CSE 291B 1/22

Nonnegative matrix factorization (NMF) (review)

- How to approximate $X \approx VY$ for $X, V, Y \geq 0$?
- Approximation error:

$$\varepsilon_{KL} = \sum_{in} \left[X_{in} \log \frac{X_{in}}{(VY)_{in}} - X_{in} + (VY)_{in} \right]$$

- Auxiliary function: (p_{ia} for Jensen's inequality)

$$G(v,w) = \sum_{in} \left[X_{in} \log \frac{X_{in}}{(WY)_{in}} - X_{in} + (VY)_{in} - X_{in} \sum_{a} \left(\frac{w_{ia} Y_{an}}{(WY)_{in}} \right) \log \frac{v_{ia}}{w_{ia}} \right]$$

- Properties:
  (i) $G(v,v) = \varepsilon_{KL}(v)$
  (ii) $G(v,w) \geq \varepsilon_{KL}(v)$
- Update:

$$v_{t+1} = \underset{v}{\operatorname{argmin}}\, G(v, v_t)$$

- Jensen's inequality
  - interchanges log-sum and sum-log operations
  - replaces $\log(\sum_a v_{ia} Y_{an})$ in cost function by $\sum_a p_a \log v_{ia}$ in auxiliary function
  - simplifies optimization
    $\partial \varepsilon_{KL} / \partial v_{ia} = 0$ cannot be solved in closed form.
    $\partial G / \partial v_{ia} = 0$ can be solved in closed form
  - also used to derive EM in discrete BNs

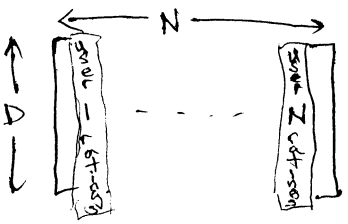
Binary matrix factorization

Ex: movie-user data

$X_{ij} \in \{0,1\}$ binary rating

$i = 1..D$ movies

$j = 1..N$ users



How to fill in missing matrix elements?

Netflix challenge: $1 million

Matrix has some missing elements:

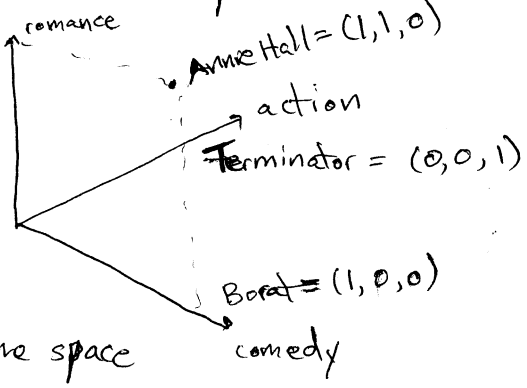
$\eta_{ij} \in \{0,1\}$ rating is absent / present

Logistic PCA

- Intuition

Represent movies as points in $d << D$ dimensional space.

$\vec{V}_i \in \mathbb{R}^d$ (i^{th} movie) Ex: $d=3$



Represent user preferences as points in same space

$\vec{V}_j \in \mathbb{R}^d$ (j^{th} user)

$\vec{V}_j \cdot \hat{e}_\alpha$ = user's preference for movie genre α ($\alpha = 1$ comedy, $= 2$ romance, $= 3$ action)

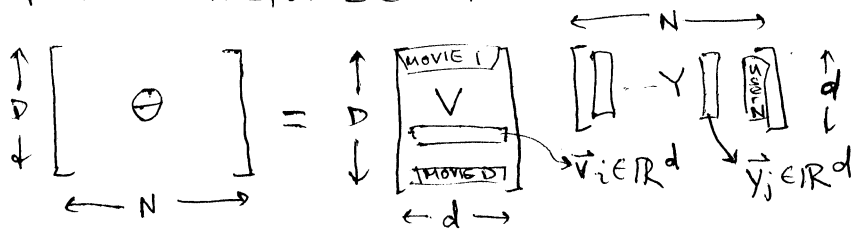
Model ratings as binary random variables: $X_{ij} \in \{0,1\}$

~~Fitting~~ Fit ratings by logistic model:

$P(X_{ij} = 1) = \sigma(\vec{V}_i \cdot \vec{V}_j)$ where $\sigma(z) = \frac{1}{1+e^{-z}}$

Use predictions to fill in matrix elements.

