INTRODUCTION TO MACHINE LEARNING

MARKOV MODELS

* The contents are adapted from Dr. Jean Gao at UT Arlington

Mingon Kang, Ph.D.
Department of Computer Science @ UNLV
Bayes Theorem

\[
Pr(x \mid y) = \frac{Pr(y \mid x) Pr(x)}{Pr(y)} = \frac{Pr(y \mid x) Pr(x)}{\sum_{x} Pr(y \mid x) Pr(x)}
\]

- this theorem is extremely useful
- there are many cases when it is hard to estimate \( Pr(x \mid y) \) directly, but it's not too hard to estimate \( Pr(y \mid x) \) and \( Pr(x) \)
MDs usually aren’t good at estimating $\Pr(\text{Disorder} \mid \text{Symptom})$

they’re usually better at estimating $\Pr(\text{Symptom} \mid \text{Disorder})$

if we can estimate $\Pr(\text{Fever} \mid \text{Flu})$ and $\Pr(\text{Flu})$ we can use Bayes’ Theorem to do diagnosis

$$\Pr(\text{flu} \mid \text{fever}) = \frac{\Pr(\text{fever} \mid \text{flu}) \Pr(\text{flu})}{\Pr(\text{fever} \mid \text{flu}) \Pr(\text{flu}) + \Pr(\text{fever} \mid \neg \text{flu}) \Pr(\neg \text{flu})}$$
Example

X: 35 years old customer with an income of $4,000 and fair credit rating

H: Hypothesis that the customer will buy a computer
- $P(H \mid X)$: probability that the customer will buy a computer given that we know his age, credit rating and income (posterior)
- $P(H)$: probability that the customer will buy a computer regardless of age, credit rating, and income
- $P(X \mid H)$: probability that the customer is 35 year old, have fair credit rating and earns $40,000, given that he has bought our computer (likelihood)
- $P(X)$: probability that a person from our set of customers is 35 year old, have fair credit rating and earns $40,000.
Maximum Likelihood Estimation (MLE)

- Strategy of estimating the parameters of a statistical model given observations
- Choose a value that maximizes the probability of observed data
- $\theta^* = \text{argmax } P(x|\theta)$, where $x$ is observation and $\theta$ is a parameter
Maximum a posteriori estimation (MAP)

- Estimating unobserved population (or making a decision)
- Choose a value that is most probable given observed data and prior belief:
- \( \theta^* = \text{argmax} \ P(y|x) \)
A Markov Chain Model

Transition probabilities:
\[
\begin{align*}
&\Pr(x_i = a \mid x_{i-1} = g) = 0.16 \\
&\Pr(x_i = c \mid x_{i-1} = g) = 0.34 \\
&\Pr(x_i = g \mid x_{i-1} = g) = 0.38 \\
&\Pr(x_i = t \mid x_{i-1} = g) = 0.12
\end{align*}
\]
Markov Chain Models

- A Markov chain model (a triplicate) 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
  - Initial state probabilities
Markov Chain Models

- given some sequence $x$ of length $L$, we can ask how probable the sequence is given based on our model
- for any probabilistic model of sequences, we can write this probability as

$$\Pr(x) = \Pr(x_L, x_{L-1}, \ldots, x_1)$$

$$= \Pr(x_L | x_{L-1}, \ldots, x_1) \Pr(x_{L-1} | x_{L-2}, \ldots, x_1) \ldots \Pr(x_1)$$

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

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

$$= \Pr(x_1) \prod_{i=2}^{L} \Pr(x_i | x_{i-1})$$
The Probability of a Sequence for a Given Markov Chain Model

\[
\Pr(cggt) = \Pr(c) \Pr(g \mid c) \Pr(g \mid g) \Pr(t \mid g)
\]
Markov Chain Models

- can also have an end state; allows the model to represent
  - a distribution over sequences of different lengths
  - preferences for ending sequences with certain symbols

