Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Matematica

PROCESSI MARKOVIANI NASCOSTI

Relatore:
Chiar.mo Prof.
MIRKO DEGLI ESPOSTI
Correlatore:
Dott.
GIAMPAOLO CRISTADORO

Presentata da:
ALEXANDRA ZELENANSKA

Sessione III
Anno Accademico 2013/2014

Index

Introduzione

Nella tesi vengono trattati alcuni algoritmi utilizzati nelle applicazioni di modelli markoviani nascosti. In particolare vengono esaminate le loro applicazioni nel campo della biologia.

All' inizio vengono definite le catene di Markov, ossia meccanismi probabilistici con determinate proprietà. Le catene markoviane sono le sequenze di variabili aleatorie $\{X_n\}_{n=0}^{\infty}$ caratterizzate dal fatto che lo stato futuro della catena, X_{n+1}, dipende solo da quello attuale, X_n, e non da tutti gli stati precedenti, $X_0, ..., X_n$. Gli elementi di base di una catena di Markov sono le probabilità di transizione degli stati, che vengono esposti nella matrice di transizione e la sua distribuzione iniziale di probabilità.

Questi processi, detti anche processi di Markov osservabili, sono difficilmente utilizzabili nelle applicazioni. Viene quindi esposto l'argomento delle catene di Markov nascoste. Queste sono dei processi probabilistici più complicati che includono due meccanismi: una catena di Markov che è *nascosta* e un altro processo aleatorio i cui risultati sono osservabili. È definita la matrice di transizione degli stati nascosti e le probabilità di emissione di un' osservazione che dipende dallo stato particolare della catena nascosta. Però gli stati stessi della catena di Markov non sono osservabili.

Il modello in generale è una descrizione semplificata di un fenomeno di cui riusciamo a studiare meglio le caratteristiche determinandone le proprietà, le pos-

sibili incertezze e i parametri. Di conseguenza derivano tre domande principali riguardanti i modelli nascosti, che nascono dalla necessità di poter essere utilizzati nelle applicazioni.

Tali domande sono:
La prima, dato un modello ed una sequenza di osservazioni o prodotti da un processo random, qual è la probabilità che il modello abbia prodotto la sequenza delle osservazioni?
La seconda, dato un modello ed una sequenza di osservazioni o, qual è la più probabile sequenza degli stati nascosti s?
Il terzo problema è quello di migliorare le proprietà ed i parametri del modello, affinché questo riesca a descrivere meglio le proprietà del fenomeno che stiamo esaminando.

I modelli markoviani nascosti hanno un ampio utilizzo in diversi campi, uno dei primi è stato quello di *speech recognition*.

Per quanto riguarda le applicazioni nel campo della biologia, i modelli markoviani nascosti vengono usati maggiormente nella modellizzazione della sintesi delle proteine dal DNA e nel determinare certe proprietà delle sequenze geniche. Le proteine sono le macromolecole essenziali per il funzionamento della cellula e sono composte da sequenze di amminoacidi definite dal DNA. La sequenza degli amminoacidi dipende dalla successione delle quattro basi azotate nella parte del gene. Conoscendo questa, riusciamo a determinare la catena degli amminoacidi con le sue proprietà e la sua conformazione spaziale. Tuttavia, conoscendo le proprietà delle proteine, non è chiaro da quale catena di basi azotate esse provengano. Quindi può essere conveniente usare gli algoritmi basati sui modelli markoviani per determinarle. Questi modelli risultano spesso essere più efficaci rispetto alla ricerca sperimentale. Gli argomenti di biologia vengono esaminati nel Capitolo 4.

Le sequenze geniche del DNA sono composte dalle parti che codificano l'informazione

per la sintesi proteica e quelle che non codificano informazioni particolari, cioè le parti *coding* e *non coding*. Le parti non codificanti vengono escluse durante il processo. Siccome le parti *coding* e *non coding* hanno alcune proprietà diverse, anche qui vengono usati i modelli markoviani nascosti per determinare l'informazione non osservabile, cioè determinare quali sono gli stati che hanno la più alta probabilità di essere *coding* o *non coding*.

Nel determinare la soluzione di questi problemi si utilizza l'algoritmo di Viterbi. Questo è stato introdotto da Andrew Viterbi nel 1967. L'algoritmo cerca di determinare la sequenza più probabile di stati nascosti, data una sequenza di osservazioni ed un modello. L'idea è quella di considerare tutte le possibili sequenze di stati nascosti e scegliere quella con la probabilità più alta. Siccome questo non è conveniente computazionalmente, Viterbi propose di determinare le probabilità di alcuni tipi particolari di sottosequenze per determinare quella più verosimile mediante l'algoritmo ricorsivo. L'algoritmo viene descritto più in dettaglio nel Capitolo 5.

Si conclude con la descrizione di un algoritmo simile a quello di Viterbi che a differenza di quest' ultimo, dato un modello, viene utilizzato per determinare la probabilità di ottenere una sequenza di osservazioni o. Questo viene chiamato algoritmo *forward*. L'algoritmo mediante la ricorsione e le sequenze parziali calcola la probabilità che il modello produca la sequenza osservata.

Chapter 1

Markov chains

1.1 Definitions

We will first introduce a few general notations that characterise probabilistic experiments and their results.

The **stochastic process** X is a family $\{X_a : a \in A\}$ of random variables indexed in the set A. Stochastic processes can be divided into a few categories. One possible classification can be done according to the cardinality of the set A. One can refer to A as either a discrete-time or a continuous-time process depending on whether A is discrete or continuous, respectively.

A family of random variables might be seen as a probabilistic experiment, in which the values of the random variables are the outcomes of the experiment. Let S be the set of the values of the random variables X_a. We will refer to S as the **state space**.

Now we will assign a restrictive condition to the general definition of a stochastic process. We will obtain a new family of stochastic processes, Markov chains.

Let us consider a sequence of random variables $\{X_n\}_{n=0}^{\infty}$ which takes values in

the countable set S. The sequence $\{X_n\}_{n=0}^{\infty}$ is called a **Markov chain** if

$$P(X_{n+1} = j_{n+1}|X_0 = j_0, X_1 = j_1, ..., X_n = j_n) \\ = P(X_{n+1} = j_{n+1}|X_n = j_n)$$

for all $n \geq 0$ and $j_0, ..., j_{n+1}$ in S.

This condition is called the Markov property.

This means that a sequence of random variables satisfying this property has the probability of the future event $X_{n+1} = j_{n+1}$ conditioned only by its present state $X_n = j_n$ and not by all its past (and present) states $X_0 = j_0, ..., X_n = j_n$.

Homogeneous chains are a particular type of Markov chains. A chain is said to be homogeneous if its transition probabilities are stationary, which means that they do not depend on time. The probabilities only depend on j and i and not on n,

$$P(X_{n+1} = j|X_n = i) = P(X_1 = j|X_0 = i).$$

This means that the probability distribution does not change over time.

For example, flipping a coin does not change the coin and the probabilities of obtaining heads or tails always remain the same.

Hereafter we will only consider homogeneous Markov chains, unless specified otherwise.

1.2 Transition probabilities

When describing the behaviour of a Markov chain, we must define the probability of its state at the beginning, the initial distribution, and the probabilities of the

state i transiting to the next state j. If we know these things, the chain is characterized completely.

The probabilities of transiting from the state i to j are called the **transition probabilities**:

$$p_{ij} = P(X_{n+1} = j | X_n = i).$$

The transition probabilities can be displayed in a **transition matrix** $P = (p_{ij})$. This is an $|S| \times |S|$ matrix of the conditional probabilities. It is a stochastic matrix, which means that the entries of the matrix are greater or equal than 0 and the sum of the rows of the matrix is always equal to 1, e.g., $p_{ij} \geq 0$ and $\Sigma_j p_{ij} = 1$.

1.3 Examples

An example of a Markov chain is a **simple random walk**.

The state space is defined to be $S = \{0, \pm 1, \pm 2, ...\}$ and the transition probabilities are given by $P(X_{n+1} = j+1 | X_n = j) = p$, $P(X_{n+1} = j-1 | X_n = j) = 1-p$ and 0 otherwise.

This means that the probability of moving forward is p and the probability of moving one step back is $1 - p$. No other transition probabilities are defined, so we can assume these to be zero, in particular the probability of remaining at the same position is zero as well.

