PCA

From [U,S,V] = svd(Sigma) we get: $U = [u^{(1)}\ u^{(2)}\ \cdots\ u^{(n)}] \in R^{n\times n}$

取 k 个

$x \in R^n \xrightarrow{\text{reduce}} z \in R^k$

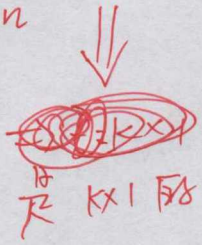
$z = [u^{(1)}\ u^{(2)}\ \cdots\ u^{(k)}]_{n\times k}$

$z = z^T \cdot x = \begin{bmatrix} -(u^{(1)})^T- \\ -(u^{(2)})^T- \\ \vdots \\ -(u^{(k)})^T- \end{bmatrix} \cdot x_{n\times 1}$

$k \times n$

$Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}\cdot(x^{(i)})^T = \frac{1}{m}\cdot X^T\cdot X$

$X = \begin{bmatrix} -(x^{(1)})^T- \\ -(x^{(2)})^T- \\ \vdots \\ -(x^{(m)})^T- \end{bmatrix}$



$U_{reduce} = U(:, 1:k)$

$z = U_{reduce}^T \cdot x$

mathematics.

* our goal: retain as much as possible info. after dimensional reduction
* info: $S_x = Sigma = \frac{1}{n}\cdot X^T\cdot X$ ... covariance matrix
* solve.
* what is a point? $x \in R^D$, $z \in R^M$, $M << D$

linear dimensionality reduction

$X = [x_1, \cdots\ x_n]^T \quad X \in R^{N\times D}$

$z = [z_1, \ldots z_n]^T \quad z \in R^{n\times m}$

$z = U\cdot X$

* : problem: $\max_U\ U^T\cdot S_x\cdot U \quad s.t.\ U^T\cdot U = I.$

即 $S_z = \frac{1}{n}\cdot z^T\cdot z \in R^{M\times M}$.

$\Leftrightarrow \max S_Z = \max \frac{1}{N} \cdot Z^T \cdot Z \Leftrightarrow \max \frac{1}{N} \cdot (Xu)^T (Xu) \Leftrightarrow \max u^T X^T X u$

$\Updownarrow$

$\max u^T \cdot S_X \cdot u$
$s.t.\ u^T \cdot u = I$

$\frac{\delta L}{\delta X} = 0$ 用 Lagrain. $L(u, \lambda) = u^T S_X u + \lambda(I - u^T u)$

即 $S_X \cdot u = \lambda u$ ★

$\Downarrow$

$S_X u_i = \lambda_i u_i$

* SVD Total var $= \sum_{i=1}^{N} \lambda_i$.

Retained var $= \sum_{i=1}^{D} \lambda_i$

$\Rightarrow$ % info $= \frac{\sum_{i=1}^{D} \lambda_i}{\sum_{i=1}^{N} \lambda_i}$ 90%-95%

取 $\lambda_1 > \lambda_2 > \cdots > \lambda_D$ 中的 top M $\Rightarrow$ $u = [u_1 \ldots u_M]^T$

kernel PCA

$\phi: R^D \to R^M$

feature mapping

$x \in R^D$

$\phi(x) \in R^M$

M << D.

max var retained we get

$S_\phi \cdot u = \lambda \cdot u$

$\frac{1}{N} \cdot \phi^T \phi u = \lambda u$.

$u = \frac{1}{\lambda N} \cdot \phi^T \phi u$

2 properties of kernel:

① Symmetry: $k(X_m, X_n) = k(X_n, X_m)$

② PSD: $\sum_m \sum_n V_m V_n \cdot k(X_m, X_n) \geq 0$

$\phi^T(X_m) \phi^T(X_n) = k(X_m, X_n)$

$z = \phi u = \phi \cdot \frac{1}{\lambda w} \cdot \phi^T \phi u = \frac{1}{\lambda w} (\phi \phi^T)(\phi u)$

$z = \frac{1}{\lambda w} \cdot kz \Rightarrow kz = (\lambda w \cdot) z$ 即 $kz_i = \lambda_i z_i$ just diag k.

