EECS 445

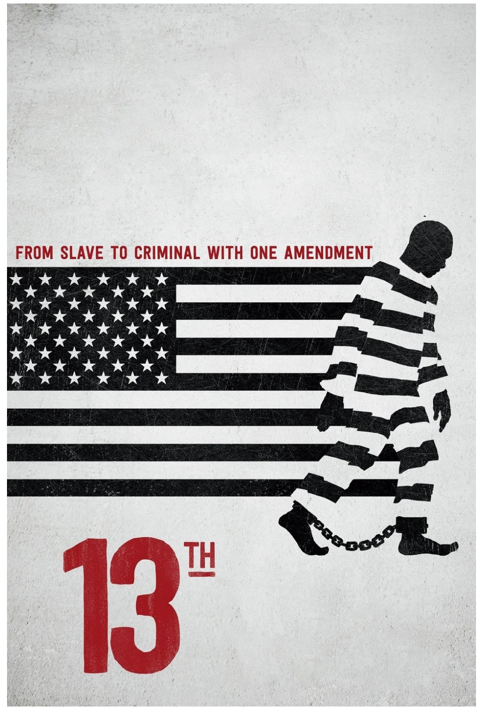
Introduction to Machine Learning

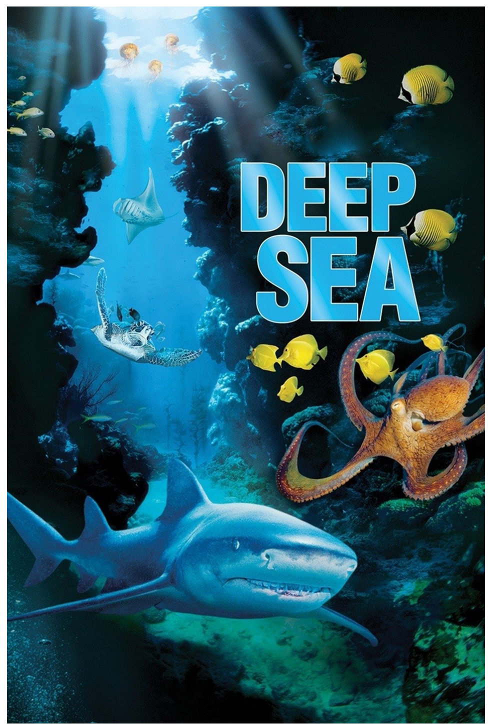
Collaborative Filtering (UV Decomp) and Generative Models

Prof. Kutty

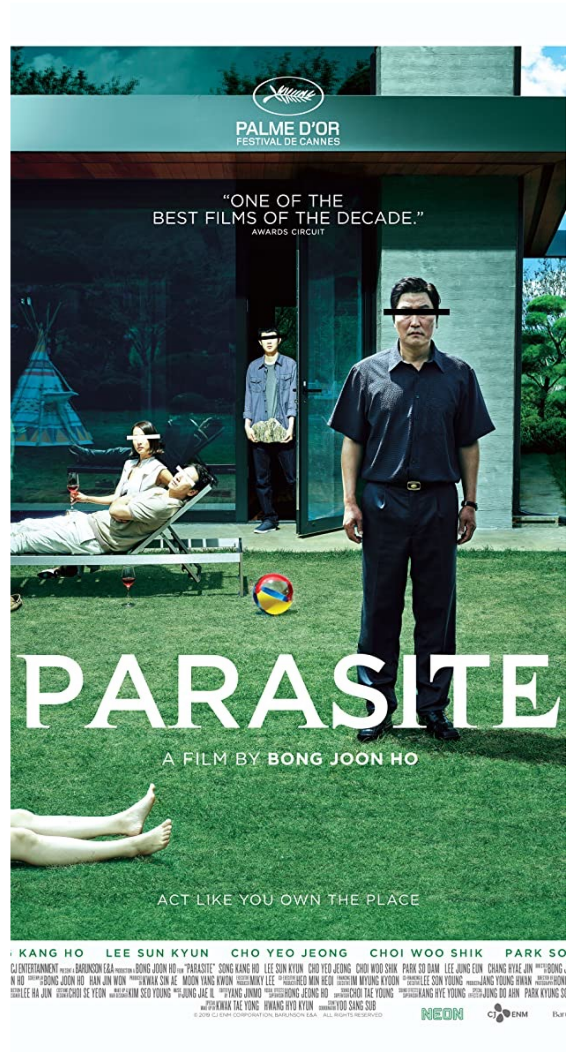
* announcement: no alternate finals
↳ 7pm Apr 25

if you liked





you might like



<u>Steps</u>:

- generate predictions
- pick movies to present to user

Recommendations as Matrix Completion

m items

n users

	5				4
		2	3		
	4				
				4	
1					
	2		3		
	5	1			3

call this the utility (or user-item) matrix Y

How to solve for the missing ratings?

