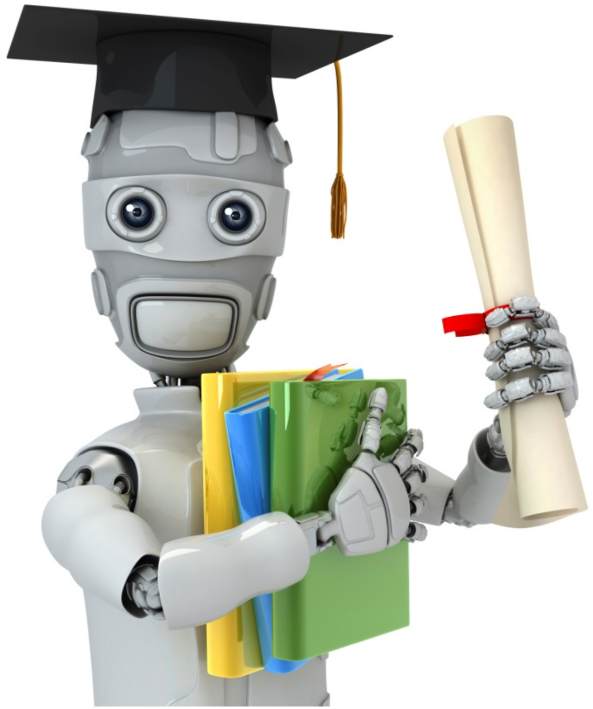


Machine Learning

Recommender Systems

Problem formulation

Example: Predicting movie ratings

- User rates movies using ~~one~~ zero to five stars

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	? 4.5	? 0	0
Cute puppies of love	? 5	4	0	? 0
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	? 4

$n_u = 4$ $n_m = 5$

n_u = no. users

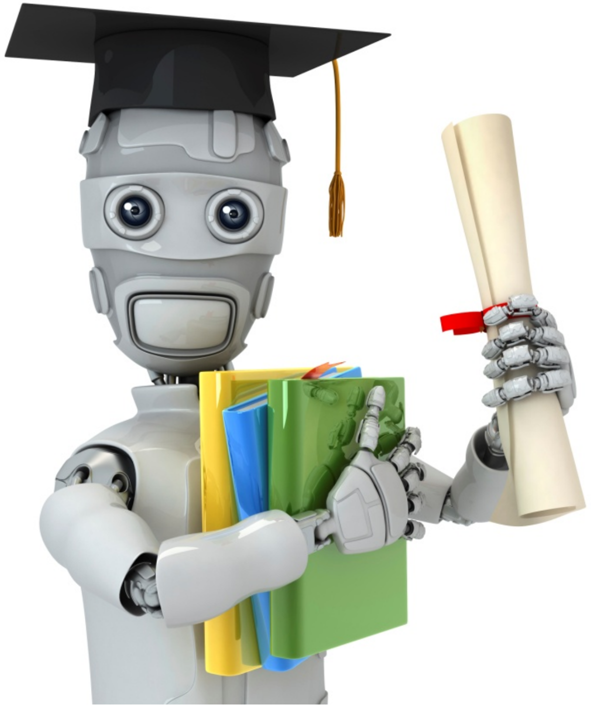
n_m = no. movies

$r(i,j)$ = 1 if user j has rated movie i

$y^{(i,j)}$ = rating given by user j to movie i (defined only if $r(i,j) = 1$)

$0, \ldots, 5$

Andrew Ng



Machine Learning

Recommender Systems

Content-based recommendations

Content-based recommender systems

$n_u = 4$, $n_m = 5$

$\boxed{x_0 = 1}$

$x^{(1)} = \begin{bmatrix} 1 \\ 0.9 \\ 0 \end{bmatrix}$

Movie	Alice (1) $\theta^{(1)}$	Bob (2) $\theta^{(2)}$	Carol (3) $\theta^{(3)}$	Dave (4) $\theta^{(4)}$
$x^{(1)}$ Love at last 1	5	5	0	0
$x^{(2)}$ Romance forever 2	5	?	?	0
$x^{(3)}$ Cute puppies of love 3	? 4.95	4	0	?
Nonstop car chases 4	0	0	5	4
$x^{(5)}$ Swords vs. karate 5	0	0	5	?