- Matrix factorization



$$\Theta = VY \quad \text{(exact factorization)}$$

$$P(X_{ij}=1) = \sigma(\Theta_{ij}) \quad \text{(indirection)}$$

X is generated from Θ

- Log-likelihood

$$\mathcal{L} = \sum_{ij} r_{ij} \left[X_{ij} \log \sigma(\Theta_{ij}) + (1-X_{ij}) \log \left[1-\sigma(\Theta_{ij}) \right] \right]$$

(r_{ij}: non-missing ratings; $\sigma(-z) = 1-\sigma(z)$)

$$= \sum_{ij} r_{ij} \left[X_{ij} \log \sigma(\vec{v}_i \cdot \vec{y}_j) + (1-X_{ij}) \log \sigma(-\vec{v}_i \cdot \vec{y}_j) \right]$$

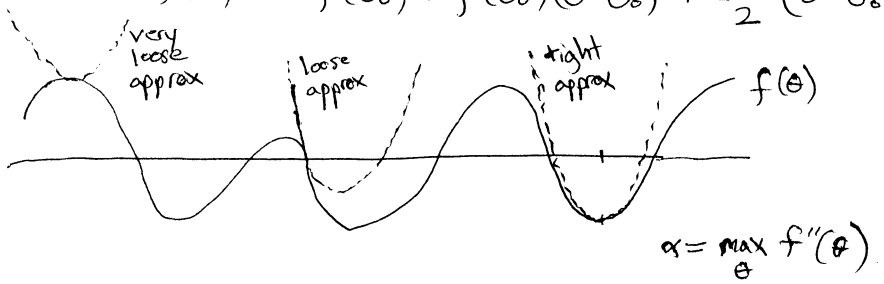
How to maximize w.r.t. $\vec{v}_i$ and $\vec{y}_j$?

- Alternating maximization
  - Optimize $\{\vec{v}_i\}_{i=1}^D$ for fixed $\{\vec{y}_j\}_{j=1}^N$
  - Optimize $\{\vec{y}_j\}_{j=1}^N$ for fixed $\{\vec{v}_i\}_{i=1}^D$
  - Converge to local maximum of $\mathcal{L}$
  - subproblems are logistic regression
  - simpler update rules? Look for auxiliary function.

- Simple inequality

Lemma: for any twice differentiable function $f(\theta)$, let $\alpha \triangleq \max_\theta f''(\theta)$. Then:

$$f(\theta) \le f(\theta_0) + f'(\theta_0)(\theta-\theta_0) + \frac{\alpha}{2}(\theta-\theta_0)^2$$



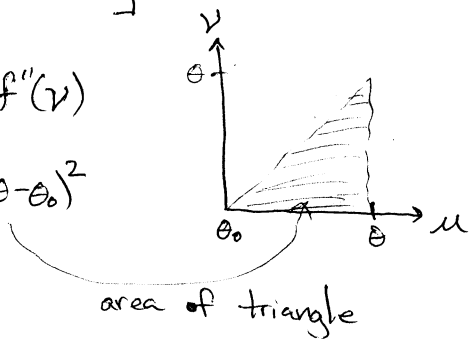
Proof:

$$f(\theta) = f(\theta_0) + \int_{\theta_0}^{\theta} f'(u)\,du$$

$$= f(\theta_0) + \int_{\theta_0}^{\theta} du \left[f'(\theta_0) + \int_{\theta_0}^{u} dv\, f''(v) \right]$$

$$= f(\theta_0) + f'(\theta_0)(\theta-\theta_0) + \int_{\theta_0}^{\theta} du \int_{\theta_0}^{u} dv\, f''(v)$$

$$\le f(\theta_0) + f'(\theta_0)(\theta-\theta_0) + \left[\max_v f''(v)\right] \cdot \frac{1}{2}(\theta-\theta_0)^2$$

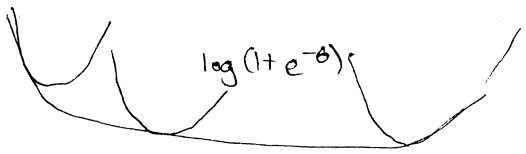


- Apply to negative log-sigmoid function:

$$-\log\sigma(\theta) = \log[1+e^{-\theta}]$$

$$-\frac{d}{d\theta}\log\sigma(\theta) = -\sigma(-\theta)$$

$$-\frac{d^2}{d\theta^2}\log\sigma(\theta) = \sigma(\theta)\sigma(-\theta) \le [\sigma(0)]^2 = \frac{1}{4}$$



$$-\log\sigma(\theta) \le -\log\sigma(\theta_0) - \sigma(-\theta_0)(\theta-\theta_0) + \frac{1}{8}(\theta-\theta_0)^2$$

- Simple auxiliary function for logistic PCA

$$F(v) = -\sum_{ij} r_{ij}\left[x_{ij}\log\sigma(\vec{v}_i\cdot\vec{y}_j) + (1-x_{ij})\log\sigma(-\vec{v}_i\cdot\vec{y}_j)\right]$$

$$G(v,w) = F(w) + \sum_i \frac{\partial F}{\partial \vec{w}_i}(\vec{v}_i-\vec{w}_i) + \frac{1}{8}\sum_{ij} r_{ij}\left[(\vec{v}_i-\vec{w}_i)\cdot\vec{y}_j\right]^2$$

Property (i): $G(v,v) = F(v)$ b/c last terms vanish at $v=w$

Property (ii): $G(v,v) \ge F(v)$ from lemma (and algebra)

- Simpler update rule:

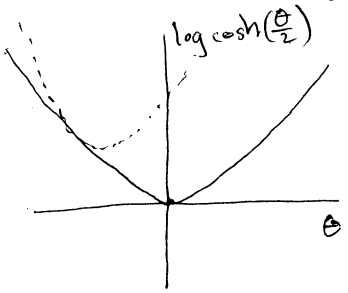
$$v_{t+1} = \underset{v}{\operatorname{argmin}}\, G(v, v_t)$$

Now $G(v,w)$ is quadratic in first argument.

