CMPT-413: Computational Linguistics

HMM6: Deriving HMM updates using Lagrange Multipliers

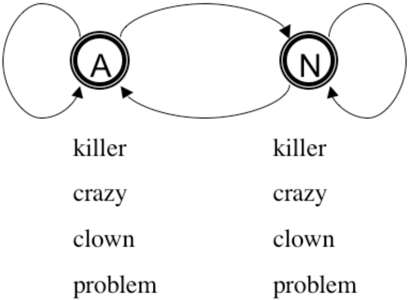
Anoop Sarkar
http://www.cs.sfu.ca/∼anoop

February 28, 2013

Hidden Markov Model

$$\text{Model } \theta = \begin{cases} \pi_i & \text{probability of starting at state } i \\ a_{i,j} & \text{probability of transition from state } i \text{ to state } j \\ b_i(o) & \text{probability of output } o \text{ at state } i \end{cases}$$

$$\text{Constraints}: \sum_i \pi_i = 1, \sum_j a_{i,j} = 1, \sum_o b_i(o) = 1$$



Learning from Fully Observed Data

$$L(\theta) = \sum_{\ell=1}^{m}$$
$$\sum_i f(i, x_\ell, y_\ell) \log \pi_i + \sum_{i,j} f(i, j, x_\ell, y_\ell) \log a_{i,j} + \sum_{i,o} f(i, o, x_\ell, y_\ell) \log b_i(o)$$

- $\theta = (\pi, a, b)$
- $L(\theta)$ is the log probability of the labeled data $(x_1, y_1), \ldots, (x_m, y_m)$
- We want to find a θ that will give us the maximum value of $L(\theta)$
- Find the θ such that $\frac{dL(\theta)}{d\theta} = 0$

Learning from Fully Observed Data

$$L(\theta) = \sum_{\ell=1}^{m}$$
$$\sum_i f(i, x_\ell, y_\ell) \log \pi_i + \sum_{i,j} f(i, j, x_\ell, y_\ell) \log a_{i,j} + \sum_{i,o} f(i, o, x_\ell, y_\ell) \log b_i(o)$$

- Find the θ such that $\frac{dL(\theta)}{d\theta} = 0$ and $\theta = (\pi, a, b)$
- Split up $L(\theta)$ into $L(\pi), L(a), L(b)$
- Let $\nabla L = \forall i, j, o : \frac{\partial L(\pi)}{\partial \pi_i}, \frac{\partial L(a)}{\partial a_{i,j}}, \frac{\partial L(b)}{\partial b_i(o)}$
- We must also obey constraints:
  $\sum_k \pi_k = 1, \sum_k a_{i,k} = 1, \sum_o b_i(o) = 1$

Learning from Fully Observed Data

$$L(\pi) = \sum_{\ell=1}^{m} \sum_{i} f(i, x_\ell, y_\ell) \log \pi_i$$

- Let us focus on $\nabla L(\pi)$ (the other two: a and b are similar)
- For the constraint $\sum_k \pi_k = 1$ we introduce a new variable into our search for a maximum:

$$L(\pi, \lambda) = L(\pi) + \lambda(1 - \sum_k \pi_k)$$

- λ is called the Lagrange multiplier
- λ penalizes any solution that does not obey the constraint
- The constraint ensures that π is a probability distribution

Learning from Fully Observed Data

$$\frac{\partial L(\pi)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \quad \underbrace{\sum_{\ell=1}^{m} f(i, x_\ell, y_\ell) \log \pi_i}_{\text{the only part with variable } \pi_i} \quad + \underbrace{\sum_{\ell=1}^{m} \sum_{j:j \neq i} f(j, x_\ell, y_\ell) \log \pi_j}_{\text{no } \pi_i \text{ so derivative is } 0}$$

