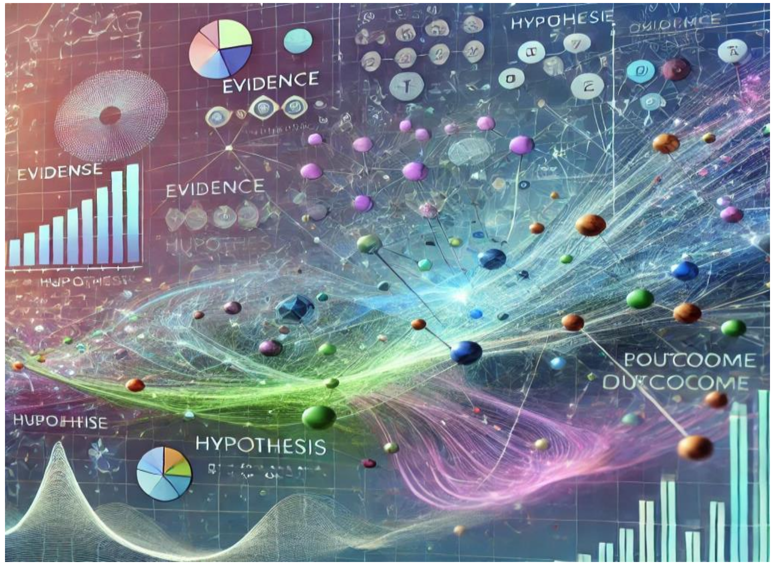
CHAPTER V

Probabilistic Reasoning



Probabilistic reasoning

- **Definition:** Probabilistic reasoning is a method of reasoning and decision-making that deals with **uncertainty** by using **probabilities** to represent and manage uncertain information

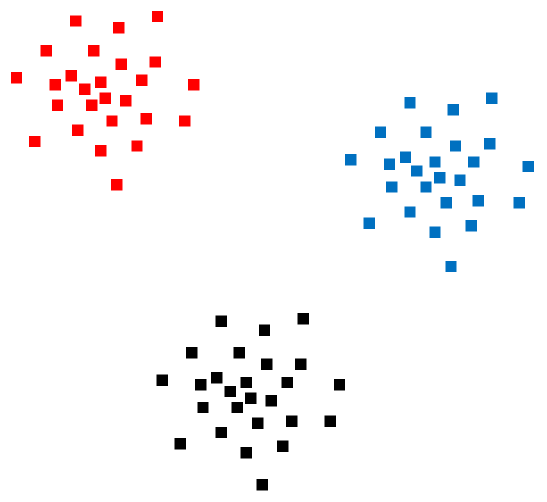
- **Principle:** Probabilistic reasoning **quantifies** uncertainty instead of ignoring it, providing a systematic framework for making predictions or decisions based on **incomplete** or **ambiguous** data

Probabilistic reasoning

- **Comparison to Deterministic reasoning:**
  - **Deterministic Reasoning**: Operates on certain, fixed outcomes.
  - **Probabilistic Reasoning**: Works with uncertainty, offering probabilities of outcomes rather than fixed answers.

Probabilistic reasoning

- **Key techniques:**
  - Naïve Bayes
  - Markov Models
  - Bayesian Networks
  - Monte Carlo Method



Bayes' Method

Definition

- Bayes' method is a probabilistic reasoning technique. It is based on **Bayes' Theorem**, which provides a mathematical framework for updating probabilities as new evidence is observed.

- **Bayes' Theorem** describes the probability of a hypothesis given evidence

- This technique is widely used in fields such as statistics, machine learning, and artificial intelligence to model uncertainty and make decisions based on incomplete or changing information

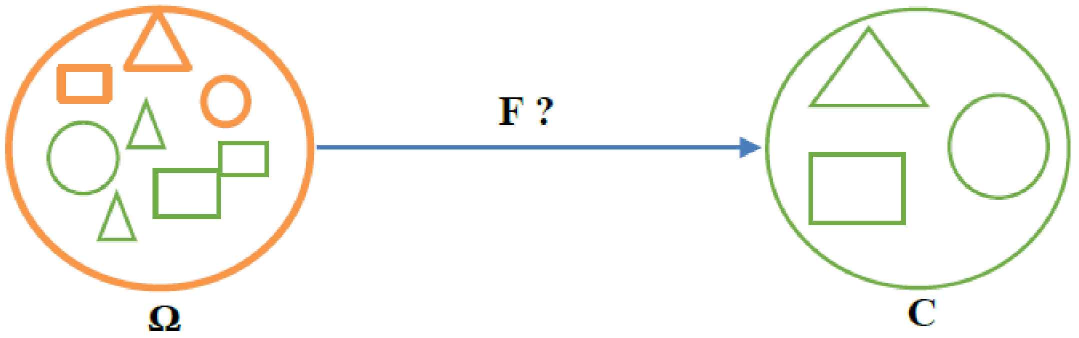
Bayes' Method

Classification problem

- A classification problem can, in some cases, be likened to a diagnosis problem, which involves making a decision based on certain parameters.

- **For example,** in the medical field, making a diagnosis means being able to associate the name of a **disease** with a certain number of **symptoms** presented by **patients**. Three essential elements can be identified in this problem: the patients, the diseases, and the symptoms.