The transition matrix is infinite and has zeros on the diagonal, p above the diagonal (the elements p_{01}, p_{12}, p_{23}, ... of the matrix), $1 - p$ under the diagonal (the elements p_{10}, p_{21}, p_{32}, ...) and zeros otherwise.

We can represent this graphically as in Figure 1.1.

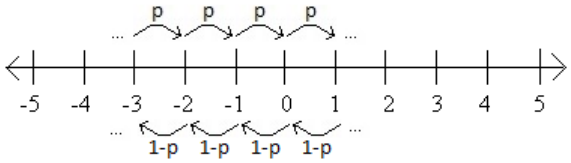

Figure 1.1: A simple random walk.

A **binary Markov chain** is a mechanism that produces zeros and ones. Its state space is given by $S = \{0, 1\}$ and its transition matrix is $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$.

When the chain is at state 0, the probability of returning to state 0 on the next step is $1 - p$, and the probability of discovering the chain in state 1 in the next step is p. Respectively, for state 1, the probability of returning to 1 on the next step is $1 - q$ and to gain 0 is q.

The graphical representation follows in Figure 1.2.

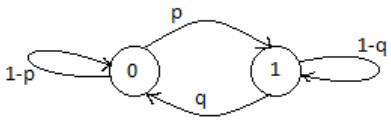

Figure 1.2: A binary Markov chain.

Another example.
Let us take the sequence $\{X_n\}_{n=0}^{\infty}$ of independent and identically distributed random variables such that $P(X_k = 1) = p$ and $P(X_k = 0) = 1 - p$.

Let us define $Y_n = X_1 + ... + X_n$ and $Y_0 = 0$.

How can we prove whether the sequence $\{Y_n\}_{n=1}^{\infty}$ given by Y_n is a Markov chain or not? We shall check whether the Markov property holds.

First of all, we observe that Y_{n-1} is function of $X_{n-1}, X_{n-2}, ..., X_0$ and that these are supposed to be independent. We now use this observation in the statement that follows.

$P(Y_n = y_n|Y_{n-1} = y_{n-1})$
$= P(Y_{n-1} + X_n = y_n|Y_{n-1} = y_{n-1})$
$= P(X_n = y_n - y_{n-1}|Y_{n-1} = y_{n-1})$
$= P(X_n = y_n - y_{n-1})$ [$\leftarrow$ by the observation]

Now let us observe that
$P(Y_n = y_n|Y_{n-1} = y_{n-1}, ..., Y_0 = y_0)$
$= P(Y_{n-1} + X_n = y_n|Y_{n-1} = y_{n-1}, ..., Y_0 = y_0)$
$= P(X_n = y_n - y_{n-1})$

We see that these two values are equal. This means that the Markov property holds and, accordingly, $\{Y_n\}_{n=1}^{\infty}$ is a Markov chain.

A counterexample.
We want to find out when a random process is not a Markov chain.

Let us consider the same notation and distribution from the example above. Now we set $S_n = \sum_{i=1}^{n} X_i$ and $S_0 = X_0$. Let also $Y_n = S_0 + S_1 + ... + S_n$.

Is $\{Y_n\}_{n=1}^{\infty}$ a Markov chain?

Proceeding as in the previous example, we see that

$$P(Y_n = y_n|Y_{n-1} = y_{n-1}, ..., Y_0 = y_0) = P(X_n = y_n - y_{n-1} - X_{n-1} - ... - X_0|Y_{n-1} = y_{n-1}, ..., Y_0 = y_0)$$

but since this is not independent from y_{n-1}, we cannot use the observation we have seen and, thus, this might not be a Markov chain.

Let us try to find a specific counterexample.

We consider a sequence of X_i such that
$X_0 = 0$, $X_1 = 1$, $X_2 = 0$, $X_3 = 0$,$X_4 = 0$ $X_5 = 1$.
In this case the S_i is following:
$S_0 = 0$, $S_1 = 1$, $S_2 = 1$, $S_3 = 1$, $S_4 = 1$, $S_5 = 2$.
Respectively, for the Y_i,
$Y_0 = 0$, $Y_1 = 1$, $Y_2 = 2$, $Y_3 = 3$, $Y_4 = 4$, $Y_5 = 6$.

Now, let us observe that
$P(Y_5 = 6|Y_4 = 4, Y_3 = 3)$
$= P(X_5 = y_5 - y_4 - X_4 - X_3 - X_2 - X_1 - X_0|Y_4 = 4, Y_3 = 3)$
$= P(X_5 = 6 - 4 - 1|Y_4 = 4, Y_3 = 3)$
$= P(X_5 = 1) = p.$

But if $\{Y_n\}_{n=1}^{\infty}$ was a Markov chain, it would also have been true that $P(Y_5 = 6|Y_4 = 4) = p$.

Let us compute this.
$P(Y_5 = 6|Y_4 = 4)$
$= P(X_5 = y_5 - y_4 - X_4 - X_3 - X_2 - X_1 - X_0|Y_4 = 4)$
$= P(X_5 = 6 - 4 - 2|Y_4 = 4)$
$= P(X_5 = 0) = 1 - p.$

This is obviously not the same as $P(X_5 = 1) = p$ for every p between 0 and 1, so $\{Y_n\}_{n=1}^{\infty}$ is not a Markov chain.

1.4 Conditional probability Joint probability distribution

How does the probability of event A change if we know that another event B occurs? This intuitive question might be answered using the definition of conditional probability.

We define the **conditional probability** of event A, given the probability of an event B as

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

for $P(B) > 0$.

Thus, the conditional probability of event A, given the occurrence of event B, is the ratio between the probability of the intersection between these two events and the probability of B itself. Of course, this makes sense only if $P(B)$ is greater than 0.

We want to see how we can express $P(X_0 = j_0, X_1 = j_1, ..., X_{n-1} = j_{n-1}, X_n = j_n)$ using this identity and the Markov property.

$$\begin{aligned}
&P(X_0 = j_0, X_1 = j_1, ..., X_{n-1} = j_{n-1}, X_n = j_n) \\
&= P(X_n = j_n | X_0 = j_0, ..., X_{n-1} = j_{n-1}) P(X_0 = j_0, ..., X_{n-1} = j_{n-1}) \\
&= P(X_n = j_n | X_{n-1} = j_{n-1}) P(X_0 = j_0, ..., X_{n-1} = j_{n-1}) \\
&= p_{j_{n-1} j_n} P(X_0 = j_0, ..., X_{n-1} = j_{n-1}) \\
&= ... \\
&= p_{j_{n-1} j_n} p_{j_{n-2} j_{n-1}} ... p_{j_0 j_1} p_{j_0}^{(0)}
\end{aligned}$$

where $p_{j_0}^{(0)}$ is the initial probability of j_0, $p_{j_0}^{(0)} = P(X_0 = j_0)$.

This is called the **joint probability distribution**. Eventually,

$$P(X_0 = j_0, X_1 = j_1, ..., X_{n-1} = j_{n-1}, X_n = j_n) = p_{j_0}^{(0)} \prod_k p_{j_{k-1}j_k}.$$

1.5 K-th order Markov chains Chapman-Kolmogorov equation

There are sequences of random variables whose future event does not only depend upon the present state but that might have also some other restrictions, that go back to the past. Let us see how it is possible to apply the concept of Markov chains to this class of sequences of random variables.

Let us consider the sequence of random variables $\{X_n\}_{n=0}^{\infty}$ taking values in a countable set S. The sequence $\{X_n\}_{n=0}^{\infty}$ is called a **k-th order Markov chain** if

$$\begin{aligned} &P(X_{n+1} = j_{n+1} | X_0 = j_0, X_1 = j_1, ..., X_n = j_n) \\ &= P(X_{n+1} = j_{n+1} | X_{n+1-k} = j_{n+1-k}, ..., X_n = j_n) \end{aligned}$$

for all $n \geq 0$, $j_0, ..., j_n$ in S, and k a positive integer. Thus, the future event X_{n+1} depends on the last k states of the chain. An ordinary chain defined in the past section is then referred to as a first order Markov chain.

Let us denote

$$p_{ij}(n) = P(X_{m+n} = j | X_m = i)$$

for all $n \geq 1$ and $j_0, ..., j_n$ in S. This is the representation of the state of the chain on its n-th step and thus these are called **the n-step transition probabilities**. One can observe that these quantities do not depend on m, as the chain is homogeneous.

Similarly, $p_{ij}(m+n) = P(X_{m+n} = j | X_0 = i)$.