steps

① determine eignvectors

② sort by eignvectors

③ select top M vectors for z

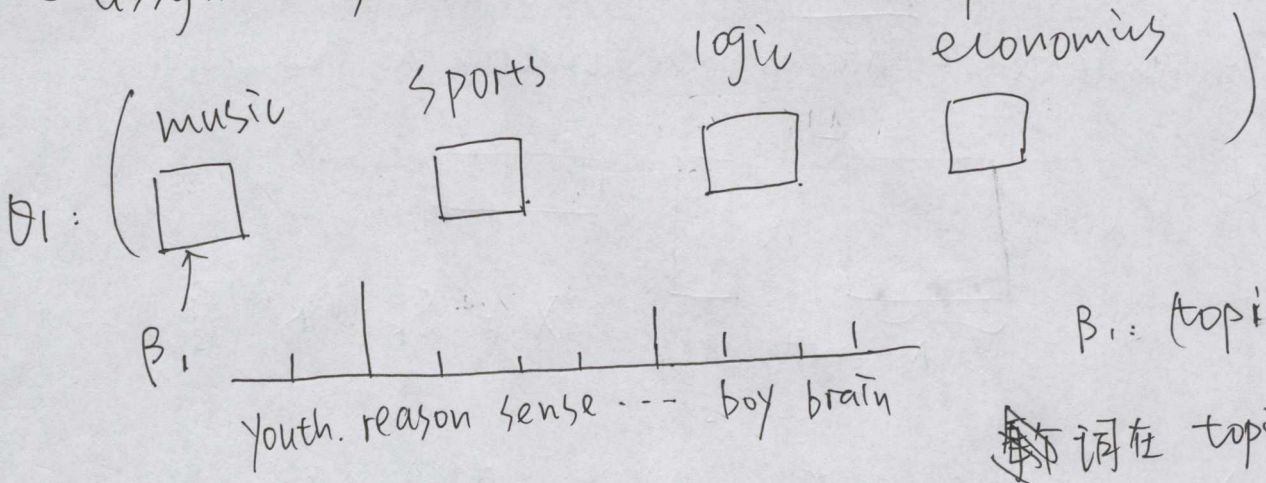
Topic modeling

- a prob topic model:

① learn dist of words ⟶ topics 每个topic是一个词簇

② learn dist of topics for each doc → 学习每个doc的topic分布

③ assign every word in a doc to a topic.



β_1: (topic) = music

~~词~~词在 topic 为 music 上有一个 dist.

θ_1: 一个 document. 含有若干 topics 的 dist.

LDA: 1) generate each topic $\beta_k \sim$ Dirichlet (γ), $k = 1 \cdots k$

2) generate dist on topic for each doc.

$\theta_d \sim$ Dirichlet (α), $d = 1 \cdots D$

3) for the n-th word in the d-th doc:

a) allocate the word to a topic b) generate the word

Dirichlet dist: a continuous dist on discrete prob vector.

Let β_k be a prob vector and γ a positive para vector.

$$p(\beta_k|\gamma) = \frac{\Gamma(\sum_v \gamma_v)}{\prod_{v=1}^{V}\Gamma(\gamma_v)} \prod_{v=1}^{V} \beta_{k,v}^{\gamma_v - 1}$$

$\gamma \uparrow$, var $\downarrow$.

对LDA来说，会选得比较小的 γ、α，来 focus on 很小的 subset of word topics.

LDA output ① topics. ② 每个doc的 topic dist.

Q: for a particular doc, what is $p(X_{dn}=i|\beta, \theta_d)$

A: $p(X_{dn}=i|\beta,\theta) = \sum_{k=1}^{K} p(X_{dn}=i, C_{dn}=k|\beta,\theta_d)$ 某个word分到第k个topic

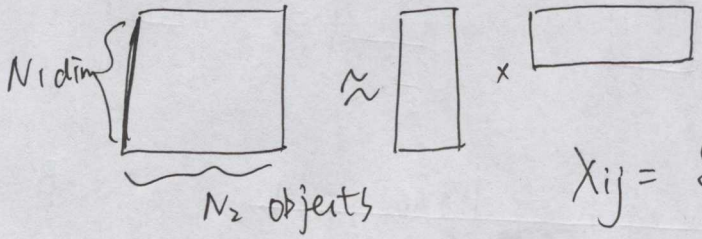
$C_{dn} \sim D(\theta_d)$

generate the word.

$X_{dn} \sim D(\beta_{c_{dn}})$

NMF

- LDA is an instance of non-negative matrix factorization.
- NMF 用处:



$$X_{ij} = \sum_k W_{ik} \cdot H_{kj}$$

- data

text: X_{ij}: # times word i appears in document j.

image: ~~X_{ij}~~: put each vectorized $N \times M$ image of a face on a col of X

- obj function

① $\|X - WH\|^2 = \sum_i \sum_j (X_{ij} - (WH)_{ij})^2$

② $D(X\|WH) = -\sum_i \sum_j [X_{ij} \ln (WH)_{ij} - (WH)_{ij}]$

W, H Non negative values