Lecture 13

Parameter Estimation of Markov Chains

$$p_\theta(y) = \left(\prod_{i=2}^{N+1} p_\theta(y_i|y_{i-1})\right) p(y_1)$$

$$\log p_\theta(y) = \sum_{i=2}^{N+1} \log p_\theta(y_i|y_{i-1}) + c$$

$$\hat{\theta}_{ML} = \arg\max_\theta \left\{ \sum_{i=2}^{N+1} \log p_\theta(y_i|y_{i-1}) \right\}$$

Example) y discrete as before

$$p_\rho(y) = (1-\rho)^{N-K} \rho^K (1/2)$$

$$\log p_\rho(y) = (N-K)\log(1-\rho) + K\log\rho + c$$

$$\frac{d}{d\rho}\log p_\rho(y) = \frac{N-K}{1-\rho}(-1) + \frac{K}{\rho} = 0$$

$$(N-K)\rho = K(1-\rho)$$

$$N\rho = K$$

$$\boxed{\rho = \frac{K}{N}}$$

Generalization

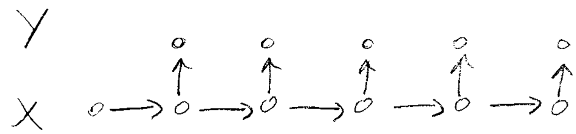
$$p(x_n \mid x_i \; i<n) = p(x_n \mid x_{n-1}, x_{n-2}, \ldots x_{n-p})$$

Defn x_n is Markov with order p

Define $$y_n = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_{n-p+1} \end{bmatrix}$$

Then y_n is Markov with order $p=1$.

<u>Application</u> - Hidden Markov Models (HMM)



X - Unobserved Markov Chain

Y - Observed Signal

X - as in previous examples

$$P\{X_0 = 0\} = P\{X_0 = 1\} = \tfrac{1}{2} \qquad p(x_n \mid x_{n-1}) = \begin{cases} \rho & x_n \neq x_{n-1} \\ 1-\rho & x_n = x_{n-1} \end{cases}$$

$$p_y(y_n \mid x_n) = \begin{cases} N(0, \sigma_1^2) & x_n = 0 \\ N(0, \sigma_2^2) & x_n = 1 \end{cases}$$

$$p(x_n \mid x_{n-1}) = \delta(x_n, x_{n-1})(1-\rho) + (1 - \delta(x_n, x_{n-1}))\rho$$

$$\log p_{x,y}(x, y) =$$

$$\sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma_{x_i}^2} y_i^2 - \frac{1}{2}\log(\sigma_{x_i}^2) + \log\left[\delta(x_i, x_{i-1})(1-\rho) + (1 - \delta(x_i, x_{i-1}))\rho\right] \right.$$

$$+ C$$

Define: $t(x_i, x_{i-1}) = 1 - \delta(x_i, x_{i-1})$

Notice

$$\log\left[\delta(x_i, x_{i-1})(1-\rho) + (1-\delta(x_i, x_{i-1}))\rho\right]$$

$$= t(x_i, x_{i-1}) \log\left(\frac{\rho}{1-\rho}\right) + \log(1-\rho)$$

Defn

$$\beta \triangleq \log\left(\frac{1-\rho}{\rho}\right) = -\log\left(\frac{\rho}{1-\rho}\right)$$

$$l(y|k) \triangleq \frac{1}{2\sigma_k^2} y^2 + \frac{1}{2}\log(\sigma_k^2)$$

$$\log p_{x,y}(x,y) = \sum_{i=1}^{N} \{ -l(y_i | x_i) - t(x_i, x_{i-1})\beta \} + c''$$

likelihood of data

transition probability

$$\hat{X}_{MAP} = \arg\max_{X} \left\{ \sum_{i=1}^{N} -l(y_i | x_i) - t(x_i, x_{i-1})\beta \right\}$$

$$\hat{X}_{ML} = \arg\max_{X} \left\{ \sum_{i=1}^{N} -l(y_i | x_i) \right\}$$

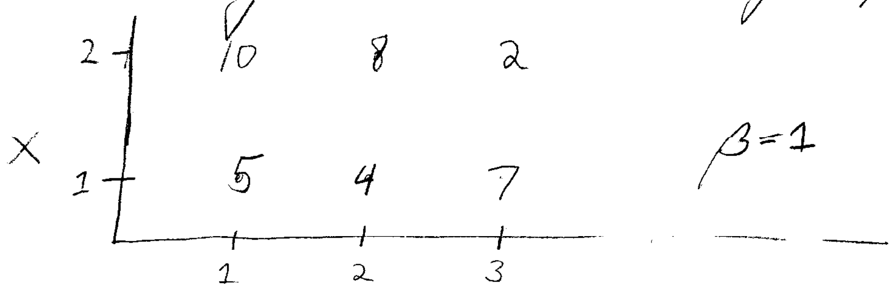
$$(\hat{X}_{ML})_i = \arg\max_{x_i} \{ -l(y_i | x_i) \}$$

Lecture 14

$$\hat{X}_{MAP} = \underset{X}{\arg\max} \left\{ \underbrace{\sum_{i=1}^{N} -l(y_i \mid x_i)}_{\substack{\text{data fitting} \\ \text{term}}} \underbrace{- t(x_i, x_{i-1})\beta}_{\substack{\text{prior model} \\ \text{term}}} \right\}$$

Notice

1) Choosing X is like choosing a "path"



2) $l(y_i \mid x_i)$ is the "cost" of each node
3) $t(x_i, x_{i-1})\beta$ is the "cost" of each transition

A) Dynamic Programming

Defn

$L(K,n)$ - Cost of best path starting at n with value K.

$L(1,3) = 7 \quad 2$
$L(2,3) = 2$

$L(K,n) = \underset{l}{\operatorname{argmin}}\{L(l,n+1) + t(K,l)\} + l(y_n|K)$

$L(1,2) = \min\{L(1,3), L(2,3)+1\} + 4$
$= \min\{7, 3\} + 4$
$= 7 \quad (2)$

$L(2,2) = \min\{L(1,3)+1, L(2,3)\} + 8$
$= \min\{8, 2\} + 8$
$= 10 \quad (2)$

$L(1,1) = \min\{L(1,2), L(2,2)+1\} + 5$
$= \min\{7, 11\} + 5$
$= 12 \quad (1)$

$L(2,1) = \min\{L(1,2)+1, L(2,2)\} + 10$
$= \min\{8, 10\} + 10$
$= 18 \quad (1)$

$L(1,1) < L(2,1) \Rightarrow$ Best path $1, 1, 2$

B. Coordinate decent,
<u>Gauss-Seidel / ICM</u>

$$\hat{x}_n = \arg\min_{x_n} \left\{ \sum_{i=1}^{N} l(y_i|x_i) + t(x_i, x_{i-1})\beta \right\}$$

$$= \arg\min_{x_n} \left\{ l(y_n|x_n) + \beta\left(t(x_n, x_{n-1}) + t(x_{n+1}, x_n) \right) \right\}$$

- Iterative application converges to local minimum

- Does not generally converge to global minimum

C. Gradient decent $\Rightarrow$ doesn't work!
X discrete!