Hidden Markov Models

based on chapters from the book
Durbin, Eddy, Krogh and Mitchison
Biological Sequence Analysis
via Shamir’s lecture notes
music recognition

deal with variations in
- pitch
- timing
- timbre
- …
• Actual Value versus Forecasted Value for Tata Steel in Rupees over the period 5-9 2009 – 23-9 2011.
• Variations of value over time.
application: gene finding

deal with variations in
- actual sound → actual base (match/substitutions)
- timing → insertions/deletions
Basic Questions

Given:
• A sequence of “observations”
• A probabilistic model of our “domain”

Questions:
• Does the given sequence belong to a certain family?
  – Markov chains
  – Hidden Markov Models (HMMs)

• Can we say something about the internal structure of the sequence? (indirect observations)
  – Hidden Markov Models (HMMs)
**Introduction Markov Chain Model**

**Characteristics**
- Discrete time
- Discrete space
- No state History
  - Present state only
- States and transitions

**Notations:**
- $P(X)$: probability for event $X$
- $P(X,Y)$: event $X$ and event $Y$
- $P(X|Y)$: event $X$ given event $Y$
A Markov chain\textsuperscript{[1]} model is defined by

- a set of states
  - some states emit symbols
  - other states (e.g., the begin state) are silent
- a set of transitions with associated probabilities
  - the transitions emanating from a given state define a distribution over the possible next states (\textit{i.e.}, all positive, and sum equals 1)

\textsuperscript{[1]} Марков А. А., Распространение закона больших чисел на величины, зависящие друг от друга. — Известия физико-математического общества при Казанском университете. — 2-я серия. — Том 15. (1906) — С. 135—156
Markov Model $M = (Q, P, T)$, with

- $Q$ the set of states
- $P$ the set of initial probabilities $p_x$ for each state $x$ in $Q$
- $T = (t_{xy})$ the transition probabilities matrix/graph, with $t_{xy}$ the probability of the transition from state $x$ to state $y$.

This is a first order Markov Model:
no history is modeled

An observation $X$ is a sequence of states:
$X = x_1 x_2 \ldots x_n$

The probability of an observation $X$ given the model $M$ is equal to:

$$P(X|M) = p_{x_1} t_{x_1 x_2} t_{x_2 x_3} \ldots t_{x_{n-1} x_n} = p_{x_1} \cdot \prod_{i=2}^{n} t_{x_{i-1} x_i}$$
A Markov Chain Model Example

Transition probabilities

- \( \Pr(x_i=a|x_{i-1}=g)=0.16 \)
- \( \Pr(x_i=c|x_{i-1}=g)=0.34 \)
- \( \Pr(x_i=g|x_{i-1}=g)=0.38 \)
- \( \Pr(x_i=t|x_{i-1}=g)=0.12 \)

\[ \sum \Pr(x_i \mid x_{i-1} = g) = 1 \]

over all neighbors \( x_i \)
The Probability of a Sequence for a Markov Chain Model

\[ \Pr(CGGT) = \Pr(C) \Pr(G|C) \Pr(G|G) \Pr(T|G) \]
Markov Chains: Another Example

\[ Q = \{ A, B, C \} \]

\[ P = (1, 0, 0) \]

unique starting state A

\[ T = \begin{pmatrix}
A & B & C \\
0.7 & 0.3 & 0 \\
0.4 & 0.2 & 0.8 \\
0.5 & 0 & 0.6
\end{pmatrix} \]

\[ P( AABBCC | M_1 ) = 1 \cdot 0.7 \cdot 0.3 \cdot 0.2 \cdot 0.8 \cdot 0.6 \cdot 10^{-6} = 1.2 \times 10^{-2} \]

\[ P( AABBCC | M_2 ) = 1 \cdot 0.6 \cdot 0.4 \cdot 0.6 \cdot 0.6 \cdot 10^{-6} = 1.1 \times 10^{-2} \]
• Given some sequence $x$ of length $L$, we can ask how probable the sequence is given our model $M$

