Part 3

Markov Chain Modeling
Markov Chain Model

- Stochastic model
- Amounts to sequence of random variables
  \[ X_1, X_2, \ldots, X_t \]
- Transitions between states
- State space
  \[ S = \{ s_1, s_2, \ldots, s_m \} \]
Markov Chain Model

- Stochastic model
- Amounts to sequence of random variables $X_1, X_2, ..., X_t$
- Transitions between states
- State space $S = \{s_1, s_2, ..., s_m\}$
Markovian property

• Next state in a sequence only depends on the current one

• Does not depend on a sequence of preceding ones

\[ P(X_{t+1} = s_j | X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) = \]

\[ P(X_{t+1} = s_j | X_t = s_{i_t}) = p_{i,j} \]
Transition matrix

$$\begin{pmatrix}
p_{1,1} & p_{1,2} & \cdots & p_{1,j} \\
p_{2,1} & p_{2,2} & \cdots & p_{2,j} \\
\vdots & \vdots & \ddots & \vdots \\
p_{i,1} & p_{i,2} & \cdots & p_{i,j}
\end{pmatrix}$$

Rows sum to 1

$$\sum_j p_{ij} = 1$$

Single transition probability

$$p_{i,j} = p(s_j | s_i)$$
Likelihood

- Transition probabilities are parameters

\[ D = x_1, x_2, x_3, \ldots, x_n \]

\[
P(D|\theta) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2})\ldots p(x_2|x_1)p(x_1)
= p(x_1) \prod_i \prod_j p_{i,j}^{n_{i,j}}
\]

Sequence data
MC parameters
Transition count
Transition probability
Maximum Likelihood Estimation (MLE)

- Given some sequence data, how can we determine parameters?
- MLE estimation: count and normalize transitions

\[
\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta)) = \log \left( p(x_1) \prod_i \prod_j p_{i,j}^{n_{i,j}} \right) = \log p(x_1) + \sum_i \sum_j n_{i,j} \log(p_{i,j})
\]

\[
p_{i,j} = \frac{n_{i,j}}{\sum_j n_{i,j}}
\]

See ref [1]

[Singer et al. 2014]
Example

Training sequence depends on
Example

Transition counts

Transition matrix (MLE)

\[
\begin{array}{cc}
\text{Transition counts} & \text{Transition matrix (MLE)} \\
\begin{array}{cc}
5 & 2 \\
2 & 1 \\
\end{array} & \begin{array}{cc}
5/7 & 2/7 \\
2/3 & 1/3 \\
\end{array}
\end{array}
\]
Example

Transition matrix (MLE)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>Blue</td>
<td></td>
</tr>
<tr>
<td>5/7</td>
<td>2/7</td>
<td></td>
</tr>
<tr>
<td>2/3</td>
<td>1/3</td>
<td></td>
</tr>
</tbody>
</table>

Likelihood of given sequence

\[
(\frac{5}{7})^5 \times (\frac{2}{7})^2 \times (\frac{2}{3})^2 \times (\frac{1}{3})^1 = 0.002248
\]

\[
5 \times \ln(\frac{5}{7}) + 2 \times \ln(\frac{2}{7}) + 2 \times \ln(\frac{2}{3}) + 1 \times \ln(\frac{1}{3}) = -6.0974
\]

We calculate the probability of the sequence with the assumption that we start with the yellow state.
Reset state

- Modeling start and end of sequences
- Specifically useful if many individual sequences

[Chierichetti et al. WWW 2012]
Properties

● **Reducibility**
  - State j is accessible from state i if it can be reached with non-zero probability
  - Irreducible: All states can be reached from any state (possibly multiple steps)

● **Periodicity**
  - State i has period k if any return to the state is in multiples of k
  - If k=1 then it is said to be aperiodic

● **Transcience**
  - State i is transient if there is non-zero probability that we will never return to the state
  - State is recurrent if it is not transient

● **Ergodicity**
  - State i is ergodic if it is aperiodic and positive recurrent

● **Steady state**
  - Stationary distribution over states
  - Irreducible and all states positive recurrent → one solution
  - Reverting a steady-state [Kumar et al. 2015]
Higher Order Markov Chain Models

- Drop the memoryless assumption?
- Models of increasing order
  - 2\textsuperscript{nd} order MC model
  - 3\textsuperscript{rd} order MC model
  - ...

\[ P(X_{t+1} = s_j | X_1 = s_{i_1}, \ldots, X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) = P(X_{t+1} = s_j | X_t = s_{i_t}, X_{t-1} = s_{i_{t-1}}, \ldots, X_{t-k+1} = s_{i_{t-k+1}}) \]
Higher Order Markov Chain Models

- Drop the memoryless assumption?
- Models of increasing order
  - 2\textsuperscript{nd} order MC model
  - 3\textsuperscript{rd} order MC model
  - ...

