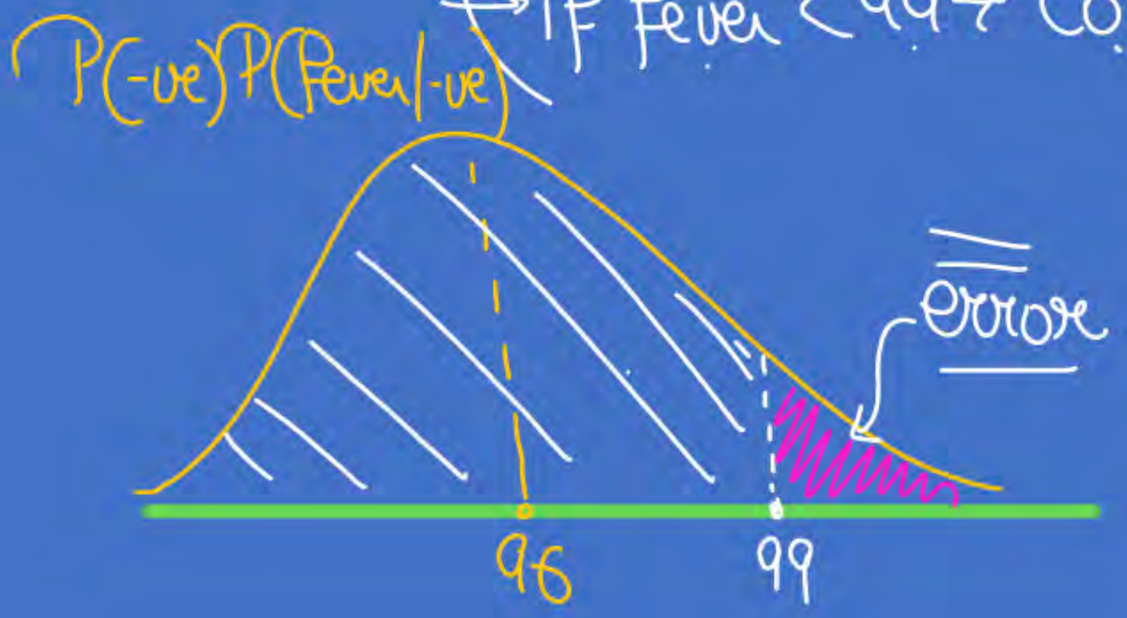
The probability of error in MAP case ...

