Speech Recognition

Lecture 8: Expectation-Maximization Algorithm, Hidden Markov Models.

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.com

This Lecture

- Expectation-Maximization (EM) algorithm
- Hidden-Markov Models

Latent Variables

- Definition: unobserved or hidden variables, as opposed to those directly available at training and test time.
  - example: mixture models, HMMs.
- Why latent variables?
  - naturally unavailable variables: e.g., economics variables.
  - modeling: latent variables introduced to model dependencies.

ML with Latent Variables

- Problem:
  - with fully observed variables, ML is often straightforward.
  - with latent variables, log-likelihood contains a sum (harder to find the best parameter values):

$$L(\theta, x) = \log p_\theta[x] = \log \sum_z p_\theta(x, z) = \log \sum_z p_\theta[z|x] p_\theta[x].$$

- Idea: use current parameter values to estimate latent variables and use those to re-estimate parameter values $\longrightarrow$ EM algorithm.

EM Theorem

- **Theorem**: for any x, and parameter values θ and θ',

$$L(\theta', x) - L(\theta, x) \geq \sum_z p_\theta[z|x] \log p_{\theta'}[x, z] - \sum_z p_\theta[z|x] \log p_\theta[x, z].$$

- **Proof**:

$$\begin{aligned}
L(\theta', x) - L(\theta, x) &= \log p_{\theta'}[x] - \log p_\theta[x] \\
&= \sum_z p_\theta[z|x] \log p_{\theta'}[x] - \sum_z p_\theta[z|x] \log p_\theta[x] \\
&= \sum_z p_\theta[z|x] \log \frac{p_{\theta'}[x, z]}{p_{\theta'}[z|x]} - \sum_z p_\theta[z|x] \log \frac{p_\theta[x, z]}{p_\theta[z|x]} \\
&= \sum_z p_\theta[z|x] \log \frac{p_\theta[z|x]}{p_{\theta'}[z|x]} + \sum_z p_\theta[z|x] \log p_{\theta'}[x, z] - \sum_z p_\theta[z|x] \log p_\theta[x, z] \\
&= D(p_\theta[z|x] \| p_{\theta'}[z|x]) + \sum_z p_\theta[z|x] \log p_{\theta'}[x, z] - \sum_z p_\theta[z|x] \log p_\theta[x, z] \\
&\geq \sum_z p_\theta[z|x] \log p_{\theta'}[x, z] - \sum_z p_\theta[z|x] \log p_\theta[x, z].
\end{aligned}$$

EM Idea

- Maximize expectation:

$$\sum_z p_\theta[z|x] \log p_{\theta'}[x, z] = \mathop{\mathrm{E}}_{z \sim p_\theta[\cdot|x]} \Big[\log p_{\theta'}[x, z] \Big].$$

- Iterations:
  - compute $p_\theta[z|x]$ for current value of θ.
  - compute expectation and find maximizing θ'.

EM Algorithm

(Dempster, Laird, and Rubin, 1977; Wu, 1983)

- Algorithm: maximum-likelihood meta algorithm for models with latent variables.
  - E-step: $q^{t+1} \leftarrow p_{\theta^t}[z|x]$.
  - M-step: $\theta^{t+1} \leftarrow \operatorname*{argmax}_{\theta} \sum_{z} q^{t+1}(z|x) \log p_{\theta}[x, z]$.
- Interpretation:
  - E-step: posterior probability of latent variables given observed variables and current parameter.
  - M-step: maximum-likelihood parameter given all data.

EM Algorithm - Notes

- Applications:
  - mixture models, e.g., Gaussian mixtures. Latent variables: which model generated points.
  - HMMs (Baum-Welsh algorithm).
- Notes:
  - positive: each iteration increases likelihood. No parameter tuning.
  - negative: can converge to local optima. Note that likelihood could converge but not parameter. Dependence on initial parameter.

EM = Coordinate Ascent

- **Theorem**: EM = coordinate ascent applied to

$$A(q,\theta) = \sum_z q[z|x] \log \frac{p_\theta[x,z]}{q[z|x]}$$

$$= \sum_z q[z|x] \log p_\theta[x,z] - \sum_z q[z|x] \log q[z|x].$$

  - E-step: $q^{t+1} \leftarrow \underset{q}{\operatorname{argmax}}\, A(q,\theta^t)$.
  - M-step: $\theta^{t+1} \leftarrow \underset{\theta}{\operatorname{argmax}}\, A(q^{t+1},\theta)$.

