Hidden Markov Models

Recitation for ML 10-701
Pulkit Bhuwalka

Introduction

- Time Series Models
- State Machines
- Markov Property
- Simple Graphical Models

Motivation

- Speech Recognition
- Activity Recognition
- Financial Forecasting
- POS tagging
- basically, any time series data

Markov Models

- Markov Property $p(X) = p(Xn|Xn-1)$
- Chapman Kolmogorov $A(m+n) = A(m)A(n)$
- Stationary distribution
- Stochastic Properties (sum(a_ij) = 1,a_ij >=0)

Markov Models - example

- Weather Example
- $p(O|Model)$
- How long will it stay in state d for exactly t time durations?

HMM

- Coin Toss
- Dishonest Casino
- Classic Urn and Ball

Model

- Number of states (N)
- Number of observation symbols
- Prior Probabilities
- Transition Probabilities
- Emission Probabilities

$$\lambda = (A, B, \pi)$$

Problems

1. Evaluation (Useful for model selection)
2. Decoding (No correct state. Best solution based on criterion)
    a. Find most probable state given observation
    b. Find most probable sequence given observation
3. Estimation (Model training)

Problems

1. Evaluation - P(O|\lambda)
2. Decoding
   a. argmax P(S_t|O, \lambda)
   b. argmax P(S|O, \lambda)
3. Estimation (Model training) - P(\lambda|O)

Naive Solution - Evaluation

- Enumerate every possible state
- To calculate p(O|\lambda)
- Problem N^T
- N=5, T = 100 approx. 10^72

$$P(O|Q,\lambda) = b_{q1}(O_1)b_{q2}(O_2)...$$
$$P(Q|\lambda) = \pi_{q1}a_{q1q2}a_{q2q3}...$$
$$P(O,Q|\lambda) = P(O|Q,\lambda).P(Q|\lambda)$$

Forward procedure

- Initialize $$\alpha_1(i) = \pi_i b_i(O_1), 1 \leq i \leq N$$
- Induction $$\alpha_{t+1}(i) = [\sum_{j=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}), 1 \leq j \leq N$$
- Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

$$\alpha_t(i) = P(O_1, ..., O_t, q_t = S_i|\lambda)$$

Backward Procedure

- Initialize $$\beta_T(i) = 1, \quad 1 \le i \le N$$

- Induction

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

$$1 \le i \le N, \quad t = T-1, T-2, \ldots 1$$

$$\beta_t(i) = P(O_{t+1} \ldots O_T | q_t = S_i, \lambda)$$

Decoding - Best state

- Forward Backward
- Doesn’t give best sequence (why?)

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

$$P(q_t = S_i | O, \lambda) = \frac{P(O_1, \ldots O_t, q_t = S_i | \lambda) P(O_{t+1}, \ldots O_T | q_t = S_i, \lambda)}{P(O|\lambda)}$$

$$P(q_t = S_i | O, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}$$

$$q_t = argmax[\gamma_t(i)]$$

$$1 \leq i \leq N, \quad 1 \leq t \leq T$$

Viterbi

- Similar to Forward
- At each step, calculate max
- Backtrack at the end

Baum-Welch

- Local Maxima, Can't solve analytically

$$P(\lambda|O) = argmax\ P(O|\lambda)$$

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}).\beta_{t+1}(j)}{P(O|\lambda)}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

Baum Welch

- Expected Transitions from state i $\sum_{t=1}^{T-1} \gamma_t(i)$

- Expected Transitions from i to j $\sum_{t=1}^{T-1} \xi_t(i,j)$

$$\hat{\pi}_i = \gamma_1(i)$$

$$\hat{a_{i,j}} = \frac{\sum_{i=1}^{T-1} \xi_t(i,j)}{\sum_{i=1}^{T-1} \gamma_t(i)}$$

$$\hat{b_{j,k}} = \frac{\sum_{i=1,O_t=v_k}^{T-1} \gamma_t(j)}{\sum_{i=1}^{T-1} \gamma_t(j)}$$

- Speech Recognition
- Activity Recognition