• For any probabilistic model of sequences, we can write this probability as

$$\Pr(x) = \Pr(x_Lx_{L-1}...x_1)$$

$$= \Pr(x_L \mid x_{L-1}...x_1)\Pr(x_{L-1} \mid x_{L-2}...x_1)...\Pr(x_1)$$

• Key property of a (1st order) Markov chain: the probability of each $x_i$ depends only on the value of $x_{i-1}$

$$\Pr(x) = \Pr(x_L \mid x_{L-1})\Pr(x_{L-1} \mid x_{L-2})...\Pr(x_2 \mid x_1)\Pr(x_1)$$

$$= \Pr(x_1)\prod_{i=2}^{L}\Pr(x_i \mid x_{i-1})$$
Markov Model: Underflow Problem

- initial state $x_0$ fixed
  - $\sim$ initial probabilities
- final state [not in this picture]

$X = x_1 x_2 \ldots x_n$

$P(X|M) = \prod_{i=1}^{n} t_{x_{i-1}x_i}$

small values: underflow

$\log P(X|M) = \sum_{i=1}^{n} \log t_{x_{i-1}x_i}$
Markov Model: Comparing Models

Given:
\[ X = x_1 x_2 \ldots x_n \]
\[ P(X|M) = \prod_{i=1}^{n} t^{x_{i-1} x_i} \]

Question: X best explained by which model?

We can calculate: \[ P(X | M_1) \text{ vs. } P(X | M_2) \]

We want to know: \[ P(M_1 | X) \text{ vs. } P(M_2 | X) \]

Bayes Rule: \[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]

\[
\frac{P(M_1|X)}{P(M_2|X)} = \frac{P(X|M_1) \cdot P(M_1)}{P(X|M_2) \cdot P(M_2)}
\]
bases are not random
• There are many cases in which we would like to represent the statistical regularities of some class of sequences
  – genes
  – various regulatory sites in DNA (e.g., where RNA polymerase and transcription factors bind)
  – proteins in a given family
• Markov models are well suited to this type of task
• **CpG islands**
  – CG di-nucleotides are *rarer* in eukaryotic genomes than expected given the marginal probabilities of C and G
  – but the regions upstream of genes *(reading is from 5’ to 3’)* are *richer* in CG di-nucleotides than elsewhere – so called CpG islands
  – useful evidence for finding genes

• **Application: Predict CpG islands with Markov chains**
  – a Markov chain to represent CpG islands
  – a Markov chain to represent the rest of the genome
Markov Chains for Discrimination

• Suppose we want to distinguish CpG islands from other sequence regions

• Given sequences from CpG islands, and sequences from other regions, we can construct
  – a model to represent CpG islands
  – a null model to represent the other regions

• We can then score a test sequence $X$ by:

$$score(X) = \log \frac{Pr(X \mid CpGModel)}{Pr(X \mid nullModel)}$$
Markov Chains for Discrimination

• Why can we use

\[ \text{score}(X) = \log \frac{\Pr(X \mid \text{CpG Model})}{\Pr(X \mid \text{null Model})} \]

• Again, according to Bayes’ rule:

\[
\Pr(\text{CpG} \mid X) = \frac{\Pr(X \mid \text{CpG}) \Pr(\text{CpG})}{\Pr(X)}
\]

\[
\Pr(\text{null} \mid X) = \frac{\Pr(X \mid \text{null}) \Pr(\text{null})}{\Pr(X)}
\]

• If we are not taking into account prior probabilities (\(\Pr(\text{CpG})\) and \(\Pr(\text{null})\)) of the two classes, then from Bayes’ rule it is clear that we just need to compare \(\Pr(X\mid\text{CpG})\) and \(\Pr(X\mid\text{null})\) as is done in our scoring function \text{score}().
### Markov Chain Application: CpG islands

#### observed frequencies

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>island</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>non island</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
</tbody>
</table>

#### diagrams

- **island**
  - A → C
  - C → G
  - G → T
  - T → A
- **non island**
  - A → C
  - C → G
  - G → T
  - T → A

Consecutive CG pairs (CG → CG) are mostly **rare**, although ‘islands’ occur in signal (e.g.) promoter regions.
basic questions

- observation: DNA sequence
- model 1: CpG islands
- model 2: non-islands

• does this sequence belong to a certain family?
  Markov chains
  is this a CpG island (or not)?