1) Matrix factorization
2) Nearest neighbor prediction

Collaborative Filtering (kNN)

review

Approach 2: Nearest Neighbor Prediction

Key idea:

Suppose user ***a*** has not rated movie ***i***

To predict the rating

- compute *similarity* between user ***a*** and all other users in the system
- find the *k* 'nearest neighbors' of user ***a*** who have rated movie ***i***
- compute a prediction based on these users' ratings of ***i***

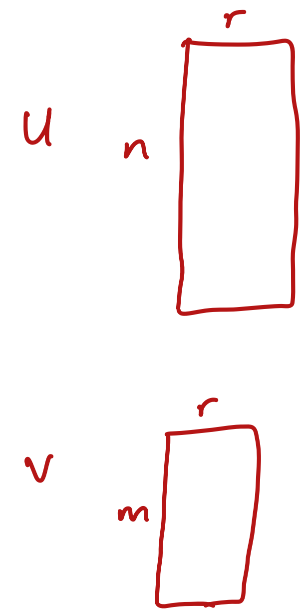
Collaborative Filtering

UV Decomposition

How to solve for the missing ratings?

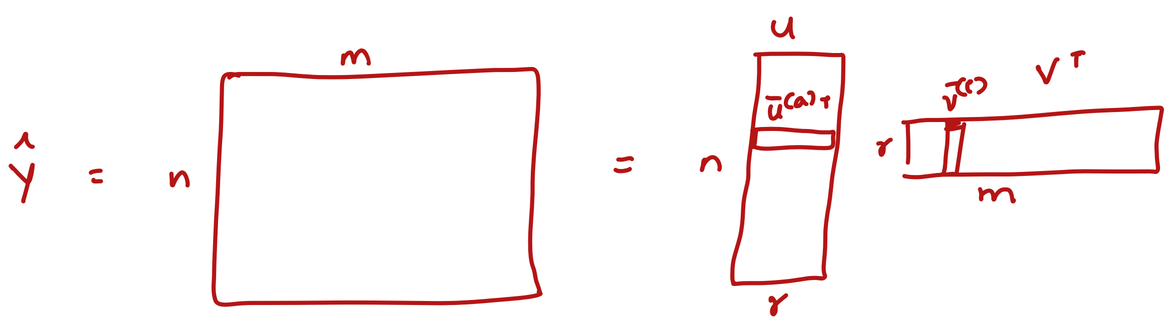
1) Matrix factorization
2) Nearest neighbor prediction