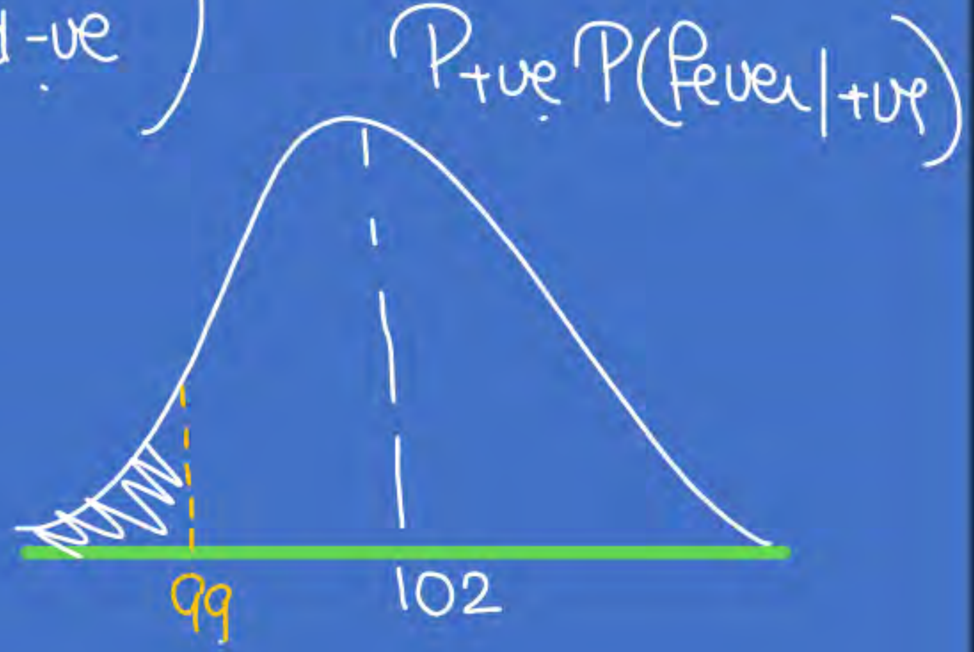
ex

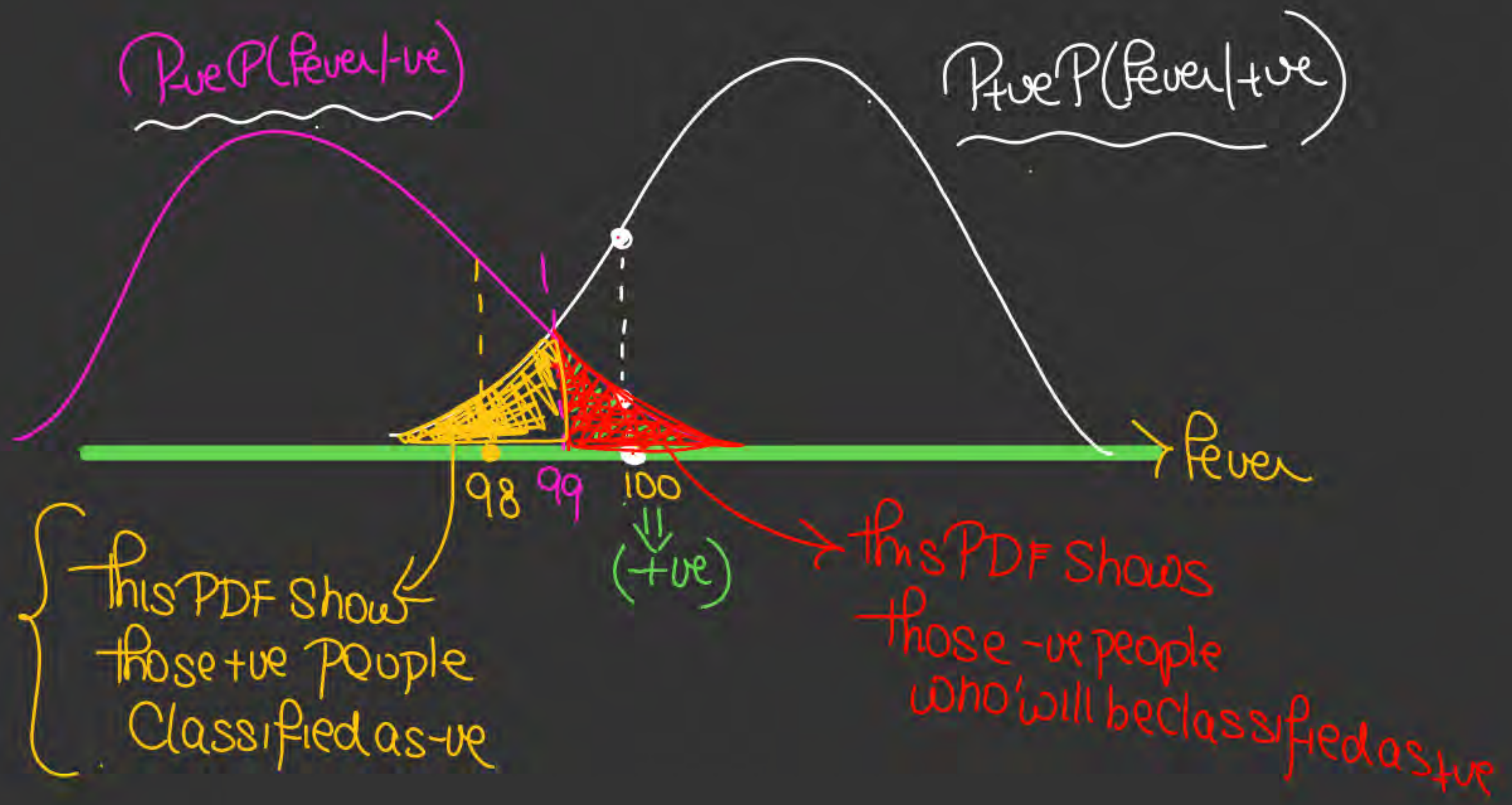
threshold ⇒ 99

→ if fever > 99 → Covid +ve

→ if fever < 99 → Covid -ve






(P-ve P(fever|-ve))
P+ve P(fever|+ve)
fever
98
99
100
(+ve)
This PDF show those +ve people classified as -ve
This PDF shows those -ve people who will be classified as +ve

- hypothesis ⇒ * in the previous class using the sample of data we try to predict the whole PDF of the data
  * in ML we try to create models of the data using various algo.