![Diagram of Markov Chain Models]
Markov Chain Notation

- the transition parameters can be denoted by

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

- similarly we can denote the probability of a sequence \( x \) as

\[ a_{Bx_1} \prod_{i=2}^{L} a_{x_{i-1}x_i} = \Pr(x_1) \prod_{i=2}^{L} \Pr(x_i \mid x_{i-1}) \]

where represents the transition from the \textit{begin} state
Example of a Grad Student

- Grad Student come in two flavors:
  - Happy vs Depressed about research

- Each type of grad student has two Markov chain models
  - Three locations we can observe the grad students at:
    - Lab
    - Coffee shop
    - Bar

Ref: https://faculty.coe.drexel.edu/gailr/ECE-S690-503/markov_models.ppt.pdf
Happy Grad Student Markov Chain

Observations:
Lab, Coffee, Lab, Coffee, Lab, Lab, Bar, Lab, Coffee,…
Depressed about research
Example Application

- CpG islands
  - CpG is a pair of nucleotides C and G, appearing successively in this order, along one DNA strand.
  - CpG dinucleotides are rarer in eukaryotic genomes than expected given the marginal probabilities of C and G
  - but the regions upstream of genes (promoter regions) are richer in CpG dinucleotides than elsewhere – CpG islands
  - useful evidence for finding genes
- could predict CpG islands with Markov chains
  - one to represent CpG islands
  - one to represent the rest of the genome
Estimating the Model Parameters

- given some data (e.g. a set of sequences from CpG islands), how can we determine the probability parameters of our model?

- one approach: maximum likelihood estimation
  - given a set of data \( D \)
  - set the parameters \( \theta \) to maximize

\[
\Pr(D \mid \theta)
\]

- i.e. make the data \( D \) look likely under the model
Maximum Likelihood Estimation

- suppose we want to estimate the parameters $\Pr(a)$, $\Pr(c)$, $\Pr(g)$, $\Pr(t)$
- and we’re given the sequences
  - accgcgctta
  - gcttagtgac
  - tagccgtttac
- then the maximum likelihood estimates are

  $\Pr(a) = \frac{6}{30} = 0.2$  \hspace{1cm} $\Pr(g) = \frac{7}{30} = 0.233$

  $\Pr(c) = \frac{9}{30} = 0.3$  \hspace{1cm} $\Pr(t) = \frac{8}{30} = 0.267$
Maximum Likelihood Estimation

- suppose instead we saw the following sequences
  
  gccgcgccttg
  gcttggtggc
tggccgttgc

- then the maximum likelihood estimates are

\[
\begin{align*}
\text{Pr}(a) &= \frac{0}{30} = 0 \\
\text{Pr}(c) &= \frac{9}{30} = 0.3 \\
\text{Pr}(g) &= \frac{13}{30} = 0.433 \\
\text{Pr}(t) &= \frac{8}{30} = 0.267
\end{align*}
\]

do we really want to set this to 0?
A Bayesian Approach

- instead of estimating parameters strictly from the data, we could start with some prior belief for each
- for example, we could use Laplace estimates

\[
\Pr(a) = \frac{n_a + 1}{\sum_i (n_i + 1)}
\]

where \( n_i \) represents the number of occurrences of character \( i \)

- using Laplace estimates with the sequences
  
  \[
  \begin{align*}
  \text{gccgcgttg} & \quad \Pr(a) = \frac{0 + 1}{34} \\
  \text{gcttggtggc} & \quad \Pr(c) = \frac{9 + 1}{34}
  \end{align*}
  \]
A Bayesian Approach

- A more general form: *m-estimates*

\[
\Pr(a) = \frac{n_a + p_a m}{\sum_i n_i + m}
\]

- Prior probability of \(a\)
- Number of "virtual" instances

- With \(m=8\) and uniform priors

  \[
  \text{gccgcgcgttg, gcttggtggc, tggccgtttgc}
  \]

\[
\Pr(c) = \frac{9 + 0.25 \times 8}{30 + 8} = \frac{11}{38}
\]
Estimation for 1st Order Probabilities