- We want a value of π_i such that $\frac{\partial L(\pi,\lambda)}{\partial \pi_i} = 0$

$$\frac{\partial}{\partial \pi_i} \sum_{\ell=1}^{m} \left(f(i, x_\ell, y_\ell) \log \pi_i + \lambda(1 - \sum_k \pi_k) \right) = 0$$

$$\frac{\partial}{\partial \pi_i} \sum_{\ell=1}^{m} \left(\underbrace{f(i, x_\ell, y_\ell) \log \pi_i}_{\frac{\partial}{\partial \pi_i} = \frac{f(i, x_\ell, y_\ell)}{\pi_i}} + \lambda - \underbrace{\lambda \pi_i}_{\frac{\partial}{\partial \pi_i} = \lambda} - \lambda \sum_{j:j \neq i} \pi_j) \right) = 0$$

Learning from Fully Observed Data

$$\frac{\partial L(\pi)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \quad \underbrace{\sum_{\ell=1}^{m} f(i, x_\ell, y_\ell) \log \pi_i}_{\text{the only part with variable } \pi_i} \quad + \underbrace{\sum_{\ell=1}^{m} \sum_{j:j \neq i} f(j, x_\ell, y_\ell) \log \pi_j}_{\text{no } \pi_i \text{ so derivative is } 0}$$

- From Eqn (1) we can obtain a value of π_i wrt λ:

$$\frac{\partial L(\pi, \lambda)}{\partial \pi_i} = \underbrace{\sum_{\ell=1}^{m} \frac{f(i, x_\ell, y_\ell)}{\pi_i} - \lambda}_{\text{see previous slide}} = 0 \tag{1}$$

$$\pi_i = \frac{\sum_{\ell=1}^{m} f(i, x_\ell, y_\ell)}{\lambda} \tag{2}$$

- Combine π_is from Eqn (3) with constraint $\sum_k \pi_k = 1$

$$\lambda = \sum_k \sum_{\ell=1}^{m} f(k, x_\ell, y_\ell)$$

Learning from Fully Observed Data

$$\frac{\partial L(\pi)}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \quad \underbrace{\sum_{\ell=1}^{m} f(i, x_\ell, y_\ell) \log \pi_i}_{\text{the only part with variable } \pi_i} \quad + \underbrace{\sum_{\ell=1}^{m} \sum_{j:j \neq i} f(j, x_\ell, y_\ell) \log \pi_j}_{\text{no } \pi_i \text{ so derivative is } 0}$$

- The value of π_i for which $\frac{\partial L(\pi,\lambda)}{\partial \pi_i} = 0$ is Eqn (3) which can be combined with the value of λ from Eqn (4).

$$\pi_i = \frac{\sum_{\ell=1}^{m} f(i, x_\ell, y_\ell)}{\lambda} \tag{3}$$

$$\lambda = \sum_{k} \sum_{\ell=1}^{m} f(k, x_\ell, y_\ell) \tag{4}$$

$$\pi_i = \frac{\sum_{\ell=1}^{m} f(i, x_\ell, y_\ell)}{\sum_k \sum_{\ell=1}^{m} f(k, x_\ell, y_\ell)} \tag{5}$$

Learning from Fully Observed Data

$$L(\theta) = \sum_{\ell=1}^{m}$$

$$\sum_i f(i, x_\ell, y_\ell) \log \pi_i + \sum_{i,j} f(i, j, x_\ell, y_\ell) \log a_{i,j} + \sum_{i,o} f(i, o, x_\ell, y_\ell) \log b_i(o)$$

- The values of $\pi_i, a_{i,j}, b_i(o)$ that maximize $L(\theta)$ are:

$$\begin{aligned}
\pi_i &= \frac{\sum_\ell f(i, x_\ell, y_\ell)}{\sum_\ell \sum_k f(k, x_\ell, y_\ell)} \\
a_{i,j} &= \frac{\sum_\ell f(i, j, x_\ell, y_\ell)}{\sum_\ell \sum_k f(i, k, x_\ell, y_\ell)} \\
b_i(o) &= \frac{\sum_\ell f(i, o, x_\ell, y_\ell)}{\sum_\ell \sum_{o' \in V} f(i, o', x_\ell, y_\ell)}
\end{aligned}$$