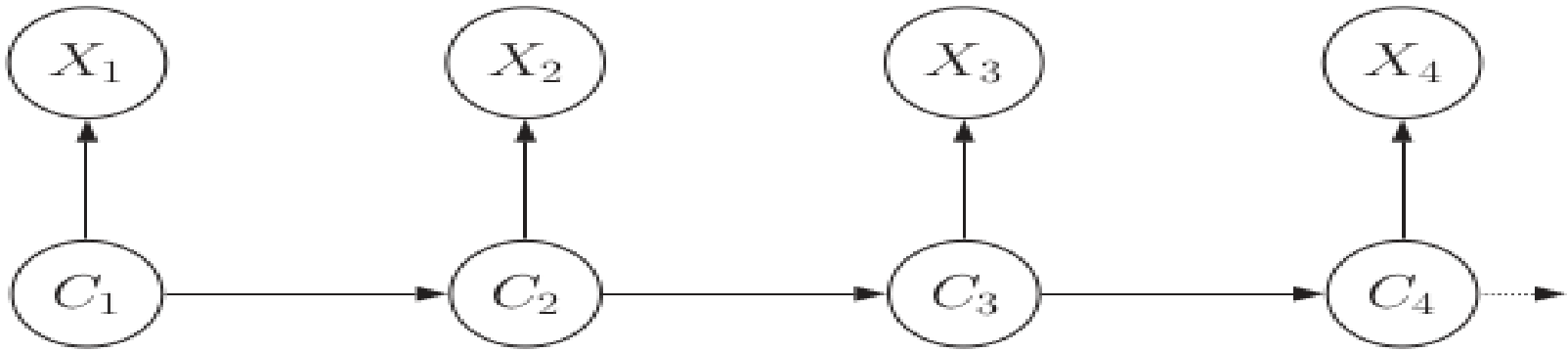
Hidden Markov Models

Sweta Mahajan

Under the supervision of Dr. Anirvan Chakraborty
Indian Institute of Science Education and Research,Kolkata

November 27, 2019

Notations



States take values in $\{s_1, s_2, \cdots, s_m\}$ and observations take values in $\{o_1, o_2, \cdots, o_n\}$.

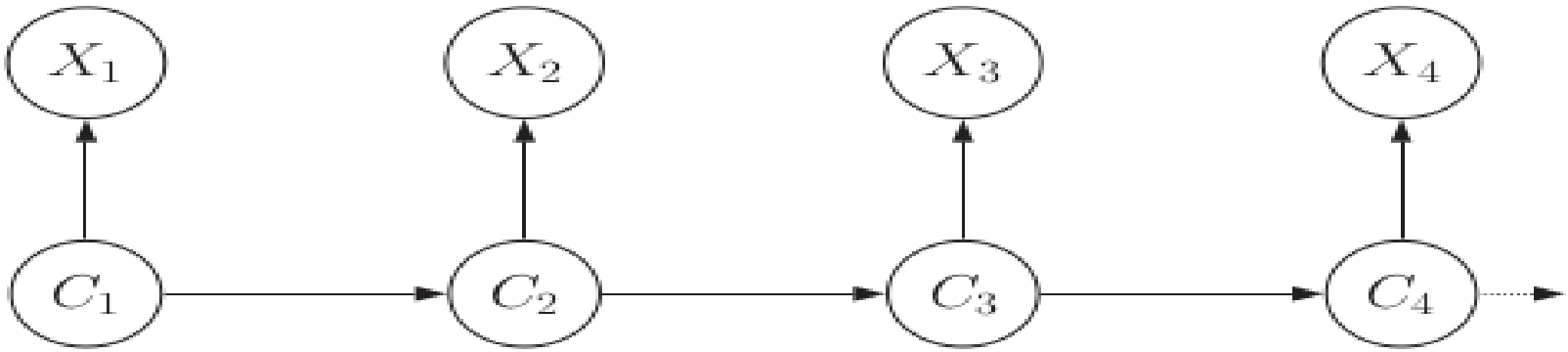
- $\gamma_{ij} = P(c_{t+1} = j | c_t = i)$
- $p_j(k) = P(x_t = k | c_t = j)$
- $\delta_i = P(c_1 = i)$
- $\delta_i(t) = P(c_t = i)$
- $c_1^t = (c_1, c_2, \cdots, c_t)$

Apply HMM to gene finding

AIM: Finding coding region(CR) and non-coding region(NCR) in the DNA.
DNA has four nucleotides A,T,G,C.