$n = 2$

For each user j, learn a parameter $\underline{\theta^{(j)} \in \mathbb{R}^3}$. Predict user j as rating movie i with $\underline{(\theta^{(j)})^T x^{(i)}}$ stars.

$\theta^{(j)} \in \mathbb{R}^{n+1}$

$x^{(3)} = \begin{bmatrix} 1 \\ 0.99 \\ 0 \end{bmatrix} \longleftrightarrow \theta^{(1)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$

$(\theta^{(1)})^T x^{(3)} = 5 \times 0.99$

$= 4.95$

Andrew Ng

Problem formulation

$\rightarrow$ $r(i,j) = 1$ if user j has rated movie i (0 otherwise)

$\rightarrow$ $y^{(i,j)} =$ rating by user j on movie i (if defined)

$\rightarrow$ $\theta^{(j)}$ = parameter vector for user j

$\rightarrow$ $x^{(i)}$ = feature vector for movie i

$\rightarrow$ For user j, movie i, predicted rating: $\underline{\underline{(\theta^{(j)})^T(x^{(i)})}}$

$\theta^{(j)} \in \mathbb{R}^{\overset{\swarrow}{n+1}}$

$\rightarrow$ $\underline{m^{(j)}}$ = no. of movies rated by user j

To learn $\underline{\theta^{(j)}}$:

$$\min_{\theta^{(j)}} \frac{1}{2\cancel{m^{(j)}}} \underline{\sum_{i:r(i,j)=1}} \underline{\left((\theta^{(j)})^T(x^{(i)}) - y^{(i,j)}\right)^2} + \frac{\lambda}{2\cancel{m^{(j)}}} \sum_{\underset{\uparrow}{k=1}}^{n} (\theta_k^{(j)})^2$$

Optimization objective:

To learn $\underline{\theta^{(j)}}$ (parameter for user j):

$$\rightarrow \min_{\theta^{(j)}} \underbrace{\frac{1}{2} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2} + \underbrace{\frac{\lambda}{2} \sum_{k=1}^{n} (\theta_k^{(j)})^2}$$

To learn $\underline{\theta^{(1)}}, \underline{\theta^{(2)}}, \ldots, \underline{\theta^{(n_u)}}$:

$$\min_{\theta^{(1)},\ldots,\theta^{(n_u)}} \frac{1}{2} \boxed{\sum_{j=1}^{n_u}} \underbrace{\sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^{n} (\theta_k^{(j)})^2}_{\Theta^{(1)}, \ldots, \Theta^{(n_u)}}$$

Andrew Ng

Optimization algorithm:

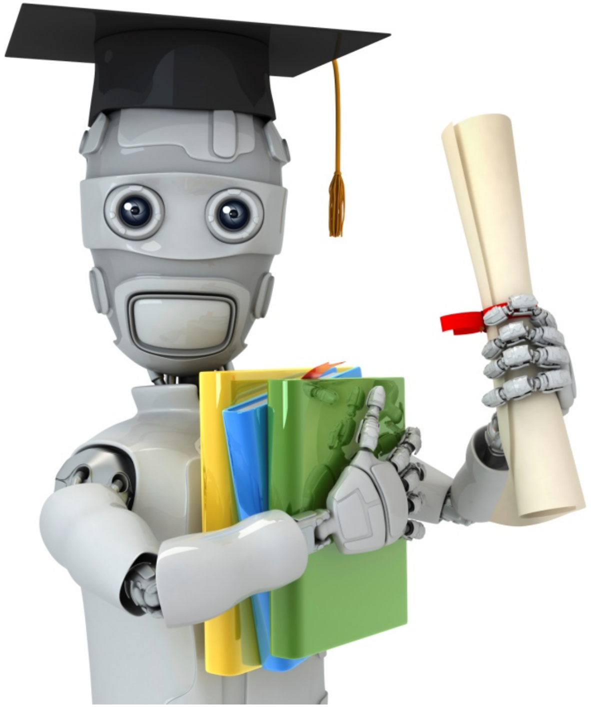
$$\min_{\theta^{(1)},\dots,\theta^{(n_u)}} \underbrace{\frac{1}{2}\sum_{j=1}^{n_u}\sum_{i:r(i,j)=1}\left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2}_{J(\theta^{(1)},\dots,\theta^{(n_u)})}$$