- Population = Patients
- Classes = Diseases
- Features = Symptoms (Descriptions)



F → Associates a disease with a list of symptoms

Bayes’ Method

Classification problem

Formalization

Ω : The population

D : The set of descriptions (Features)

C : The set of classes

X : Ω → **D** is the function that associates a description to all population individuals

Y : Ω → **C** is the function that associates a class to all population individuals

F : D → **C** is the function that associates a class to all decriptions

$$\begin{tikzcd} \Omega \arrow[rr, "X"] \arrow[dr, "Y"'] & & D \arrow[dl, "F"] \\ & C & \end{tikzcd}$$

→ Classifying is related to finding **F?**

Note: Bayes’ method is said to be Naïve sine it assumes that all features are **conditionally independent** given the class label

Bayes' Method

Classification problem

Probabilities

Let's assume that the set **Ω** is probabilized (labeled with Priors) and that the set **D** is discrete

Let's consider P the probability defined on the population Ω, we can define the following probabilities:

- **P(d):** the probability that an element of Ω has d as description.
- **P(k):** the probability that an element of Ω belongs to class k.
- **P(d/k):** the probability that an element of class k has d as description.
- **P(k/d):** the probability that an element with d as description belongs to class k.

Bayes formula:

$$p(k/d) = \frac{p(d/k) \times p(k)}{p(d)}$$

This formula assumes that we can evaluate the probabilities P(d/k), P(k), and P(d).

Bayes' Method

Classification problem

Example: Let **Ω** be the population of a country, and we have a representative sample of the population of this country.

We describe individuals by a logical attribute "**iPhone**" which is '**True**' if the individual owns an iPhone and '**False**' otherwise.

The feature space is therefore **D = {iPhone, No iPhone}.**

We wish to classify individuals into two classes: '**Wealthy**' for individuals with an income above the average, and '**Non-wealthy**' for the others.

We have the following information:

- 40% of the population has an income above the average.
- 80% of wealthy people own an iPhone, while 45% of the remaining population owns an iPhone

Class K	Wealthy	Non-wealthy
P(k)	0.4	0.6
P(iPhone/k)	0.8	0.45
P(No iPhone/k)	0.2	0.55

Bayes' Method

Classification problem

Class K	Wealthy	Non-wealthy
P(k)	0.4	0.6
P(iPhone/k)	0.8	0.45
P(No iPhone/k)	0.2	0.55

- **Choice of the classification function F :**

  - **First rule (Majority Class):**

Assign each description (feature) to the **majority class** (i.e., the class for which **P(k) is maximum**). The function F (called $\mathbf{F_{maj}}$ in this case) will assign the majority class (**Non-wealthy**) with a probability of **0.6** to every individual, regardless of whether they own an iPhone or not

- **Drawback:** The main disadvantage of this rule is that it does not take the description into account at all.

Bayes' Method

Classification problem

Class K	Wealthy	Non-wealthy
P(k)	0.4	0.6
P(iPhone/k)	0.8	0.45
P(No iPhone/k)	0.2	0.55

- ### Choice of the classification function F :

  - **Second Rule (Maximum Likelihood):**

If '**d**' is observed, choose the class for which this observation is the **most likely** (i.e., the class for which **P(d/k) is maximum**). This rule is called the maximum likelihood rule.

The classification function F ($\mathbf{F_{likelihood}}$) will assign the class (**wealthy: 0.8**) to any individual owning an iPhone and the class (**Non-wealthy**) to everyone else.

It is evident that this classification function is more refined than the previous one and corresponds more closely to what we would intuitively expect.

Bayes' Method

Classification problem

- **Choice of the classification function F :**

Class K	Telecom	Doctor	Laborer
P(k)	0.3	0.2	0.5
P(iPhone/k)	1	0.65	0.1
P(No iPhone/k)	0	0.35	0.9

  - **Second Rule (Maximum Likelihood):**

The main drawback of this classification function appears in the following example:

Let's assume three classes (Telecom engineer, Doctor, Laborer) and assume that the probability of a Telecom engineer owning an iPhone is equal to **1**.

The **maximum likelihood** rule will then **assign** the class **'Telecom engineer'** to every individual owning a smartphone, without taking into account the proportions of the different classes within the population.

Bayes' Method

Classification problem

Class K	Wealthy	Non-wealthy
P(k)	0.4	0.6
P(iPhone/k)	0.8	0.45
P(No iPhone/k)	0.2	0.55

- ### Choice of the classification function F:
  - **Third rule (Bayes function) :**

This rule involves assigning to a description '**d**' the class **k** that maximizes the probability **P(k/d)**, using Bayes' formula and noting that P(d) is constant $P(d) = P(d/k) \, . \, P(k) + P(d/k') \, . \, P(k')$

Note: It is therefore sufficient to choose the class k that maximizes the product $[P(d/k) \cdot P(k)]$

- P(iPhone/Wealthy) × P(Wealthy) = 0.8 × 0.4 = <u>0.32</u>
- P(No iPhone/ Wealthy) × P(Wealthy) = 0.2 × 0.4 = 0.08
- P(iPhone/Non-wealthy) × P(Non-wealthy) = 0.45 × 0.6 = 0.27
- P(No iPhone/Non-wealthy) × P(Non-wealthy) = 0.55 × 0.6 = <u>0.33</u>

The function $\mathbf{F_{Bayes}}$ will assign the class 'wealthy' to anyone owning an iPhone and the class 'Non-wealthy' to anyone not owning an iPhone. (in this example, $\mathbf{F_{Bayes}=F_{Likelihood}}$ but this is not always the case.)