- $P(A|CR) = 0.05, P(T|CR) = 0.05, P(G|CR) = 0.5, P(C|CR) = 0.4$
- $P(A|NCR) = 0.45, P(T|NCR) = 0.35, P(G|NCR) = 0.15, P(C|NCR) = 0.05$
- $P(CR|NCR) = 0.2, P(NCR|NCR) = 0.8, P(CR|CR) = 0.6, P(NCR|CR) = 0.4$
- $P(A) = 0.2, P(T) = 0.35, P(G) = 0.25, P(C) = 0.2$

Assumptions



$$P(c_t|c^{(t-1)}) = P(c_t|c_{t-1})$$
$$P(x_t|x^{(t-1)}, c^{(t)}) = P(x_t|c_t)$$

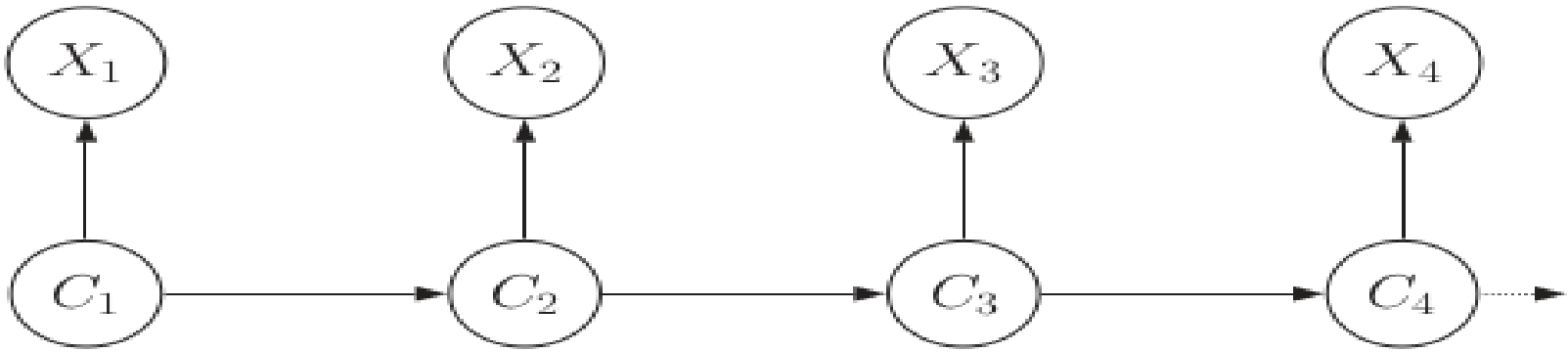
Note: Conditional independence

$$P(x_1^T|c_t) = P(x_1^t|c_t)P(x_t^T|c_t)$$

Bivariate Distributions

$$
\begin{aligned}
&P(x_t, x_{t+k}) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} P(x_t, x_{t+k}, c_t = i, c_{t+k} = j) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} P(x_t, x_{t+k} | c_t = i, c_{t+k} = j) P(c_{t+k}, c_t) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} P(x_t | c_t = i, c_{t+k} = j) P(x_{t+k} | c_t = i, c_{t+k} = j) P(c_{t+k}, c_t) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} P(c_t) P(x_t | c_t) P(c_{t+k} | c_t) P(x_{t+k} | c_{t+k}) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} \delta_i(t) p_i(x_t) \gamma_{ij} p_j(x_{t+k}) \\
&= \boldsymbol{\delta}(t) \mathbf{P}(x_t) \boldsymbol{\Gamma}^k \mathbf{P}(x_{t+k}) \mathbf{1}'
\end{aligned}
$$