- **Proof**: the E-step follows the two identities

$$A(q,\theta) \leq \log \sum_z q[z|x] \frac{p_\theta[x,z]}{q[z|x]} = \log \sum_z p_\theta[x,z] = \log p_\theta[x] = L(\theta,x),$$

$$A(p_\theta[z|x],\theta) = \sum_z p_\theta[z|x] \log p_\theta[x] = \log p_\theta[x] = L(\theta,x).$$

EM = Coordinate Ascent

- **Proof** (cont.): finally, note that

$$\operatorname*{argmax}_{\theta} A(q^{t+1}, \theta) = \operatorname*{argmax}_{\theta} \sum_{z} q^{t+1}[z|x] \log p_{\theta}[x, z],$$

coincides with the M-step of EM.

EM - Extensions and Variants

(Dempster et al., 1977; Jamshidian and Jennrich, 1993; Wu, 1983)

- Generalized EM (GEM): maximization is not necessary at all steps since any (strict) increase of the auxiliary function guarantees an increase of the log-likelihood.
- Sparse EM: posterior probabilities computed at some points in E-step.

This Lecture

- Expectation-Maximization (EM) algorithm
- Hidden-Markov Models

Motivation

- Data: sample of m sequences over alphabet Σ drawn i.i.d. according to some distribution D :

$$x_1^i, x_2^i, \ldots, x_{k_i}^i, \quad i = 1, \ldots, m.$$

- Problem: find sequence model that best estimates distribution D.

- Latent variables: observed sequences may have been generated by model with states and transitions that are hidden or unobserved.

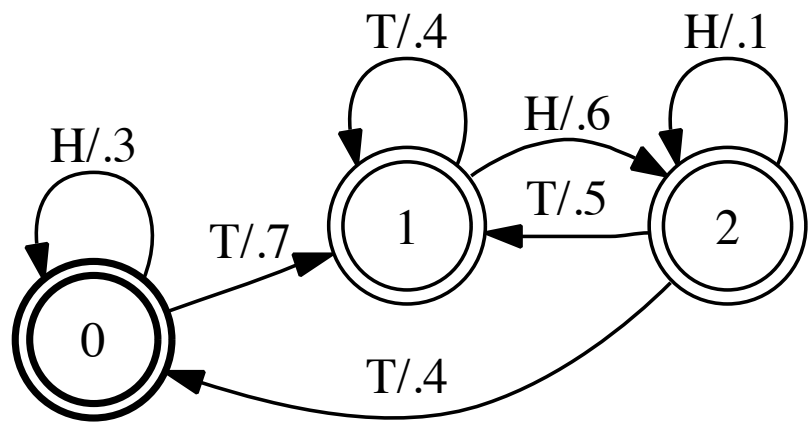
Example

- Observations: sequences of Heads and Tails.

observed sequence: $H, T, T, H, T, H, T, H, H, H, T, T, \ldots, H.$

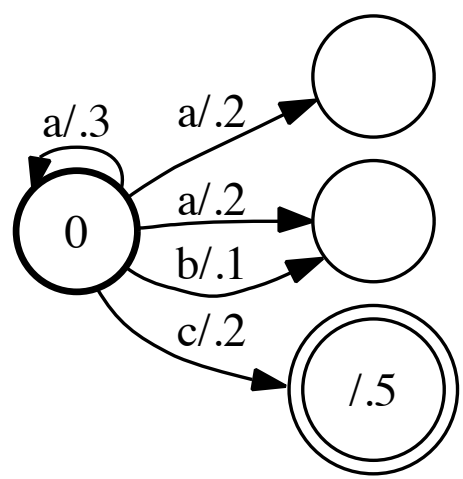
unobserved paths: $(0, H, 0), (0, T, 1), (1, T, 1), \ldots, (2, H, 2).$

- Model:



Hidden Markov Models

- Definition: probabilistic automata, generative view.
  - discrete case: finite alphabet Σ.
  - transition probability: $\Pr[\text{transition } e]$.
  - emission probability: $\Pr[\text{emission } a|e]$.
  - simplification: for us, $w[e] = \Pr[a|e] \Pr[e]$.
- Illustration:



Other Models

- Outputs at states (instead of transitions): equivalent model, dual representation.
- Non-discrete case: outputs in $\mathbb{R}^N$.
  - fixed probabilities replaced by distributions, e.g., Gaussian mixtures.
  - application to acoustic modeling.

Three Problems

(Rabiner, 1989)

- Problem 1: given sequence $x_1, \ldots, x_k$ and hidden Markov model p_θ compute $p_\theta[x_1, \ldots, x_k]$.
- Problem 2: given sequence $x_1, \ldots, x_k$ and hidden Markov model p_θ find most likely path $\pi = e_1, \ldots, e_r$ that generated that sequence.
- Problem 3: estimate parameters θ of a hidden Markov model via ML: $\theta_\star = \operatorname*{argmax}_\theta p_\theta[x]$.

