M.S. Ramaiah Institute of Technology
(Autonomous Institute, Affiliated to VTU)
Department of Computer Science and Engineering

Artificial Intelligence and Machine Learning (CS52)

UNIT - 4

OUTLINE

Artificial Neural Networks - Introduction, Neural Network Representation, Appropriate problems for Neural Network Learning, Perceptrons, Multilayer Networks and the Backpropagation algorithm.

Bayesian Learning - Introduction, Bayes theorem, Naive Bayes Classifier, The EM Algorithm.

Chapter 4 and 6(6.1,6.2,6.9,6.12) of TextBook2

Define the Bayesian theorem? What are the relevance and features of the Bayesian theorem? Explain the practical difficulties of the Bayesian theorem.

Introduction

Bayesian learning methods are relevant to study of machine learning for two different reasons.

- First, Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems
- The second reason is that they provide a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities.

Introduction

Features of Bayesian Learning Methods

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.
- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
- Bayesian methods can accommodate hypotheses that make probabilistic predictions
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
- Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

Introduction

Practical difficulty in applying Bayesian methods

- One practical difficulty in applying Bayesian methods is that they typically require initial knowledge of many probabilities. When these probabilities are not known in advance they are often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.

- A second practical difficulty is the significant computational cost required to determine the Bayes optimal hypothesis in the general case. In certain specialized situations, this computational cost can be significantly reduced.

Conditional Probability

Is defined as the probability of an event A, given that another event B has already occurred (i.e. A conditional B).

This is represented by P(A|B) and we can define it as:

$$\underbrace{P(A \mid B)}_{\text{Probability of } A \text{ given } B} = \frac{\overbrace{P(A \cap B)}^{\text{Probability of } A \text{ and } B}}{\underbrace{P(B)}_{\text{Probability of } B}}$$

Example:

Susan took two tests. The probability of her passing both tests is 0.6. The probability of her passing the first test is 0.8. What is the probability of her passing the second test given that she has passed the first test?

Solution:

$$P(\text{second} \mid \text{first}) = \frac{P(\text{first and second})}{P(\text{first})} = \frac{0.6}{0.8} = 0.75$$

