DS-GA 3001.009 Modeling Time Series Data Lab 5

Artie Shen | Center for Data Science

NYU

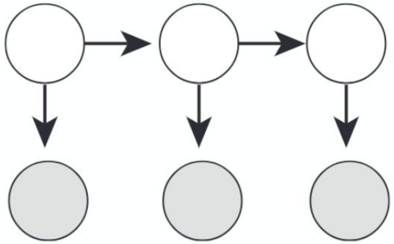
Agenda

- Recap
  - Hidden Markov Model (Bishop, Ch 13.2 & 13.3)
  - Viterbi Algorithm

- Programming - POS Tagging
  - Sampling
  - Decoding
  - Learning

Formulation

Observed States $O = \{o_1, ..., o_T\}, o_i \in \{1, 2, ..., H\}$



Latent States $Q = \{q_1, ..., q_T\}, q_i \in \{1, 2, ..., K\}$

Transition Probability Matrix $A \in \mathbb{R}^{K\times K}$ where $a_{i,j} = p(q_t = j | q_{t-1} = i), \sum_{j=1}^{K} a_{i,j} = 1, \forall i.$

Emission Probability Matrix $B \in \mathbb{R}^{K\times H}$ where $B_{i,j} = p(o_t = j | q_t = i), \sum_{j=1}^{H} b_{i,j} = 1, \forall i.$

Initial State Probability $q_1 \sim D$. In our case, we set D to the Categorical Distribution with probability $\pi \in \mathbb{R}^K$ and $\pi_i = p(q_1 = i)$.

Likelihood Computation

Goal Given an HMM $\theta = \{A, B, \pi\}$ and an observation sequence $O = \{o_1, ..., o_T\}$, find the likelihood $P(O|\theta)$.

Challenge Because iid assumption does not hold, brute force algorithm takes $O(K^T)$ time to solve.

Observations

- First of all, we can take advantage of the conditional independence property of HMM, $P(Q) = \prod_{i=1}^{T} p(q_i|q_{i-1})$ with the assumption that $p(q_1|q_0)$ is specified by π.
- Suppose we know the full sequence of latent states Q, then $P(O|Q) = \prod_{i=1}^{T} p(o_i|q_i)$.

Likelihood Computation

- Use Bayes Rule, we have $P(O,Q) = P(O|Q) \times P(Q) = \prod_{i=1}^{T} P(o_i|q_i) \times \prod_{i=1}^{T} P(q_i|q_{i-1})$
- We can obtain the marginalized probability $P(O) = \sum_Q P(O,Q) = \sum_Q P(O|Q)P(Q)$
- Let $\alpha_t(j)$ denotes the probability of being in state j after seeing the first t observations.
- $\alpha_t(j) = P(o_1, ..., o_t, q_t = j) = \sum_{i=1}^{K} \alpha_{t-1}(i) a_{ij} b_j(o_t)$

The Forward Algorithm

- Initialize $\alpha_1(i) = \pi_i b_i(o_1)$
- Recursively compute $\alpha_t(j) = \sum_{i=1}^{K} \alpha_{t-1}(i) a_{ij} b_j(o_t); 1 \le j \le N, 1 \le t \le T$
- $P(O|\theta) = \sum_{i=1}^{K} \alpha_T(i)$

Decoding

Goal Given an HMM $\theta = \{A, B, \pi\}$, and a sequence of observations $O = \{o_1, ..., o_T\}$, find the most likely sequence of latent states $Q = \{q_1, ..., q_T\}$.

Viterbi Algorithm

- $v_t(j) = max_{q_1,...,q_{t-1}} P(o_1, ..., o_t, q_1, ..., q_{t-1}, q_t = j)$ denotes the probability that the HMM is in state j after seeing the first t observations and passing through the most probable latent state sequence $q_1, q_2, ..., q_{t-1}$.
- Initialization $v_1(j) = \pi_j b_j(o_1)$
- Recursively update v_t: $v_t(j) = max_{i=1}^{K} v_{t-1}(i) a_{ij} b_j(o_t)$
- Store the best previous state $b_t(j) = argmax_{i=1}^{K} v_{t-1}(i) a_{ij} b_j(o_t)$
- The best state for the last latent space $B_T = argmax_j v_T(j)$, $B_t = b_t(B_{t+1})$

Learning

Goal Given an observation sequence O and optionally the ground truth for latent sequence Q, learn the HMM parameters A, B, and π.

Supervised Learning When Q in the training data is given learning is simple:

- $\hat{a}_{ij} = \frac{num\ \ transitions\ \ from\ \ i\ \ to\ \ j}{num\ \ transitions\ \ from\ \ i} = \frac{C(q_t{=}i\ \&\ q_{t+1}{=}j)}{C(q_t{=}i)}$
- $\hat{b}_{ij} = \frac{num\ of\ times\ state\ i\ emits\ j}{num\ of\ state\ i} = \frac{C(q_t{=}i\ \&\ o_t{=}j)}{C(q_t{=}i)}$
- $\hat{\pi}_i = \frac{num\ of\ chains\ start\ with\ i}{total\ num\ of\ chains}$

Unsupervised Learning When Q is not given:

- Use Baum-Welch EM Algorithm.
- In E step, we calculate expected number of counts for the quantities above.
- In M step, we update the parameters A, B, and π.

POS-tagging

- **Part of Speech**
  - A category of words that have same grammatical property
  - “There are 70 children there.”
  - DT JJ CD NNS RB.
  - Disambiguation: book VB/NN?
- **Dataset**
  - WSJ POS subset
  - **39815** sentences for training, **1700** sentences for test

Programming

- **Github:**
    - **https://github.com/charlieblue17/timeseries2018**
- **Due Date 03/08/2018 06:45 pm on NYU Classes**
- **Please rename your submission to net_id.ipynb**