What is a hypothesis ?

A hypothesis is a conjecture or proposed explanation that is based on insufficient facts or assumptions. It is only a conjecture based on certain known facts that have yet to be confirmed.

we have some Samples, using that we try to make predictions from whole data

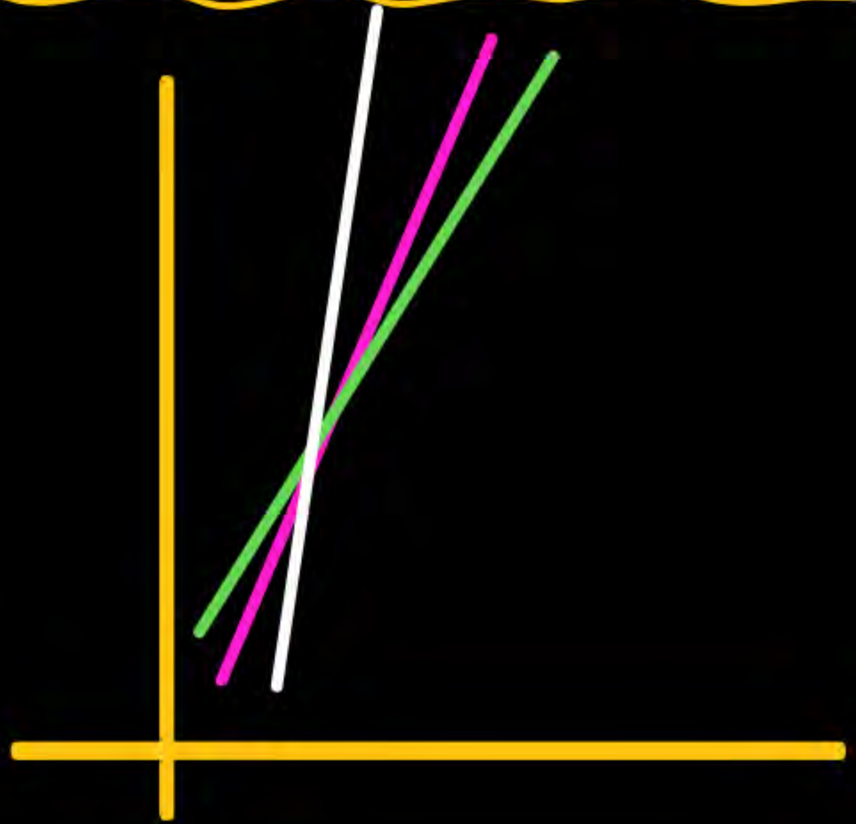
What is a hypothesis space

So if we divide the sample into subsets and then create the predictions of the distribution of the whole data then we will get many different distributions....