Bayes' Method

Classification problem

Exercise: We consider two attributes to determine an individual's nationality. The attribute "**height**" which can take the values "**tall**" or "**short**" and the attribute "**hair color**" can take the values "**brown**" or "**blonde**" The possible nationalities are **French** and **Swedish**.

We assume that the French and Swedish populations are distributed as follows:

	Swedish	French
Short,Brown	10	25
Short,Blonde	20	25
Tall,Brown	30	25
Tall,Blonde	40	25

- In an assembly consisting of 60% Swedish and 40% of French, describe:

a. Majority decision rule**?**

b. Maximum likelihood**?** (P(d/k)↑)

c. Bayes' rule**?** (P(d/k).P(k) ↑)

Bayes' Method

Classification problem

	Swedish	French
Short,Brown	10	25
Short,Blonde	20	25
Tall,Brown	30	25
Tall,Blonde	40	25
P(k)	**60%**	**40%**

A. Majority decision rule:

Each individual, regardless of their height and hair color, is **assigned** to the "**Swedish**" class, which is the majority (**60%** of the population).

Bayes' Method

Classification problem

	Swedish	French
Short,Brown	10	25
Short,Blonde	20	25
Tall,Brown	30	25
Tall,Blonde	40	25
P(k)	**60%**	**40%**

B. Maximum likelihood:

An individual with a description **d(height, color)** is assigned to the nationality for which this description is the most probable, i.e., where **P(d/k)is maximum**. Thus, any individual with:

- (Short, Brown) will be assigned to French,
- (Short, Blonde) will be assigned to French,
- (Tall, Brown) will be assigned to Swedish,
- (Tall, Blonde) will be assigned to Swedish.

Bayes' Method

Classification problem

C. Bayes' rule :

- P(Short,Brown/Swedish) x P(Swedish)=0.10 x 0.6=0.06
- P(Short,Brown /French) x P(French)=0.25 x 0.4=<u>0.10</u>
- P(Short,Blonde/Swedish) x P(Swedish)=0.2 x 0.6=<u>0.12</u>
- P(Short,Blonde/French) x P(French)=0.25 x 0.4=0.1
- P(Tall,Brown/ Swedish) x P(Swedish)=0.3 x 0.6=<u>0.18</u>
- P(Tall,Brown/ French) x P(French)=0.25 x 0.4=0.1
- P(Tall, Blonde/ Swedish) x P(Swedish)=0.40 x 0.6=<u>0.24</u>
- P(Tall, Blonde/ French) x P(French)=0.25 x 0.4=0.1

Thus, any individual with:

- (Short, Brown) will be assigned to French,
- (Short, Blonde) will be assigned to Swedish,
- (Tall, Brown) will be assigned to Swedish,
- (Tall, Blonde) will be assigned to Swedish

	Swedish	French
Short,Brown	10	25
Short,Blonde	20	25
Tall,Brown	30	25
Tall,Blonde	40	25
P(k)	**60%**	**40%**

Origins

- Hidden Markov Models (HMM) were introduced by **Baum** in the **1970s**; this model is inspired by **probabilistic automata**

- A **probabilistic automata** is defined by a structure composed of **states** and **transitions** and by a set of **probability** distributions over the transitions. Each transition is associated with a symbol from a finite **alphabet**. This symbol is generated each time the transition is taken

Definition

- An **HMM** is also defined by a structure composed of **states** and **transitions** and by a set of **probability** distributions over the transitions

- The essential **difference** with probabilistic automata is that the **generation of symbols occurs at the states rather than on the transitions**. Additionally, each **state** is associated not with a single symbol but with a **probability distribution over the symbols** of the alphabet

Applications

HMMs are used in the following fields:

- Speech recognition
- Handwritten text recognition
- DNA sequence recognition
- Information extraction
- POS tagging, etc.

Formalization

An HMM is defined by a quadruplet (S, ∑, T, G)

- H=(S, ∑, T, G)
- S : a set of N states, it contains two particular states : Start et End indicating the beginning and end of a sequence
- ∑ : an Alphabet composed of M symbols.
- T : a matrix that indicates the probabilities of transition between states
  - T = S-{end} x S-{start} → [0,1]
- G : a matrix that indicates the probabilities of emission for states
  - G : S-{start,end} x ∑ → [0,1]

Formalization

- Consider **P(o/s)**, the probability of generating the symbol **o** by the state **s.**
- We do associate to each state **s** :
  - a distribution of transition probabilities :