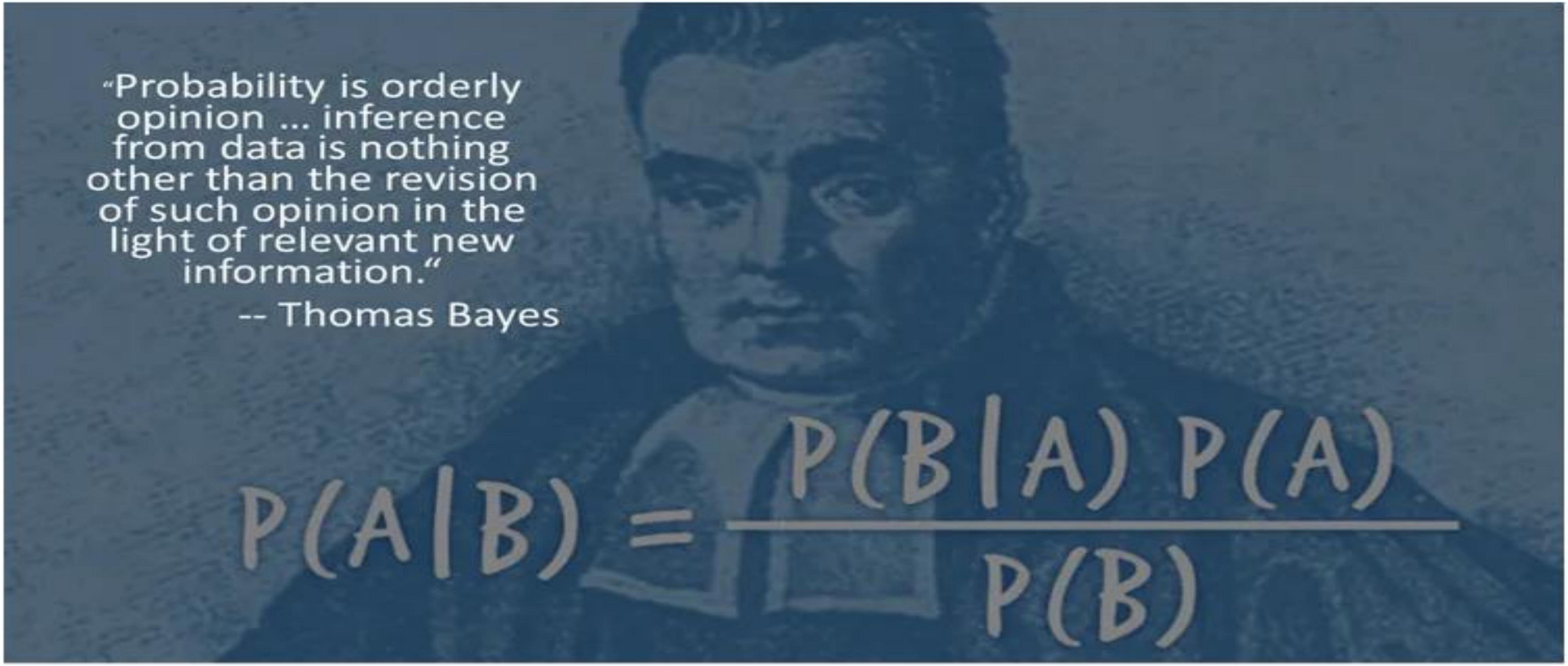
Probability of A and B

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

Probability of A given B

Probability of B

What is Bayes' Theorem?



Consider that A and B are any two events from a sample space S

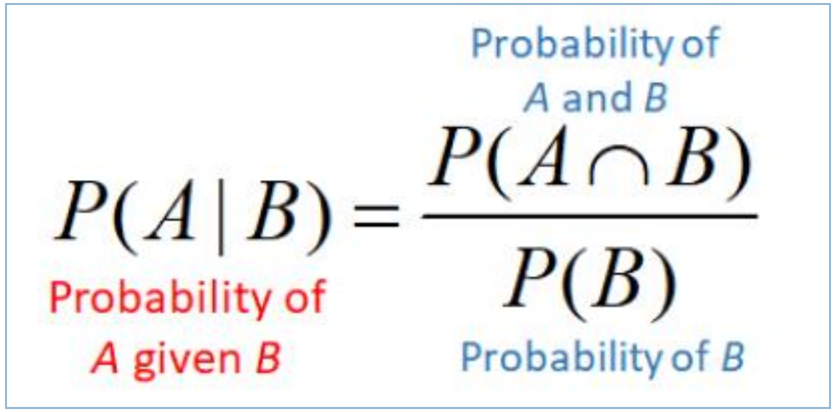
Using our understanding of conditional probability, we have:

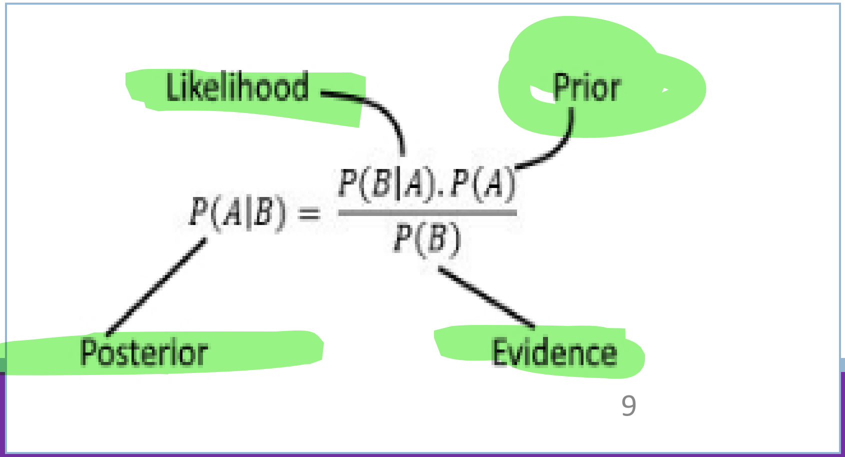
$P(A|B) = P(A \cap B) / P(B)$

$P(B|A) = P(A \cap B) / P(A)$

It follows that $P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$

Thus, $\boxed{P(A|B) = P(B|A)*P(A) / P(B)}$





RAMAIAH
Institute of Technology

Explain the concept of Bayesian Learning and the Naive Bayes Classifier.

BAYES THEOREM

Bayes theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself.

Notations

- P(h|D) posterior probability of h, reflects confidence that h holds after D has been observed
- P(h) initial or prior probability that hypothesis h holds, before we have observed the training data.
- P(D) prior probability that training data D will be observed
- P(D|h) probability of observing data D given a world in which hypothesis h holds

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$P(h)$ = prior probability of hypothesis h

$P(D)$ = prior probability of training data D

$P(h|D)$ = probability of h given D

$P(D|h)$ = probability of D given h

RAMAIAH Institute of Technology

BAYES THEOREM

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$$\begin{aligned} P(h) &= \text{prior probability of hypothesis } h \\ P(D) &= \text{prior probability of training data } D \\ P(h|D) &= \text{probability of } h \text{ given } D \\ P(D|h) &= \text{probability of } D \text{ given } h \end{aligned}$$

P(h|D) increases with ***P(h)*** and with ***P(D|h)*** according to Bayes theorem.

P(h|D) decreases as ***P(D)*** increases, because the more probable it is that ***D*** will be observed independent of ***h*,** the less evidence ***D*** provides in support of ***h*.**

Define is Maximum a Posteriori (MAP), Maximum Likelihood (ML) Hypothesis. Derive the relation for hMAP and hML using the Bayesian theorem.

BAYES THEOREM

Maximum a posteriori (MAP) hypothesis

The learner considers some set of **candidate hypotheses H** and 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.

Bayes theorem to calculate the posterior probability of each candidate hypothesis is $\mathbf{h_{MAP}}$ is a MAP hypothesis provided

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

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

discard

P(D) can be dropped, because it is a constant independent of h

BAYES THEOREM

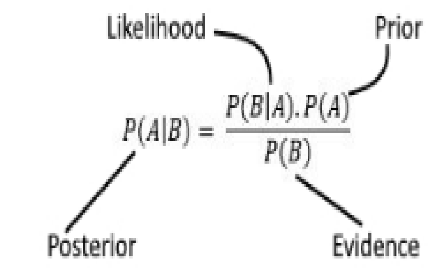
Maximum Likelihood (ML) Hypothesis

In some cases, it is assumed that every hypothesis in H is equally probable a priori **($P(h_i) = P(h_j)$** for all $\mathbf{h_i}$ and $\mathbf{h_j}$ in H).

In this case the below equation can be simplified and need only consider the term **P(D|h)** to find the most probable hypothesis

$$h_{MAP} = \arg\max_{h \in H} P(D|h)P(h)$$

$$h_{MAP} = \arg\max_{h \in H} P(D|h)\cancel{P(h)}$$



$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Bayes Theorem:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(D|h) is often called the **likelihood** of the data **D** given **h**, and any hypothesis that maximizes **P(D|h)** is called a **maximum likelihood** (ML) hypothesis

RAMAIAH Institute of Technology

BAYES THEOREM - Example

Consider a medical diagnosis problem in which there are two alternative hypotheses:

(1) Patient has a particular form of cancer (+)

(2) Patient does not have any form of cancer (-)

A patient takes a lab test and the results comes positive.

The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present.

Furthermore, 0.008 of the entire population have this cancer.