\[
P(X_{t+1} = s_j|X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) = P(X_{t+1} = s_j|X_t = s_{i_t}, X_{t-1} = s_{i_{t-1}}, ..., X_{t-k+1} = s_{i_{t-k+1}})
\]
Higher order to first order transformation

- Transform state space
- 2nd order example – new compound states
Higher order to first order transformation

- Transform state space
- 2\textsuperscript{nd} order example – new compound states
- Prepend (nr. of order) and append (one) reset states
Example
Example

1st order parameters
Example

1st order parameters
Example

1\textsuperscript{st} order parameters:

- 5/8
- 2/8
- 1/8
- 2/3
- 1/3
- 0/3
- 1/1
- 0/1
- 0/1

2\textsuperscript{nd} order parameters:

- 3/5
- 1/5
- 1/5
- 1/2
- 1/2
- 0
- 0
- 1/1
- 0
- 0

R
Example

\[ \ln(P(D|\theta_1)) = -9.11 \]

1\textsuperscript{st} order parameters:
- 5/8
- 2/3
- 1/1

2\textsuperscript{nd} order parameters:
- 3/5
- 1/2
- 1/1
- 1/1

\[ \ln(P(D|\theta_2)) = -7.52 \]
Example

\[ \ln(P(D|\theta_1)) = -9.11 \]

\[
\begin{array}{ccc}
5/8 & 2/8 & 1/8 \\
2/3 & 1/3 & 0/3 \\
1/1 & 0/1 & 0/1 \\
\end{array}
\]

18 free parameters

\[ \ln(P(D|\theta_2)) = -7.52 \]

6 free parameters
Model Selection

• Which is the “best” model?
• 1\textsuperscript{st} vs. 2\textsuperscript{nd} order model
• Nested models $\rightarrow$ higher order always fits better
• Statistical model comparison
• Balance goodness of fit with complexity
Model Selection Criteria

- **Likelihood ratio test**
  - Ratio between likelihoods for order m and k: \( k \eta_m = -2(\mathcal{L}(P(D|\theta_k)) - \mathcal{L}(P(D|\theta_m))) \)
  - Follows chi2 distribution with dof: \( (|S|^m - |S|^k)(|S| - 1) \)
  - Nested models only

- **Akaike Information Criterion (AIC)**
  \[
  AIC(k) = 2 \ast (|S|^k)(|S| - 1) - 2(\mathcal{L}(P(D|\theta_k)))
  \]

- **Bayesian Information Criterion (BIC)**
  \[
  BIC(k) = (|S|^k)(|S| - 1) \ast ln(n) - 2(\mathcal{L}(P(D|\theta_k)))
  \]

- **Bayes factors**

- **Cross Validation**

[Singer et al. 2014], [Strelioff et al. 2007], [Anderson & Goodman 1957]
Bayesian Inference

- Probabilistic statements of parameters
- Prior belief updated with observed data

\[ P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)} \]

posteriors

likelihood

prior

marginal likelihood
Bayesian Model Selection

• Probability theory for choosing between models
• Posterior probability of model $M$ given data $D$

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

$$P(\theta|D, M) = \frac{\text{posterior}}{\text{marginal likelihood}} = \frac{\text{likelihood} \cdot \text{prior}}{P(D|M)}$$
Bayes Factor

- Comparing two models

\[
P(M|D) = \frac{P(D|M)P(M)}{P(D)}
\]

\[
\frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(\theta_1|M_1)P(D|\theta_1, M_1)d\theta}{\int P(\theta_2|M_2)P(D|\theta_2, M_2)d\theta}
\]

- Evidence: Parameters marginalized out

- Automatic penalty for model complexity

- Occam's razor

- Strength of Bayes factor: Interpretation table

[Kass & Raftery 1995]
Example

$P(D|M_1) = -13.43$

$P(D|M_2) = -14.31$
Hands-on jupyter notebook
Methodological extensions/adaptations

- Variable-order Markov chain models
  - Example: AAABCAAABC
  - Order dependent on context/realization
  - Often huge reduction of parameter space
  - [Rissanen 1983, Bühlmann & Wyner 1999, Chierichetti et al. WWW 2012]


- Markov Random Field [Li 2009]

- MCMC [Gilks 2005]
Some applications

- Sequence of letters [Markov 1912, Hayes 2013]
- Weather data [Gabriel & Neumann 1962]
- Computer performance evaluation [Scherr 1967]
- Speech recognition [Rabiner 1989]
- Gene, DNA sequences [Salzberg et al. 1998]
- Web navigation, PageRank [Page et al. 1999]
What have we learned?

• Markov chain models

• Higher-order Markov chain models

• Model selection techniques: Bayes factors
Questions?
References 1/2