Let us introduce the following identities:

1) the law of total probability:

$$P(X) = \sum_k P(X|A_k)P(A_k)$$

2) the formula of conditional probabilities:

$$P(A \cap B|C) = P(A|B \cap C)P(B|C).$$

Using these and the Markov property,

$$\begin{aligned}
&p_{ij}(m+n) = P(X_{m+n} = j|X_0 = i) \\
&= \sum_k P(X_{m+n} = j, X_m = k|X_0 = i) \\
&= \sum_k P(X_{m+n} = j|X_m = k, X_0 = i)P(X_m = k|X_0 = i) \\
&= \sum_k P(X_{m+n} = j|X_m = k)P(X_m = k|X_0 = i) \\
&= \sum_k p_{kj}(n)p_{ik}(m) \\
&= \sum_k p_{ik}(m)p_{kj}(n).
\end{aligned}$$

The equation

$$p_{ij}(m+n) = \sum_k p_{ik}(m)p_{kj}(n)$$

is called the **Chapman-Kolmogorov equation**. It expresses the probability of transiting from i to j in $m+n$ steps as the sum of all the possible transitions from state i to states k in m steps and from k to j in n steps.

It can also be represented using the matrix notation as

$$P(m+n) = P(m)P(n).$$

By induction, since $P^0 = I$, $P^1 = P$, $P^2 = PP$, ..., one observes that

$$P(n) = P^n.$$

Thus, the Chapman-Kolmogorov equation can be expressed as

$$P^{m+n} = P^m P^n.$$

1.6 Invariant distribution

A chain is described completely by its transition probabilities and the probability of its initial state. The transition probabilities have been defined and discussed above. Let us now define the **initial distribution** as

$$\pi(0) = (P(X_0 = 1), ..., P(X_0 = J)) = (p_1^{(0)}, ..., p_J^{(0)}).$$

Similarly, we set

$$\pi(n) = (p_1^{(n)}, ..., p_J^{(n)}) = (P(X_n = 1), ..., P(X_n = J)).$$

This notation represents the vector of probabilities of finding the chain at time n in state j, for $j = 1, ..., J$.

We will now introduce two other definitions.

First, a Markov chain is said to be **stationary** if $p_j^{(n)}$ does not depend on n, which means that it is independent from time.

Second, a distribution $\pi = (\pi_1, ..., \pi_J)$ is called **invariant** if it holds that $p_j^{(0)} = \pi_j$ implies $p_j^{(1)} = \pi_j$.

Let us now make a few observations about invariant distributions.

First of all, for any distribution, it is true that

$$\pi(n) = \pi(n-1)P.$$

In fact, the equality holds, as $p_j^{(n)} = \sum_k p_k^{(n-1)} p_{kj}$, applying the Chapman-Kolmogorov equation.

Second, if π is invariant, then

$$\pi = \pi P.$$

If π is invariant, then the result is $\pi = \pi(0) = \pi(1)$, and as $\pi(n) = \pi(n-1)P$, $\pi(1) = \pi(0)P = \pi(0)$, and thus $\pi = \pi P$.
Vice versa, if $\pi = \pi P$, then $\pi(1) = \pi(0)P = \pi(0)$ and then the distribution is invariant.

Lastly, every Markov chain with a finite state space has at least one invariant distribution.

In fact, let $\pi = (1/J, ..., 1/J)$.
Than

$$P\pi = \begin{pmatrix} p_{11} & \cdots & p_{1J} \\ \vdots & \ddots & \vdots \\ p_{J1} & \cdots & p_{JJ} \end{pmatrix} \begin{pmatrix} 1/J \\ \vdots \\ 1/J \end{pmatrix} = \begin{pmatrix} 1/J \sum_j p_{1j} \\ \vdots \\ 1/J \sum_j p_{nj} \end{pmatrix} = \begin{pmatrix} 1/J \\ \vdots \\ 1/J \end{pmatrix} = \pi$$

since $\sum_j p_{ij} = 1$ for every i.

We now want to discuss what the conditions are for **the uniqueness** of such a distribution. First of all, we study the long-term behaviour of the chain.

1.7 Ergodic, irreducible, aperiodic chains

We want to examine how the chain behaves over the long term. We now seek the position in which we find the chain after n steps and try to understand whether the chain converges at a specific distribution. Let us see whether particular conditions can be set such that every chain possessing these conditions will tend to a unique invariant distribution.

A chain is said to be **ergodic** if there is a distribution $a = (a_1, ..., a_n)$ such that a is the limiting distribution for a distribution $\pi(n)$, e.g.,

$$\lim_{n\to\infty} \pi(n) = a.$$

If a is a limiting distribution, then it is also the invariant distribution:
$a = \lim_{n\to\infty} \pi(n)$
$= \lim_{n\to\infty} \pi(n+1)$
$= \lim_{n\to\infty} (\pi(n)P)$
$= (\lim_{n\to\infty} \pi(n))P$ [$\leftarrow$ since we suppose the state space to be finite]
$= aP$.

We note that this holds for any initial distribution $\pi(0)$.

It is important to understand whether some states of the chain have anything in common and examine some further properties of the states. We will try to find out what these smaller parts of the chain are and how they can help us to study the whole chain.

We say that state i can be **reached** from state j if there is an n such that $p_{ij}(n) > 0$.

The two states i, j are said to be **communicating** if the state i can be reached from j and the state j can be reached from i.

A matrix of the chain is called **irreducible** if all the states are communicating.

We define a **period** $d(i)$ of a state i the greatest common divisor of the times when the chain returns to the state i, starting from i,

$$d(i) = gcd\{n > 0 : p_{ii}(n) > 0\}.$$

A state is said to be **periodic** or **aperiodic** if $d(i) > 1$ or $d(i) = 1$ respectively. We note that all the communicating states have the same period.

An example: The simple random walk has the period 2.

Let us consider a Markov chain with **finite state space** S **and stationary transition probabilities**, e.g., independent of n. If $p_{ij} > 0$ for all i,j, then

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j$$

which means that the n-step transition probabilities converge at the stationary distribution π for n tending to infinity. The rows of the n-step transition matrix converge to the vector of the stationary distribution.

Let us consider an **aperiodic irreducible Markov chain on a finite state space**. Then there is a unique distribution π such that

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j$$

for all i,j.

An example

Let us consider a Markov chain with transition matrix $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ with state space $S = \{0, 1\}$. The distribution

$$\pi = \left(\frac{q}{q+p}, \frac{p}{q+p} \right)$$

is an invariant distribution for P.

Now we want to see how we may compute it. The first step is to calculate P^n and then let $n \to \infty$. The easiest way to compute P^n is to calculate the eigenvalues of P.

The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = 1 - p - q$. Using the n-th power matrix properties, this implies that a general element $p_{ij}^{(n)}$ of P^n is

$$p_{ij}^{(n)} = a1^n + b(1-p-q)^n$$

for some a and b. We must calculate these now.

For p_{11} one has it that $p_{11}^{(0)} = 1$ since the probability to get back to the state 1 if we start from 1 in time 0 is 1. From the general equation obtained above we see that

$p_{11}^{(0)} = a + b = 1$, by setting $n = 0$. Furthermore, $p_{11}^{(1)} = 1 - p = a + b(1 - p - q)$, and so we get that

$$a = \frac{q}{p+q} \text{ and } b = \frac{p}{p+q}.$$

By resolving similar equations for p_{12}, p_{21} and p_{22} we obtain

$$P^n = \begin{pmatrix} p_{11}^{(n)} & p_{12}^{(n)} \\ \\ p_{21}^{(n)} & p_{22}^{(n)} \end{pmatrix} = \frac{1}{p+q} \begin{pmatrix} q & p \\ q & p \end{pmatrix} + \frac{(1-p-q)^n}{p+q} \begin{pmatrix} p & -p \\ -q & q \end{pmatrix}$$

and in the long term for n tending to infinity

$$P^n \xrightarrow{n\to\infty} \frac{1}{p+q} \begin{pmatrix} q & p \\ q & p \end{pmatrix}.$$

This implies that π, the invariant distribution, is

$$\pi = \left(\frac{q}{p+q}, \frac{p}{p+q} \right).$$

By direct computation one can establish that

$$\pi = \pi P.$$

This is the long-term behaviour of the chain. It tends to the unique invariant distribution we computed above. When we let $n \to \infty$, π is the representation of the percentage of time that the chain spends in each state.