Determine whether the patient has Cancer or not using MAP hypothesis

BAYES THEOREM - Example

Two alternative hypotheses

- The patient has a particular form of cancer (denoted by ***cancer***)
- The patient does not (denoted by **¬ *cancer***)

The available data is from a particular laboratory with two possible outcomes:

+ (positive) and - (negative)

$$P(cancer) = .008 \qquad P(\neg cancer) = 0.992$$

1 - 0.008 = 0.992

$$P(\oplus|cancer) = .98 \qquad P(\ominus|cancer) = .02$$

2 /100 = 0.02

$$P(\oplus|\neg cancer) = .03 \qquad P(\ominus|\neg cancer) = .97$$

3 /100 = 0.03

RAMAIAH
Institute of Technology

BAYES THEOREM - Example

Suppose a new patient is observed for whom the lab test returns a positive (+) result.

Should we diagnose the patient as having cancer or not?

$$P(h|D) = \frac{P(D|h)P(h)}{\cancel{P(D)}}$$

$$P(cancer) = .008 \qquad P(\neg cancer) = 0.992$$
$$P(\oplus|cancer) = .98 \qquad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03 \qquad P(\ominus|\neg cancer) = .97$$

$$P(cancer|+) = P(+|cancer) * P(cancer) = 0.98 * 0.008 = 0.0078$$

$$P(\neg cancer|+) = P(+|\neg cancer) * P(\neg cancer) = 0.03 * 0.992 = 0.0298$$

$$h_{MAP} = \neg cancer$$

Hence, the new patient with lab test positive is not having cancer

RAMAIAH
Institute of Technology

BAYES THEOREM

- Suppose we now observe a new patient for whom the lab test returns a **negative** result.
- Should we diagnose the patient as having cancer or not?

$$P(h|D) = \frac{P(D|h)P(h)}{\cancel{P(D)}}$$

$P(cancer) = .008$	$P(\neg cancer) = 0.992$
$P(\oplus|cancer) = .98$	$P(\ominus|cancer) = .02$
$P(\oplus|\neg cancer) = .03$	$P(\ominus|\neg cancer) = .97$

$$P(cancer|-) = P(-|cancer) * P(cancer) = 0.02 * 0.008 = 0.00016$$

$$P(\neg cancer|-) = P(-|\neg cancer) * P(\neg cancer) = 0.97 * 0.992 = 0.96224$$

$$h_{MAP} = \neg cancer$$

BAYES THEOREM

For any propositions a and b, we have

Conditional probability can be written in a different form called the Product rule:

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)},$$

$$P(a \wedge b) = P(a \mid b) P(b)$$

BAYES THEOREM

- *Product rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

- *Sum rule*: probability of a disjunction of two events A and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Bayes theorem*: the posterior probability $P(h|D)$ of h given D

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

TABLE 6.1
Summary of basic probability formulas.

Explain the concept of the Naive Bayes Classifier.

RAMAIAH Institute of Technology

NAIVE BAYES CLASSIFIER

One highly practical Bayesian learning method is the naive Bayes learner, often called the **naive Bayes classifier.**

The naive Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f (x) can take on any value from some finite set V.

RAMAIAH
Institute of Technology

NAIVE BAYES CLASSIFIER

A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values (al, a2.. .a,).

The learner is asked to predict the target value, or classification, for this new instance.

RAMAIAH Institute of Technology

NAIVE BAYES CLASSIFIER

The Bayesian approach to classifying the new instance is to assign the most probable target value, V_{MAP}, given the attribute values $<a_l, a_2 \ldots a_n>$ that describe the instance.

$$v_{MAP} = \operatorname*{argmax}_{v_j \in V} P(v_j | a_1, a_2 \ldots a_n)$$

We can use Bayes theorem to rewrite this expression as

$$v_{MAP} = \operatorname*{argmax}_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \operatorname*{argmax}_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

Bayes Theorem:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

NAIVE BAYES CLASSIFIER

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n)$$

We can use Bayes theorem to rewrite this expression as

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

Naive Bayes classifier:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$$

Bayes Theorem:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

RAMAIAH
Institute of Technology

NAIVE BAYES CLASSIFIER - Example

The following table gives data set about target concept PlayTennis. Using Naïve Bayes classifier classify the following novel instance:

(Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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
D15	Sunny	Cool	High	Strong	?

