Lecture 6: Markov models

Statistical Methods for Natural Language Processing
Fredrik Engström

February 16, 2011
Summary of lecture 5

- \( I(X; Y) = H(X) - H(X|Y) \)
- The cross-entropy between \( p \) and \( q \) is
  \[
  \sum_x p(x) \log \frac{1}{q(x)}.
  \]
- MLE: Maximize \( P(x_1, \ldots, x_k|\theta) \)
- Needs smoothing with sparse data.
- Bayesian: Maximize \( P(\theta|x_1, \ldots, x_k) \). Same as maximizing \( P(x_1, \ldots, x_k|\theta)P(\theta) \)
Markov models

Intuition: Some random process changing over time with the following properties:

- **Memoryless**, meaning that only the present state and **not** the past affects the future states.
- **Stationary**, meaning that it is time homogeneous.

Examples: Bigram analysis. Fia med knuff. Trigram(?). Cards(?).
Definition

A **Markov model** is an (infinite) sequence of random variables $X_1, X_2, \ldots$ such that

- $P(X_{k+1} = y | X_1 = x_1, \ldots X_k = x_k) = P(X_{k+1} = y | X_k = x_k)$
- $P(X_{k+1} = y | X_k = x) = P(X_2 = y | X_1 = x)$

Alternative definition as a **finite state machine**:

- Set of states $S$.
- **Initial state probabilities** $\pi_i$ for $i \in S$.
- **State transition probabilities** $a_{ij}$ for $i, j \in S$. 
Example: Weather
Hidden Markov model

- States: Start, Rainy, Sunny, Walk, Shop, Clean
- Transitions:
  - Start to Rainy: 0.6
  - Start to Sunny: 0.4
  - Rainy to Sunny: 0.3
  - Sunny to Rainy: 0.4
  - Rainy to Walk: 0.7
  - Rainy to Shop: 0.1
  - Rainy to Clean: 0.1
  - Sunny to Walk: 0.6
  - Sunny to Shop: 0.5
  - Sunny to Clean: 0.1
  - Walk to Shop: 0.4
  - Walk to Clean: 0.3
  - Shop to Clean: 0.3

Lecture 6: Markov models
Statistical Methods for Natural Language Processing
Fredrik Engström
Hidden Markov model

Definition of a (state-emitting) hidden Markov model:

- Set of states $S$.
- Set of outputs $K$.
- Initial state probabilities $\pi_i$ for $i \in S$.
- State transition probabilities $a_{ij}$ for $i, j \in S$.
- Output emission probabilities $b_{ik}$ for $i \in S, k \in K$. 
Hidden Markov model

Definition of a (arc-emitting) hidden Markov model:

• Set of states $S$.
• Set of outputs $K$.
• Initial state probabilities $\pi_i$ for $i \in S$.
• State transition probabilities $a_{ij}$ for $i, j \in S$.
• Output emission probabilities $b_{ijk}$ for $i, j \in S, k \in K$. 
Inferences

The three inferences:

1. Given a HMM $\mu$ what is the probability of a certain output sequence $O$, i.e., what is $P(O|\mu)$?

2. Given an output sequence $O$ and a HMM $\mu$ what is the best guess at a state sequence explaining the output $O$, i.e. which state sequence $X$ maximizes $P(X|O, \mu)$?

3. Given an output sequence $O$ what is the best guess at a HMM explaining the output?

Next we give solutions to 1 and 2.
The forward algorithm

Given a HMM $\mu$ what is the probability of a certain output sequence $O$, i.e., what is $P(O|\mu)$?
Let $\alpha_i(t)$ be $P(o_1, \ldots, o_t, X_t = s_i)$.

- $\alpha_i(1) = P(X_1 = s_i)b_{i0} = \pi_i b_{i0}$.
- $\alpha_i(t + 1) = P(o_1, \ldots, o_{t+1}, X_{t+1} = s_i) =$
  \[ b_{i0} \sum_{j=1}^{N} P(o_1, \ldots, o_t, X_t = s_j)a_{ji} = b_{i0} \sum_{j=1}^{N} \alpha_j(t)a_{ji} \]

Example: $O = (\text{walk'}, \text{shop'}, \text{clean'})$, $s_1 = \text{Rainy}$ $s_2 = \text{Sunny}$

<table>
<thead>
<tr>
<th></th>
<th>t=1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>$\alpha_1(1)$</td>
<td>$\alpha_1(2)$</td>
<td>$\alpha_1(3)$</td>
</tr>
<tr>
<td>$s_2$</td>
<td>$\alpha_2(1)$</td>
<td>$\alpha_2(2)$</td>
<td>$\alpha_2(3)$</td>
</tr>
</tbody>
</table>

\[
\begin{array}{c|ccc}
& t=1 & 2 & 3 \\
\hline
s_1 & .06 & .0552 & .02904 \\
s_2 & .24 & .0486 & .004572 \\
\end{array}
\]

$P(\text{walk'}, \text{shop'}, \text{clean'}) = .02904 + .004572 = .033612$
The Viterbi algorithm

Given an output sequence $O$ and a HMM $\mu$ what is the best guess at at a state sequence explaining the output $O = o_1, \ldots, o_T$, i.e. which state sequence $X$ maximizes $P(X|O, \mu)$?

Let $\delta_i(t)$ be

$$\max_{x_1, \ldots, x_{t-1}} P(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = s_i, o_1, \ldots, o_t).$$

- $\delta_i(1) = P(X_1 = s_i)b_{io_1} = \pi_i b_{io_1}$.
- $\psi_i(1) = 0$
- $\delta_i(t + 1) = \max_{x_1, \ldots, x_t} P(X_1 = x_1, \ldots, X_t = x_t, o_1, \ldots, o_{t+1}) = b_{i o_{t+1}} \max_j (\delta_j(t) a_{ji})$
- $\psi_i(t + 1) = \arg\max_j \delta_j(t) a_{ji}$
- $\hat{X}_T = \arg\max_i \delta_T(i)$
- $\hat{X}_t = \psi_{\hat{X}_{t+1}}(t + 1)$
Summary

- Markov models
- Hidden (state / arc emitting) Markov models
- Forward algorithm
- Viterbi algorithm