2021-22

MATH F424

Stochastic Processes and their Applications

Assignment-1

Text Generation through Markov Chains

Suchismita Tripathy, 2019A7PS0554P
Yash Bansal, 2019A7PS0484P

29 April, 2022

Contents

1 Introduction

Markov chains have a number of uses in a wide range of fields. Pertaining to text and language itself, they are used for speech recognition, information retrieval, handwriting recognition, NLG etc. One particular such use is its application to the field of information theory, revolutionised by Claude Shannon, in his paper A Mathematical Theory of Communication. Apart from dealing with the mechanisms of encoding information using bit strings, and the capacity limits for the same in the face of noise and attenuation, Shannon realised that the process of information transfer and communication can be eased by probabilistically modelling the information source and noise source. Obtaining a complete probabilistic characterisation of the information source can help decide how to encode the strings transmitted and the minimum average length that can be achieved for the binary encodings of each string. Claude Shannon thus proposed that a piece of text can be statistically modeled using Markov Chains. With this as a basis, Markov chains can be used for letter prediction, text generation and even spell-check. In this study, the process of text generation using a first order Markov chain model is explored.

2 Basic Theory

A Markov process is of course characterised by the fact that the next state depends only on the current state i.e.

$$P(X_n = i | X_{n-1} = j, ...X_1 = k) = P(X_n = i | X_{n-1} = j)$$

where
X_n : nth state
i, j, k : Valid states part of the state set of the process

A Markov process is defined by its transition probability matrix or the 1 step transition matrix, which records the probabilities of going from one state to another for all valid states of the process i.e.

$$T[i, j] = P(X_{n+1} = j | X_n = i)$$

where
T represents the transition probability matrix
This matrix can be multiplied by itself n times to give the n-step transition matrix i.e.

$$T^{(k)}[i, j] = P(X_{n+k} = j | X_n = i)$$

where
$T^(k)$ represents the k-step transition matrix and

$$T^{(k)} = T^k$$

3 Modelling Text as a Markov process

3.1 Using words as the basic unit

Since a piece of text is just a string of words, the probability of word 1 appearing after word 2 can be calculated by observing the number of times word 1 is present after word 2 in the text. The probability will thus be :

$$P(\text{word 1 appears after word 2}) = \frac{\text{Number of times word 1 appears after word 2}}{\text{Number of times word 2 appears}}$$

For our research we use the Python library Markovify generate novel sentences using Markov Chains on a corpus of text, using words as the base unit. The fundamental notion behind Markov Chains and text is that you obtain a corpus of text and build a parser that will scan through every set of k+1 words in the text, where k is the number of words the probability of a next word depends on. More accurately we want to build a mapping of key-word pairs where key is the first k words in the set mapped to the $k + 1^{th}$ word. All possible words for a key and repetitions of words for a key are noted to give the following probability for a word to be selected:

$$P(\text{word}) = \frac{\text{Number of times word appears after key set}}{\text{Number of times key set appears}}$$

For text corpora, Markovify begins the process of building its Markov models by splitting each corpus into a series of sentences, and each sentence into a series of ”tokens” (i.e., words, plus placeholder tokens for the beginning and end of a sentence). Then, walking sentence-by-sentence, it identifies every sequence of k (size of key) tokens, and calculates how many times any other token comes immediately afterward. This dictionary of corpus[(token_a, token_b, ...)][next_token]: count frequencies comprise the ”transition matrix” of the resulting Markov model.

Thus, ”Janice walked to the park. After that, Janice walked to the zoo.” for a key size of k=2 becomes:

```
{
 ('___BEGIN__', '___BEGIN__'): {'After': 1, 'Janice': 1},
 ('___BEGIN__', 'After'): {'that,': 1},
 ('___BEGIN__', 'Janice'): {'walked': 1},
 ('After', 'that,'): {'Janice': 1},
 ('Janice', 'walked'): {'to': 2},
 ('that,', 'Janice'): {'walked': 1},
     ('the', 'park.'): {'___END__': 1},
 ('the', 'zoo.'): {'___END__': 1},
 ('to', 'the'): {'park.': 1, 'zoo.': 1},
 ('walked', 'to'): {'the': 2}}
}
```

Numbering these keys from 1-10 we get the following transition probability matrix and creating a common state 11 for keys containing second token as '___END__'

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	0	0.5	0.5	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0	0
4	0	0	0	0	0	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0.5	0.5	0	0	0
10	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	0	0	0	0	0	1

Table 1: Transition Matrix for "Janice walked to the park. After that, Janice walked to the zoo."

When Markovify attempts to generate a new sentence, it randomly chooses each successive token using these frequency-based probabilities. This allows Markovify to potentially generate a different sentence even if given the same starting word (given that there exists more than one choice of words for any of the keys occurring in the sentence generation).

Hence, given the starting word Janice the following sentences can be generated:

1. Janice walked to the park.
2. Janice walked to the zoo.

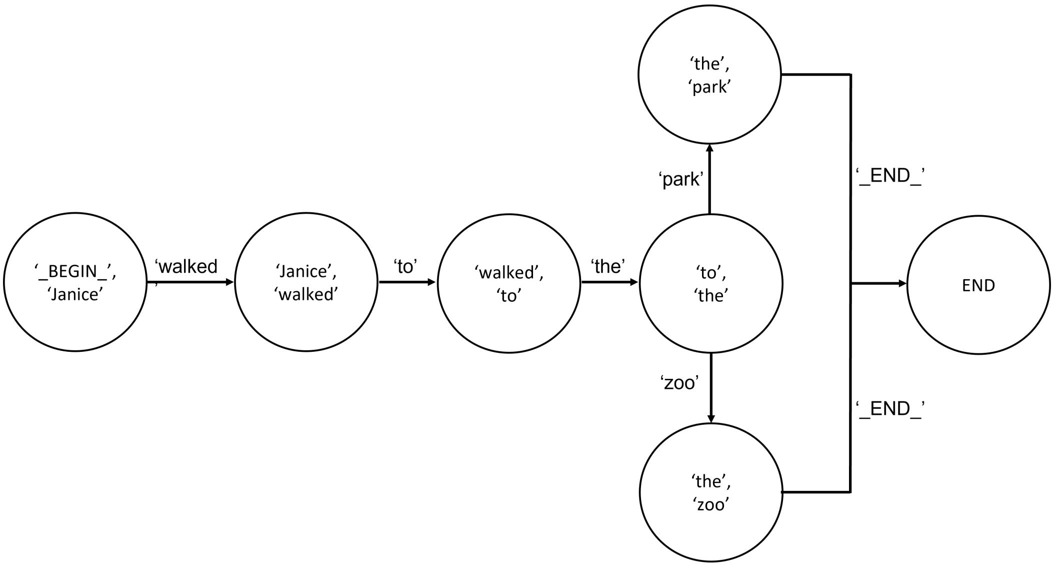

Figure 1: Transition diagram for above sentences

At the state ('to', 'the'), we can get the word 'park' or 'zoo' with equal probability and hence the above sentences can be generated with equal probability. We can also generate sentences without a start word and we are then by default in state 1 : ('___BEGIN__', '___BEGIN__') from which 'After' or 'Janice' can be used as the first word of the sentence for generation.

As an added bonus, Markovify does multiple runs for the initial keys to try and generate sentences with minimal overlap to our original corpus and so the the sentence ”Janice walked to the park” would be the final output of our library function with start word ”Janice”.

3.2 Using letters as the basic unit

Increasing the granularity, this can be extended to using characters instead of words as the base unit.

k-grams of words can also be used instead of letters. This involves using the current k-gram to predict the next k-gram. Since the next k-gram will share k - 1 letters in common with the current k-gram (else the order of the Markov chain will be more than 1), this process is essentially the same as letter prediction. For example, let k = 3, i.e. every 3 letters forms a state. If the word "weather" is predicted starting with the state "wea", the following shows the states reached and the letter predictions along the arrows:

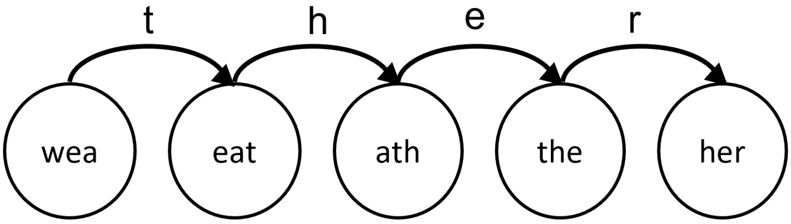

Figure 2: States reached as the word "weather" is predicted

Thus, to model a piece of text with Markov chains, all unique k-grams of the text are recorded and the probabilities of going from one k-gram to the next are calculated as part of the 1-step transition matrix. Then, starting with an initial state i.e. one k-gram, the rest of the text can be generated. For this, a probability value between 0 and 1 can be generated from a random distribution, and the probabilities out of the current state can be compared with this to predict the next k-gram i.e. the next letter. This way, n-step transition matrices are not required and the same 1-step probability matrix can be used at each state. The same was described with words instead of k-grams in section 3.1, where the entire vocabulary is first recorded and then the transition probability matrix is built and used for word prediction.

Letter prediction also has uses in gene/protein prediction, where given an initial gene/protein, and a Markov model based on related sequences, the following characters in the sequence can be predicted. The Markov chain thus models not only the kind of sequences that are characteristic of the input group of sequences but also the different mutations that are common for the group. The comparison of 2 gene/protein query sequences to assess their similarity and possibly predict common ancestry also depends on modelling text as a Markov process. Here, sequences that are related to each other by a given factor (a family) are analysed, and the probabilities of going from one nucleotide to the next or one protein to the next are calculated by studying all the sequences given and a transition probability matrix is built. It is assumed that these probabilities are independent of the position of the character in the sequence (hence making it markovian). These probabilities are then used to score different alignments of the 2 query sequences, to see if the probability of getting a particular alignment through mutations and other natural changes characteristic of the family is more than the probability of getting the alignment by chance. This helps classify the unknown query sequence and build an evolutionary tree to trace the sequence's ancestry.

4 Examples

The following are some sentences generated from the plot synopsis of Titanic the movie given on Titanic's Wikipedia page:

When Jack and Rose return to the stern rail.
Rose is saved by a returning lifeboat, keeping her promise, and Jack urge Rose to board a lifeboat.
Alone on the ship.
As passengers fall to their deaths, Jack and Rose return to the boat deck.
The upended ship breaks in half and the ship's stern is rising as the flooded bow sinks.
In 1912 Southampton, 17-year-old Rose DeWitt Bukater, her wealthy fiancé Caledon Cal Hockley, and Rose's insulting note left inside his safe, along with the necklace.
On the forward deck, they witness the ship's stern is rising as the Heart of the Ocean necklace.
Cal then puts the necklace tucked inside the cargo hold.
After setting sail, Rose, distraught over her loveless engagement, climbs over the stern of Keldysh, Rose takes out the Heart of the Ocean necklace.
Cal discovers Jack's sketch and Rose's widowed mother, Ruth, board the Titanic.
Cal grabs Lovejoy's pistol and chases Rose and Jack urge Rose to board a lifeboat.
The lifeboats have departed and the bow section dives downward.
On the forward deck, they witness the ship's stern is rising as the flooded bow sinks.
In 1912 Southampton, 17-year-old Rose DeWitt Bukater, her wealthy fiancé Caledon Cal
Hockley, and Rose's insulting note left inside his safe, along with the necklace.
After setting sail, Rose, distraught over her loveless engagement, climbs over the stern of Keldysh, Rose takes out the Heart of the Ocean.
Cal grabs Lovejoy's pistol and chases Rose and Jack urge Rose to board a lifeboat.

The following are some sentences generated from The Critique of Pure Reason by Immanuel Kant:

The physico-theologians have therefore no objective proof, and although we admit it, we are driven, in our reason, subjectively considered as occupying all time and space, infinite.
For, let us take the opposite be itself false; and, consequently, that the predicate of another thing in this case, I find that the practical interest arising from it, and consequently a formula like the present, we may content ourselves with the synthesis of the world, from the constitution of our personality through all the rights and claims respect.
The very essence of the practical reference is either empirical or pure intuitions.
Both are transcendental, not merely from everything empirical, but also a given conditioned, were perfectly homogeneous.
For, when I have got so far above us, that is, of free action.
That it may appear, lies open to it—the path of experience sufficiently ample for our conception of the external boundaries and the incomplete exposition must precede the act.
But this extension of the objects of our volition.
Thus God and a future life, unless—since it could not entirely put a stop to the empirical conjunction of a thing in itself.
For as existence does not represent the conditions of sensuous intuition—and, for this reason all the limitations of it; and the advantage of novelty, against as illusory grounds of proof.
And just in this system, because they relate to a something of which reason continually strives.
But although pure speculative reason, is bound to discover the clue to the warnings of philosophy, and can be given along with the pure conception of such a system of astronomy, such as lie without the aid of experience—which presents to it absolute and permanent future for Project Gutenberg-tm electronic works.
But, let me form any judgement respecting them; and therefore in vain; as, indeed, we have asserted; and we consequently do not really deduce anything from experience towards the attainment of which we perceive the moon and then the moon; and for that need of reason.
It has hitherto remained in so far as the existence of a science, for which the subject a predicate to thought, we cannot think away those through which everything else it may most fully establish its claims and the same time, in-

dicate à priori propositions, they may try to discover truth by means of an opponent as proofs of the soul.

Gaining, as they would still be considered as mere forms of the philosopher for maintaining that they fall when their supports are taken away, must have the power, by calling up the use of the categories suggests considerations of interest, and thus their testimony is invalid.

The conditions of phenomena we penetrate into the idea of metaphysics, we derive a surprising result, and one which, to given conceptions of the functions of reason.

The specious error which leads to groundless assertions, against which others equally specious can always connect the phenomena of the world as per se determined in regard to the pure conceptions of the categories; we cannot and do not exclude identity of the ideality of space and time, but even the smallest empirical element of thought, or of the faculty of sense is commonly held in especial respect.

These, in so far as it forms merely the duty of the categories in this place.

The former appertain absolutely and for this reason, the second its determination, both in kind and in the unity of phenomena it does not exist.

The Antinomy of Pure Reason By the term exposition—a more modest expression, which the sophist devises for the false, and it is, with the laws of the hope of ever reaching a state of rest after motion, but we cannot know any determinate object.

The second takes no account of our understandings, that it is undoubtedly of great importance, we consider merely the empirical character; and, when we regard it as valid in the aid of rational psychology.

5 References

1. `https://www.quantamagazine.org/how-claude-shann` ons-information-theory-invented-the-future-20201222/

2. `https://pypi.org/project/markovify/`

3. `https://en.wikipedia.org/wiki/Titanic_(1997_film)`

4. `https://www.gutenberg.org/`