Counterexamples.

If the chain is periodic, there is no convergence, as the next simple example shows.

On $S = \{0, 1\}$ with $P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $P^{2k+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $P^{2k} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

If the chain is reducible, there might not be the uniqueness of a stationary distribution.

On $S = \{1, 2, 3\}$ with $P = \begin{pmatrix} a & b & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ every distribution has to satisfy the equation $\pi = \pi P$ to be stationary. But if a distribution satisfies the equation, it is such that $\pi_1 = a\pi_1$, $\pi_2 = b\pi_1 + \pi_2$ and $\pi_3 = c\pi_1 + \pi_3$. This means that every distribution $\tilde{\pi} = (0, p, 1-p)$, for $0 \leq p \leq 1$, is stationary, and thus there is more than one stationary distributions.

Chapter 2

Learning of Markov chains

2.1 Introduction

A **model** is a simplified description of a phenomenon. It represents data or information observable in the real world. Since the real world is not perfectly describable by the models, we must express a particular range of the uncertainties for each model. Likewise, all the possible parameters of the model must be examined.

To completely describe a phenomenon, one must define all the possible models and learn the parameters and the uncertainties. After having studied these, one can find out what the best possible model for the phenomenon is. The operation of defining the best model is called learning. We want to see how this is applicable in the field of Markov models.

As we have seen before, a phenomenon described by a Markov chain, to be defined completely, must be specified by a transition matrix and an initial distribution. Let us consider the transition matrix $\Theta = (\theta_{ij})$ and an initial distribution $p_{j_0}^{(0)} = P(X_0 = j_0)$. We call $P(x|\Theta)$ the **family of models** for a training sequence $x = (j_0 j_1 ... j_n)$, with $x \in S^{n+1}$, $j_0, j_1, ..., j_n$ outcomes of a Markov chain $\{X_n\}_{n=0}^{\infty}$, given the transition matrix Θ.

Next, we will discuss how to find the best model of the defined family. We will start with the definition of the likelihood function.

2.2 Likelihood function

The **probability function** predicts unknown outcomes based on known parameters. We have seen before that

$$P(x|\Theta) = P(X_0 = j_0, X_1 = j_1, ..., X_{n-1} = j_{n-1}, X_n = j_n) = p^{(0)}_{j_0} \prod_k \theta_{j_{k-1}j_k}.$$

Let us now introduce the **likelihood function** $L(\Theta|x)$

$$L(\Theta|x) := P(x|\Theta).$$

According to this definition, in contrast to the probability function, this is a function that predicts unknown parameters based on known outcomes.

Note: The likelihood function $L(\Theta|x)$ is often written just as $L(\Theta)$.

One can omit the initial distribution $p^{(0)}_{j_0}$ and consider it a part of the problem. By this assumption,

$$L(\Theta) = \prod_k \theta_{j_{k-1}j_k}.$$

It is often more convenient to use the addition notation, and when the numbers are too small, it is more advantageous to do the computations using the logarithm functions.

Thus, we define the **log-likelihood function** as

$$l(\Theta) = \ln(\prod_k \theta_{j_{k-1}j_k}) = \sum_k \ln(\theta_{j_{k-1}j_k}).$$

The likelihood function predicts the unknown parameters of the model θ_{ij} based on the known outcome of a probabilistic experiment x. We do not know the exact

value of θ_{ij} but we can estimate its **"true value"**. Let us denote the true value as θ_0. One of the possible methods of the estimation is maximizing the likelihood of x.

2.3 Maximum likelihood

Let us consider the sequence $x = (j_0 j_1 ... j_n)$, the outcomes of an experiment. We will now establish a few notations that will be used subsequently.

Let us first define N_{ij} the number of l such that $j_{l-1} = i, j_l = j$, $1 \leqslant l \leqslant n$.

We also set N_i the number of l such that $j_l = i$, $1 \leqslant l \leqslant n-1$.

The meaning of these two numbers is as follows: N_{ij} is the number of passages from state i to state j in x and N_i is the number of visits to the state i, excluding the final instance.

Using the notation introduced above, the likelihood function can then be written as

$$L(\Theta) = \prod_k \theta_{j_{k-1} j_k} = \prod_i \prod_j \theta_{ij}^{N_{ij}}.$$

Similarly for the log-likelihood,

$$l(\Theta) = \sum_i \sum_j N_{ij} \ln(\theta_{ij}).$$

The **maximum likelihood estimate** is denoted by $\hat{\theta}_{ij}$. It is defined as

$$\hat{\theta}_{ij} = \arg\max_{0 \leq \theta \leq 1} P(x|\Theta = \theta).$$

It can be proved that

$$\hat{\theta}_{ij} = \frac{N_{ij}}{N_i}, \text{ for all } i, j.$$

Let us see this better in detail. The transition matrix is defined as $\Theta = (\theta_{ij})$. Let us set

$$\theta_i = (\theta_{i1}, ..., \theta_{in}).$$

As this is a stochastic matrix, it is necessary that $\sum_j \theta_{ij} = 1$. Now let us calculate the maximum likelihood estimate for θ_i.

By the definition,

$$\hat{\theta}_i = \arg\max_{\theta_{ij}} P(x|\theta_i) = \arg\max_{\theta_{ij}} \theta_{i1}^{N_{i1}} ... \theta_{in}^{N_{in}}$$

where $j = 1, ..., n$.

To make the computations clearer, let us set

$$\theta_i = t,\ \theta_{ij} = t_j,\ N_i = M \text{ and } N_{ij} = M_j$$

for one particular i. Note that in these conditions, $\sum_j t_j = 1$. Now, the equation above becomes

$$\hat{t} = \arg\max_{t_j} P(x|t) = \arg\max_{t_j} t_1^{M_1} ... t_n^{M_n}$$

for $j = 1, ..., n$.

By the definition of the log-likelihood

$$l(t) = l(t_1, ..., t_n) = \ln P(x|t).$$

We now want to find the maximum of this value. This will be the solution of the problem.

It can be solved by partially differentiating the log-likelihood. But, as the constraint $\sum_j t_j = 1$ is known, only $n - 1$ of the variables are free. Thus, we can set

$$\tilde{l}(t_1, ..., t_{n-1}) = l(t_1, ..., t_{n-1}, 1 - (t_1 + ... + t_{n-1})).$$

By the definition

$$l(t_1, ..., t_{n-1}, 1 - (t_1 + ... + t_{n-1})) =$$
$$M_1 \ln(t_1) + ... + M_{n-1} \ln(t_{n-1}) + M_n \ln(1 - (t_1 + ... + t_{n-1})).$$

We can take the partial derivatives of this value and set them to 0, as we want to find the maximum:

$$\frac{\partial \tilde{l}}{\partial t_j}(t_1, ..., t_{n-1}) = \frac{M_j}{t_j} - \frac{M_n}{1 - (t_1 + ... + t_{n-1})} = 0.$$

By resolving a similar equation for all j from 1 to n, we obtain

$$\frac{M_1}{t_1} = \frac{M_2}{t_2} = ... = \frac{M_n}{1 - (t_1 + ... + t_{n-1})} =: \lambda.$$

We call the common value λ. Then

$$t_j = \frac{M_j}{\lambda}$$

for $j = 1, ..., n$. But since we know that $\sum_j t_j = 1$, then

$$\sum_j \frac{M_j}{\lambda} = 1$$

this is

$$\frac{\sum_j M_j}{\lambda} = 1$$

and so

$$\lambda = M.$$

This means that $\hat{t_j} = \frac{M_j}{M}$ and $\hat{t} = (\hat{t_1}, ..., \hat{t_n}) = (\frac{M_1}{M}, ..., \frac{M_n}{M})$.

Note: To check that this is a maximum one needs to take the second order derivatives.

Returning to the original notation from the beginning where $\theta_{ij} = t_j$,

$$\hat{\theta}_i = (\hat{\theta}_{i1}, ..., \hat{\theta}_{in}) = (\hat{t}_1, ..., \hat{t}_n) = (\frac{M_1}{M}, ..., \frac{M_n}{M}) = (\frac{N_{i1}}{N_i}, ..., \frac{N_{in}}{N_i}).$$

The procedure of calculating the maximum likelihood is the method of learning the most likely value of the parameter Θ. This procedure estimates the unknown parameters of the probability distribution from the given data. There are also other methods of estimating the distribution of Θ, such as model averaging.

