Hidden Markov Models

Notes on Spoken Language Processing

Chia-Ping Chen

Department of Computer Science and Engineering

National Sun Yat-Sen University

Kaohsiung, Taiwan ROC

Stochastic Processes

- A stochastic process is a collection of random variables (e.g. physical quantities) indexed by time.
  - discrete-time vs. continuous-time
  - discrete vs. continuous
- Let $\{X_1, X_2, \ldots X_n\}$ be a discrete-time stochastic process. Then

$$
\begin{aligned}
p(x_1, \ldots, x_n) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \ldots p(x_n|x_1, \ldots, x_{n-1}) \\
&= \prod_{t=1}^{n} p(x_t|x_{1:t-1})
\end{aligned}
$$

Discrete-Time Markov Chain

- Let $\mathcal{X} = \{1, 2, \ldots, N\}$ be the set of values that the X_t's can assume.
- (first-order) Markov assumption

$$p(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) = p(x_t|x_{t-1})$$

- It follows that

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{t=2}^{n} p(x_t|x_{t-1})$$

Time-Invariant Markov Chain

- A Markov chain is time-invariant if the transition probability does not vary with time

$$p(x_t = j | x_{t-1} = i) = a_{ij} \quad \forall\, t.$$

- a_{ij} is called the state transition probability, satisfying

$$a_{ij} \geq 0, \ \ \forall\, i, j$$

$$\sum_{j=1}^{N} a_{ij} = 1, \ \ \forall\, i.$$

- The initial probability is

$$\pi_i = p(x_1 = i).$$

An Example of Markov Chain

- DJI
- state space

$$\mathcal{X} = \{1 = \text{up}, 2 = \text{down}, 3 = \text{flat}\}.$$

- transition probability matrix and initial probability

$$P = \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}, \ \pi = (0.5,\ 0.2,\ 0.3).$$

- What is the probabilty that DJI is up for the first 5 days?

Hidden Markov Models

- In a hidden Markov model,
  - there is an unobserved Markov chain, and
  - the output observation at a time is a random variable whose distribution depends on the state of the hidden Markov chain.
- For the DJI example, one can imagine there are three market states (such as bull, bear), leading to different probability distributions of up, down and flat.
- To describe an HMM, additional output probabilities for each state have to be specified.

Output Probability

- Let X_t be the observation at time t and S_t be the hidden state, the output probability is defined by

$$b_i(k) = p(X_t = o_k | S_t = i).$$

- It satisfies that, if $\{o_1, \ldots, o_M\}$ is the alphabet for observation, then

$$\sum_{k=1}^{M} b_i(k) = 1, \ \forall\, i.$$

Basic Problems in HMM

- (evaluation problem) Given the observations $\mathbf{o}$ and the model parameters λ, evaluate
$$p(\mathbf{o}|\lambda).$$
- (decoding problem) Given $\mathbf{o}$ and λ, determine the optimal state sequence $\mathbf{s}^*$
$$\mathbf{s}^* = \arg\max_{\mathbf{s}} P(\mathbf{o}, \mathbf{s}|\lambda).$$
- (estimation problem) Given $\mathbf{o}$, decide λ^* by
$$\lambda^* = \arg\max_{\lambda} P(\mathbf{o}|\lambda).$$

Probability Evaluation

- brute-force method

$$p(\mathbf{o}|\lambda) = \sum_{\mathbf{s}} p(\mathbf{o}, \mathbf{s}|\lambda) = \sum_{\mathbf{s}} p(\mathbf{o}|\mathbf{s}, \lambda) p(\mathbf{s}|\lambda)$$

$$= \sum_{s_1, \ldots, s_n} p(s_1) p(o_1|s_1) \prod_{t=2}^{n} p(o_t|s_t) a_{s_{t-1} s_t}$$

- $O(nN^n)$ time complexity.

Forward Algorithm

- Define the forward probability

$$\alpha_t(i) \triangleq p(o_1, \ldots, o_t, s_t = i|\lambda)$$

- The data likelihood is given by

$$p(\mathbf{o}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

- $O(N^2 n)$ time complexity.

Decoding Problem

- We look at the optimal state sequence with the maximum posterior probability given $\mathbf{o}$.

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} P(\mathbf{s}|\mathbf{o}, \lambda) = \arg\max_{\mathbf{s}} P(\mathbf{o}, \mathbf{s}|\lambda).$$

- Viterbi algorithm: denote $\mathbf{s}_t = s_1, \ldots, s_t$, and define

$$\begin{aligned}
\delta_t(i) &\triangleq \max_{\mathbf{s}_{t-1}} \; p(\mathbf{s}_{t-1}, s_t = i, \mathbf{o}_t|\lambda) \\
\Rightarrow \delta_{t+1}(j) &\triangleq \max_{\mathbf{s}_t} \; p(\mathbf{s}_t, s_{t+1} = j, \mathbf{o}_t, o_{t+1}|\lambda) \\
&= \max_{\mathbf{s}_t} \; p(\mathbf{s}_t, \mathbf{o}_t|\lambda) \, p(o_{t+1}, s_{t+1} = j|\mathbf{s}_t, \mathbf{o}_t, \lambda) \\
&= \max_{i} \max_{\mathbf{s}_{t-1}} \; p(\mathbf{s}_{t-1}, s_t = i, \mathbf{o}_t|\lambda) \, a_{ij} \, p(o_{t+1}|s_{t+1} = j) \\
&= \max_{i} \; \delta_t(i) \, a_{ij} \, p(o_{t+1}|s_{t+1} = j)
\end{aligned}$$

Estimation Problem

- By far the most difficult one of the three problems.
- The basic principle is the EM algorithm that iteratively increases the data likelihood.
- To compute the posterior probability of states, the Baum-Welch algorithm, a.k.a. forward-backward algorithm, is called for.

Forward-Backward Algorithm

- Define the backward probability

$$\beta_t(i) \triangleq p(o_{t+1}, \ldots, o_n | s_t = i, \lambda)$$

Speech as HMMs

- The state space corresponds to a set of lingustic units which should exhaust all possible speech.
- Speech waveform is segmented into speech windows. Each window of speech is processed into speech features as the observations.
- For example, phone states + MFCC features.
- The more data we have, the more detailed the models become.