• can we say something about the internal structure?
  Markov Chains: windowing
  where are the CpG islands?
**application: CpG islands**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>island</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>non island</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>island</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td>non island</td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>island</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>non island</td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>island</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
<tr>
<td>non island</td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>

**score**

\[
\frac{P(X|\text{ island})}{P(X|\text{ non})} = \frac{\prod_{i=1}^{n} t_{x_{i-1}x_i}}{\prod_{i=1}^{n} t_{x_{i-1}x_i}}
\]

\[X = \text{ACGT} \quad A\rightarrow C \quad C\rightarrow G \quad G\rightarrow T\]

\[
\frac{0.274 \cdot 0.274 \cdot 0.125}{0.205 \cdot 0.078 \cdot 0.208} = 2.82
\]
application: CpG islands

LLR = Log-Likelihood Ratio

$$\log \left( \frac{t_{xy}^+}{t_{xy}^-} \right)$$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.74</td>
<td>0.42</td>
<td>0.58</td>
<td>-0.80</td>
</tr>
<tr>
<td>C</td>
<td>-0.91</td>
<td>0.30</td>
<td>1.81</td>
<td>-0.69</td>
</tr>
<tr>
<td>G</td>
<td>-0.62</td>
<td>0.46</td>
<td>0.33</td>
<td>-0.73</td>
</tr>
<tr>
<td>T</td>
<td>-1.17</td>
<td>0.57</td>
<td>0.39</td>
<td>-0.68</td>
</tr>
</tbody>
</table>

‘bits’ (log\(_2\))

log-score (log\(_2\))

$$\log \frac{P(X| \text{island})}{P(X| \text{non})} = \log \frac{\prod_{i=1}^{n} t_{x_{i-1}x_i}^+}{\prod_{i=1}^{n} t_{x_{i-1}x_i}^-} = \sum_{i=1}^{n} \log \left( \frac{t_{x_{i-1}x_i}^+}{t_{x_{i-1}x_i}^-} \right)$$

X = ACGT

$$\log_2 \frac{0.274 \cdot 0.274 \cdot 0.125}{0.205 \cdot 0.078 \cdot 0.208} = 0.42 + 1.81 - 0.73 = 1.50$$
CpG Log-Likelihood Ratio

\[ \log \left( \frac{t_{xy}^+}{t_{xy}^-} \right) \]

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.74</td>
<td>0.42</td>
<td>0.58</td>
<td>-0.80</td>
</tr>
<tr>
<td>C</td>
<td>-0.91</td>
<td>0.30</td>
<td>1.81</td>
<td>-0.69</td>
</tr>
<tr>
<td>G</td>
<td>-0.62</td>
<td>0.46</td>
<td>0.33</td>
<td>-0.73</td>
</tr>
<tr>
<td>T</td>
<td>-1.17</td>
<td>0.57</td>
<td>0.39</td>
<td>-0.68</td>
</tr>
</tbody>
</table>

\[ \text{LLR(ACGT)} = 0.42 + 1.81 - 0.73 = 1.50 \]

(0.37 ‘bits’ per base)

\[ 1.5/4 = 0.375 \]

- is a (short) sequence a CpG island?
- compare with observed data (normalized for length)
- where (in long sequence) are CpG islands?
  - first approach: sliding window

- What would be the length of window?
empirical data

• is a (short) sequence a CpG island?
  compare with observed data (normalized for length)

Figure 3.2 The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.
where (in long sequence) are CpG islands?  
first approach: *sliding window*

Detection of regions of genomic sequences that are rich in the CpG pattern is important because such regions are resistant to methylation and tend to be associated with genes which are frequently switched on. Regions rich in the CpG pattern are known as CpG islands. The function of the program **CpGplot** is to plot CpG rich areas, and **cpgreport** to report all CpG rich regions.

The nuclear genomes of vertebrates are mosaics of isochores, very long stretches of DNA that are homogeneous in base composition and are compositionally correlated with the coding sequences that they embed. Isochores can be partitioned in a small number of families that cover a range of GC levels. Program **isochore** plots GC content over a sequence.