An example.
Let us consider a sequence of outcomes of a probabilistic experiment - in this case, a coin flip - with possible results H for heads or T for tails. Suppose that the tosses are independent. Let us suppose the result of the coin toss is the sequence $(HTTHH)$. The probability distribution is given by

$$P(X_i = H) = \theta = 1 - P(X_i = T).$$

The coin is not necessarily fair, so the probabilities to obtain H or T are not equal and θ varies between 0 and 1.

What is the most likely value for θ, given the outcomes $(HTTHH)$? First, we calculate the probability of the sequence of the coin tosses. We can compute it as

$$P(HTTHH) = \theta(1-\theta)(1-\theta)\theta\theta.$$

Thus,

$$L(\Theta|x) = P(x|\Theta) = \theta(1-\theta)(1-\theta)\theta\theta.$$

To obtain the maximum likelihood, we need to take the derivative of the quantity obtained above. Taking the derivative of $L(\Theta|x)$ one gets

$$\hat{\theta} = 0.6.$$

This is the maximum likelihood estimation for the unknown parameter θ, given the observable outcomes, the sequence $(HTTHH)$.

2.4 Consistency of the maximum likelihood

We will now discuss when the maximum likelihood computation is useful. Does an estimate of the likelihood always converge at the value of θ_0? How can we use

the value maximum likelihood to obtain the *true* value of an unknown parameter?

We define an estimate $\hat{\theta}$ of the likelihood to be **consistent** if $\hat{\theta} \to \theta_0$ in probability.

To derive some other results, we first state the **law of large numbers**. This is an important aspect of probability theory that deals with sums of random variables. Let us take the independent and identically distributed random variables $X_1, ..., X_n$ such that $|\mathbb{E}(X_1)| < \infty$. Then

$$\bar{X}_n = \frac{X_1 + ... + X_n}{n} \to \mathbb{E}(X_1) \text{ in probability,}$$

or equivalently,

$$\forall \varepsilon > 0 \; P(|\bar{X}_n - \mathbb{E}(X_1)| > 0) \to 0 \text{ as } n \to \infty.$$

Let us now make an observation: It can be proved that the law of large numbers holds for ergodic Markov chains.

Now, let $\{X_n\}_{n=0}^{\infty}$ be an ergodic Markov chain taking values in $S = \{1, 2, ..., J\}$, with the invariant distribution $\pi = (\pi_1, ..., \pi_J)$. If ϕ is a measurable, bounded, real valued function on $S \times S$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \phi(X_{j-1}, X_j) = \mathbb{E}_\Theta[\phi(X_0, X_1)] = \sum_i \sum_j \phi(i,j)\pi_i\theta_{ij}.$$

.

Now we want to discuss when the value of the maximum likelihood converges to the *true value* we want to identify. We have seen before that the estimation of the maximum likelihood is $\hat{\theta}_{ij} = \frac{N_{ij}}{N_i}$. Applying the law of large numbers for ergodic Markov chains with a suitable function ϕ we get that

$$\hat{\theta}_{ij} \to \theta_0.$$

Thus, we have found out that the condition for a chain to have the value of the maximum likelihood $\hat{\Theta}$ converging to the *true value* Θ_0 is that it must be an ergodic Markov chain.

Chapter 3

Hidden Markov Models

3.1 Introduction

Up until this point, we were speaking about processes in which all the outputs corresponded to an actual state that was recognizable. These are referred to as observable Markov models. Since the conditions set above might be too restrictive for real applications, we want to introduce a new family of more complex models.

A **hidden Markov model** is a stochastic process generated by two probabilistic mechanisms: a finite state Markov chain and a set of random functions, each associated to a state. When considering a hidden Markov model, one can only observe the output of random functions, while the states of the Markov chain are observable directly only through another set of processes.

3.2 Characteristics

When studying a hidden Markov model, we need to know some of its characteristics. These are:

1) **hidden Markov chain**

This is a homogeneous Markov chain $\{X_n\}_{n=0}^{\infty}$ taking values in a finite state space $S = \{1, 2, ..., J\}$. The transition probabilities of the chain are defined by

$$p_{ij} = P(X_{n+1} = j | X_n = i)$$

and these give the transition matrix $P = (p_{ij})$, with $p_{ij} \geq 0$ and $\Sigma_j p_{ij} = 1$. The initial distribution

$$\pi(0) = (p_1^{(0)}, ..., p_J^{(0)})$$

is also specified. The Markov chain is said to be hidden because though we know the probabilities of transiting from one state to another, we cannot see in which state the chain is at a specific time.

2) **observable random process**

This is a random process $\{Y_n\}_{n=0}^{\infty}$ taking values in a finite state space $O = \{o_1, o_2, ..., o_K\}$. K does not necessarily equal J. The process is observable, we can recognize the outputs of the functions. The conditional probabilities are given by

$$b_j(O_k) = P(Y_n = O_k | X_n = j)$$

and these make part of the so called emission probability matrix $B = (b_{jk})$. This is also a stochastic matrix, so $b_j(O_k) \geq 0$ and $\Sigma_k b_j(O_k) = 1$.

3) **conditional independence**

We assume that the emitted symbols are independent, given $\{X_n\}_{n=0}^{\infty}$. Thus,

$$P(Y_0 = O_0, ..., Y_n = O_n | X_0 = j_0, ..., X_n = j_n, B) = \prod_l b_{j_l}(l).$$

3.3 Examples

Coin toss.
Let us consider a sequence of coin tosses of two coins - a fair one (F) and a biased one (B). This is a hidden Markov model.

The transition of the coins is the hidden Markov chain with the state space $S = \{F, B\}$. Its transition matrix is known. For example, if we know that on the n-th toss X_n we use the fair coin, we know the probability p that it will be used also on the $n+1$-st toss and the probability $1-p$ that the biased coin will be used next.

We can observe the results of the tosses - heads (H) and tails (T). This is an observable process where $O = \{H, T\}$. Since one of the coins is biased, the sequence observed will not have the long term proportion between Hs and Ts equal to 0.5. Based on these observations it is not clear what exactly the sequence of the states is, e.g., when the fair coin was used and when the biased the coin was used. This is the hidden part of the model, only by seeing the results of the coin tosses, it is not clear what the sequence of coins that have produced it is.

This model is demonstrated in the following example. Let $P = \begin{pmatrix} 0.1 & 0.9 \\ 0.3 & 0.7 \end{pmatrix}$ be the transition matrix of the coin states. Let also $B = \begin{pmatrix} 0.5 & 0.5 \\ 0.8 & 0.2 \end{pmatrix}$ be the emission probabilities matrix.

Therefore, if the coin is fair, the probabilities to obtain heads or tails are equal, $b_F(H) = 0.5$ and $b_F(T) = 0.5$, while if the coin is biased, we have $b_B(H) = 0.8$ and $b_B(T) = 0.2$. The probability of observing heads and tails clearly depends on the coin used.

If we also set the initial distribution $\pi(0) = (\pi_F(0), \pi_B(0))$, we have character-

ized completely the model. The reason why this is called a hidden Markov model is that the states of the coins are not observable.

If we only consider a single coin, there is just one unknown parameter, namely the probability of obtaining heads or tails. It is clear that when increasing the number of coins, the number of unknown parameters increases as well. For example, if we use two biased coins, we have 2^2 unknown parameters, for a system with three biased coins it is 3^2 and so on. We can extend the example of the coin toss to a more general case.

Urn and ball model.
Let us suppose to have J different urns, numbered $1, 2, ..., J$, and K different colours of balls. In each urn there are K balls, all of them of different colours. The emission probability $b_{number}(colour)$ is given for each urn and each colour.

By a defined random process, we choose a ball from an urn. The process is the hidden Markov chain, so we do not know which urn has been chosen. Then the colour of the ball is registered and the ball is replaced to the same urn from which it was selected.

Afterwards, according to the defined probabilistic mechanism, we pick another urn, choose a ball and repeat the procedure. Thus, we obtain a sequence of observed colours, $o = (O_1O_2...) = (yellow, green, ...)$. The colour of the ball is the observable random process. According to the colour sequence, one can use a specific algorithm and determine the most probable sequence of the numbers of the urns from which the balls have been chosen. However, there is no such thing as the *correct* state sequence.

3.4 Influence Diagram

A hidden Markov model can be expressed by an **influence diagram** as shows Figure 3.1.