RAMAIAH Institute of Technology

NAIVE BAYES CLASSIFIER - Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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

$$P(PlayTennis = yes) = 9/14 = .64$$

$$P(PlayTennis = no) = 5/14 = .36$$

RAMAIAH
Institute of Technology

P(SUNNY/YES) = P(YES/SUNNY) * P(YES)

P(YES/SUNNY) = P(SUNNY/YES) / P(YES)
= (2/14) / (9/14) = 2/9

NAIVE BAYES CLASSIFIER - Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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

$$P(PlayTennis = yes) = 9/14 = .64$$

$$P(PlayTennis = no) = 5/14 = .36$$

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5

Windy		
Strong	3/9	3/5
Weak	6/9	2/5

NAIVE BAYES CLASSIFIER - Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j)$$

$$\cdot P(Humidity = high|v_j) P(Wind = strong|v_j)$$

NAIVE BAYES CLASSIFIER - Example

$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$

$$v_{NB} = \underset{v_j \in \{yes, no\}}{\text{argmax}}\; P(v_j) \prod_i P(a_i|v_j)$$

$$v_{NB}(yes) = P(yes)\; P(sunny|yes)\; P(cool|yes)\; P(high|yes)\; P(strong|yes)$$

$= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053$

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5
Windy		
Strong	3/9	3/5
Weak	6/9	2/5

Probability that we can play the game.

> P(Outlook=Sunny | Play=Yes) = 2/9
> P(Temperature=Cool | Play=Yes) = 3/9
> P(Humidity=High | Play=Yes) = 3/9
> P(Wind=Strong | Play=Yes) = 3/9
> P(Play=Yes) = 9/14

RAMAIAH
Institute of Technology

NAIVE BAYES CLASSIFIER - Example

$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

$$v_{NB}(no) = P(no)\, P(sunny|no)\, P(cool|no)\, P(high|no)\, P(strong|no)$$

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5
Windy		
Strong	3/9	3/5
Weak	6/9	2/5

$= (3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0.0206$

Probability we cannot play a game:

> P(Outlook=Sunny | Play=No) = 3/5

> P(Temperature=Cool | Play=No) = 1/5

> P(Humidity=High | Play=No) = 4/5

> P(Wind=Strong | Play=No) = 3/5

> P(Play=No) = 5/14

Thus, the naive Bayes classifier assigns the target value **PlayTennis = no to this** new instance, based on the probability estimates learned from the training data.

RAMAIAH
Institute of Technology

NAIVE BAYES CLASSIFIER - Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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

Calculate the conditional probability that the target value is **no, given the observed attribute** values.

$$= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053$$

$$= (3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0.0206$$

$$v_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = 0.795$$

NORMALIZATION $\frac{.0206}{.0206+.0053} = .795$

NAIVE BAYES CLASSIFIER - Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
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

$$= (2/9) * (3/9) * (3/9) * (3/9) * (9/14) = 0.0053$$

$$= (3/5) * (1/5) * (4/5) * (3/5) * (5/14) = 0.0206$$

$$v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes) + v_{NB}(no)} = 0.205$$

NAIVE BAYES CLASSIFIER - Example

The following table gives data set about stolen vehicles. Using Naïve bayes classifier classify the new data (Red, SUV, Domestic).