Gradient descent update:

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \sum_{i:r(i,j)=1}((\theta^{(j)})^T x^{(i)} - y^{(i,j)})x_k^{(i)} \quad \underline{(\text{for } k = 0)}$$

$$\theta_k^{(j)} := \theta_k^{(j)} - \alpha \underbrace{\left(\sum_{i:r(i,j)=1} \underline{((\theta^{(j)})^T x^{(i)} - y^{(i,j)})x_k^{(i)}} + \underline{\lambda\theta_k^{(j)}}\right)}_{\frac{\partial}{\partial\theta_k^{(j)}} J(\theta^{(1)},\dots,\theta^{(n_u)})} \quad \underline{(\text{for } k \neq 0)}$$

$$\cancel{\frac{1}{m^{(j)}}}$$



Machine Learning

Recommender Systems

Collaborative filtering

Problem motivation

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)	x_1 (romance)	x_2 (action)
Love at last	5	5	0	0	0.9	0
Romance forever	5	?	?	0	1.0	0.01
Cute puppies of love	?	4	0	?	0.99	0
Nonstop car chases	0	0	5	4	0.1	1.0
Swords vs. karate	0	0	5	?	0	0.9

Problem motivation

$x_0 = 1$

Movie	Alice (1) $\theta^{(1)}$	Bob (2) $\theta^{(1)}$	Carol (3) $\theta^{(3)}$	Dave (4) $\theta^{(4)}$	x_1 (romance)	x_2 (action)
$x^{(1)}$ ~~Love at last~~	5	5	0	0	~~?~~ 1.0	~~?~~ 0.0
Romance forever	5	?	?	0	?	?
Cute puppies of love	?	4	0	?	?	?
Nonstop car chases	0	0	5	4	?	?
Swords vs. karate	0	0	5	?	?	?

$$x^{(1)} = \begin{bmatrix} 1 \\ 1.0 \\ 0.0 \end{bmatrix}$$

$x^{(1)}$

$$\theta^{(1)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}, \quad \theta^{(2)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}, \quad \theta^{(3)} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}, \quad \theta^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}$$

$\theta^{(j)}$

$(\theta^{(1)})^T x^{(1)} \approx 5$

$(\theta^{(2)})^T x^{(1)} \approx 5$

$(\theta^{(3)})^T x^{(1)} \approx 0$

$(\theta^{(4)})^T x^{(1)} \approx 0$

Andrew Ng

Optimization algorithm

Given $\underline{\theta^{(1)},\dots,\theta^{(n_u)}}$, to learn $\underline{x^{(i)}}$:

$$\rightarrow \quad \min_{x^{(i)}} \underbrace{\frac{1}{2} \underbrace{\sum_{j:r(i,j)=1}}_{} \underbrace{\left(\left(\underline{(\theta^{(j)})^T x^{(i)}} - \underline{y^{(i,j)}}\right)^2\right.}_{\uparrow}}_{} + \underbrace{\frac{\lambda}{2}\sum_{k=1}^{n} (x_k^{(i)})^2}_{} \quad \leftarrow$$

Given $\theta^{(1)},\dots,\theta^{(n_u)}$, to learn $\underline{x^{(1)},\dots,x^{(n_m)}}$:

$$\min_{x^{(1)},\dots,x^{(n_m)}} \frac{1}{2} \underset{\uparrow}{\sum_{i=1}^{n_m}} \sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2} \underset{\uparrow}{\sum_{i=1}^{n_m}} \sum_{k=1}^{n} (x_k^{(i)})^2$$

Andrew Ng

Collaborative filtering

$r^{(i,j)}$

$y^{(i,j)}$

Given $\underline{x^{(1)}, \ldots, x^{(n_m)}}$ (and movie ratings),

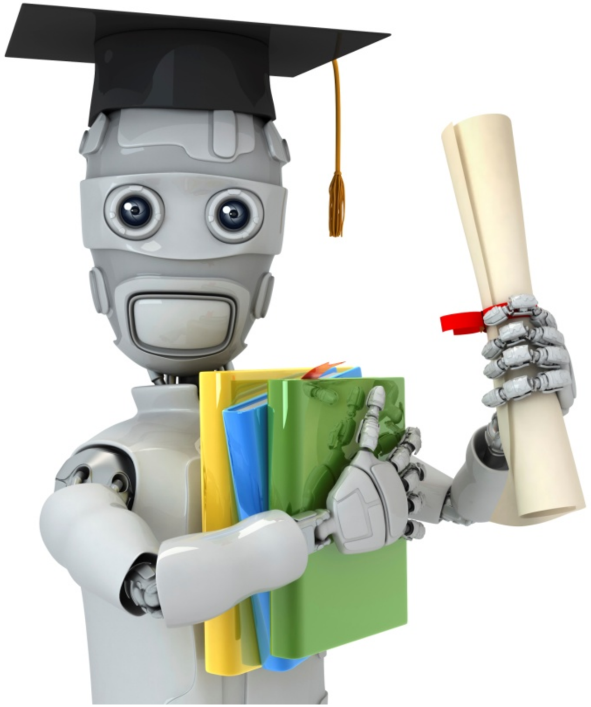
can estimate $\underline{\theta^{(1)}, \ldots, \theta^{(n_u)}}$

Given $\underline{\theta^{(1)}, \ldots, \theta^{(n_u)}}$,

can estimate $\underline{x^{(1)}, \ldots, x^{(n_m)}}$

Guess $\theta \to x \to \theta \to x \to \theta \to x \to \ldots$

Andrew Ng



Machine Learning

Recommender Systems

Collaborative filtering algorithm

Collaborative filtering optimization objective

$(i,j) : r(i,j)=1$

$\cancel{x_0 = 1}$ $\cancel{\theta_0}$ $x \in \mathbb{R}^n$ $\theta \in \mathbb{R}^n$ $\cancel{x \in \mathbb{R}^{n+1}}$ $x_1 = 1$

Given $x^{(1)}, \ldots, x^{(n_m)}$, estimate $\theta^{(1)}, \ldots, \theta^{(n_u)}$:

$$\min_{\theta^{(1)},\ldots,\theta^{(n_u)}} \frac{1}{2}\sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2$$

Given $\theta^{(1)}, \ldots, \theta^{(n_u)}$, estimate $x^{(1)}, \ldots, x^{(n_m)}$:

$$\min_{x^{(1)},\ldots,x^{(n_m)}} \frac{1}{2}\sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(i)})^2$$

Minimizing $x^{(1)}, \ldots, x^{(n_m)}$ and $\theta^{(1)}, \ldots, \theta^{(n_u)}$ simultaneously:

$$J(x^{(1)}, \ldots, x^{(n_m)}, \theta^{(1)}, \ldots, \theta^{(n_u)}) = \frac{1}{2}\sum_{(i,j):r(i,j)=1} \left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(i)})^2 + \frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2$$

$$\min_{\substack{x^{(1)},\ldots,x^{(n_m)} \\ \theta^{(1)},\ldots,\theta^{(n_u)}}} J(x^{(1)}, \ldots, x^{(n_m)}, \theta^{(1)}, \ldots, \theta^{(n_u)})$$

$\theta \rightarrow x \rightarrow \theta \rightarrow x \rightarrow \ldots$

Andrew Ng

~~$x_0 = 1$~~ $\underset{\uparrow}{x} \in \mathbb{R}^n$, $\underset{\uparrow}{\theta} \in \mathbb{R}^n$

Collaborative filtering algorithm

→ 1. Initialize $x^{(1)}, \ldots, x^{(n_m)}, \theta^{(1)}, \ldots, \theta^{(n_u)}$ to small random values.

~~θ_0~~ θ_1 ⋮ θ_n

→ 2. Minimize $J(x^{(1)}, \ldots, x^{(n_m)}, \theta^{(1)}, \ldots, \theta^{(n_u)})$ using gradient descent (or an advanced optimization algorithm). E.g. for every $j = 1, \ldots, n_u, i = 1, \ldots, n_m$:

2

$$x_k^{(i)} := x_k^{(i)} - \alpha \underbrace{\left(\sum_{j:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) \theta_k^{(j)} + \lambda x_k^{(i)} \right)}_{\frac{\partial}{\partial x_k^{(i)}} J(\ldots)} \leftarrow$$

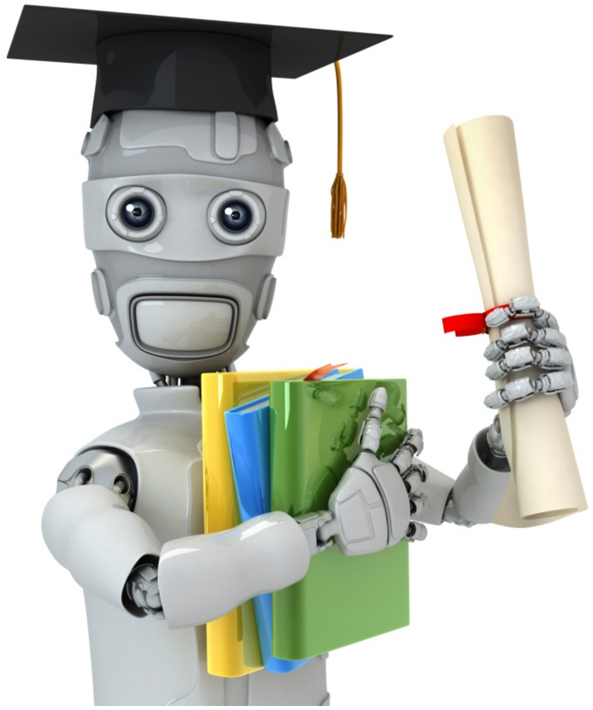
3

$$\theta_k^{(j)} := \underline{\theta_k^{(j)}} - \alpha \underbrace{\left(\sum_{i:r(i,j)=1} ((\theta^{(j)})^T x^{(i)} - y^{(i,j)}) x_k^{(i)} + \lambda \theta_k^{(j)} \right)} \leftarrow$$

3. For a user with parameters $\underline{\theta}$ and a movie with (learned) features $\underline{x}$, predict a star rating of $\underline{\theta^T x}$.

$(\theta^{(j)})^T (x^{(i)})$

Andrew Ng



Machine Learning

Recommender Systems

Vectorization: Low rank matrix factorization

Collaborative filtering

$n_m = 5$
$n_u = 4$

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 \\ 5 & ? & ? & 0 \\ ? & 4 & 0 & ? \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 5 & 0 \end{bmatrix}$$

$\uparrow$ $\uparrow$

$y^{(i,j)}$

Andrew Ng

Collaborative filtering

$X\Theta^T \leftarrow$

$(\Theta^{(j)})^T(x^{(i)})$

$(i,j) \rightarrow$

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 \\ 5 & ? & ? & 0 \\ ? & 4 & 0 & ? \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 5 & 0 \end{bmatrix}$$

Predicted ratings:

$$\rightarrow \underbrace{\begin{bmatrix} (\theta^{(1)})^T(x^{(1)}) & (\theta^{(2)})^T(x^{(1)}) & \dots & (\theta^{(n_u)})^T(x^{(1)}) \\ (\theta^{(1)})^T(x^{(2)}) & (\theta^{(2)})^T(x^{(2)}) & \dots & (\theta^{(n_u)})^T(x^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ (\theta^{(1)})^T(x^{(n_m)}) & (\theta^{(2)})^T(x^{(n_m)}) & \dots & (\theta^{(n_u)})^T(x^{(n_m)}) \end{bmatrix}}$$

$$\rightarrow X = \begin{bmatrix} - (x^{(1)})^T - \\ - (x^{(2)})^T - \\ \vdots \\ - (x^{(n_m)})^T - \end{bmatrix} \qquad \rightarrow \Theta = \begin{bmatrix} - (\Theta^{(1)})^T - \\ - (\Theta^{(2)})^T - \\ \vdots \\ - (\Theta^{(n_u)})^T - \end{bmatrix}$$