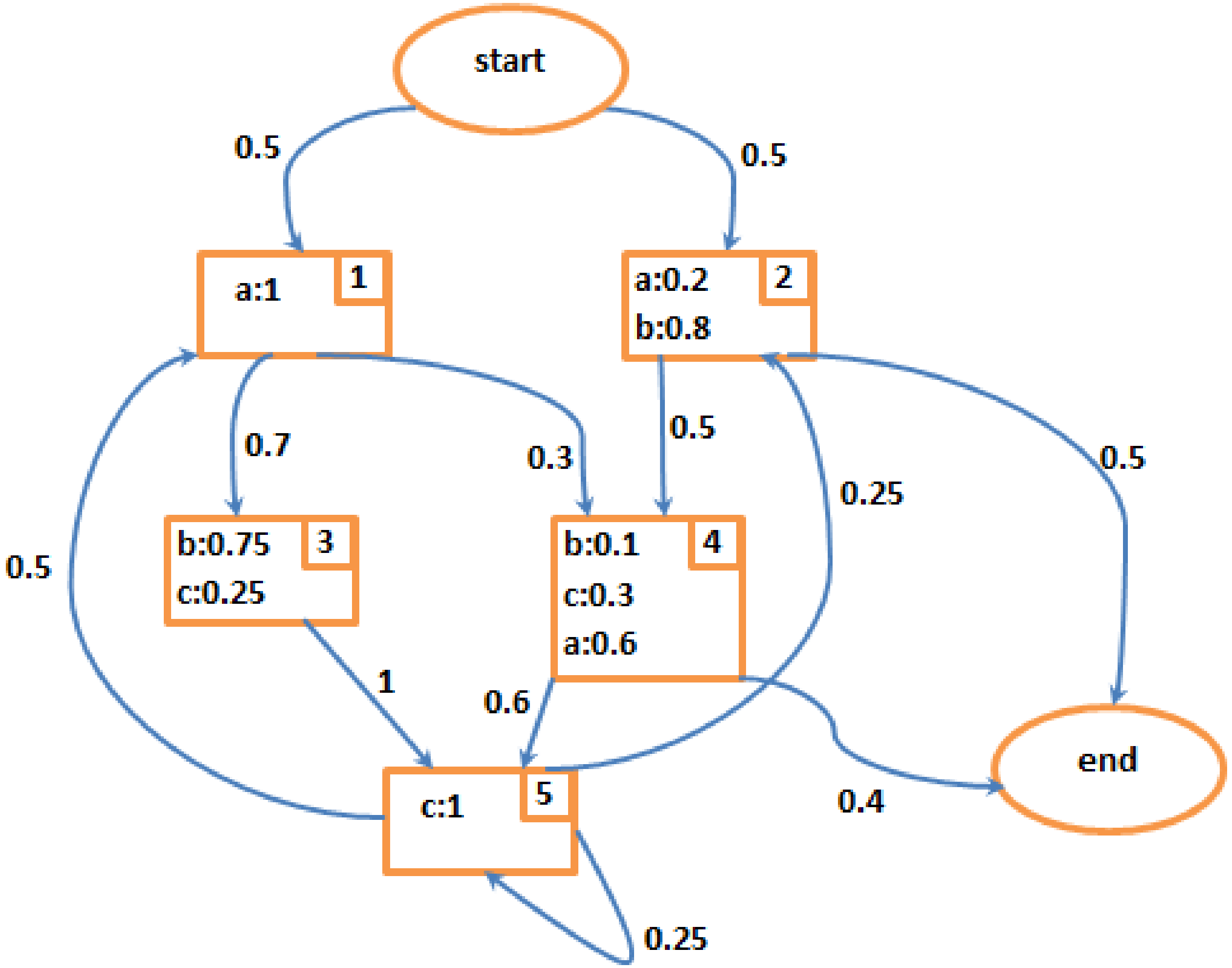
$$\sum_{s' \epsilon S} P(s \rightarrow s') = 1$$

  - a distribution of emission probabilities :

$$\sum_{o' \in \Sigma} P(o' \;/\; s) = 1$$

Example

- The figure shows an example of HMM with 7 states and 11 transitions :