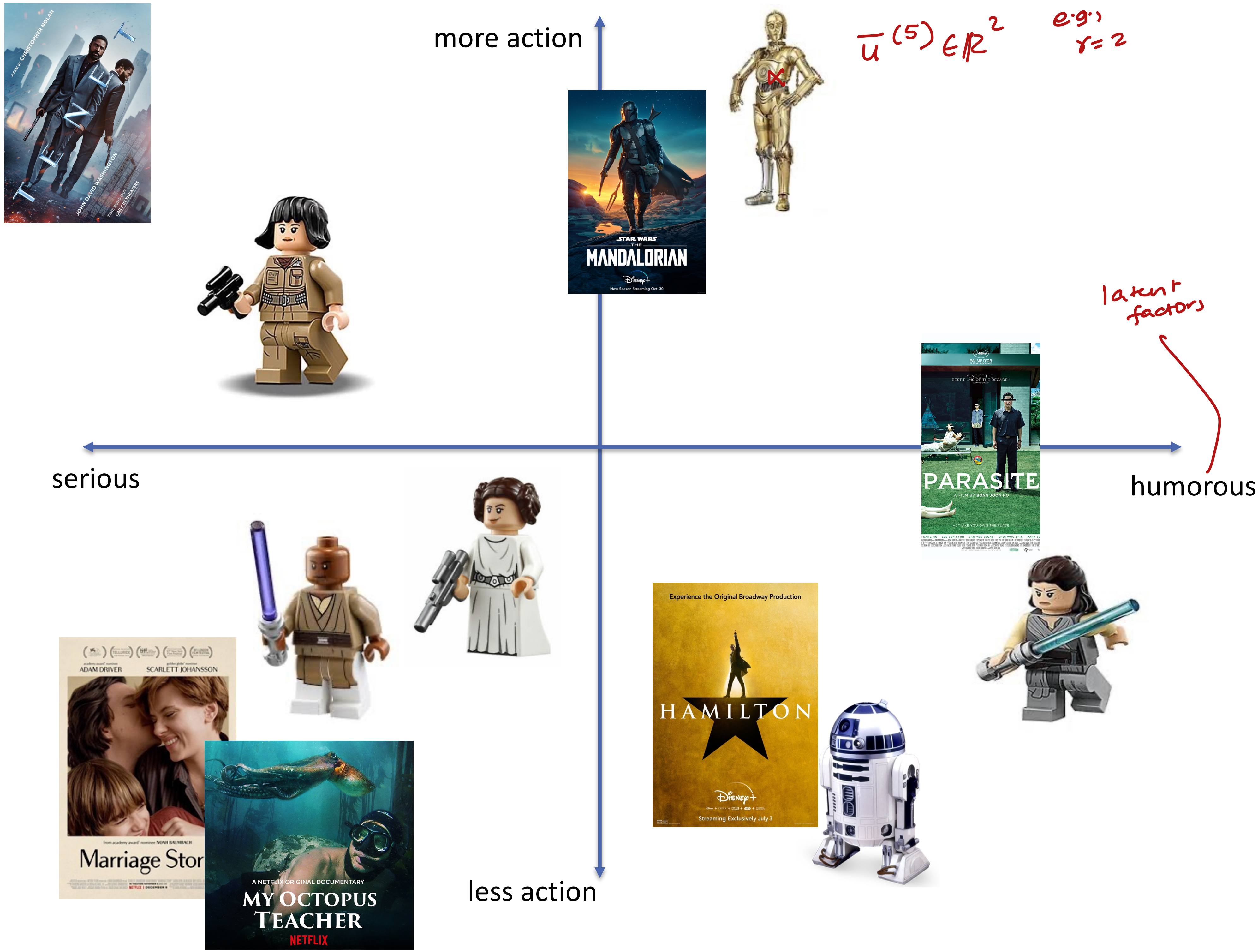
Given $\overset{n\times m}{Y}$ with empty cells



Construct $\underbrace{\textbf{low rank}}_{r}$ $\underbrace{\hat{Y}}_{n\times m} = \underbrace{U}_{n\times r}\,\underbrace{V^T}_{r\times m}$

$r \in \{1, \dots, \min(m,n)\}$




more action
less action
serious
humorous
$\bar{u}^{(5)} \in \mathbb{R}^2$
e.g., r = 2
latent factors
TENET
STAR WARS THE MANDALORIAN
PARASITE
Experience the Original Broadway Production
HAMILTON
Streaming Exclusively July 3
ADAM DRIVER
SCARLETT JOHANSSON
Marriage Stor
A NETFLIX ORIGINAL DOCUMENTARY
MY OCTOPUS TEACHER
NETFLIX

Low-Rank Factorization: example

COMEDY and ACTION are the *latent* factors

R

		The Big Sick	Can You Ever Forgive Me?	The Adam Project	Captain America: Civil War	Black Panther	Captain Marvel
COMEDY	1	1	1	1			
	2	1	1	1			
	3	1	1	1			
BOTH	4	1	1	1	1	1	1
ACTION	5	-1	-1	-1	1	1	1
	6	-1	-1	1	1	1	1
	7	-1	-1	-1	1	1	1

$\approx$

rows are user vectors

U

	COMEDY	ACTION
1	1	0
2	1	0
3	1	0
4	1	1
5	-1	1
6	-1	1
7	-1	1

X

V^T

	The Big Sick	Can You Ever Forgive Me?	The Adam Project	Captain America: Civil War	Black Panther	Captain Marvel
COMEDY	1	1	1	0	0	0
ACTION	0	0	1	1	1	1

columns are item vectors

Example of rank-2 matrix factorization

approximated by

based on example from Aggarwal 2016

Matrix Rank

- Column rank of a matrix $\hat{Y} \in \mathbb{R}^{n \times m}$ is the size of the largest subset of columns of $\hat{Y}$ that constitute a linearly independent set.