$\rightarrow$ Low rank matrix factorization

Andrew Ng

Finding related movies

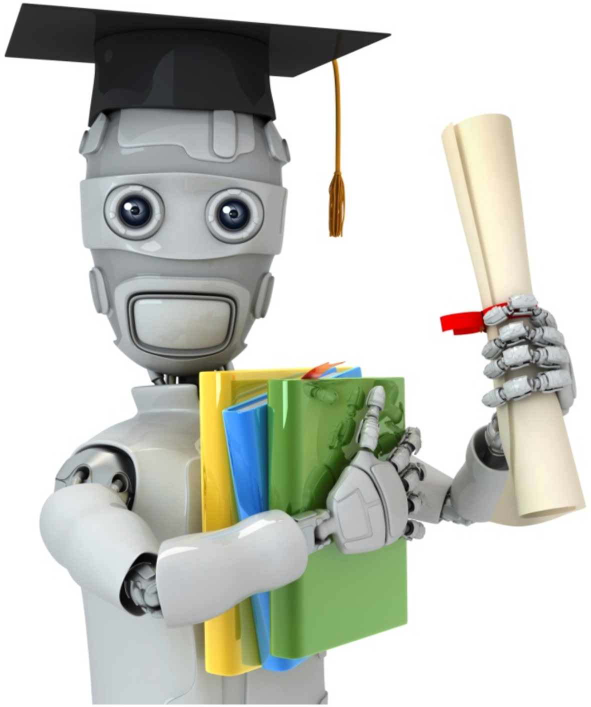
For each product i, we learn a feature vector $\underline{x^{(i)}} \in \mathbb{R}^n$.

$\rightarrow$ x_1 = romance, x_2 = action, x_3 = comedy, x_4 = ...

How to find $\underline{\text{movies } j}$ related to $\underline{\text{movie } i}$?

small $\|x^{(i)} - x^{(j)}\|$ $\rightarrow$ movie j and i are "similar"

5 most similar movies to movie i:
Find the 5 movies j with the smallest $\|x^{(i)} - x^{(j)}\|$.



Machine Learning

Recommender Systems

Implementational detail: Mean normalization

Users who have not rated any movies

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)	Eve (5)	
Love at last	5	5	0	0	?	0
Romance forever	5	?	?	0	?	0
Cute puppies of love	?	4	0	?	?	0
Nonstop car chases	0	0	5	4	?	0
Swords vs. karate	0	0	5	?	?	0

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{bmatrix}$$

$$\min_{\substack{x^{(1)},\dots,x^{(n_m)} \\ \theta^{(1)},\dots,\theta^{(n_u)}}} \underbrace{\frac{1}{2}\sum_{(i,j):r(i,j)=1}\left((\theta^{(j)})^T x^{(i)} - y^{(i,j)}\right)^2}_{} + \frac{\lambda}{2}\sum_{i=1}^{n_m}\sum_{k=1}^{n}(x_k^{(i)})^2 + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{n_u}\sum_{k=1}^{n}(\theta_k^{(j)})^2}_{}$$

$n=2$ $\quad \underline{\theta^{(5)}} \in \mathbb{R}^2 \quad \theta^{(5)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\quad \frac{\lambda}{2}\left[(\theta_1^{(5)})^2 + (\theta_2^{(5)})^2\right]$

$(\underline{\theta^{(5)}})^T \underline{x^{(i)}} = 0$

Andrew Ng

Mean Normalization:

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{bmatrix} \begin{matrix} \leftarrow 2.5 \\ \leftarrow 2.5 \\ \leftarrow 2 \\ \\ \\ \end{matrix} \qquad \mu = \begin{bmatrix} 2.5 \\ 2.5 \\ 2 \\ 2.25 \\ 1.25 \end{bmatrix} \rightarrow \underline{Y} = \begin{bmatrix} 2.5 & 2.5 & -2.5 & -2.5 & ? \\ 2.5 & ? & ? & -2.5 & ? \\ ? & 2 & -2 & ? & ? \\ -2.25 & -2.25 & 2.75 & 1.75 & ? \\ -1.25 & -1.25 & 3.75 & -1.25 & ? \end{bmatrix}$$

$\downarrow$ learn $\underline{\Theta^{(j)}}, \underline{x^{(i)}}$

For user j, on movie i predict:

$$\rightarrow (\Theta^{(j)})^T (x^{(i)}) + \mu_i$$

User 5 (Eve):

$$\underline{\Theta^{(5)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}} \qquad \underbrace{(\Theta^{(5)})^T (x^{(i)})}_{\rightarrow 0} + \boxed{\mu_i}$$