HMM Part-of-Speech Tagging

Khoa Huu Tran

University of Information Technology - VNUHCM

Ngày 25 tháng 12 năm 2022

Table of Contents

1 INTRODUCTION

2 Markov chain

3 Hidden Markov Model

Introduction

Classic sequence labeling algorithm Hidden Markov Model (HMM).

Sequence labeler: model whose job is to assign a label to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels of the same length.

Introduction

Classic sequence labeling algorithm Hidden Markov Model (HMM).

Sequence labeler: model whose job is to assign a label to each unit in a sequence, thus mapping a sequence of observations to a sequence of labels of the same length.

HMM is a classic model that introduces many of the key concepts of sequence modeling.

HMM is a probabilistic sequence model.

Table of Contents

1 INTRODUCTION

2 Markov chain

3 Hidden Markov Model

Markov Chains

The HMM is based on augmenting the Markov chain.
A Markov chain is a model that tells us something about the probabilities of sequences of random variables, *states*.

Markov Chains

The HMM is based on augmenting the Markov chain.
A Markov chain is a model that tells us something about the probabilities of sequences of random variables, *states*.

Very strong assumption: We only need the current state to predict the future in the sequence.

Markov Chains

The HMM is based on augmenting the Markov chain.
A Markov chain is a model that tells us something about the probabilities of sequences of random variables, *states*.

Very strong assumption: We only need the current state to predict the future in the sequence.

It's as if to predict tomorrow's weather you could examine today's weather but you weren't allowed to look at yesterday's weather.

Markov chain

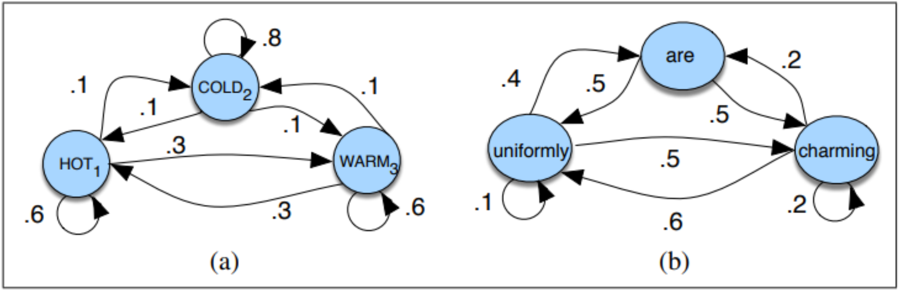


Figure 8.8 A Markov chain for weather (a) and one for words (b), showing states and transitions. A start distribution π is required; setting $\pi = [0.1, 0.7, 0.2]$ for (a) would mean a probability 0.7 of starting in state 2 (cold), probability 0.1 of starting in state 1 (hot), etc.

Markov assumption

More formally, consider a sequence of state variables $q_1, q_2, ..., q_i$.
A Markov model embodies the Markov assumption on the probabilities of this sequence:

$$P(q_i = a | q_1 ... q_{i-1}) = P(q_i = a | q_{i-1}) \tag{8.3}$$

Markov assumption

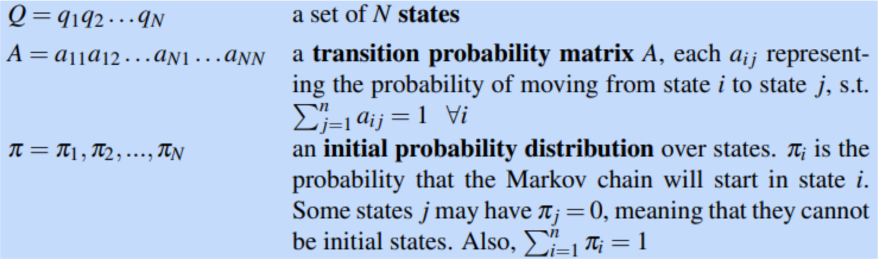
More formally, consider a sequence of state variables $q_1, q_2, ..., q_i$.
A Markov model embodies the Markov assumption on the probabilities of this sequence:

$$P(q_i = a|q_1...q_{i-1}) = P(q_i = a|q_{i-1}) \tag{8.3}$$

=> Markov chain represents a bigram language model.

Markov assumption

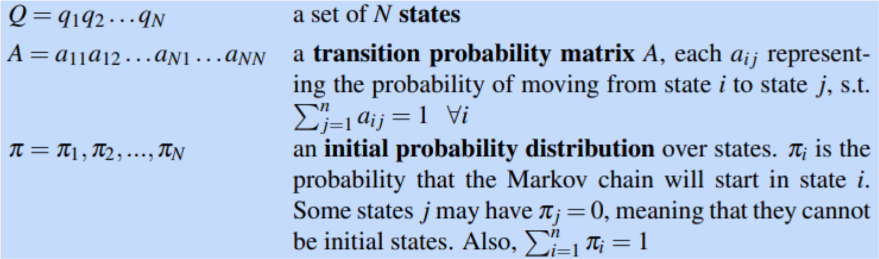
Formally, a Markov chain is specified by the following components:



$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, ..., \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Markov assumption

Formally, a Markov chain is specified by the following components:



$Q = q_1 q_2 \ldots q_N$ a set of N **states**

$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$ a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$

$\pi = \pi_1, \pi_2, \ldots, \pi_N$ an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Ex : Use the sample probabilities in Fig 8.8a (with $\pi = [0.1, 0.7, 0.2]$) to compute the probability of each of the following sequences:

(8.4) hot hot hot hot
(8.5) cold hot cold hot

Markov assumption

$$\pi = [0.1, 0.7, 0.2]$$

(8.4) hot hot hot hot
P(HHHH) $=$

(8.5) cold hot cold hot
P(CHCH) $=$

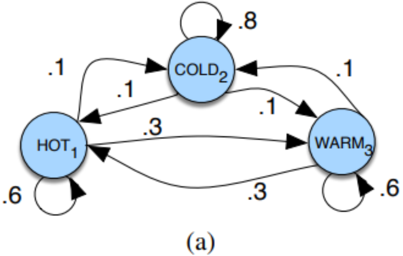


Fig.8.8a

Markov assumption

$$\pi = [0.1, 0.7, 0.2]$$

(8.4) hot hot hot hot
P(HHHH) = P(H) * P(H|H) * P(H|H) * P(H|H)
= 0.1 * 0.6 * 0.6 * 0.6 = 0.0216

(8.5) cold hot cold hot
P(CHCH) = P(C) * P(H|C) * P(C|H) * P(H|C)
= 0.7 * 0.1 * 0.1 * 0.1 = 0.0007

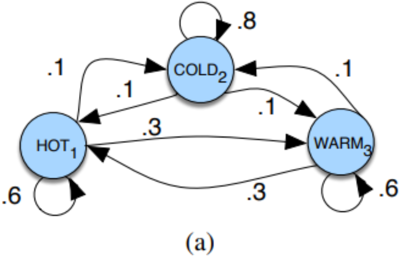


Fig.8.8a

Table of Contents

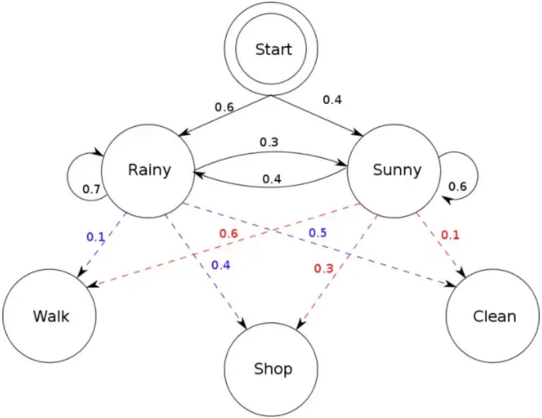
1 INTRODUCTION

2 Markov chain

3 Hidden Markov Model

Hidden Markov Model

Sometimes, what we want to predict is a sequence of states that aren't directly observable in the environment.

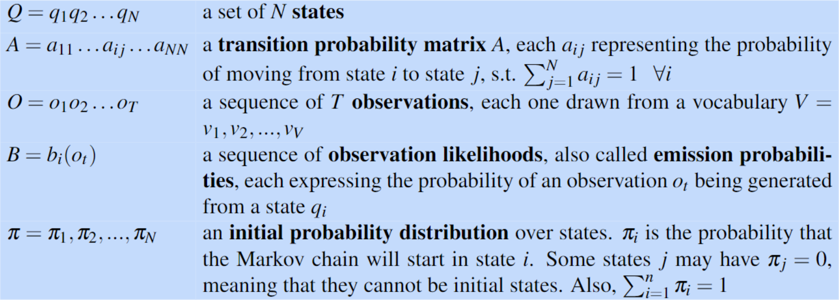


Hidden Markov Model

Why do we need HMM for POS Tagger?

If you notice closely, we can have the words in a sentence as **Observable States (given to us in the data) but their POS Tags as Hidden states and hence we use HMM for estimating POS tags**.

It must be noted that we call **Observable states 'Observation' Hidden states 'States'**.



$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state q_i
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

Hidden Markov Model

A first-order hidden Markov model instantiates two simplifying assumptions. First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:

Markov Assumption:

$$P(q_i|q_1,...,q_{i-1}) = P(q_i|q_{i-1}) \tag{8.6}$$

Second, the probability of an output observation o_i depends only on the state that produced the observation q_i and not on any other states or any other observations:

Output Independence:

$$P(o_i|q_1,...q_i,...,q_T,o_1,...,o_i,...,o_T) = P(o_i|q_i) \tag{8.7}$$