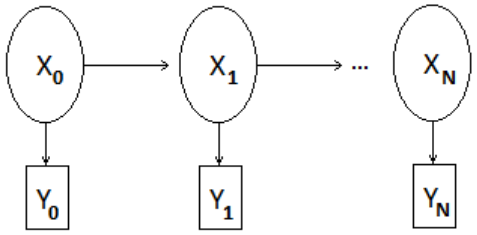

Figure 3.1: Influence diagram.

The nodes Y_i represent the observations we made. For a standard model, these are independent. The only assumption we make is represented by the arrows in the figure. This is the Markov property for $\{X_n\}_{n=0}^{\infty}$, so that the state X_{j+1} depends only upon X_j.

We can only see the observations Y_i while the states X_j are hidden. We can, though, try to reveal them using the algorithms that will be discussed in the following sections.

3.5 Problems

Given a standard hidden Markov model, there are a few basic problems we want to solve so that the model can be useful in applications.

The first question is the so-called likelihood problem. We can formulate the question as follows: Given a sequence and a model, what is **the probability that the model has generated the sequence**? To resolve this question, we can calculate the probability of all the possible sequences that might have produced the

observations and than choose the one with the highest probability. This might not be easy for long sequences with lots of possible states, and we will see algorithms that might facilitate this task later on.

This question can also be viewed as a scoring problem. The question would be: How well does the model describe the phenomenon? We can give a score to each of the models and according to this we can choose the best model for the phenomenon we are modelling.

The second task is to try to understand what is **the optimal hidden state sequence** that produces the observations $o = (O_0...O_n)$, given a sequence of symbols-observations. The most frequent way to do this involves finding the sequence $(j_0^*...j_n^*)$ that maximizes $P(X_0 = j_0, ..., X_n = j_n, Y_0 = O_0, ..., Y_n = O_n|P, B, \pi(0))$. It is impossible to say what state the system is in when looking at the output, since there are many sequences of the states that can generate the sequence of symbols. Considering that all these have different probabilities, we want to find the most likely one.

The difficulty here is to set the right criteria concerning what the optimal sequence actually is. We might be looking for the sequence of the most probable single individual states, for the whole sequence of states with the highest probability or set some other criteria.

Lastly, one needs to find the **model that is the most likely to produce the sequence** $o = (O_0...O_n)$. Given some data, we want to create the best model that represents a phenomenon. We can improve the properties of the model so that it has greater capability to model the sequences. This cannot be resolved analytically. Instead, we use a so-called training sequence. Based on this, we re-estimate the parameters of the model and continue iteratively until we obtain the desired parameter's quantity.

It is important to reveal only the main features of the modelled data, not to describe every single detail - otherwise, the model might become over-fitted and not be able to generalise the data well enough.

Later, the first and the second problem will be discussed in detail.

Chapter 4

Biology

4.1 The structure of the nucleic acids

A cell is the basic structural, functional, and biological unit of all known living organisms. The biological information of all the organisms is contained in the cell. There are two different types of genetic material: deoxyribonucleic acid and ribonucleic acid.

The molecule of deoxyribonucleic acid (DNA) consists of two strands of nucleotides. A nucleotide is the basic element of the acid. It is made of a nitrogenous base, a sugar and a phosphate group. There are two different types of nitrogenous bases: the purines, adenine (A) and guanine (G), and the pyrimidines, cytosine (C) and thymine (T). The bases of the two strands form hydrogen bonds between themselves - namely, guanine forms three hydrogen bonds with cytosine and adenine forms two hydrogen bonds with thymine. These bonds are the cause of the characteristic natural structure of DNA: the double helix. The bases are orientated towards the inside of the helix, with the phosphate group and the sugar bone outside. Two strands of DNA run antiparallel and are held together by the hydrogen bonds between the bases. The strands are complementary, so it is sufficient to have the representation of just one of them to know the other one. The order of the nucleotides in the strand is dependent on the way the genetic information is

saved in the cell. Different organisms have different sequences of DNA in their cells.

RNA, ribonucleic acid, is a chain of nucleotides that forms a unique strand. The bases here are adenine, guanine, cytosine and uracil (U). There are different types of RNA, all having different functions inside the cell.

A gene is a sequence of DNA that can be transcribed into RNA and translated into a string of amino acids called a polypeptide. Genes are made of two different types of segments, exons and introns. The exons are the parts of the gene that are later transformed into a polypeptide. The introns are the non-coding parts.

4.2 Proteins

Proteins are important cell macromolecules with a wide variety of functions vital for the functioning of the cell, and they are involved in all intracellular phenomena. They are composed of strings of amino acids, the polypeptides. There are 20 different amino acids in nature, and they can form thousands of different proteins.

Proteins are synthesised from DNA through the process of transcription and translation. First, the information encoded in DNA is transcribed onto mRNA with the respective complementary bases. Later, the non-coding part of the sequence of the gene is removed. A triplet, three bases of mRNA, is called a codon. On one side, tRNA carries another triplet of bases. This is called an anticodon. The bases of the anticodon are complementary to the bases of the codon on mRNA. The other end of the tRNA carries an amino acid. The amino acid is specific for each codon. In the process of translation, the amino acids are joined together to form a protein. The amino acids are labelled by three letters, depending on their names.

The nature and the properties of the proteins change depending on the sequence of amino acids. Also, the sequence influences the 3D form of the protein. There

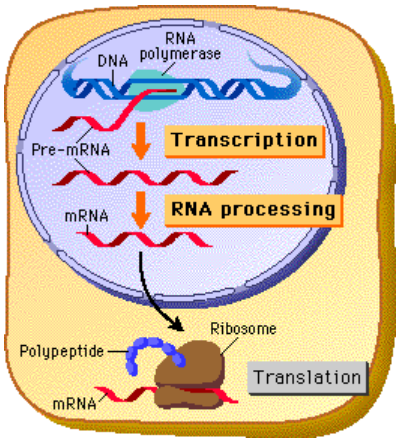

Figure 4.1: The translation and transcription process.

are different levels of organisation of the molecule. The first level is a simple linear sequence of amino acids. The sequence folds further according to what amino acids it is made of, as these form bonds among themselves, and forms a 3D secondary structure. An example of a secondary structure is an alpha helix or a beta sheet.

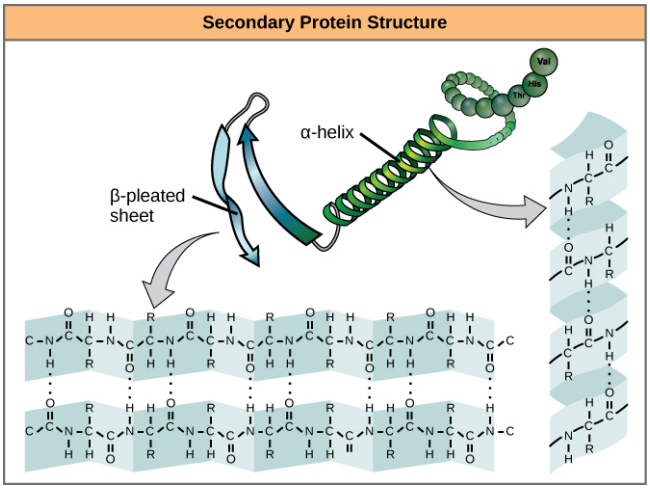

Figure 4.2: Secondary structure of a protein.

There are 4 different nitrogenous bases, and they can form $4^3 = 64$ different triplets, but there only are 20 amino acids, so for each amino acid there are different codons, as shown in Figure 4.3. There is one codon that initiates the protein

synthesis and three possible codons to end it.

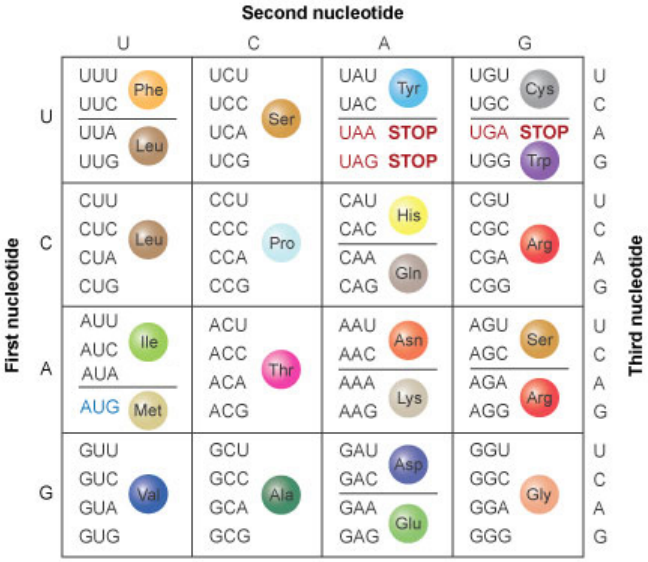

Figure 4.3: Aminoacids and codons.

