Statistical Sequence Recognition

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- With DTW, local distortions (distance) between acoustic frames are integrated temporally to compute the global distance between two templates.
- For fast computation, DP can be used for DTW.
- The notion of distance can be generalized to a statistical framework: a large distance signals a small probability.

Statistical Speech Recognition

- We assume that the speech features are generated according to the probability models of the underlying linguistic units.
- The model parameters are learned from labelled data during the **training** phase.
- Once learned, they are used to find the hypothesis with the maximum a posteriori (MAP) probability given the speech features during the **testing** phase.

Bayes Rule and MAP

- The fundamental equation for pattern recognition is the MAP criterion and the Bayes rule:

$$M^* = \arg\max_M P(M|X) = \arg\max_M P(X|M)P(M)$$

- The problem is solved in principle if
  - we have accurate models for $P(X|M)$ ad $P(M)$.
  - we have a way to search M^*.

Markov Models

- Markov model is a set of states, an initial distribution, and a transition probability matrix.
- We impose two model assumptions. The first is that the probability of state q_t given q_{t-1} is independent of any state prior to $t-1$.

$$p(q_t|q_{t-1}, q_{t-2}, \ldots, q_1) = p(q_t|q_{t-1})$$

The second assumption is that the transition probability does not vary with t, i.e.,

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

An Example of Markov Model

- A starter W of New York Yankees
- State space $\mathcal{X} = \{1 = \text{A}, 2 = \text{B}, 3 = \text{T}\}$.
- Transition probability

$$A = \begin{bmatrix} 0.6 & 0.1 & 0.3 \\ 0.4 & 0.3 & 0.3 \\ 0.5 & 0.3 & 0.2 \end{bmatrix}$$

- Initial probability is uniform.
- $p(W$ leaves with lead in the first 10 games$) = \ ?$

Hidden Markov Models

- We may want to say something about the pitcher's performance given the team's record for T consecutive games.
- In an HMM, the state identities are *hidden* and the *observed* sequence depends probabilistically on the state sequence.
- In addition to the components required in a Markov model, in HMM there are the observation likelihoods, denoted by $b_i(o_t)$, representing the probability of observing o_t when the state $q_t = i$.

Coin-toss Models

- Suppose there are a number of coins, each with its own bias.
- One of the coins, coin q_t, is randomly selected. The selection probability is dependent on the identity of the previous coin, q_{t-1}.
- Coin q_t is tossed and the outcome (head or tail) o_t is recorded, *but not the coin*.
- The probability is

$$p(q_1^T, o_1^T) = p(q_1)p(o_1|q_1)\prod_{t=2}^{T} p(q_t|q_{t-1})p(o_t|q_t)$$

Urn-and-ball Models

- Suppose there are N urns, each consisting of a distinct composition of colored balls.
- One of the urns, urn q_t, is randomly selected. The selection probability is dependent on the identity of the previous urn, q_{t-1}.
- A ball is picked from the selected urn q_t and the color of the ball is recorded as o_t.
- As we only observe the colors of balls, do we really know how many urns there are?

Basic Problems in HMM

- Probability evaluation: Given the observations O and the model parameters λ, compute the data likelihood $p(O|\lambda)$.
- Optimal state sequence: Given the observations O and λ, determine the optimal state sequence Q^*

$$Q^* = \arg\max_Q p(O, Q|\lambda)$$

- Parameter estimation: Given the observations O, choose the model parameters λ to maximize the data-likelihood

$$\lambda^* = \arg\max_\lambda p(O|\lambda)$$

Forward-Backward Algorithm

- Denote the parameters in HMM by
  - the initial probability $\pi_i = a_{1i}$
  - the transition probability a_{ij}
  - the observation likelihood $b_i(o_t)$
- Given these parameters, the data likelihood can be computed via the forward-backward algorithm.
- With data likelihood, many things can be computed.

Forward Probability

Define the forward probability α as

$$\alpha_i(t) = p(o_1, \dots, o_t, q_t = i).$$

Then

$$\alpha_j(1) = a_{1j} b_j(o_1),$$

$$\alpha_j(t) = \sum_{i=2}^{N-1} \alpha_i(t-1) a_{ij} b_j(o_t),$$

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN}.$$

Backward Probability

Similarly, define the backward probability β as

$$\beta_i(t) = p(o_{t+1}, \ldots, o_T | q_t = i).$$

Then

$$\beta_i(T) = a_{iN},$$

$$\beta_i(t) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}) \beta_j(t+1),$$

$$\beta_1(1) = \sum_{j=2}^{N-1} a_{1j} b_j(o_1) \beta_j(1).$$

Data Likelihood

- The joint probability of $q_t = j$ and O is

$$p(O, q_t = j) = \alpha_j(t)\beta_j(t).$$

- The data likelihood is

$$p(O) = \sum_j p(O, q_t = j) = \sum_j \alpha_j(t)\beta_j(t).$$

- Alternatively,

$$p(O) = \alpha_N(T) = \beta_1(1)$$

Viterbi Approximation

- Best-path approximation to $p(O)$ is

$$p(O) \triangleq \sum_Q p(Q, O) \sim \max_Q p(Q, O) \triangleq \bar{p}(O).$$

- Define $\delta_j(t) \triangleq \max_{q_1^{t-1}} p(q_1^{t-1}, q_t = j, o_1^t)$. Then

$$\begin{aligned} \delta_j(t) &= \max_i \max_{q_1^{t-2}} p(q_1^{t-2}, q_{t-1} = i, o_1^{t-1}) \; a_{ij} \; p(o_t | q_t = j) \\ &= \max_i \delta_i(t-1) \; a_{ij} \; b_j(o_t). \end{aligned}$$

- Taking logarithm, this is similar to DTW.