→ models

→ hypothesis

the Combination of this is Called hypothesis space



Sample

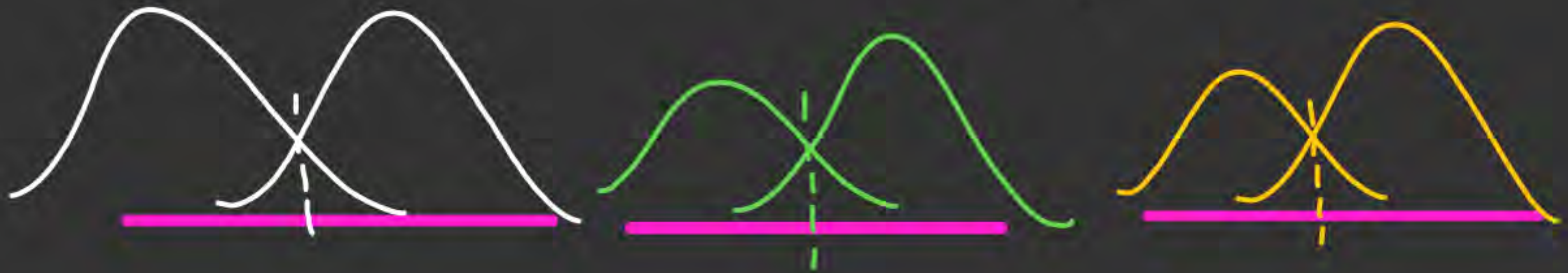
hypethesis space inbayesian learning $\Rightarrow$

H_1 Subset 1	H_2 Subset 2	H_3 Subset 3
$P_{c_1} \Rightarrow$	$P_{c_1} \Rightarrow$	$P_{c_1} \Rightarrow$
$P_{c_2} \Rightarrow$	$P_{c_2} \Rightarrow$	$P_{c_2} \Rightarrow$
$P(x \mid c_1)$	$P(x \mid c_1)$	$P(x \mid c_1)$
$P(x \mid c_2)$	$P(x \mid c_2)$	$P(x \mid c_2)$

MAP

all 3 will have some variation

$(H_1, H_2, H_3) \Rightarrow$ hypothesis space



So we have a hypothesis space...

- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

We can have two tasks

not Imp ✓✓.

1. Most probable hypothesis
2. Most probable classification ✓

   for any new point

Bayesian Decision Theory

Only for Knowledge

How to find the most probable hypothesis

- **P(h)** is ***prior probability of hypothesis h***
  - P(h) to denote the initial probability that hypothesis h holds, before observing training data.
  - P(h) may reflect any background knowledge we have about the chance that h is correct. If we have no such prior knowledge, then each candidate hypothesis might simply get the same prior probability. ← generally P(h) ⇒ equal = $\frac{1}{\text{Number of hypothesis}}$
- **P(D)** is ***prior probability of training data D***
  - The probability of D given no knowledge about which hypothesis holds
- **P(h|D)** is ***posterior probability of h given D***
  - P(h|D) is called the ***posterior probability*** of ***h***, because it reflects our confidence that ***h*** holds after we have seen the training data ***D***.
  - The posterior probability P(h|D) reflects the influence of the training data ***D***, in contrast to the prior probability P(h), which is independent of D.
- **P(D|h)** is ***posterior probability of D given h***
  - The probability of observing data ***D*** given some world in which hypothesis ***h*** holds.
  - Generally, we write ***P(x|y)*** to denote the probability of **event x** given **event y**.

Only for Knowledge

How to find the most probable hypothesis

- So to find the best hypothesis/model. we simply take model or hypothesis and check its accuracy on whole data
- Now the hypothesis which gives min error will be the best.