Hidden Markov Models

Example

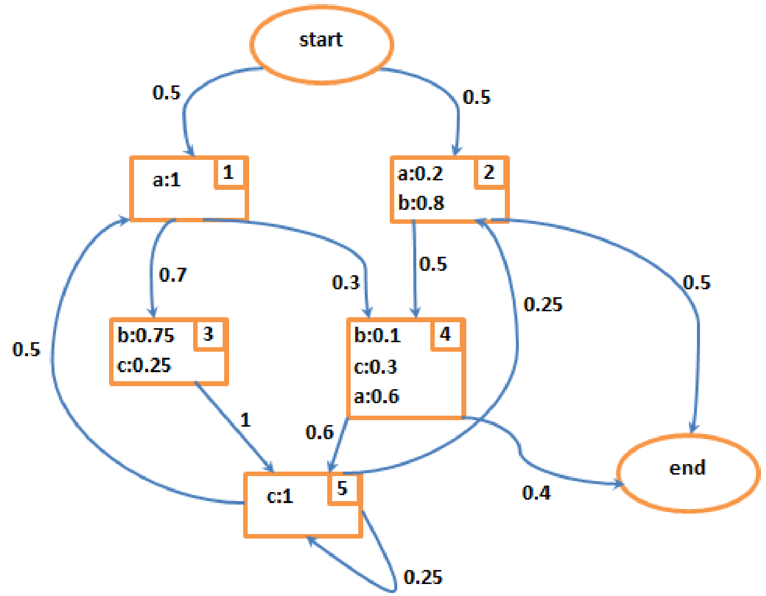
- S={start,1,2,3,4,5,end}
- ∑={a,c,b}

- T : Transition matrix

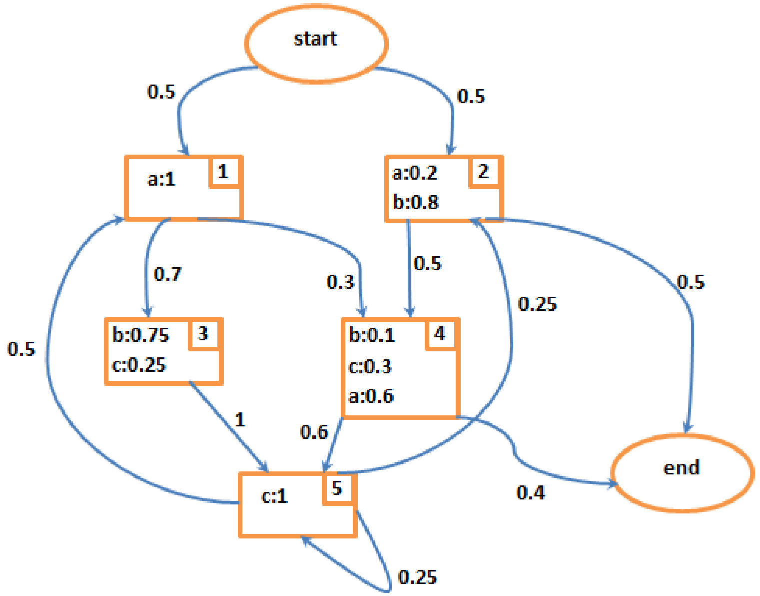
	1	2	3	4	5	end
start	0.5	0.5				
1			0.7	0.3		
2				0.5		0.5
3					1	
4					0.6	0.4
5	0.5	0.25			0.25	

- G : Emission matrix

	a	b	c
1	1		
2	0.2	0.8	
3		0.75	0.25
4	0.6	0.1	0.3
5			1



Example



This HMM allows to generate the following observable sequences:

abca, aacb, ab,…etc.

To these observable sequences correspond the following hidden sequences:

1-3-5-2, 1-4-5-2, 2-4

Each observable sequence could be generated by lot of possible paths.

For example, the sequence **abccb** could be generated by:

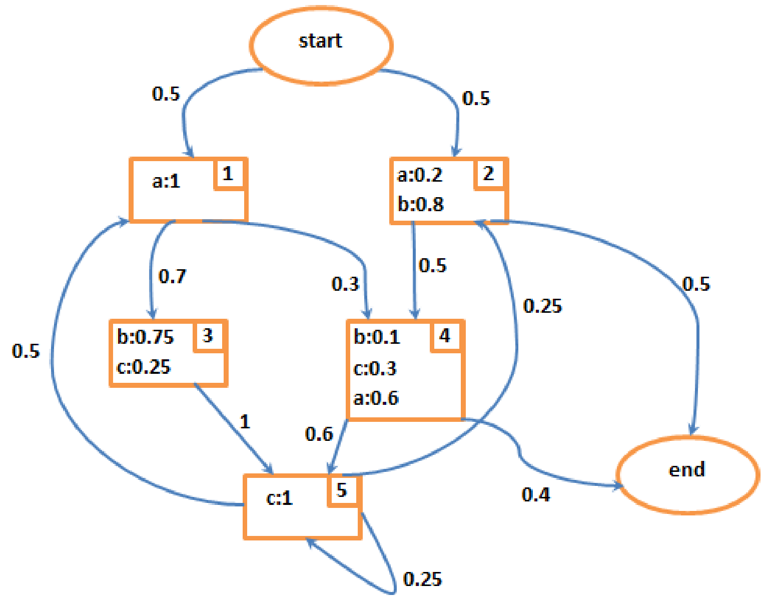
Path 1 : start-1-3-5-5-2-end

Path 2 : start-1-4-5-5-2-end

Path 3 : start-2-4-5-5-2-end

Example

What will be the probability of generating **abccb** by this HMM**?**



Path 1 : start-1-3-5-5-2-end

Path 2 : start-1-4-5-5-2-end

Path 3 : start-2-4-5-5-2-end

P(path 1) = $(0.5 \times 1) \times (0.7 \times 0.75) \times (1 \times 1) \times (0.25 \times 1) \times (0.25 \times 0.8) \times (0.5) = 6.5 \times 10^{-3}$

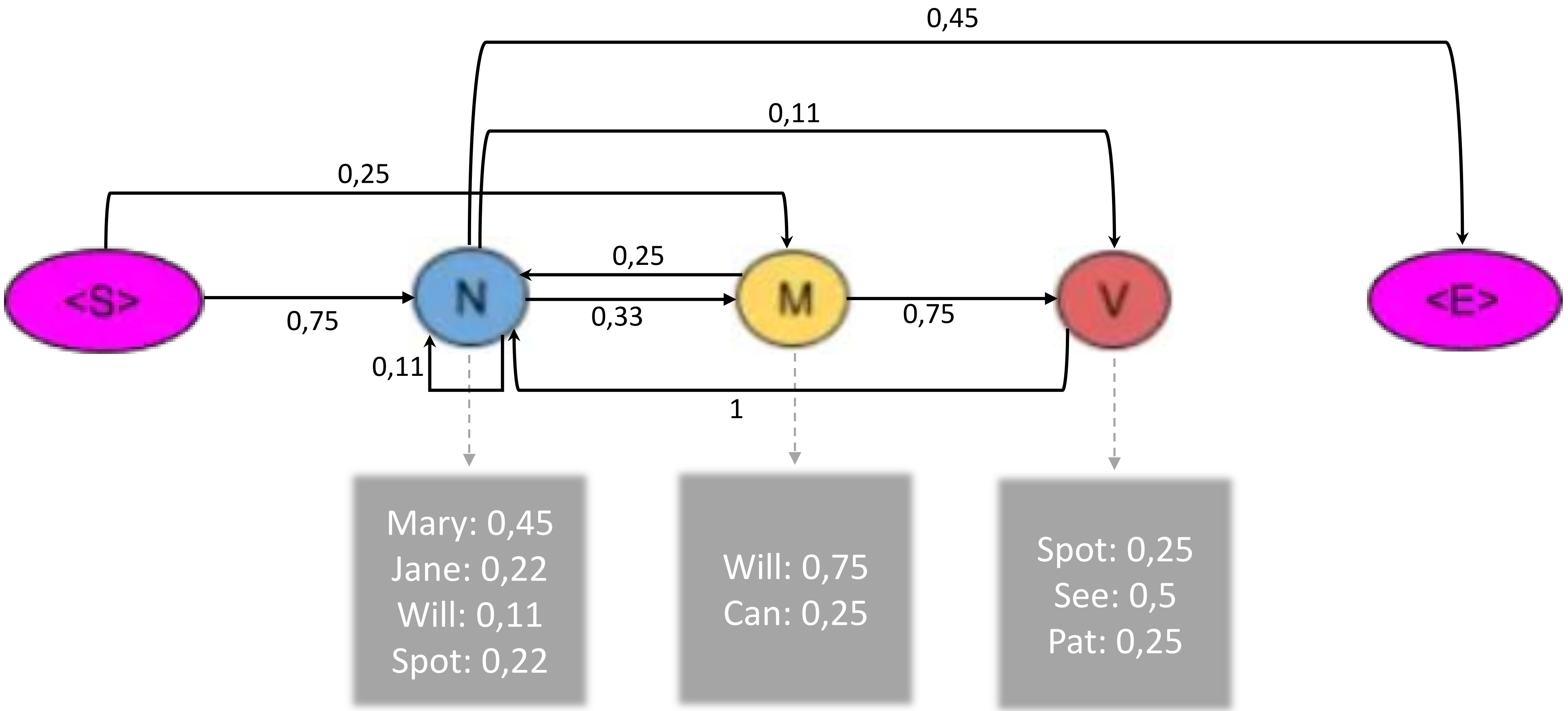
P(path 2) = $(0.5 \times 1) \times (0.3 \times 0.1) \times (0.6 \times 1) \times (0.25 \times 1) \times (0.25 \times 0.8) \times (0.5) = 2.2 \times 10^{-3}$

P(path 3) = $(0.5 \times 0.2) \times (0.5 \times 0.1) \times (0.6 \times 1) \times (0.25 \times 1)\times(0.25 \times 0.8) \times (0.5) = 0.75 \times 10^{-3}$