Note: If the sequence of the bases is known, we are able obtain the sequence of the amino acids and understand the structure and the function of the protein. If only the structure of the protein is known, we do not precisely know the sequence of the bases.

4.3 Sequences

For biological purposes, it is important to understand whether two sequences of proteins, DNA or RNA have some structural or evolutionary similarities. Therefore, we want to align two sequences and find out whether there are any regions that might share the same ancestor. If two sequences have some analogies, we say they are homologous. We can compare two sequences by finding the number of steps needed to transform one sequence to the other one.

We align two sequences and see whether the corresponding symbols match, mismatch or cannot be compared. This is translated into biological terminology as a possible mutation, insertion or deletion of a part of a sequence.

We want to find the method of aligning the sequences with the lowest possible number of steps needed to transform one to another and then evaluate whether they are homologous or not.

It is important to note that in biology, if two symbols of two sequences do not match, it does not necessarily mean they are completely different. It is necessary to establish a scoring system that assigns certain values to symbols that are not equal but that might have analogies anyway.

4.4 Hidden Markov models

How is the concept of hidden Markov models applicable in the field of biology? Let us make a short and rather simplified description of the problems we are concerned about.

Regarding the genome sequences, we have seen that a gene is made of introns and exons. These have certain properties that differ from the others - namely, the frequency of the occurrence of the four nitrogenous bases. We want to determine which parts of the gene participate actively in the protein synthesis and use algorithms to reveal the sections of the gene that are most likely to be introns and the segments that might be exons. Hidden models are also useful in predicting the protein secondary structure, in modelling families that have related DNA or protein sequences and other problems.

Chapter 5

Viterbi algorithm

5.1 The algorithm

We will now discuss the solution to the second problem. The question we want to answer is: Given a sequence of symbols, what is **the optimal hidden state sequence**?

We want to find the sequence that represents the best the observed symbols. There is no *correct* sequence that solves this problem; however, we can set a few parameters, according to a few of our criteria, to be able to define which one is the best solution, and based on these we can attempt to uncover the hidden sequence.

This can be done in many different ways - there are more possibilities of what exactly can be meant by *the optimal* sequence. For our purposes, it makes sense to look for the most probable *sequence* of states.

One other option is to seek a sequence of single states having the highest probability. In this case, the whole sequence might not be very likely to occur, as the probability p_{ij} that one state follows another might be zero or very low. Thus, we will now look for the solution for the problem defined in the previous paragraph, trying to find the most probable sequence as a whole.

The solution to this is the Viterbi algorithm, proposed by Andrew Viterbi in 1967.

The **idea** is to consider all the possible sequences and evaluate the probabilities that these are generated by the model we defined. Eventually, one can choose the one with the highest probability. This is the simplest possible idea. Since it is computationally not very reasonable, one can proceed in the following way.

For every sequence of hidden states $x = (j_0...j_n)$ we evaluate the probability for the subsequence ending at the position i, for $i = 0, ..., n$, at a state l with a symbol O_i by

$$p_l(O_i) = b_l(O_i) \cdot \max_k\{p_k(O_{i-1}) \cdot p_{kl}\}$$

where

p_{kl} is the probability of transiting from state k to l, by the hidden Markov chain with the state space $S = \{1, 2, ..., J\}$, as defined above,

$p_k(O_{i-1})$ is the probability for the path ending at the position $i-1$ at state k with a symbol O_{i-1},

and $b_l(O_i)$ is the emission probability of symbol O_i at the state l.

It is important to notice that it is often more convenient to use the logarithm function and the addition notation for calculations, as probability values might be too small for the calculus.

There are J possible states at the position $i-1$ and, therefore J possible paths, of which one has a higher probability. Thus, at the state i, there are J^2 possible states, since there are J states up to the step $i-1$ and other J possibilities for the passage to the step i.

Upon reaching the final state n, we look for the state with the highest probability from the previous step $n-1$. Since there is only one such sequence, we can return back to the state $n-1$ to reveal the position at the state, and so on, up to the state 0, tracking back the entire the sequence.

The **algorithm** consists of four steps.

It begins with the initialization:

$$p_l(O_0) = b_l(O_0) \cdot p_l^{(0)}$$

for all the states l from 1 to J. Since we also want to remember the best possible sequence, that will be back-traced at the end of the procedure, a new array is introduced:

$$\psi_1(0) = 0.$$

The next step is to continue the operations recursively:

$$p_l(O_i) = b_l(O_i) \cdot \max_k\{p_k(O_{i-1}) \cdot p_{kl}\}$$

for all the states l from 1 to J, and all the positions i from 1 to n. We also continue recursively recalling the ψ function:

$$\psi_i(l) = \arg\max_k\{p_k(O_{i-1}) \cdot p_{kl}\}$$

for l from 1 to J, for i from 1 to n.

Then the termination follows. We set

$$P^* = \max_l\{p_l(O_n)\},$$

this is the highest probability at the last position,
and

$$Q_i^* = \arg\max_l\{p_l(O_n)\},$$

this is the state with the highest probability.

Lastly, we want to reverse the selected sequence. We do this using the $\psi_i(l)$ array we introduced at the beginning:

$$Q_i^* = \psi_{i+1}(Q_{i+1}^*)$$

for the positions i back from $n-1, n-2, ...$ to the beginning 0.

5.2 An example

Let us consider a hidden Markov model as in Figure 5.1.

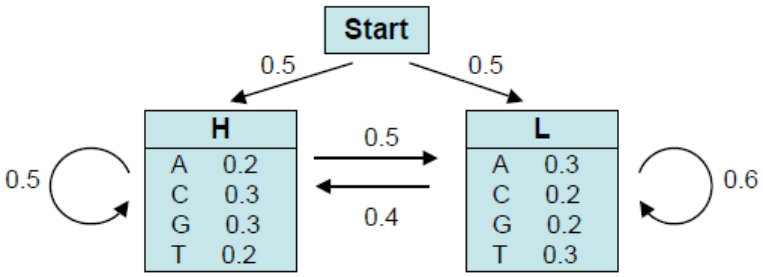

Figure 5.1: Hidden Markov Model.

According to the figure, the state space is made of the states $S = \{L, H\}$ and the random process, which in this case might be a DNA sequence, takes values in $O = \{C, G, T, A\}$.

Let us consider a sequence of the observed symbols $o = (GCAGGATA)$. Now, we know that there might be many possible sequences of the states hidden under this observation.

One of them, for example, might be $s = (LLHHLHHH)$.

First of all, what is the probability that the model has produced the sequence o through s? We can compute the probability, using the notations defined above, as

$$P = p_L^{(0)} \cdot b_L(G) \cdot p_{LL} \cdot b_L(C) \cdot p_{LH} ... = 0.5 \cdot 0.2 \cdot 0.6 \cdot 0.2 \cdot 0.4 ...$$

Any other path of hidden states producing the sequence o will have a different probability.

Then, to describe the most likely path, we use the Viterbi algorithm.

For example, using the formula for the probability of a subsequence introduced above, one can calculate the probability of the most likely path ending at state H with the observation A on the 3^{rd} position:

$$p_H(A[3]) = b_H(A) \cdot \max\{p_L(C[2]) \cdot p_{LH}, p_H(C[2]) \cdot p_{HH}\}.$$

One can compute these probabilities for both states H and L for all the positions. At the last position, we seek the hidden state with the highest probability. Considering that in the algorithm the states are recalled through the ψ function, we can reverse the sequence from the final state to the beginning using ψ.

Note: There are more efficient ways to calculate the probabilities then those described above.

Chapter 6

Forward-backward algorithm

6.1 Forward algorithm

We will now try to find a solution to the first problem. That is, given a model, what is **the probability of seeing the sequence of observations** $o = (O_0...O_n)$?

Let us first suggest the simplest possible answer to the question. It would be easy to take every possible sequence of the hidden states $x = (j_0...j_n)$ of length $n + 1$ and evaluate all the probabilities of all these sequences.