P1: Evaluation

- Algorithm: let X represent a finite automaton representing the sequence $x = x_1 \cdots x_k$ and let H denote the current hidden Markov model.
  - Then, $p_\theta[x] = \sum_{\pi \in X \circ H} w[\pi]$.
  - Thus, it can be computed using composition and a shortest-distance algorithm in time $O(|X||H|)$.

P2: Most Likely Path

- Algorithm: shortest-path problem in $(\max, \times)$ semiring.
  - any shortest-path algorithm applied to the result of the composition $X \circ H$. In particular, since the automaton is acyclic, with a linear-time shortest-path algorithm, the total complexity is in $O(|X \circ H|) = O(|X||H|)$.
  - a traditional solution is to use the Viterbi algorithm.

P3: Parameter Estimation

(Baum, 1972)

- Baum-Welsh algorithm: special case of EM applied to HMMs.
  - Here, θ represents the transitions weights $w[e]$.
  - The latent or hidden variables are paths π.
  - M-step:

$$\operatorname*{argmax}_{\theta} \quad \sum_{\pi} p_{\theta'}[\pi|x] \log p_{\theta}[x, \pi]$$

$$\text{subject to} \quad \forall q \in Q, \sum_{e \in E[q]} w_{\theta}[e] = 1.$$

P3: Parameter Estimation

- Using Lagrange multipliers $\lambda_q, q \in Q$, the problem consists of setting the following partial derivatives to zero:

$$\frac{\partial}{\partial w_\theta[e]}\Big[\sum_\pi p_{\theta'}[\pi|x]\log p_\theta[x,\pi] - \sum_q \lambda_q \sum_{e\in E[q]} w_\theta[e]\Big] = 0$$

$$\Leftrightarrow \sum_\pi p_{\theta'}[\pi|x]\frac{\partial \log p_\theta[x,\pi]}{\partial w_\theta[e]} - \lambda_{\mathrm{orig}[e]} = 0.$$

- $p_\theta[x,\pi] \neq 0$ only for paths π labeled with x: $\pi \in \Pi(x)$.
- In that case, $p_\theta[x,\pi] = \prod_e w_\theta[e]^{|\pi|_e}$, where $|\pi|_e$ is the number of occurrences of e in π.

P3: Parameter Estimation

- Thus, the equation with partial derivatives can be rewritten as

$$\sum_{\pi \in \Pi(x)} p_{\theta'}[\pi|x] \frac{|\pi|_e}{w_\theta[e]} = \lambda_{orig(e)}$$

$$\Leftrightarrow w_\theta[e] = \frac{1}{\lambda_{orig(e)}} \sum_{\pi \in \Pi(x)} p_{\theta'}[\pi|x]|\pi|_e$$

$$\Leftrightarrow w_\theta[e] = \frac{1}{\lambda_{orig(e)} p_{\theta'}[x]} \sum_{\pi \in \Pi(x)} p_{\theta'}[x, \pi]|\pi|_e$$

$$\Leftrightarrow w_\theta[e] \propto \sum_{\pi \in \Pi(x)} p_{\theta'}[x, \pi]|\pi|_e.$$

P3: Parameter Estimation

- But, $\sum_{\pi \in \Pi(x)} p_{\theta'}[x, \pi] |\pi|_e = \sum_{i=1}^{k} \alpha_{i-1}(\theta') w_{\theta'}[e] \beta_i(\theta')$.
- Thus,

$$w_\theta[e] \propto \sum_{i=1}^{k} \alpha_{i-1}(\theta') w_{\theta'}[e] \beta_i(\theta')$$

Expected number of times through transition e.

with $\alpha_i(\theta') = p_{\theta'}[x_1, \ldots, x_i \wedge q = orig[e]]$

$\beta_i(\theta') = p_{\theta'}[x_i, \ldots, x_k | dest[e]]$.

Can be computed using forward-backward algorithms, or, for us, shortest-distance algorithms.

References

- L. E. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation for Probalistic Functions of Markov Processes. *Inequalities*, 3:1-8, 1972.
- Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin . Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, No. 1. (1977), pp. 1-38..
- Jamshidian, M. and R. I. Jennrich: 1993, Conjugate Gradient Acceleration of the EM Algorithm. Journal of the American Statistical Association 88(421), 221-228.
- Lawrence Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of IEEE*, Vol. 77, No. 2, pp. 257, 1989.
- O'Sullivan. Alternating minimzation algorithms: From Blahut-Arimoto to expectation-maximization. Codes, Curves and Signals: Common Threads in Communications, A. Vardy, (editor), Kluwer, 1998.
- C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. The *Annals of Statistics*, Vol. 11, No. 1 (Mar., 1983), pp. 95-103.