**Program**

<table>
<thead>
<tr>
<th>Program</th>
<th>Window</th>
<th>Step</th>
<th>Obs/Exp</th>
<th>MinPC</th>
<th>Length</th>
<th>Reverse</th>
<th>Complement</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpgplot</td>
<td>100</td>
<td>1</td>
<td>0.6</td>
<td>50</td>
<td>50</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Enter or Paste a nucleic acid Sequence (at least 100bp) in any format:

```
ACCGATACGATGAGAATGAGCAATGTAGTGAATCGTTTCAGCTACTCTCTATCGTAGCATTACTATGCAGTCAGTGATGCGCGCTAGCCGCGTAGCTCGCGGTCGCATCGCTGGCCGTAGCTGCGTACGATCTGCTGTACGCTGATCGGAGCGCTGCATCTCAACTGACTCATACTCATATGTC
```

Upload a file: [Bladeren...]

Run  Reset
observed vs. expected

Islands of unusual CG composition
EMBOSS_001 from 1 to 286
Observed/Expected ratio > 0.60
Percent C + Percent G > 50.00
Length > 50
Length 114 (51..164)
Some Notes on: Higher Order Markov Chains

- The Markov property specifies that the probability of a state depends only on the probability of the previous state.
- But we can build more “memory” into our states by using a higher order Markov model.
- In an $n$-th order Markov model

$$
\Pr(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_1) = \Pr(x_i \mid x_{i-1}, \ldots, x_{i-n})
$$

The probability of the current state depends on the previous $n$ states.
Selecting the Order of a Markov Chain Model

- But the number of parameters we need to estimate for an \( n \text{-th} \) order Markov model grows exponentially with the order
  - for modeling DNA we need \( O(4^{n+1}) \) parameters (# of state transitions) for an \( n \text{-th} \) order model

- The higher the order, the less reliable we can expect our parameter estimates to be
  - estimating the parameters of a \( 2^{\text{nd}} \) order Markov chain from the complete genome of E. Coli (5.44 x 10^6 bases), we would see each (length 3) word ~ 85,000 times on average (divide by 4^3)
  - estimating the parameters of a \( 9^{\text{th}} \) order chain, we would see each (length 10) word ~ 5 times on average (divide by 4^{10} ~ 10^6)
• An $n$-th order Markov chain over some alphabet $A$ is equivalent to a first order Markov chain over the alphabet of $n$-tuples: $A^n$

• Example: a 2nd order Markov model for DNA can be treated as a 1st order Markov model over alphabet

  AA, AC, AG, AT
  CA, CC, CG, CT
  GA, GC, GG, GT
  TA, TC, TG, TT

  Transition probabilities: $P(A|AA), P(A|AC)$, etc.
A Fifth Order Markov Chain Equivalent

\[ \Pr(GCTACA) = \Pr(GCTAC) \Pr(A \mid GCTAC) \]
• **where (in long sequence) are CpG islands?**
  first approach: Markov Chains + windowing
  second approach: *hidden Markov model*
Given observed sequence AGGCT, which state emits which item?
What is a hidden Markov model? Sean R Eddy
Example: weather

- **Initial probabilities**
  - \( p_H = 0.4 \)
  - \( p_M = 0.2 \)
  - \( p_L = 0.4 \)

- **Transition probabilities**
  - From `H` to `H`: 0.6
  - From `H` to `M`: 0.3
  - From `H` to `L`: 0.1
  - From `M` to `H`: 0.4
  - From `M` to `M`: 0.2
  - From `M` to `L`: 0.5
  - From `L` to `H`: 0.1
  - From `L` to `M`: 0.4
  - From `L` to `L`: 0.4

- **Emission probabilities**
  - \( P(\text{rain}) = 0.1 \)
  - \( P(\text{cloudy}) = 0.2 \)
  - \( P(\text{sunny}) = 0.7 \)
  - \( P(\text{rain}) = 0.3 \)
  - \( P(\text{cloudy}) = 0.4 \)
  - \( P(\text{sunny}) = 0.3 \)
  - \( P(\text{rain}) = 0.6 \)
  - \( P(\text{cloudy}) = 0.3 \)
  - \( P(\text{sunny}) = 0.1 \)