Let us try to solve this for one particular sequence. The probability of seeing the observations is given by

$$P(o|x) = \prod_{i=0}^{n} P(Y_i = O_i|X_i = j_i) = b_{j_0}(O_0)...b_{j_n}(O_n).$$

The probability of obtaining the sequence of the hidden states x is

$$P(x) = p_{j_0}^{(0)} p_{j_0 j_1} p_{j_1 j_2} ... p_{j_{n-1} j_n}.$$

Using the formula of the conditional probability

$$P(o, x) = P(o|x)P(x).$$

Then, considering all the possible sequences x,

$$P(o) = \sum_x P(o|x)P(x)$$

and this can be written as

$$P(o) = \sum_{j_i} p^{(0)}_{j_0} b_{j_0}(O_0) p_{j_0 j_1} b_{j_1}(0_1) p_{j_1 j_2} ... b_{j_n}(O_n).$$

This is the sum of

the initial distribution probability $p^{(0)}_{j_0}$,

the transition probabilities form one state to another, p_{ij},

the emission probabilities of the symbol O_i at the state j_k, $b_{j_k}(O_i)$.

Thus, we are summing up the initial probability, the probability of transiting from the first state to the second one, the probability of observing the first symbol at the second state, and so on.

This is the formal description of the intuitive computation of the probability of obtaining a sequence of symbols, given a sequence of states, that we made in the previous section. But this is not an efficient computation - it requires an order of $(n+1)J^{(n+1)}$ calculations, as at every time from 0 to n there are J possible states and for every state we need an order of $n+1$ calculations.

This is the reason we need **a more efficient algorithm**. We will now discuss such an algorithm: The forward-backward algorithm.

Similar to our investigation of the Viterbi algorithm, we consider a partial sequence of observations from time 0 up to time i. Let us introduce a new variable

$$\alpha_l(O_i)$$

to be the probability of observing a subsequence of x up to the time i being at state l at the time i. One can compute $\alpha_k(O_{i+1})$ by induction, using the "forward"

part of the algorithm.

It consists of 3 steps that are similar to those introduced before.

First of all, the initialization,

$$\alpha_l(O_0) = \pi_l(0) b_l(O_0).$$

For all the states l from 1 to J, this is the probability of finding the chain on the first step at the state l and to observe the first symbol of the sequence at the state l.

Then the recursion follows,

$$\alpha_l(O_{i+1}) = (\sum_k \alpha_k(O_i) \cdot p_{kl}) \cdot b_l O_{i+1}$$

for all the states l from 1 to J and all the times from 1 up to $n-1$. Here, as usual,

p_{kl} is the transition probability of the hidden states of the chain,

$b_l(O_{i+1})$ is the emission probability of the symbol O_{i+1} at state l,

$\alpha_k(O_i)$ is the probability to obtain the subsequence $(O_0...O_i)$

and by multiplying it by p_{kl} we receive the probability of reaching state k *and* of observing the sequence $(O_0...O_i)$.

By summing over k the α_ks we get the probability of reaching state k at time $i+1$.

The final multiplication gives us the probability of observing the symbol O_{i+1} at time $i+1$, given that the chain is at the state l.

The last step is the termination,

$$P(o) = \sum_l \alpha_l(O_n).$$

This algorithm is computationally more reasonable than its simpler version from the beginning in that it is of the order of $(n+1)J^2$ - for every time i there are J possible states, no matter now long the sequence of the observations is. Using this algorithm, we can obtain the probability that the model has produced the observed sequence. This is the solution to the first problem.

6.2 Backward algorithm

The solution to the second problem is the computation of the most likely sequence of hidden states. Let us now formulate a different criterion of the definition of *the optimal sequence*. We will now calculate the sequence of the single most probable *states*.

To calculate the probability of a hidden state, we do not only need the observations previous to O_i, but also those following *after* the symbol O_i in the sequence, as these influence the underlying states of the Markov chain as well.

Thus, we define the **backward** algorithm similarly to the forward part. We introduce an array

$$\beta_k(O_i)$$

as the probability of observing the partial sequence $(O_{i+1}...O_n)$ given that at the time i the chain is at the state k.

The algorithm is described in what follows.

First of all,

$$\beta_k(O_n) = 1.$$

This is an arbitrary choice for any state.

The recursion steps begin from the last time and return to the first time as follows,

$$\beta_k(O_i) = \sum_l \beta_l(O_{i+1}) \cdot p_{kl} \cdot b_l(O_{i+1})$$

for all the times i from $n-1, n-2, ...$ up to 0. This works similarly, except in reverse, as the forward algorithm, and the order of the computation is also the same.

Eventually, we can define the probability of being at state k at time i given the observed sequence:

$$P(X_i = j_i|o) = \frac{\alpha_k(O_i)\beta_k(O_i)}{P(o)} = \frac{\alpha_k(O_i)\beta_k(O_i)}{\sum_k \alpha_k(O_i)\beta_k(O_i)}.$$

This is the probability of the sequence being in a particular time at a particular state. If we compute this value for every state j_i, we obtain the sequence of the most probable single states.

The difference between this result and the Viterbi algorithm solution is that these single states might have a small probability of following one another. They are the most probable states individually, but the probability that this whole state sequence has produced the observations might be very small.

Bibliography

[1] Alberts B. et al.: Molecular Biology of the Cell.
Garland Science 2002

[2] Bilingsley P.: Statistical Methods in Markov Chains.
1960

[3] Bishop C. M.: Patteren Recognition and Machine Learning.
Springer 2006

[4] Churchill G. A.: Stochastic Models for Heterogeneous DNA Sequences.
1989

[5] Durbin R. et al.: Biological Sequence Analysis.
Cambridge University Press 1998

[6] Eddy S. R.: What is a hidden Markov model.
Nature Publishing Group 2004, Nature Biotechnology, Vol. 22, No. 20, Oct 2004

[7] Ghahramani Z.: An Introduction do Hidden Markov Models and Bayesian Networks.
2001

[8] Koller D., Friedman N.: Probabilistic graphical models, principals and techniques.
Massachusetts Institute of Technology 2009, Chapter 17

[9] Koski T.: Hidden Markov Models for Bioinformatics.
Kluwer Academic Publishers 2001

[10] Kroese D.: A Short Introduction to Probability.
2009
<http://www.maths.uq.edu.au/~kroese/asitp.pdf>

[11] Krogh D.: Biology, a guide to the natural world.
2003
<http://wps.prenhall.com/esm_krogh_biology_3/0,8750,1136394-,00.html>

[12] Pascarella S., Paiardini A.: Bioinformatica: Dalla sequenza alla struttura delle proteine.
Zanichelli 2011

[13] Rabiner L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.
1989

[14] Wu B.: Invariant Probability Distributions.
2011
<http://www.math.uchicago.edu/~may/VIGRE/VIGREREU2011.html>

[15] Durham University
<http://maths.dur.ac.uk/stats/courses/ProbMC2H/Probability2H.html> Handouts 2014

[16] Massachusetts Institute of Technology
<http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/>
Topic: Properties of Maximum Likelihood Estimators

[17] Université Libre de Bruxelles
<http://homepages.ulb.ac.be/~dgonze/TEACHING/INFO_F_434.html>
Chapter: Finding motifs in sequences, viterbi.pdf

[18] University of Illinois
<http://www.stat.illinois.edu/courses/stat530/MOD2.html>
Lecture Notes: Lecture 13

[19] University of Illinois at Chicago
<https://www.evl.uic.edu/shalini/coursework.html>
Introduction to Computational Biology: Presentation on Hidden Markov Models

[20] The Hebrew University of Jerusalem
<http://www.cs.huji.ac.il/course/2004/cbio/handouts.html>
Course of Computational Methods In Molecular Biology, Title: Sequence Alignment, Class Feb 28

[21] Figure 4.1
<http://www.phschool.com/science/biology_place/biocoach/transcription/tctlpreu.html>

[22] Figure 4.2
<https://www.boundless.com/biology/textbooks/boundless-biology-textbook/biological-macromolecules-3/proteins-56/protein-structure-304-11437/>

[23] Figure 4.3
<http://www.nature.com/scitable/content/the-amino-acids-specified-by-each-mrna-6903567>

[24] Figure 5.1
<http://homepages.ulb.ac.be/~dgonze/TEACHING/INFO_F_434.html>