- to estimate a 1st order parameter, such as \( \Pr(c \mid g) \), we count the number of times that \( c \) follows the history \( g \) in our given sequences
- using Laplace estimates with the sequences

\[
\begin{align*}
\text{gccgcgcttg:} & \quad \Pr(a \mid g) = \frac{0+1}{12+4} \\
\text{gcttgtgggc:} & \quad \Pr(c \mid g) = \frac{7+1}{12+4} \\
\text{tggccgcttgcc:} & \quad \Pr(g \mid g) = \frac{3+1}{12+4} \\
\text{Pr(t \mid g):} & \quad \Pr(t \mid g) = \frac{2+1}{12+4}
\end{align*}
\]
Markov Chains for Discrimination

- Question: Given a short stretch of genomic sequence, how could we decide if it comes from a CpG island or not?
- 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
- Can then score a test sequence by:

\[
\text{score}(x) = \log \frac{\text{Pr}(x \mid \text{CpG model})}{\text{Pr}(x \mid \text{null model})}
\]
Markov Chains for Discrimination

- why use

\[
score(x) = \log \frac{\Pr(x \mid CpG)}{\Pr(x \mid null)}
\]

- Bayes’ rule tells us

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

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

- if we’re not taking into account prior probabilities of two classes (\( \Pr(CpG) \) and \( \Pr(null) \)) then we just need to compare \( \Pr(x \mid CpG) \) and \( \Pr(x \mid null) \)
where \( \Pr(x \mid CpG\text{model}) = \Pr(x_{L}, x_{L-1}, \ldots, x_{1} \mid \text{model}) \)

\[ = \Pr(x_{1}) \prod_{2}^{L} \Pr(x_{i} \mid x_{i-1}) \]

\[ = \Pr(x_{1}) \prod_{2}^{L} a_{x_{i-1}x_{i}}^{+} \]

and \( a_{st}^{+} \) is the transition probability for the CpG island model from letter \( s \) to letter \( t \), which is set as

\[ a_{st}^{+} = \frac{c_{st}^{+}}{\sum_{t'} c_{st'}^{+}} \]

and \( c_{st}^{+} \) is the number of times letter \( s \) followed by \( t \). So the score can be further written as

\[ \text{score}(x) = \log \frac{\Pr(x \mid CpG\text{model})}{\Pr(x \mid \text{null model})} = \sum_{i=1}^{L} \log \frac{a_{x_{i-1}x_{i}}^{+}}{a_{x_{i-1}x_{i}}^{-}} \]
Markov Chains for Discrimination

- parameters estimated for CpG and null models
  - human sequences containing 48 CpG islands
  - 60,000 nucleotides

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>+</strong> A</td>
<td>0.18</td>
<td>0.27</td>
<td>0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>C</td>
<td>0.17</td>
<td>0.37</td>
<td>0.27</td>
<td>0.19</td>
</tr>
<tr>
<td>G</td>
<td>0.16</td>
<td>0.34</td>
<td>0.38</td>
<td>0.12</td>
</tr>
<tr>
<td>T</td>
<td>0.08</td>
<td>0.36</td>
<td>0.38</td>
<td>0.18</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><strong>-</strong> A</td>
<td>0.30</td>
<td>0.21</td>
<td>0.28</td>
<td>0.21</td>
</tr>
<tr>
<td>C</td>
<td>0.32</td>
<td>0.30</td>
<td>0.08</td>
<td>0.30</td>
</tr>
<tr>
<td>G</td>
<td>0.25</td>
<td>0.24</td>
<td>0.30</td>
<td>0.21</td>
</tr>
<tr>
<td>T</td>
<td>0.18</td>
<td>0.24</td>
<td>0.29</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Markov Chains for Discrimination

- light bars represent negative sequences (non-CpG island)
- dark bars represent positive sequences
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,

$$\Pr(x_i \mid x_{i-1}, x_{i-2}, \ldots, x_1) = \Pr(x_i \mid x_{i-1}, \ldots, x_{i-n})$$
Markov Models in Computational Biology

- 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
References


- **Book**: Eddy & Durbin, Chapter 3.

- Jones and Pevzner: Chapter 11.