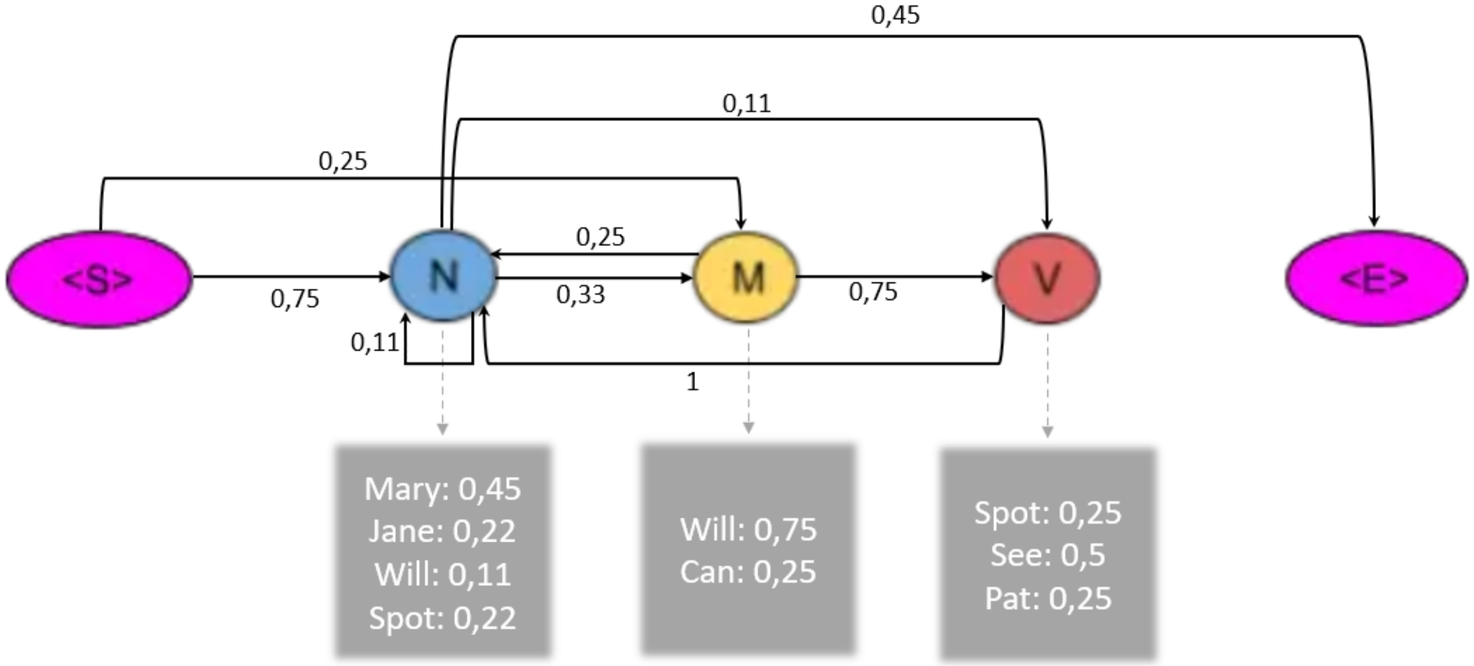
The probability of generating the sequence abccb by this HMM is:

P(abccb) = $(6.5 + 2.2 + 0.75) \times 10^{-3} = 9.45 \times 10^{-3}$

POS with HMM: Example



POS with HMM: Example



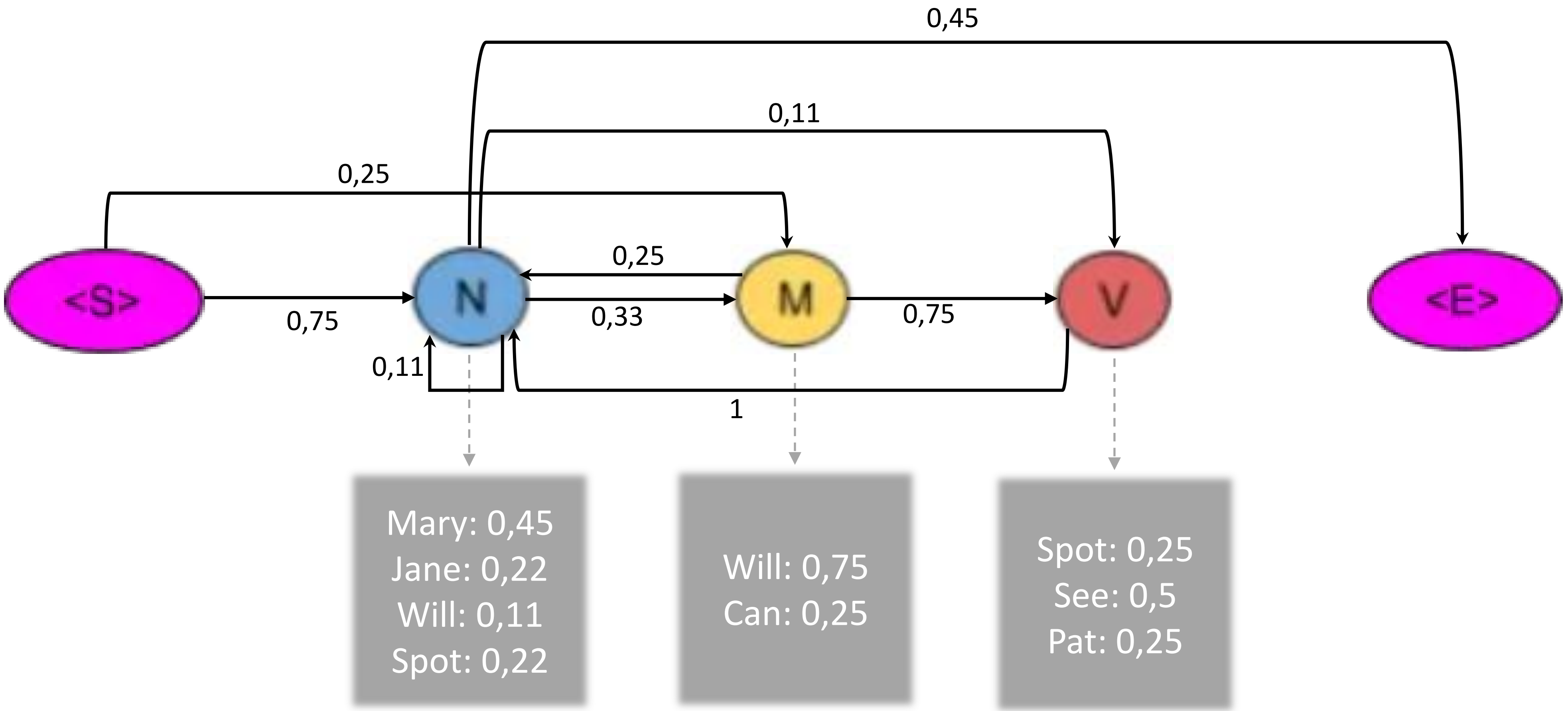
	N	M	V	<E>
<S>	0,75	0,25		
N	0,11	0,33	0,11	0,45
M	0,25		0,75	
V	1			