- Facts:
  - column rank of $\hat{Y}$ = row rank of $\hat{Y}$ = rank($\hat{Y}$)
  - rank($\hat{Y}$) ≤ min(m, n)

- If rank($\hat{Y}$) = min(m, n) then $\hat{Y}$ is said to be *full rank*

- **Theorem**: Let $\hat{Y} \in \mathbb{R}^{n \times m}$ and $\mathrm{rank}(\hat{Y}) = r$. Then there is $U \in \mathbb{R}^{n \times r}$ and $V^T \in \mathbb{R}^{r \times m}$ such that $\hat{Y} = UV^T$

UV factorization

We may think of Y as being approximated by

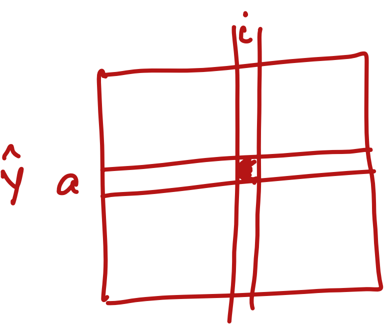
$$\hat{Y} = UV^T$$

where

U contains the relevant features of the user and
V contains the relevant features of the movie

So

$$\hat{Y}_{ai} = [UV^T]_{ai} = \left[\begin{bmatrix} \\ \boxed{\quad \bar{u}^{(a)T} \quad} \\ \\ \end{bmatrix} \begin{bmatrix} \overset{\bar{v}^{(i)}}{\boxed{\ }} & & \end{bmatrix} \right]_{ai} = \bar{u}^{(a)} \cdot \bar{v}^{(i)}$$



in this example
$\hat{Y}_{52} = -1$

1	1	0
2	1	0
3	1	0
4	1	1
5	-1	1
6	-1	1
7	-1	1

U

1	1	1	0	0	0
0	0	1	1	1	1

V^T

$\bar{u}^{(5)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ $\quad \bar{v}^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

Objective Function

D – set of observed entries

D = {(a, i) : Y_ai is non-empty}

Recall that $\hat{Y} = UV^T$

So $\hat{Y}_{ai} = [UV^T]_{ai} = \left[\left[\bar{u}^{(1)}, \ldots, \bar{u}^{(n)}\right]^T \left[\bar{v}^{(1)}, \ldots, \bar{v}^{(m)}\right]\right]_{ai} = \bar{u}^{(a)} \cdot \bar{v}^{(i)}$

$$J(U,V) = \frac{1}{2} \sum_{(a,i)\in D} \left(Y_{ai} - \bar{u}^{(a)} \cdot \bar{v}^{(i)}\right)^2 + \frac{\lambda}{2} \sum_{a=1}^{n} \left\|\bar{u}^{(a)}\right\|^2 + \frac{\lambda}{2} \sum_{i=1}^{m} \left\|\bar{v}^{(i)}\right\|^2$$

Idea: Minimize $J(U,V)$ using coordinate descent

Algorithm Overview

- Initialize “movie” features $\bar{v}^{(1)}, \ldots, \bar{v}^{(m)}$ to small (random) values
- Iterate until convergence

fix $\bar{v}^{(1)}, \ldots, \bar{v}^{(m)}$

solve for $\bar{u}^{(1)}, \ldots, \bar{u}^{(n)}$

$$\min_{\bar{u}^{(a)}} \ \frac{1}{2}\sum_{(a,i)\in D}\left(Y_{ai} - \bar{u}^{(a)} \cdot \bar{v}^{(i)}\right)^2 + \frac{\lambda}{2}\left\|\bar{u}^{(a)}\right\|^2$$

fix $\bar{u}^{(1)}, \ldots, \bar{u}^{(n)}$

solve for $\bar{v}^{(1)}, \ldots, \bar{v}^{(m)}$

$$\min_{\bar{v}^{(i)}} \ \frac{1}{2}\sum_{(a,i)\in D}\left(Y_{ai} - \bar{u}^{(a)} \cdot \bar{v}^{(i)}\right)^2 + \frac{\lambda}{2}\left\|\bar{v}^{(i)}\right\|^2$$

Ridge regression!!

$$J_{n,\lambda}(\bar{\theta}) = \lambda\frac{||\theta||^2}{2} + \frac{1}{n}\sum_{i=1}^{n}\frac{(y^{(i)} - (\bar{\theta}\cdot\bar{x}^{(i)}))^2}{2}$$

Example

Goal: Find **rank 1** $\hat{Y}$. Assume $\lambda = 1$ in the objective function.

Suppose after 1 iteration $U = [6, 2, 3, 3, 5]^T$ and $V = [4, 1, 5]^T$

$Y =$

5		7
	2	
	1	4
4		
	3	6

Fix V find new $\bar{u}^{(1)}$

https://forms.gle/ffiBvNbPjHF8ghi77

$$\min_{\bar{u}^{(a)}} \ \frac{1}{2}\sum_{(a,i)\in D}\left(Y_{ai} - \bar{u}^{(a)} \cdot \bar{v}^{(i)}\right)^2 + \frac{\lambda}{2}\left\|\bar{u}^{(a)}\right\|^2$$

Example

Goal: Find **rank 1** $\hat{Y}$. Assume $\lambda = 1$ in the objective function.

Suppose after 1 iteration $U = [6, 2, 3, 3, 5]^T$ and $V = [4, 1, 5]^T$

$Y =$

5		7
	2	
	1	4
4		
	3	6

Fix V find new $\bar{u}^{(1)}$

$$\min_{\bar{u}^{(1)}} \frac{1}{2} \sum_{(1,i)\in D} \left(Y_{1i} - \bar{u}^{(1)} \cdot \bar{v}^{(i)}\right)^2 + \frac{\lambda}{2} \left\|\bar{u}^{(1)}\right\|^2$$

$$= \min_{\bar{u}^{(1)}} \frac{1}{2}\left(Y_{11} - \bar{u}^{(1)} \cdot \bar{v}^{(1)}\right)^2 + \frac{1}{2}\left(Y_{13} - \bar{u}^{(1)} \cdot \bar{v}^{(3)}\right)^2 + \frac{\lambda}{2}\left\|\bar{u}^{(1)}\right\|^2$$

$$= \min_{\bar{u}^{(1)}} \frac{1}{2}\left(5 - 4\,\bar{u}^{(1)}\right)^2 + \frac{1}{2}\left(7 - 5\,\bar{u}^{(1)}\right)^2 + \frac{\lambda}{2}\left\|\bar{u}^{(1)}\right\|^2$$

Set partial derivative of this expression to 0 and solve for $\bar{u}^{(1)}$

$$\bar{u}^{(1)} \approx 1.3$$

Notice that error $(Y_{11} - [UV^T]_{11})^2$ goes from $(5 - 24)^2$ to $(5 - 5.2)^2$

Related ideas and issues

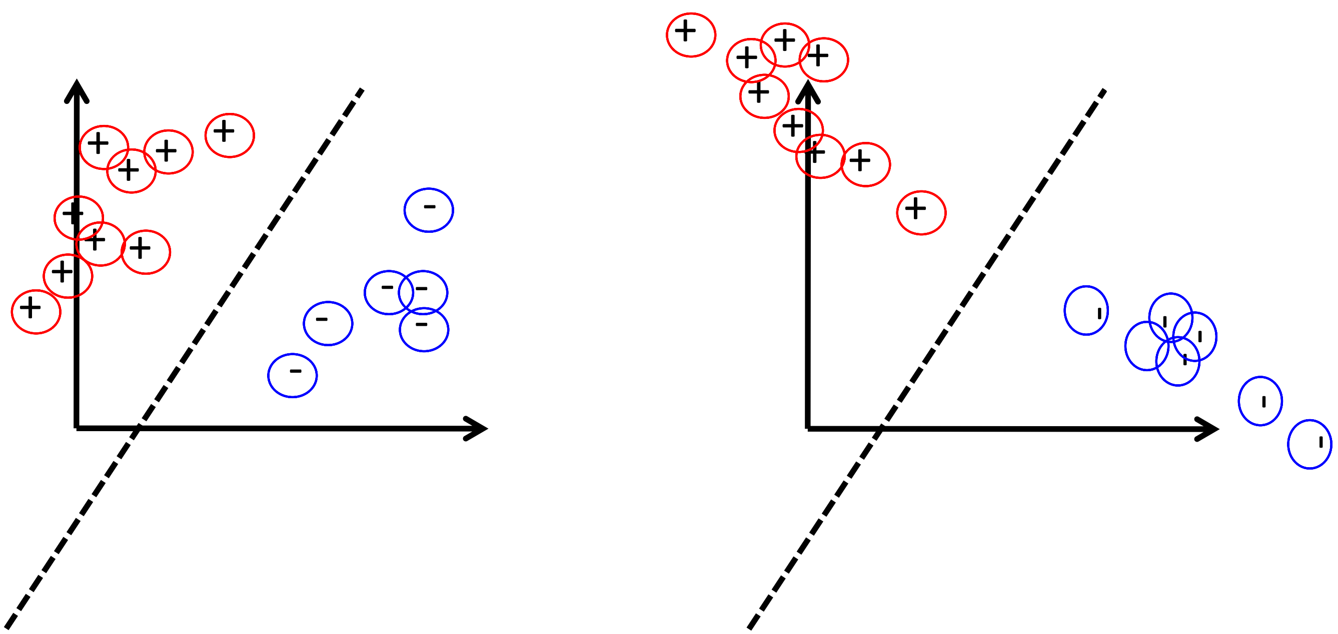
- Context-aware recommender systems
- Cold start problem
- Manipulation in recommender systems

Discriminative vs Generative Models

Discriminative Models

E.g., Classification → learned a separator to discriminate two classes

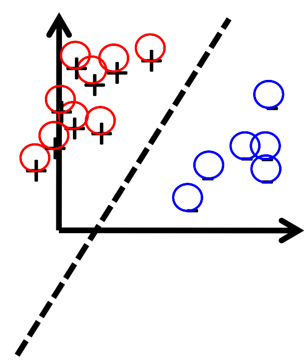
*internal structure of the classes is not captured

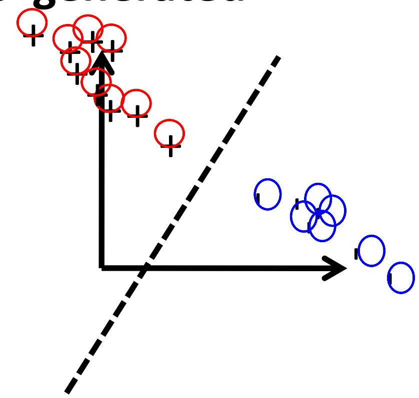


Why do we care about generative models?

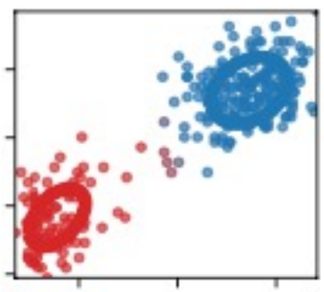
Better understanding of where our data came from; how it was 'generated'

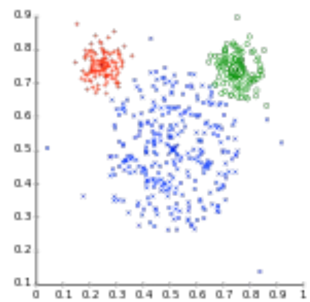
- describes internal structure of the data
- can also be used for classification



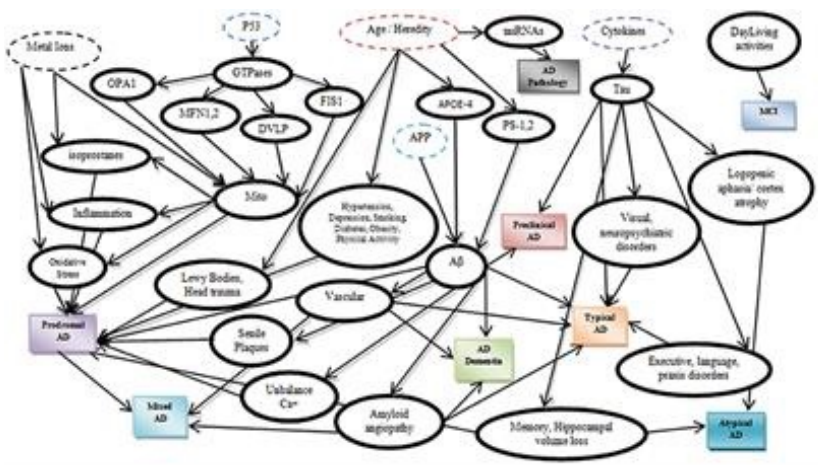


We can use this as a basis for soft clustering

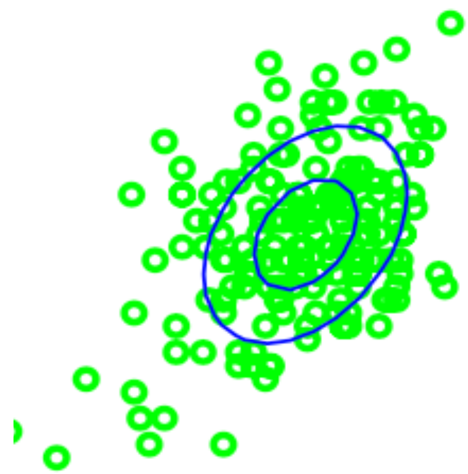




We can use this as a basis for graphical models



Maximum Likelihood Estimation (MLE)



Underlying Distribution for this (unlabeled) Dataset

$x^{(i)}$
0
1
0
0
1
0
0

$Pr(x^{(i)} = 1) = p$

$p_{MLE} = \frac{2}{7}$

We assume data are generated i.i.d. from an unknown Bernoulli distribution that has parameter p

each of these "coin flips" is with the same coin (same bias towards head) and each coin flip is independent of previous flips

generative story with i.i.d. assumption for Bernoulli

Given $S_n = \{x^{(i)}\}_{i=1}^{n}$

Assume

- each $x^{(i)} \sim \text{Bern}(x; p)$ (identically distributed)
  i.e., each $x^{(i)}$ = 1 with probability p and
  $x^{(i)}$ = 0 with probability $1 - p$

- $\forall i \neq j \quad p\left(x^{(i)}, x^{(j)}\right) = \text{Bern}(x^{(i)}; p)\ \text{Bern}(x^{(i)}; p)$ (independently distributed)
  e.g., $p\left(x^{(1)} = 1, x^{(2)} = 0, x^{(3)} = 1, x^{(4)} = 1\right) = p^3(1-p)$

$\underset{p}{p(x^{(1)}=1)}$ $\underset{(1-p)}{p(x^{(2)}=0)}$ $\underset{p}{p(x^{(3)}=1)}$ $\underset{p}{p(x^{(4)}=1)}$

$0 \leq p \leq 1$

Consequently

$$p(S_n) = \prod_{i=1}^{n} p(\bar{x}^{(i)})$$

$\frac{\partial P(S_n)}{\partial p} = 0$

$\hookrightarrow p_{MLE}$

Goal: Determine p

Underlying Distribution for this (unlabeled) Dataset

$x^{(i)}$
0.0002
1110
0.01
710
-1120.09
774.11
3.532

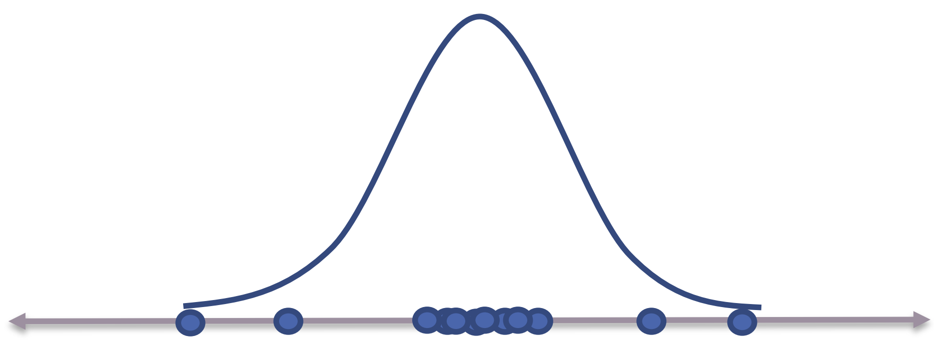
$x^{(i)} \in \mathbb{R}$

Maximum Likelihood Estimate: intuition

We assume data are generated i.i.d. from an unknown Gaussian distribution that has parameter μ, σ^2

each datapoint was drawn from the same ‘bell curve’

Use MLE to determine the *likeliest* parameter values, given the dataset

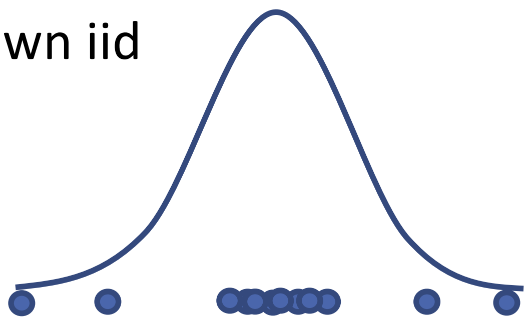


$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

examples: inches of snowfall, heights of people etc.

generative story with i.i.d. assumption for univariate Gaussian

Given $S_n = \{\bar{x}^{(i)}\}_{i=1}^{n}$ drawn iid



Assume

- each $\bar{x}^{(i)} \sim N(\bar{x}|\mu, \sigma^2)$ (identically distributed)
- $\forall i \neq j \quad p\left(\bar{x}^{(i)}, \bar{x}^{(j)}\right) = N(\bar{x}^{(i)}|\mu, \sigma^2)N(\bar{x}^{(j)}|\mu, \sigma^2)$ (independently distributed)