Only for Knowledge

How to find the most probable hypothesis

Bayes theorem

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

$$\Rightarrow \left\{ P(h/D) = \frac{P(D/h) \cdot P(h)}{P(D)} \right\}$$

So $P(h/D) \Rightarrow$ if we have data, and hypothesis h_1, h_2, h_3, h_4 - -

then $P(h_1/D) \Rightarrow$ Probab of getting hypothesis h_1 given Data

$P(h_2/D) \Rightarrow$ " " " " h_2 " "

⋮

$P(D/h) \Rightarrow$

So $P(D/h) \Rightarrow \dfrac{P(h/D) \cdot P(D)}{P(h)}$ ← fix value not in our Control

← generally if no prior knowledge of hypothesis $P(h) = \dfrac{1}{\text{No of hypothesis}}$

$P(D/h)$ → we have hypothesis What is probab to get the data

$P(h/D) \Rightarrow$ data is given what is probab of hypothesis

$\Downarrow$

Sample
- → Subset → μ_1, σ_1^2
- → Subset → μ_2, σ_2^2
- → Subset → μ_3, σ_3^2

• Conclude → $P(h)$ = Prior Probab of hypothesis

→ How to find best hypothesis

ML

→ if we donot have $P(h)$

→ we decide on basis of $P(D/h)$

→ hypothesis with max $P(D/h)$ is best hypothesis

→ MAP case

if we have $P(h)$

we use $P(D/h).P(h)$

→ $P(h/D)$

So hypothesis which has max $P(h/D)$ is best hypothesis

Which is the best hypothesis or most probable hypothesis

done

- The learner considers some set of candidate hypotheses H and it is interested in finding the ***most probable hypothesis*** h ∈ H given the observed data D
- Any such maximally probable hypothesis is called a ***maximum a posteriori (MAP) hypothesis*** h_{MAP}.
- We can determine the MAP hypotheses by using Bayes theorem to calculate the posterior probability of each candidate hypothesis.

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)}$$

$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$

Which is the best hypothesis or most probable hypothesis

done
- If we assume that every hypothesis in H is equally probable

  i.e. $P(h_i) = P(h_j)$ for all h_i and h_j in H

  We can only consider $P(D|h)$ to find the most probable hypothesis.

- $P(D|h)$ is often called the ***likelihood*** of the data ***D*** given ***h***
- Any hypothesis that maximizes $P(D|h)$ is called a ***maximum likelihood (ML) hypothesis***, h_{ML}.

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}}\, P(D|h)$$

How to find the most probable classification...

So, we can say that we have divided the sample into subsets and got many hypothesis. Then these hypothesis Are used for decision.

We use the MAP rule for final classification and decision making.

Sample $\rightarrow$ Subsets $\rightarrow$ $\{H_1, H_2, H_3 \ldots\}$ $\Rightarrow$ $P(H/D)$ for each hypothesis

Bayes optimum classifier :- any new point is tested from each hypothesis

Work on MAP Rule

$$\begin{cases} H_1 \rightarrow C_1 \checkmark \\ H_2 \rightarrow C_2 \checkmark \\ H_3 \rightarrow C_1 \checkmark \\ H_4 \rightarrow C_1 \checkmark \end{cases}$$

- the the o/p of each hypothesis is weighted by $P(h/D)$

Bayes optimal Classifier

The final result is

$$\sum_{i=1}^{m} P(h_i|D) \, C_j \rightarrow \text{Class decided by hypothesis}$$

m number of hypothesis

the class with max Coeff in expression is choosen as result.

How to find the most probable classification...

Consider a hypothesis space containing three hypotheses, hl, h2, and h3.

- Suppose that the posterior probabilities of these hypotheses given the training data are .4, .3, and .3 respectively.
- Thus, hl is the MAP hypothesis.
- Suppose a new instance x is encountered, which is classified positive by *hl,* but negative by *h2* and h3.
- Taking all hypotheses into account, the probability that x is positive is .4 (the probability associated with *h1),* and the probability that it is negative is therefore .6.
- The most probable classification (negative) in this case is different from the classification generated by the MAP hypothesis.

How to find the most probable classification...

Suppose our hypothesis space H has three functions h_1, h_2 and h_3

- $P(h_1 \mid D) = 0.4$, $P(h_2 \mid D) = 0.3$, $P(h_3 \mid D) = 0.3$
- What is the MAP hypothesis? $\boxed{h_1.}$ most probable hypothesis. i.e hypothesis with max $P(h/D)$ is Ans.
- For a new instance **x**, suppose $h_1(\mathbf{x}) = +1$, $h_2(\mathbf{x}) = -1$ and $h_3(\mathbf{x}) = -1$
- What is the most probable classification of **x**?

