Parameter Estimation for HMMs

L645 / B659
Dept. of Linguistics, Indiana University
Fall 2015

Iterative hill-climbing algorithm
- Will provably converge to a local maximum
- Operates on raw, unannotated data
- Special Case of the Expectation Maximization (EM) Algorithm
- Will provably converge to a local maximum

If we have annotated data, the task of parameter estimation can be calculated directly, from observed relative frequencies
- We are focusing on the case where we only observe signal sequences, and states are unknown

Parameter Estimation: Task and Method

Task: Given a sequence of signals
\[ \mathbf{O} = (\sigma_1, \ldots, \sigma_T) \]

Determine the set of model parameters \( \lambda = (P, A, \nu) \) which maximize the probability of this signal sequence.

Method: Baum-Welch algorithm

Baum-Welch Algorithm

- Iterative hill-climbing algorithm
- Operates on raw, unannotated data
- Will provably converge to a local maximum

Parameter Estimation for HMMs (1)

We first define the set of probabilities
\[
\epsilon_t(i, j) = \frac{P(\xi_t = s_i, \xi_{t+1} = s_j)}{P(O)} = \frac{P(O, \xi_t = s_i)}{P(O)}
\]

This is the joint probability of being in state \( s_i \) at time \( t \) and of being in state \( s_j \) at time \( t + 1 \) conditional on the entire observed signal sequence \( O \) from time 1 to time \( T \).

Parameter Estimation for HMMs (2)

Note that
\[
epsilon_t(i, j) = \frac{\alpha_t(i) \cdot p_{ij} \cdot a_{s_j}}{P(O)} \cdot \frac{\beta_{t+1}(j)}{\sum_{j'} \alpha_t(i) \cdot a_{i j'} \cdot \beta_{t+1}(j')}
\]

since
\[
P(O, \xi_t = s_i, \xi_{t+1} = s_j) = P(O, \xi_{t+1} = s_j) \cdot P(O_{[t+1:T]} | O_{[1:t]}, \xi_t = s_i) \cdot \sum_{j'} P(O_{[1:t]} | O_{[t+1:T]}, \xi_{t+1} = s_j)
\]

Parameter Estimation for HMMs (3)

Forward-Backward Variables

We can define a set of forward-backward variables:
\[
\gamma_t(i) = \frac{P(\xi_t = s_i | O)}{P(O)} = \frac{P(O, \xi_t = s_i)}{P(O)} = \frac{\alpha_t(i) \cdot \beta_t(i)}{\sum_{k=1} P(\xi_{t+1} = s_j | \xi_t = s_k)}
\]

This is the probability of being in state \( s_i \) at time \( t \) conditional on the entire observed signal sequence \( O \) from time 1 to time \( T \).

We (immediately?) see that
\[
\gamma_t(i) = \frac{P(\xi_t = s_i | O)}{P(O)} = \sum_{j=1}^n P(\xi_t = s_i, \xi_{t+1} = s_j | O) = \sum_{j=1}^n \epsilon_t(i, j)
\]
This means that the technique of Lagrangian multipliers can be used to find the optimal values of \( \nu_i, \nu_j, \gamma_{ij} \) as those values which maximize the function \( P(O \mid \lambda) \), where \( \lambda = (v, p, a) \) (subject to the above three constraints).

Putting it together:
1. Assign initial values to parameters \( v, p, a \)
2. Get values for \( \nu_i \) and \( \gamma_{ij} \), based on the observed data \( O \)
3. Re-estimate \( v, p, a \)
Reestimating the Crazy Softdrink Machine

**Parameter Estimation for HMMs**

Baum-Welch Algorithm

Example

Reestimating the Crazy Softdrink Machine (2)

- state transitions

\[
p_{CP, CP} = \frac{\sum_{i=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP) + \sum_{i=1}^{T} \alpha_i(CP) \cdot \beta_i(CP) + \alpha_i(CP) \cdot \beta_i(CP) + \alpha_i(CP) \cdot \beta_i(CP)}{\sum_{i=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP)}
\]

Reestimating the Crazy Softdrink Machine (3)

\[
p_{CP, IP} = \frac{\alpha_i(CP) \cdot p_{CP, IP} \cdot a_{CP, ice} \cdot \beta_i(CP) + \alpha_i(IP) \cdot p_{IP, CP} \cdot a_{IP, cola} \cdot \beta_i(CP)}{\sum_{i=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP) + \sum_{i=1}^{T} \alpha_i(CP) \cdot \beta_i(CP) + \alpha_i(IP) \cdot \beta_i(CP)}
\]

Reestimating the Crazy Softdrink Machine (4)

- emissions

\[
a_{CP, cola} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP) \cdot \gamma_i(CP)}{\sum_{t=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP) \cdot \gamma_i(CP)}
\]

Reestimating the Crazy Softdrink Machine (5)

\[
a_{CP, lem} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP) \cdot \gamma_i(CP)}{\sum_{t=1}^{T} \gamma_i(CP) \cdot \gamma_i(CP) \cdot \gamma_i(CP)}
\]