- **Observed weather vs. pressure**
Example: weather

Given path

\( P(\text{RCCSS} \mid \text{HHHHH}) = 1 \cdot 2 \cdot 2 \cdot 7 \cdot 7 = 196 \times 10^{-5} \)

\( P(\text{RCCSS} \mid \text{MMMMM}) = 3 \cdot 4 \cdot 4 \cdot 3 \cdot 3 = 432 \times 10^{-5} \)

\( P(\text{RCCSS}, \text{HHHHH}) = 4 \cdot 1 \cdot 6 \cdot 2 \cdot 6 \cdot 2 \cdot 6 \cdot 7 \cdot 6 \cdot 7 = 1016 \times 10^{-7} \)

\( P(\text{RCCSS}, \text{MMMMM}) = 2 \cdot 3 \cdot 2 \cdot 4 \cdot 2 \cdot 4 \cdot 2 \cdot 3 \cdot 2 \cdot 3 = 14 \times 10^{-7} \)
**hidden Markov model**

**model** $M = (\Sigma, Q, T)$
- states $Q$
- transition probabilities $t_{pq}, p, q \in Q$

**observation** $X = x_1 x_2 \ldots x_n \in \Sigma^*$

observe states *indirectly* ‘hidden’
- emission probabilities
  
  $e_{px}, p \in Q, x \in \Sigma \quad e_p(x)$

**probability**
observation given the model
? there may be *many* state seq’s
HMM main questions

observation  $X \in \Sigma^*$

- probability of this observation $X$?
- most probable state sequence?
- how to find the model? training
Given sequence X: most probable state vs. optimal path

* most probable state (over all state sequences)
posterior decoding
forward & backward probabilities

* most probable path (= single state sequence)
Viterbi
dynamic programming: probability ending in state $q$ emitting symbol $x_i$

$$f_q(i) = \sum_{p \in Q} f_p(i-1) t_{pq} e_q(x_i)$$
probability of observation $X$

probability ending in state

$$f_q(i) = P(x_1 \ldots x_i, \pi_i = q)$$

$$f_q(i) = \sum_{p\in Q} f_p(i-1) \ t_{pq} \ e_q(x_i)$$

‘forward’ probability

$$P(X) = \sum_{p \in Q} f_p(n) \ t_{p*}$$
Probability of observation: weather

Initial state:
- Remain in H: $p_H = 0.4$
- Coming from M: $p_M = 0.2$
- Coming from L: $p_L = 0.4$

Transitions:
- Remain in H
- Coming from M
- Coming from L

$$P(\text{RCCSS}) = P(\text{RC…})$$

<table>
<thead>
<tr>
<th>State</th>
<th>Action</th>
<th>Probability</th>
<th>Calculation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>R</td>
<td>0.6</td>
<td>$(4 \times 0.6 + 6 \times 0.4 + 24 \times 0.1) \times 2 = 144 \text{ (x10}^{-4})$</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>C</td>
<td>0.2</td>
<td>$(4 \times 0.3 + 6 \times 0.2 + 24 \times 0.5) \times 4 = 576 \text{ (x10}^{-4})$</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>S</td>
<td>0.1</td>
<td>$(4 \times 0.4 + 6 \times 0.4 + 24 \times 0.4) \times 3 = 372 \text{ (x10}^{-4})$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>0.1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Posterior decoding

\[ P(\pi_i = q \mid X) \]

\[ f_q(i) = P(x_1 \ldots x_i, \pi_i = q) \]
\[ b_q(i) = P(x_{i+1} \ldots x_n \mid \pi_i = q) \]

\[ P(X, \pi_i = q) = f_q(i)b_q(i) \]
\[ P(\pi_i = q \mid X) = \frac{f_q(i)b_q(i)}{P(X)} \]
HMM main questions

observation $X \in \Sigma^* \Rightarrow$ most probable state sequence

• probability of this observation?
• most probable state sequence?
• how to find the model? training
Viterbi algorithm

most probable state sequence for observation

1. *dynamic programming*: probability ending in state $q$ and emitting $x_i$

$$v_q(i) = \max_{p \in Q} v_p(i-1) \cdot t_{pq} \cdot e_q(x_i)$$
(2) *traceback*: most probable state sequence start with final maximum
HMM Example: CpG islands