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

RAMAIAH Institute of Technology

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Target class

	Color	Yes	No
Values	Red	3	2
	Yellow	2	3

$P(\text{Color} = \text{Red} \mid \text{Stolen} = \text{Yes}) = \frac{3}{5} = 0.6$

$P(\text{Color} = \text{Red} \mid \text{Stolen} = \text{No}) = \frac{2}{5} = 0.4$

$P(\text{Color} = \text{Yellow} \mid \text{Stolen} = \text{Yes}) = \frac{2}{5} = 0.4$

$P(\text{Color} = \text{Yellow} \mid \text{Stolen} = \text{No}) = \frac{3}{5} = 0.6$

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Type	Yes	No
Sport	4	2
SUV	1	3

P(Type = Sport | Stolen = yes) = 4/5 = 0.8

P(Type = Sport | Stolen = No) = 2/5 = 0.4

P(Type = SUV | stolen = yes) = 1/5 = 0.2

P(Type = SUV | Stolen = No) = 3/5 = 0.6

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Target

Value

Origin	yes	No
Domestic	2	3
Imported	3	2

$P(\text{Origin} = \text{Domestic} \mid \text{Stolen} = \text{yes}) = 2/5 = 0.4$

$P(\text{origin} = \text{Domestic} \mid \text{stolen} = \text{NO}) = 3/5 = 0.6$

$P(\text{origin} = \text{Imported} \mid \text{stolen} = \text{yes}) = 3/5 = 0.6$

$P(\text{origin} = \text{Imported} \mid \text{Stolen} = \text{no}) = 2/5 = 0.4$

RAMAIAH Institute of Technology

Classify the new data = (Red, SUV, Domestic)

* For Stolen = yes :

⇒ (color = Red | stolen = yes) * (Type = SUV | stolen = yes) * (Origin = Domestic | Stolen = yes) * P(yes)

⇒ 0.6 * 0.2 * 0.4 * 0.5

⇒ 0.024

* For Stolen = No :

⇒ (color = Red | Stolen = No) * (Type = SUV | stolen = No) * (Origin = Domestic | stolen = No) * P(NO)

⇒ 0.4 * 0.6 * 0.6 * 0.5

⇒ 0.072

So, we would classify the new data as not Stolen

RAMAIAH
Institute of Technology

NAIVE BAYES CLASSIFIER - Example

b. Estimate conditional probabilities of each attributes {colour, legs, height, smelly} for the species classes: {M, H} using the data given in the table. Using these probabilities estimate the probability values for the new instance – (Colour = Green, Legs = 2, Height = Tall and Smelly = No) **(10 Marks)**

No	Colour	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

When to use:

- Data is only partially observable
- Unsupervised clustering (target value unobservable)
- Supervised learning (some instance attributes unobservable)

Some uses:

- Train Bayesian Belief Networks
- Unsupervised clustering (AUTOCLASS)
- Learning Hidden Markov Models

The EM Algorithm

Estimation: Estimate the expectation from machine on some data to cluster

Maximization: Whatever is estimated should be maximized to find best result.

It starts with random data and repeats two steps till best result.

E Step: Cluster based on current data

M Step: Generate best theory to get best clusters.

Application:

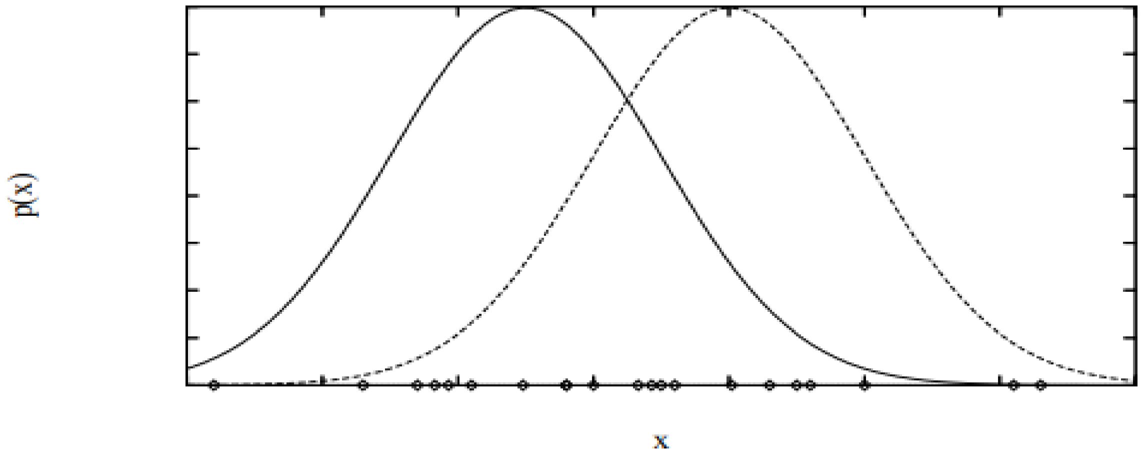
- Clustering
- Artificial Vision
- Biological areas
- NLP