T: Transition matrix

	Mary	Jane	Will	Spot	Can	See	Pat
N	0,45	0,22	0,11	0,22			
M			0,75		0,25		
V				0,25		0,5	0,25

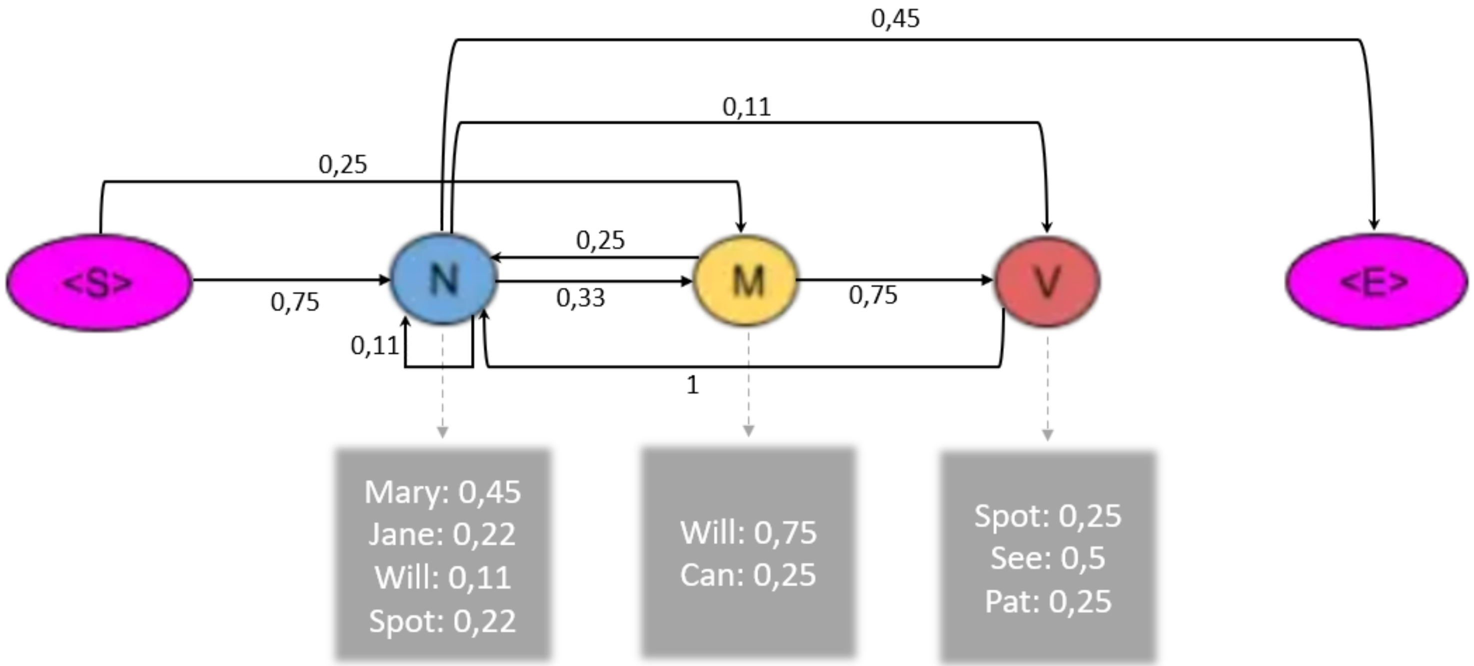
G: Emission matrix

POS with HMM: Example



POS of « Will can spot Mary » ?

POS of « Will can spot Mary » **?**



Path 1 = <S>→N→M→N→N→<E>
P(Path 1) = (0,75x0,11) x (0,33x0,25) x (0,25x0,22) x (0,11x0,45) x (0,45) = 0,0000083385

Path 2 = <S>→N→M→V→N→<E>
P(Path 2) = (0,75x0,11) x (0,33x0,25) x (0,75x0,25) x (1x0,45) x (0,45) = **0,00025842**

The probability of the second sequence is much higher

POS Tags : {Will : N, can : M, spot : V, Mary : N}

HMM challenges

Let's consider H an HMM and a given sequence of symbols $O=O_1O_2\ldots O_t$

- What is the probability of generating O with H**?**

Solution: Forward-backward algorithm

- What is the sequence of states $S=S_1S_2\ldots S_t$ in H that has the maximum probability of generating O**?**

Solution: Viterbi algorithm

- How to adjust the parameters of H (transition and emission probabilities) to best represent the sequences being processed**?**

Solution: Baum-Welch algorithm