8 states $A^+$ vs $A^-$
unique observation each state

Transition Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.180</td>
<td>0.274</td>
<td>0.426</td>
<td>0.120</td>
</tr>
<tr>
<td>C</td>
<td>0.171</td>
<td>0.368</td>
<td>0.274</td>
<td>0.188</td>
</tr>
<tr>
<td>G</td>
<td>0.161</td>
<td>0.339</td>
<td>0.375</td>
<td>0.125</td>
</tr>
<tr>
<td>T</td>
<td>0.079</td>
<td>0.355</td>
<td>0.384</td>
<td>0.182</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.300</td>
<td>0.205</td>
<td>0.285</td>
<td>0.210</td>
</tr>
<tr>
<td>C</td>
<td>0.322</td>
<td>0.298</td>
<td>0.078</td>
<td>0.302</td>
</tr>
<tr>
<td>G</td>
<td>0.248</td>
<td>0.246</td>
<td>0.298</td>
<td>0.208</td>
</tr>
<tr>
<td>T</td>
<td>0.177</td>
<td>0.239</td>
<td>0.292</td>
<td>0.292</td>
</tr>
</tbody>
</table>
HMM for Hidden Coin Tossing

Fig. 2. Three possible Markov models which can account for the results of hidden coin tossing experiments. (a) 1-coin model. (b) 2-coins model. (c) 3-coins model.

Suspect
dishonest casino dealer
dishonest casino dealer

<table>
<thead>
<tr>
<th>Rolls</th>
<th>31511</th>
<th>62464646442453211316311641521336251445436316566626566666</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLL</td>
<td></td>
</tr>
<tr>
<td>Viterbi</td>
<td>FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFULLLLL</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>65116645313265124563666463163663162326455235266666625151631</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>LLLLLLFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLTLLLLLLLLLLLLLLLLLFFFFF</td>
</tr>
<tr>
<td>Viterbi</td>
<td>LLLLLLFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLTLLLLLLLLLLLLLLLLLLLL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>222555441666566563564324364131513465146353411126414626253356</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>FFFFFFFFFFLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL</td>
</tr>
<tr>
<td>Viterbi</td>
<td>FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>366163666466232534413661661163252562462255265252266435353336</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF</td>
</tr>
<tr>
<td>Viterbi</td>
<td>LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rolls</th>
<th>23312162536441443233516324363366556246666626326666612355245242</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die</td>
<td>FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF</td>
</tr>
<tr>
<td>Viterbi</td>
<td>FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF</td>
</tr>
</tbody>
</table>
dishonest casino dealer

Observation 366163666466232534413661661163252562462255265252266435353336
Viterbi LLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Compare to:

Forward FFLLLLLLLLLLLLLLLLLLFFFFFFFFFLLLLLLLLLFLLLLLLLLLLLLLLLLLFFFFFFFFF
Posterior (total) LLLLLLLLLLLLLFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLU
Sketch: Parameter estimation

training sequences $X^{(i)}$
optimize score $\prod_{i=1}^{n} P(X^{(i)} | \Theta)$

state sequences known

- count transitions $pq$, $A_{pq}$
- count emissions $b$ in $p$, $E_p(b)$

divide by
- total transitions in $p$
- emissions in $q$

Laplace correction
Baum-Welch

state sequences unknown

Baum-Welch training
based on model
expected number of transitions, emissions
build new (better) model & iterate

\[ P(\pi_i = p, \pi_{i+1} = q \mid X, \Theta) = \]
\[
\frac{f_p(i) \cdot t_{pq} \cdot e_q(x_{i+1}) \cdot b_q(i+1)}{P(X)}
\]

\[ A_{pq} \quad \text{sum over all training sequences } X \]
\[ \text{sum over all positions } i \]

\[ E_p(b) \quad \text{sum over all training sequences } X \]
\[ \text{sum over all positions } i \text{ with } x_i = b \]
Baum-Welch training

concerns:
  • guaranteed to converge target score, not $\Theta$
  • unstable solutions!
  • local maximum

tips:
  • repeat for several initial $\Theta$
  • start with meaningful $\Theta$

Viterbi training (an alternative)
determine optimal paths recompute as if paths known
  • score may decrease!


Furthermore: