Data Mining and Machine Learning

Statistical Modelling of Sequences (1)

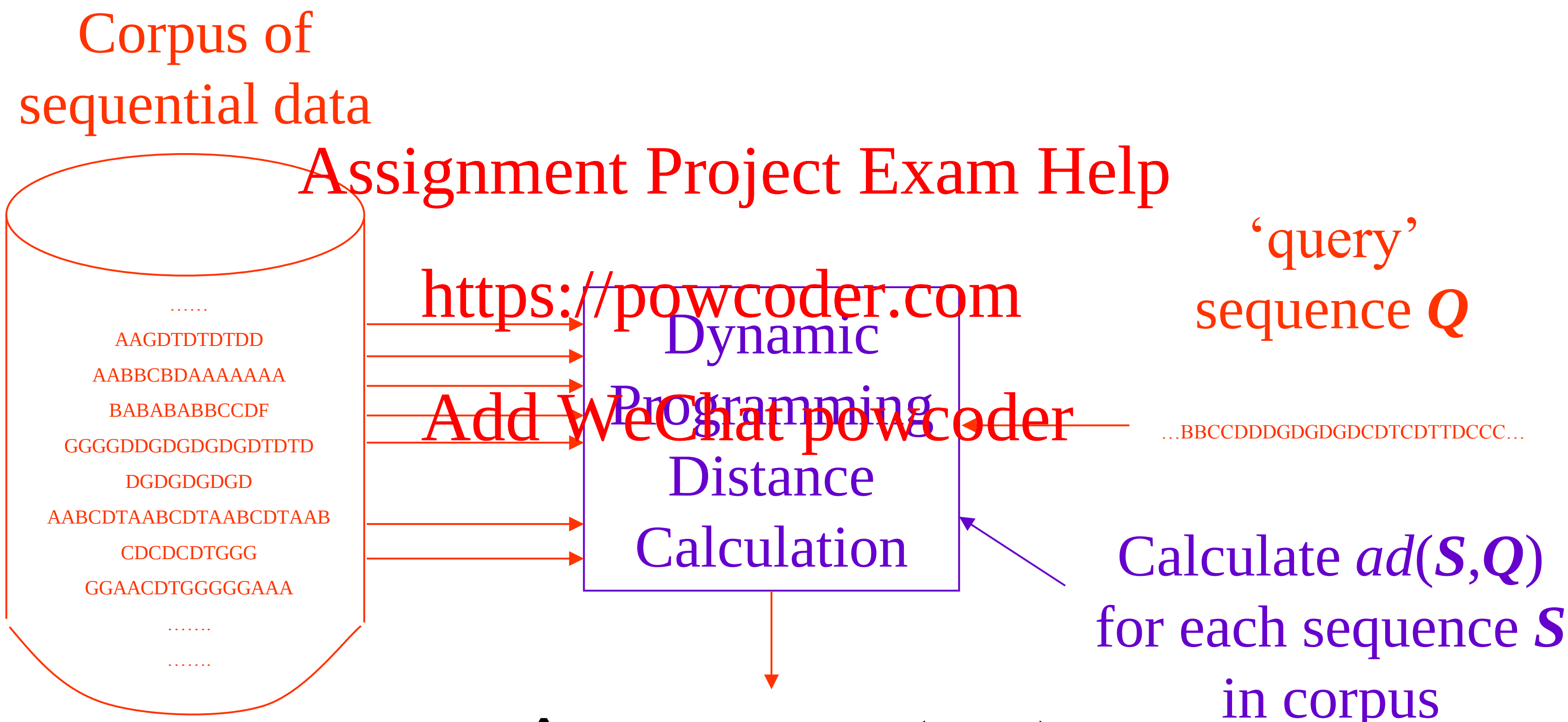
Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

- Extension of dynamic programming to statistical modelling of sequences
- Introduction to Markov models through example
- Calculation of probability of a state sequence
- State distribution
- Relationship to Page Rank

Sequence retrieval using DP



$$\hat{S} = \arg\min_{S} ad(S, Q)$$

Limitations of ‘template matching’

- This type of analysis is sometimes referred to as <u>template matching</u>

- The ‘templates’ are the sequences in the corpus

- Can think of each template as representing a ‘class’
- Problem: determine which class best fits the query

- Performance will depend on precisely which template is used to represent the class

UNIVERSITY OF BIRMINGHAM

Markov Chains

- Discussed briefly in the lecture on Page Rank
- Suppose that you want to understand the habits of shoppers on a High Street

- Suppose that:

  – There are N shops: $S_1, S_2, \ldots, S_N$

  – Probability that the next shop that a shopper visits is S_j depends only on the shop S_i that the shopper is currently visiting – this is the Markov Property

Markov Chains (continued)

- In other words, if x_n is the n^{th} shop visited:

$$P\left(x_n = S_j \mid x_{n-1} = S_i, x_{n-2}, x_{n-3}, \ldots, x_0\right) = P\left(x_n = S_j \mid x_{n-1} = S_i\right)$$

- In a Markov chain $S_1, \ldots, S_N$ are called <u>states</u>
- The behaviour of the shopper is completely described by two factors: the initial state probability vector P_0 and state transition probability matrix A:

$$P_0 = \begin{bmatrix} P_0(1) \\ P_0(2) \\ \vdots \\ P_0(N) \end{bmatrix}, A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1N} \\ a_{21} & a_{22} & \ldots & a_{2N} \\ \vdots & \vdots & a_{ij} & \vdots \\ a_{N1} & a_{N2} & \ldots & a_{NN} \end{bmatrix}$$

Markov Chains (continued)

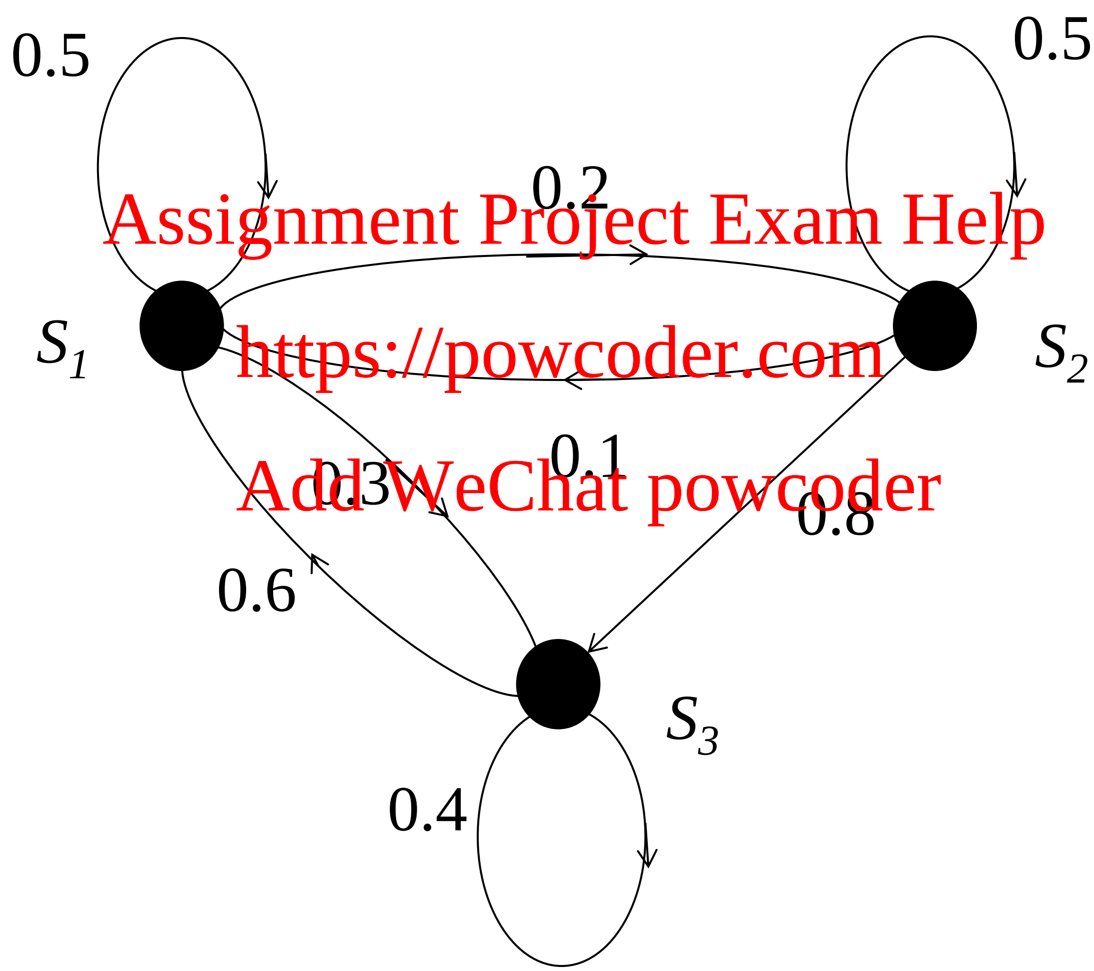
- $P_0(n)$ is the probability that the first shop that a shopper visits is S_n
- a_{ij} is the probability that the n^{th} (next) shop that is visited is S_j given that the n-1^{th} (current) shop visited is S_i
- Or (better) $a_{ij} = P(x_t = S_j | x_{t-1} = S_i)$. Note that this is independent of t.
- The Markov assumption is not really appropriate for this application – it is made because it simplifies the mathematical theory and computation