Graphical model of HMM



HMM can be converted in to a directed graphical model and joint distribution is given by:

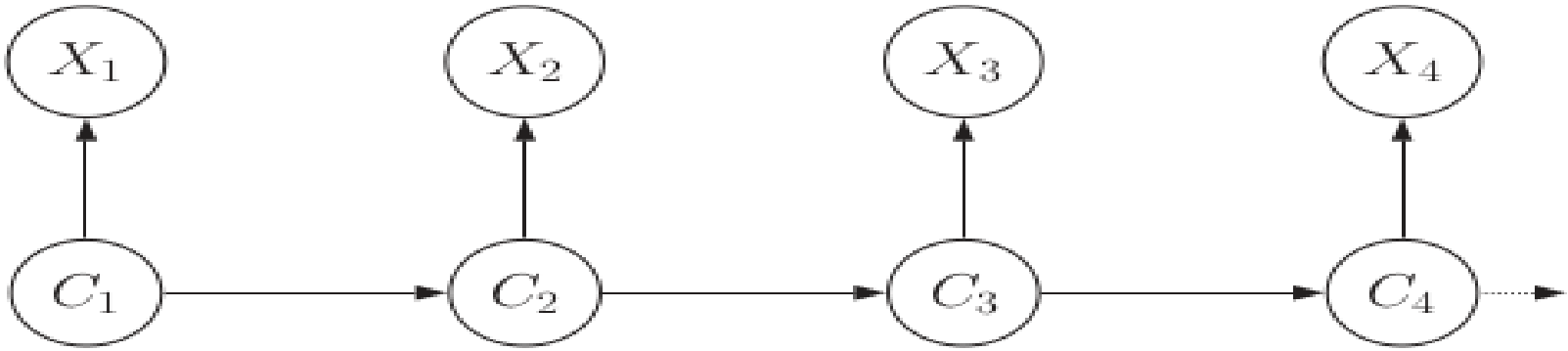
$$P(V_1, V_2, \cdots, V_n) = \prod_{i=1}^{m} P(V_i | Pa(V_i)),$$

where $Pa(V_i)$ denotes the set of all parents of V_i in the set $V_1, V_2, \cdots, V_n$.

Problems

- Efficiently evaluating the probability of observation sequence given the model.
- Determination of sequence of states which best explains the given observation sequence.
- Adjustment of model parameters so as to best account for the observed sequence.

Finding likelihood



$$
\begin{aligned}
P(X_1^T) &= \sum_{c_1,\cdots,c_T} P(X_1^T, C_1^T) \\
&= \sum_{c_1,\cdots,c_T} P(c_1)P(x_1|c_1)p(c_2|c_1)\cdots p(c_T|c_{T-1})P(x_T|c_T) \\
&= \sum_{c_1,\cdots,c_T} \delta_{c_1} p_{c_1}(x_1)\gamma_{c_1c_2} p_{c_2}(x_2)\gamma_{c_2c_3}\cdots\gamma_{c_{T-1}c_T} p_{c_T}(x_T) \\
&= \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'
\end{aligned}
$$

number of computations $\propto \mathrm{m}^T * T$

Forward Algorithm

So, to help combat the previous problem, we introduce here, the forward algorithm.Let us define,

- Definition: $\alpha_t(i) = P(x_1^t, c_t = i)$
- Initialization:
  $\alpha_1(i) = \delta_i p_i(x_1) \qquad 1 \leq i \leq m$
- Induction step:
  $$\alpha_{t+1}(j) = \left[\sum_{i=1}^{m} \alpha_t(i)\gamma_{ij}\right] p_j(x_{t+1})$$
  $1 \leq j \leq m, 1 \leq t \leq T-1$
- Final step:
  $P(X_1^T) = \sum_{i=1}^{m} \alpha_T(i) = \sum_{i=1}^{m} P(x_1^T, c_T = i)$

number of computations $\propto$m*T^2

Proof of induction step

$$
\begin{aligned}
&\alpha_{t+1}(j) \\
&= P(X_1^{t+1}, c_{t+1} = j) \\
&= \sum_{i=1}^{m} P(X_1^t, x_{t+1}, c_t = i, c_{t+1} = j) \\
&= \sum_{i=1}^{m} P(X_1^t | x_{t+1}, c_t = i, c_{t+1} = j) P(x_{t+1}, c_t = i, c_{t+1} = j) \\
&= \sum_{i=1}^{m} P(X_1^t | c_t = i) P(x_{t+1} | c_t = i, c_{t+1} = j) P(c_t = i, c_{t+1} = j) \\
&= \sum_{i=1}^{m} P(X_1^t | c_t = i) P(x_{t+1} | c_{t+1} = j) P(c_{t+1} = j | c_t = i) P(c_t = i) \\
&= \sum_{i=1}^{m} \alpha_t(i) \gamma_{ij} p_j(x_{t+1})
\end{aligned}
$$

Backward probability

In a similar fashion, we define the backward probability.

- Definition: $\beta_t(i) = P(x_{(t+1)}^T | c_t = i)$
- Initialization: $\beta_T(i) = 1 \qquad 1 \leq i \leq m$
- Induction step:
  $\beta_t(i) = \left[\sum_{j=1}^{m} \beta_{t+1}(j) \gamma_{ij} p_j(x_{t+1}) \right] \qquad 1 \leq i \leq m,$
  $1 \leq t \leq T-1$

Proof of induction step

$$
\begin{aligned}
\beta_t(i) &= P(X_{t+1}^T | c_t = i) \\
&= P(X_{t+1}^T, c_t = i) / P(c_t = i) \\
&= \sum_{j=1}^{m} P(X_{t+1}^T, c_{t+1} = j, c_t = i) / P(c_t = i) \\
&= \sum_{j=1}^{m} P(X_{t+1}^T | c_t = i, c_{t+1} = j) P(c_{t+1} = j | c_t = i) \\
&= \sum_{j=1}^{m} P(x_{t+1}, x_{t+2}^T | c_{t+1} = j) \gamma_{ij} \\
&= \sum_{j=1}^{m} P(x_{t+2}^T | c_{t+1} = j) P(x_{t+1} | c_{t+1} = j) \gamma_{ij} \\
&= \sum_{j=1}^{m} \beta_{t+1}(j) \gamma_{ij} p_j(x_{t+1})
\end{aligned}
$$

Selecting optimal individual state

$$\begin{aligned} P(c_t = i|X_1^T) &= P(c_t = i, X_1^T)/P(X_1^T) \\ &= \alpha_i(t)\beta_i(t)/L_T \end{aligned}$$

$$\begin{aligned} P(x_1^T, c_t = i) &= P(x_1^T|c_t)P(c_t) \\ &= P(x_1^t|c_t)P(x_{t+1}^T|c_t)P(c_t) \\ &= P(x_1^t, c_t)P(x_{t+1}^T|c_t) \\ &= \alpha_t(i)\beta_t(i) \end{aligned}$$

$$c_t = \underset{1 \le i \le m}{\operatorname{argmax}} P(c_t = i|X_1^T, \lambda) \qquad t = 1, \cdots, T$$

Finding best state sequence

- Initialization:

$$\pi_1(i) = \delta_i p_i(x_1) \qquad 1 \le i \le m$$
$$\psi_i = 0$$

- Induction:

$$\pi_t(j) = \left[\max_{1 \le i \le m} \pi_{t-1}(i)\gamma_{ij} \right] p_j(x_t) \qquad 1 \le i \le m, 2 \le t \le T$$
$$\psi_t(j) = \left[\operatorname*{argmax}_{1 \le i \le m} \pi_{t-1}(i)\gamma_{ij} \right] \qquad 1 \le i \le m, 2 \le t \le T$$

- Termination:

$$p = \max_{1 \le i \le m} \left[\pi_T(i) \right]$$
$$c_T = \operatorname*{argmax}_{1 \le i \le m} \left[\pi_T(i) \right]$$

- state backtracking:

$$q_t = \psi_{t+1}(c_{t+1}) \qquad 1 \le t \le T - 1$$

Baum-Welch Algorithm(EM algorithm for HMM)

$$u_j(t) = 1 \quad \text{iff} \quad c_t = j, \quad t = 1, 2, \cdots, T$$
$$v_{jk}(t) = 1 \quad \text{iff} \quad c_{t-1} = j \quad \text{and} \quad c_t = k, \quad t = 1, 2, \cdots, T$$

$$\begin{aligned}
&\log(P(X_1^T, C_1^T)) \\
&= \log(\delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1}, c_t} \prod_{t=1}^{T} p_{c_t}(x_t)) \\
&= \log(\delta_{c_1}) + \sum_{t=2}^{T} \log(\gamma_{c_{t-1}, c_t}) + \sum_{t=1}^{T} \log(p_{c_t}(x_t)) \\
&= \sum_{j=1}^{m} u_j(1) \log(\delta_j) + \sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{t=2}^{T} v_{jk}(t) \log(\gamma_{jk}) \right) \\
&+ \sum_{j=1}^{m} \sum_{t=1}^{T} u_j(t) \log(p_j(x_t))
\end{aligned}$$

Expectation step of EM algorithm

1. E step: Replace the functions of missing data by their conditional expectation given the observation and the current parameter estimate.

$$\widehat{u}_j(t) = E(c_t|X_1^T) = P(c_t|X_1^T) = \alpha(j)\beta(j)/L_T$$

and

$$\begin{aligned}\widehat{v}_{jk}(t) &= E(c_{t-1} = j, c_t = k|X_1^T) = P(c_{t-1} = j, c_t = k|X_1^T) \\ &= \alpha_{t-1}(j)\gamma_{jk}p_k(x_t)\beta_t(k)/L_T\end{aligned}$$

Proof:

$$\begin{aligned}&= P(c_{t-1} = j, c_t = k, x^{(T)})/P(x^{(T)}) \\ &= P(x^{(t-1)}, x_t^T|c_{t-1} = j, c_t = k)P(c_t = k|c_{t-1} = j)P(c_{t-1} = j)/L_T \\ &= P(x^{(t-1)}|c_{t-1} = j)P(x_t^T|c_t = k)\gamma_{jk}P(x_t^T|c_t = k)/L_T \\ &= \alpha_{t-1}(j)\gamma_{jk}p_k(x_t)\beta_t(k)/L_T\end{aligned}$$

Maximization step of EM algorithm

M step: Having replaced the $u_j(t)$ and $v_{jk}(t)$ by $\widehat{u}_j(t)$ and $\widehat{v}_{jk}(t)$, maximize the CDLL w.r.t the transition probabilities and initial distribution and the model parameters.

Analysing first summand

1. $\sum_{j=1}^{m} \hat{u}_j(1) log(\delta_j)$ w.r.t to $\boldsymbol{\delta}$.

$$\nabla\left(\sum_{j=1}^{m} \hat{u}_j(1) log(\delta_j)\right) = \lambda \nabla\left(\sum_{j=1}^{m} \delta_j\right)$$

$$\Rightarrow \hat{u}_i(1)/\delta_i = \lambda \quad i = 1, \cdots, m \quad \text{(Differentiating w.r.t } \delta_i\text{)}$$

$$\Rightarrow \lambda_i = \hat{u}_i(1)/\lambda$$

$$\Rightarrow \sum_{i=1}^{m} \delta_i = \sum_{i=1}^{m} \hat{u}_i(1)/\lambda$$

$$\Rightarrow \lambda = \sum_{i=1}^{m} \hat{u}_i(1)/\lambda$$

$$\Rightarrow \delta_i = \frac{\hat{u}_i(1)}{\sum_{i=1}^{m} \hat{u}_i(1)} \quad \text{(From 3.12)}$$

Analysing second and third summand

$\sum_{j=1}^{m}\sum_{j=1}^{m}\left(\sum_{t=2}^{T}\widehat{v}_{jk}(t)\right) log(\gamma_{jk})$ w.r.t Γ. using the lagrange multiplier we get,

$$\gamma_{jk} = \frac{f_{jk}}{\sum_{k=1}^{n} f_{jk}}, \quad \text{where} f_{jk} = \sum_{t=2}^{T}\widehat{v}_{jk}(t)$$

The maximization of the third term may be easy or difficult depending on the model assumed.

Model Checking with Pseudo Residuals

- We need to assess the general goodness of fit and
- identify outliers relative to the model

Continuous case

We know that if X is a random variable with continuous distribution function F, then $F(X) \sim U(0,1)$.
Uniform pseudo residual: The uniform pseudo residual of an observation x_t from a continuous random variable X_t is defined as

$$u_t = P(X_t \leq x_t) = F_{X_t}(x_t),$$

where the probability is found under the fitted model.
Normal Pseudo Residuals are defned as,

$$z_t = \Phi^{-1}(u_t)$$

Discrete Case

The theory of pseudo-residuals that is described in the previous section holds for continuous distributions only.

- We have to modify the previous method to allow for the discreteness.

Here, the pseudo residuals are defined as intervals. We define uniform pseudo-residual segments as,

$$[u_t^-; u_t^+] = [F_{X_t}(x_t^-); F_{X_t}(x_t)]$$

where x_t^- is the greatest realization possible that is strictly less than x_t.
And we define the normal pseudo residual segments as

$$[z_t^-; z_t^+] = [\Phi^{-1}(u_t^-); \Phi^{-1}(u_t^+)]$$

Mid pseudo residuals

This is only correct if the parameters of the fitted model are known, it still works well enough if the number of parameters is very small in comparison to the sample size. Since, we do not have qq plot for segments, we need to specify a representative element of the segments. So, we define "mid-pseudo residuals" for that purpose as

$$z_t^m = \Phi^{-1}\left(\frac{u_t^- + u_t+}{2}\right)$$

Pseudo residuals in the context of HMM

Now, having known about pseudo residuals in general. Let us see what purpose they serve in the context of HMMs:

- Ordinary Pseudo Residuals: Those residuals that are based on the conditional distribution given all other observations.
- Forecast Pseudo Residuals:Those residuals that are based on the conditional distribution given all the preceding observations.

Ordinary Pseudo Residuals

This technique points out the observations which are sufficiently extreme to suggest that they differ in nature from the rest of the data. For continuous observations, we have

$$z_t = \Phi^{-1}(P(X_t \leq x_t | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}))$$

For discrete observations we have,

$$z_t^- = \Phi^{-1}(P(X_t < x_t | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}))$$
$$z_t^+ = \Phi^{-1}(P(X_t \leq x_t | \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}))$$

Forecast Pseudo Residuals

These type of residuals measure the deviation of an observation from the median of the corresponding one step ahead forecast. If the forecast pseudo residual is extreme, then we conclude that the observation is an outlier or the model no longer describes the sequence well enough.
For continuous observations, we have

$$z_t = \Phi^{-1}(P(X_t \leq x_t | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}))$$

For discrete observations we have,

$$z_t^- = \Phi^{-1}(P(X_t < x_t | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}))$$
$$z_t^+ = \Phi^{-1}(P(X_t \leq x_t | \mathbf{X}^{(t-1)} = \mathbf{x}^{(t-1)}))$$

And.....01010100 01101000 01100001 01101110 01101011 00100000 01111001 01101111 01110101 00100001

Thank You!