Algorithm:

1. Given a set of incomplete data, consider a set of starting parameters.
2. **Expectation step (E – step):** Using the observed available data of the dataset, estimate (guess) the values of the missing data.
3. **Maximization step (M – step):** Complete data generated after the expectation (E) step is used in order to update the parameters.
4. Repeat step 2 and step 3 until convergence.

The EM Algorithm

Generating Data from Mixture of k Gaussians



Each instance x generated by

1. Choosing one of the k Gaussians with uniform probability
2. Generating an instance at random according to that Gaussian

EM for Estimating k Means

Given:

- Instances from X generated by mixture of k Gaussian distributions
- Unknown means $\langle \mu_1, \ldots, \mu_k \rangle$ of the k Gaussians
- Don't know which instance x_i was generated by which Gaussian

Determine:

- Maximum likelihood estimates of $\langle \mu_1, \ldots, \mu_k \rangle$

Think of full description of each instance as $y_i = \langle x_i, z_{i1}, z_{i2} \rangle$, where

- z_{ij} is 1 if x_i generated by jth Gaussian
- x_i observable
- z_{ij} unobservable

RAMAIAH

EM Algorithm: Pick random initial $h = \langle \mu_1, \mu_2 \rangle$, then iterate

E step: Calculate the expected value $E[z_{ij}]$ of each hidden variable z_{ij}, assuming the current hypothesis $h = \langle \mu_1, \mu_2 \rangle$ holds.

$$E[z_{ij}] = \frac{p(x = x_i | \mu = \mu_j)}{\sum_{n=1}^{2} p(x = x_i | \mu = \mu_n)} = \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{n=1}^{2} e^{-\frac{1}{2\sigma^2}(x_i - \mu_n)^2}}$$

M step: Calculate a new maximum likelihood hypothesis $h' = \langle \mu'_1, \mu'_2 \rangle$, assuming the value taken on by each hidden variable z_{ij} is its expected value $E[z_{ij}]$ calculated above. Replace $h = \langle \mu_1, \mu_2 \rangle$ by $h' = \langle \mu'_1, \mu'_2 \rangle$.

$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}] \; x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

General EM Problem

Given:

- Observed data $X = \{x_1, \ldots, x_m\}$
- Unobserved data $Z = \{z_1, \ldots, z_m\}$
- Parameterized probability distribution $P(Y|h)$, where
  - $Y = \{y_1, \ldots, y_m\}$ is the full data $y_i = x_i \cup z_i$
  - h are the parameters

Determine:

- h that (locally) maximizes $E[\ln P(Y|h)]$

Many uses:

- Train Bayesian belief networks
- Unsupervised clustering (e.g., k means)
- Hidden Markov Models

RAMAIAH
Institute of Technology

General EM Problem

Define likelihood function $Q(h'|h)$ which calculates $Y = X \cup Z$ using observed X and current parameters h to estimate Z

$$Q(h'|h) \leftarrow E[\ln P(Y|h')|h, X]$$

EM Algorithm:

Estimation (E) step: Calculate $Q(h'|h)$ using the current hypothesis h and the observed data X to estimate the probability distribution over Y.

$$Q(h'|h) \leftarrow E[\ln P(Y|h')|h, X]$$

Maximization (M) step: Replace hypothesis h by the hypothesis h' that maximizes this Q function.

$$h \leftarrow \underset{h'}{\operatorname{argmax}}\, Q(h'|h)$$

Thank you