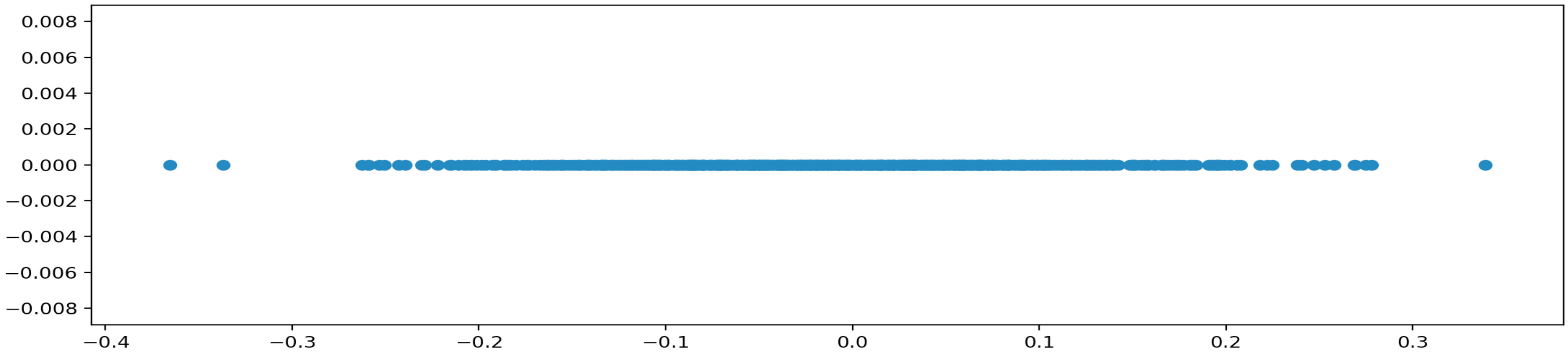
Consequently,

$$p(S_n) = \prod_{i=1}^{n} N(\bar{x}^{(i)}|\mu, \sigma^2)$$

Goal: Determine μ, σ^2

- Want to maximize $p(S_n)$ wrt μ
- Want to maximize $p(S_n)$ wrt σ^2

MLE for the univariate Gaussian



- Given $S_n = \{x^{(i)}\}_{i=1}^{n}$ drawn iid

$$p(S_n) = \prod_{i=1}^{n} p\left(x^{(i)}\right)$$

- Want to maximize $p(S_n)$ wrt μ

$$\mu_{\text{MLE}} = \sum_{i=1}^{n} \frac{x^{(i)}}{n}$$

- Want to maximize $p(S_n)$ wrt σ^2

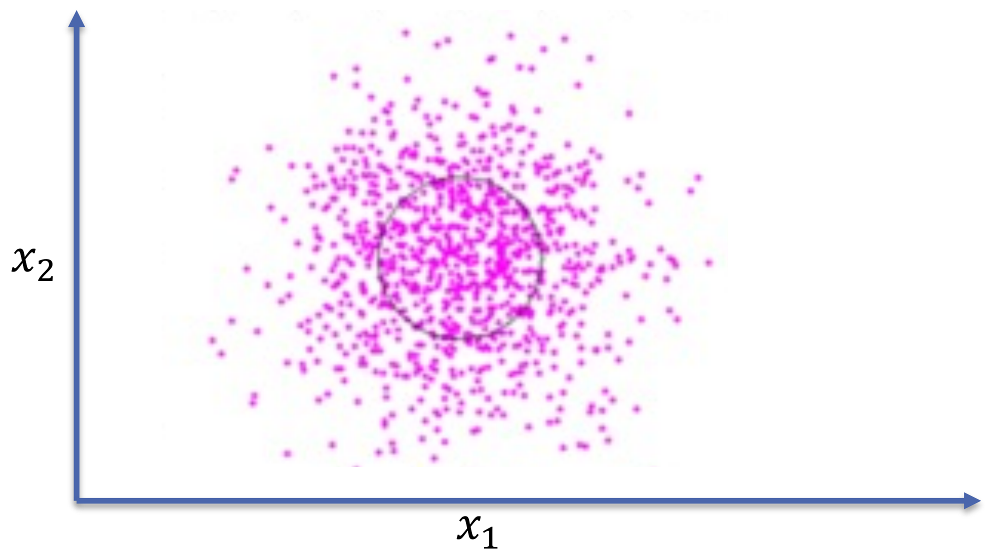
$$\sigma^2_{\text{MLE}} = \sum_{i=1}^{n} \frac{\left(x^{(i)} - \mu_{\text{MLE}}\right)^2}{n}$$

Multivariate Gaussian Distribution

Underlying Distribution for this (unlabeled) Dataset

for $\bar{x} \in \mathbb{R}^d$ $d \geq 2$

Example 1: Here $\bar{x} \in \mathbb{R}^2$



Example 2: Here $\bar{x} \in \mathbb{R}^4$

$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0.0002	10.052	8.602	227
1110	12.110	-805.1	-84.5
0.01	0.01	5292.01	837.1
710	-73610	8015.03	-2.503
-1120.09	11.01	1680	-5686
774.11	3.67	46.86	51.13
3.532	624	587.4	-3700