$\min_v G(v,w)$ reduces to least squares optimization

- Tighter inequality for $\log \sigma(\theta)$

Note: $\log \sigma(\theta) = -\log[1+e^{-\theta}]$

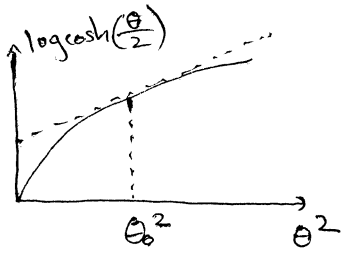


$$= -\log\{e^{-\theta/2} e^{\theta/2}[1+e^{-\theta}]\}$$

$$= \frac{\theta}{2} - \log\left(e^{\theta/2} + e^{-\theta/2}\right)$$

$$= \frac{\theta}{2} - \log\cosh\left(\frac{\theta}{2}\right) + \log 2$$

"like a smooth absolute value function"



- One can show that $\log\cosh\left(\frac{\theta}{2}\right)$ is concave function of θ^2

Bound from concavity:

$$\log\cosh\left(\frac{\theta}{2}\right) \le \log\cosh\left(\frac{\theta_0}{2}\right) + (\theta^2 - \theta_0^2)\frac{\partial}{\partial[\theta^2]}\log\cosh\left(\frac{\theta}{2}\right)\Big|_{\theta=\theta_0}$$

$$= \log\cosh\left(\frac{\theta_0}{2}\right) + (\theta^2 - \theta_0^2)\frac{\tanh(\theta_0/2)}{4\theta_0}$$

- Assembling the above:

$$-\log\sigma(\theta) \le -\frac{\theta}{2} + \log\cosh\left(\frac{\theta_0}{2}\right) + (\theta^2-\theta_0^2)\frac{\tanh(\theta_0/2)}{4\theta_0} - \log 2$$

$\therefore$ (after some algebra)

$$\boxed{-\log\sigma(\theta) \le -\log\sigma(\theta_0) - \sigma(-\theta_0)(\theta-\theta_0) + \frac{\tanh(\theta_0/2)}{4\theta_0}(\theta-\theta_0)^2}$$

Compare to previous bound:

Same except last term $\frac{1}{8}(\theta-\theta_0)^2$

Adaptive quadratic bound
$\Rightarrow$ tighter auxiliary function
$\Rightarrow$ faster convergence

Exponential family PCA

- Exponential family

  Random variable X
  Natural parameter θ

$$\log P(X|\theta) = \log P_0(X) + X\theta - G(\theta)$$

- Examples

Normal $x \in \mathbb{R}$ $P_0(X) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ $G(\theta) = \frac{\theta^2}{2}$ $P(X) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$

Bernoulli $x \in \{0,1\}$ $P_0(X) = 1$ $G(\theta) = \log(1+e^\theta)$ $P(X) = \sigma(\theta)^x \sigma(-\theta)^{1-x}$

Poisson $x \in \{0,1,2,\ldots,\infty\}$ $P_0(X) = \frac{1}{X!}$ $G(\theta) = e^\theta$ $P(X) e^{-\lambda} \left(\frac{\lambda^x}{x!}\right)$ with $\lambda = e^\theta$

(logistic PCA)

- Matrix data

$$\overset{\uparrow D \downarrow}{\left[\quad X \quad\right]}\ \overset{\leftarrow N \rightarrow}{}$$

Assume X_{ij} has exponential family distribution with natural parameter $\theta_{ij} \in \mathbb{R}$

- Matrix factorization

$$\overset{\uparrow D \downarrow}{\left[\quad \Theta \quad\right]}\ \overset{\leftarrow N \rightarrow}{} = \overset{\uparrow D \downarrow}{\left[\quad V \quad\right]}\ \overset{\leftarrow d \rightarrow}{} \left[\quad Y \quad\right] \overset{\uparrow d \downarrow}{} \qquad \theta_{ij} = \vec{v}_i \cdot \vec{y}_j$$

- Log-likelihood

$$\mathcal{L} = \sum_{ij} r_{ij} \log P(X_{ij})$$
$$= \sum_{ij} r_{ij} [\log P_0(X_{ij}) + X_{ij}\theta_{ij} - G(\theta_{ij})]$$
$$= \sum_{ij} r_{ij} [\log P_0(X_{ij}) + X_{ij}(\vec{v}_i \cdot \vec{y}_j) - G(\vec{v}_i \cdot \vec{y}_j)]$$

- Alternating maximization
  - optimize $\vec{v}_i$ for fixed $\vec{y}_j$
  - optimize $\vec{y}_j$ for fixed $\vec{v}_i$
  - repeat until convergence