Example

- Suppose:
- $N = 3$, $P_0 = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix}, A = \begin{bmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.1 & 0.8 \\ 0.4 & 0 & 0.6 \end{bmatrix}$

- This means (for example):
  – Half of all shoppers start in shop S_1
  – Half of all shoppers return to S_1 immediately after leaving it (!?)
  – Shoppers never visit S_2 immediately after S_3

Transition network representation



A simple probability calculation

- Suppose that $x = x_1,\ldots, x_T$ is a sequence of states (shops)
- Then the probability $P(x)$ is given by:

$$P(x) = P(x_T|x_{T-1}) \times P(x_{T-1}|x_{T-2}) \times \cdots \times P(x_2|x_1) \times P(x_1)$$

$$= a_{x_{T-1},x_T} \times a_{x_{T-2},x_{T-1}} \times \cdots \times a_{x_1,x_2} \times P_0(x_1)$$

- Example: given the previous 3 state Markov model, suppose

$$x = S_1S_3S_1S_1S_2$$

$$P(x) = a_{12} \times a_{11} \times a_{31} \times a_{13} \times P_0(1)$$
$$= 0.2 \times 0.5 \times 0.4 \times 0.3 \times 0.5 = 0.006$$

State distribution at time t

- Suppose that the state occupancy at time t is P_t
- i.e. $P_t(n)$ is the probability that the state at time t is state n. $P_t(n) = \text{Prob}(x_t = S_n)$
- Then $P_{t+1} = A^T P_t$
- To see this consider the 3 state example:
$$P_{t+1}(2) = P_t(1) \times a_{12} + P_t(2) \times a_{22} + P_t(3) \times a_{32}$$
- This is just the dot product of the 2nd column of A and P_t, and the result follows.

State distribution at time *t* (cont.)

- In particular $P_1 = A^T P_0$
- Hence $P_2 = A^T P_1 = A^T(A^T P_0) = (A^T)^2 P_0$
- In general $P_t = (A^T)^t P_0$
- Example: For our 3 state Markov model

$$P_0 = \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix}, P_1 = A^T P_0 = \begin{bmatrix} 0.5 & 0.1 & 0.4 \\ 0.2 & 0.1 & 0 \\ 0.3 & 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.2 \\ 0.3 \end{bmatrix} = \begin{bmatrix} 0.39 \\ 0.12 \\ 0.49 \end{bmatrix}, \text{etc}$$

Page Rank revisited

- In our discussion on Page Rank
  - N is the number of webpages
  - S_n is the n^{th} webpage
  - a_{ij} is the probability of visiting S_j if the surfer is currently at S_i
  - The Page Rank is a vector $P = \begin{bmatrix} P_1 \\ \vdots \\ P_N \end{bmatrix}$ such that $P = A^T P$, in other words an eigenvector of A^T with eigenvalue 1
  - P is an <u>equilibrium distribution</u> for the Markov process

Hidden Markov Models (HMMs)

- Let's go back to our original shopping example
- Suppose that when a shopper visits a shop he or she makes a single purchase from a set of M possible items $I_1, \ldots I_M$
- Suppose that, instead of observing the sequence of shops, we observe the sequence of purchases
- Because different shops may sell the same item it is in general not possible to know the shop sequence unambiguously from the purchase sequence
- This is an example of a Hidden Markov process

Summary

- Markov processes and Markov models
- Hidden Markov models