Bayes optimum classification...

$H = \{h1, h2, h3\}$

Training Data [D]

$P(h1|D) = .4$

$P(h2|D) = .3$

$P(h3|D) = .3$

For new datapoint

$P(c1|h1) = 0 \quad P(c2|h1) = 1 \Rightarrow h_1$ Predict c_2

$P(c1|h2) = 1 \quad P(c2|h2) = 0 \Rightarrow h_2$ Predict c_1

$P(c1|h3) = 1 \quad P(c2|h3) = 0 \Rightarrow h_3$ Predict c_1

$\Rightarrow$ find class of new data point $\Rightarrow$ (Class c_1/c_2)

find $\sum_{i=1}^{3} P(h_i/D)\, c_j \Rightarrow .4c_2 + .3c_1 + .3c_1$

$\Rightarrow .4c_2 + .6c_1 \Rightarrow c_1$

Computational complexity of Bayes optimum Classifier

The decision Rule of Bayes optimum classifier can also be written as

$$\left(\sum_{i=1}^{m} P(h_i/D)\, P(C_j/h_i)\right)$$

m number of hypothesis

Calculated for all classes

Class which has max value of this is the result.

Bayes optimum classification...

$H = \{h1, h2, h3\}$

Training Data [D]

For new datapoint

$P(c1|h1) = 0$ $P(c2|h1) = 1$ ⇒ h_1 Predict C_2

$P(c1|h2) = 1$ $P(c2|h2) = 0$ ⇒ h_2 Predict C_1

$P(c1|h3) = 1$ $P(c2|h3) = 0$ ⇒ h_3 Predict C_1

$P(h1|D) = .4$

$P(h2|D) = .3$

$P(h3|D) = .3$

for C_1 ⇒ $\sum_{i=1}^{3} P(h_i/D)\,P(C_1/h_i)$ ⇒ $.4 \times 0 + .3 \times 1 + .3 \times 1$ ⇒ $.6$

for C_2 ⇒ $\sum_{i=1}^{3} P(h_i/D)\,P(C_2/h_i)$ ⇒ $.4 \times 1 + .3 \times 0 + .3 \times 0$ ⇒ $.4$

Gibbs Algorithm

- An alternative, less optimal method is the Gibbs algorithm:

1. Choose a hypothesis ***h*** from H at random, according to the posterior probability distribution over H.
2. Use ***h*** to predict the classification of the next instance ***x***.

Naïve Bayes Algorithm

The fundamental Naive Bayes assumption is that each feature makes an:

- **Feature independence:** The features of the data are conditionally independent of each other, given the class label.
- **Features are equally important:** All features are assumed to contribute equally to the prediction of the class label.

Bayesian Decision Theory

Naïve Bayes Classifier

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

We have to calculate

....

Bayesian Decision Theory

Naïve Bayes Classifier

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		
Overcast		
Rain		

Naïve Bayes Classifier

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

Temperature	P(T/Yes)	P(T/No)
Hot		
Mild		
Cold		

Naïve Bayes Classifier

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

Humidity	P(H/Yes)	P(H/No)
High		
Normal		

Bayesian Decision Theory

Naïve Bayes Classifier

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

Wind	P(W/Yes)	P(W/No)
Weak		
Strong		

Naïve Bayes Algorithm

Outlook	Temperature	Humidity	Wind
Sunny	Cool	High	Strong

Naïve Bayes Algorithm

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

Bayesian Decision Theory

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		
Overcas		

Temperature	P(T/Yes)	P(T/No)
Hot		
Mild		

Humidity	P(H/Yes)	P(H/No)
High		
Normal		

Wind	P(W/Yes)	P(W/No)
Weak		
Strong		

Naïve Bayes Algorithm

What if the dimension are continuous in nature

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						

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 $y \in \{1,2,\dots,10\}$, 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

THANK - YOU