Problem 1

Wednesday, 22 January 2020 14:22

Problem 1: Use the SVD shown below. Suppose a new user Leslie assigns rating 3 to Alien and

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}}_{M} = \underbrace{\begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}}_{V^T}$$

rating 4 to Titanic, giving us a representation of Leslie in the 'original space' of $[0, 3, 0, 0, 4]$. Find the representation of Leslie in concept space. What does that representation predict about how well Leslie would like the other movies appearing in our example data?

$P = M \cdot V = U \cdot \Sigma$

$[0,3,0,0,4] \cdot V = [7.74, 2.84]$

look at V: sci-fi "classic"

$\Rightarrow$ preference for classic

Problem 2: Consider the latent space distribution

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$$

and a conditional distribution for the observed variable $\boldsymbol{x} \in \mathbb{R}^d$,

$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Phi})$$

where $\boldsymbol{\Phi}$ is an arbitrary symmetric, positive-definite noise covariance variable. Now suppose that we make a nonsingular linear transformation of the data variables $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a non-singular $d \times d$ matrix. If $\boldsymbol{\mu}_{\text{MLE}}$, $\boldsymbol{W}_{\text{MLE}}$, and $\boldsymbol{\Phi}_{\text{MLE}}$ represent the maximum likelihood solution corresponding to the original untransformed data, show that $\boldsymbol{A}\boldsymbol{\mu}_{\text{MLE}}$, $\boldsymbol{A}\boldsymbol{W}_{\text{MLE}}$, and $\boldsymbol{A}\boldsymbol{\Phi}_{\text{MLE}}\boldsymbol{A}^T$ will represent the corresponding maximum likelihood solution for the transformed data set. Finally, show that the form of the model is preserved if $\boldsymbol{A}$ is orthogonal and $\boldsymbol{\Phi}$ is proportional to the unit matrix so $\boldsymbol{\Phi} = \sigma^2\boldsymbol{I}$ (i.e. probabilistic PCA). The transformed $\boldsymbol{\Phi}$ matrix remains proportional to the unit matrix, and hence probabilistic PCA is covariant under a rotation of the axes of data space, as is the case for conventional PCA.

$y = Ax$ with A regular $\hat{=}$ noiseless transformation

$p(x) = \mathcal{N}(\mu, WW^T + \Phi)$

Bishop: $p(x') = \mathcal{N}(x'|\mu', \Lambda^{-1})$

$p(y'|x') = \mathcal{N}(y'|A'x' + b, L^{-1})$

$\Rightarrow p(y') = \mathcal{N}(y'|A'\mu + b, L^{-1} + A'\Lambda^{-1}A'^T)$

$p(y|x) = "\mathcal{N}(y|Ax, 0)"$

$p(y|x) = \mathcal{N}(y|Ax, 0)$

$y' = y, x' = x, \mu' = \mu, \Lambda^{-1} = WW^T + \phi, A' = A, b = 0, L^{-1} = 0$

$p(y) = \mathcal{N}(y|A\mu, A(WW^T + \phi)A^T) =$

$= \mathcal{N}(y|A\mu, AWW^TA^T + A\phi A^T)$

previously: $\mathcal{N}(x|\mu, WW^T + \phi)$

$\mathcal{N}(y|\underset{\downarrow}{A\mu_{MLE}}, \underset{\downarrow}{AW_{MLE}}\underset{\downarrow}{W_{MLE}^T}\underset{\downarrow}{A^T} + \underset{\downarrow}{A\phi_{MLE}A^T})$

$\mathcal{N}(y|\mu_{y,MLE}, W_{x,MLE}W_{y,MLE}^T + \phi_{y,MLE})$

MLE: $\mu_{y,MLE} = A\mu_{MLE}$

$W_{y,MLE} = AW_{MLE}$

$\boxed{\phi_{y,MLE} = A\phi_{MLE}A^T}$

$p_Y(As) = p_X(s) \quad s \in \mathbb{R}^d$

A orthogonal: $\phi_{MLE} = \sigma^2 I$

$\phi_{y,MLE} = A\phi_{MLE}A^T = A\sigma^2 I A^T =$

$= \sigma^2 AA^T = \sigma^2 I = \phi_{MLE}$

$p(y|z) = \mathcal{N}(y|AWz + A\mu, \sigma^2 I)$ (isotropic)

isotropic: $p(Rz) = p(z)$

Problem 3: Let the matrix $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ represent N data points of dimension $D = 10$ (samples stored as rows). We applied PCA to $\boldsymbol{X}$. By using the $K = 5$ top principal components, we transformed/projected $\boldsymbol{X}$ into $\tilde{\boldsymbol{X}} \in \mathbb{R}^{N \times K}$. We computed that $\tilde{\boldsymbol{X}}$ preserves 70% of the variance of the original data $\boldsymbol{X}$.

Suppose now we apply PCA on the following matrices:

a) $\boldsymbol{Y}_1 = \boldsymbol{XS}$ where $\boldsymbol{S} = \lambda \boldsymbol{I}$, with $\lambda \in \mathbb{R}$ and $\boldsymbol{I} \in \mathbb{R}^{D \times D}$ is the identity matrix

b) $\boldsymbol{Y}_2 = \boldsymbol{XR}$ where $\boldsymbol{R} \in \mathbb{R}^{D \times D}$ and $\boldsymbol{RR}^T = \boldsymbol{I}$

c) $\boldsymbol{Y}_3 = \boldsymbol{XP}$ where $\boldsymbol{P} = \text{diag}(+5, -5, \ldots, +5, -5)$ is a $D \times D$ diagonal matrix

d) $\boldsymbol{Y}_4 = \boldsymbol{XQ}$ where $\boldsymbol{Q} = \text{diag}(1, 2, 3, \ldots, D-1, D)$ is a $D \times D$ diagonal matrix

e) $\boldsymbol{Y}_5 = \boldsymbol{X} + \boldsymbol{1}_N \boldsymbol{\mu}^T$ where $\boldsymbol{\mu} \in \mathbb{R}^D$ and $\boldsymbol{1}_N$ is an N-dimensional column vector of all ones

f) $\boldsymbol{Y}_6 = \boldsymbol{XA}$ where $\boldsymbol{A} \in \mathbb{R}^{D \times D}$ and $\text{rank}(\boldsymbol{A}) = 5$

and obtain the projected data $\tilde{\boldsymbol{Y}}_1, \ldots \tilde{\boldsymbol{Y}}_6 \in \mathbb{R}^{N \times K}$ using the principal components corresponding to the top $K = 5$ largest eigenvalues of the respective $\boldsymbol{Y}_i$.

What fraction of variance of each $\boldsymbol{Y}_i$ will be preserved by each respective $\tilde{\boldsymbol{Y}}_i$? *Justify your answer.*

The answer "cannot tell without additional information" is also valid if you provide a justification.

a) $X^TX = U\Sigma U^T$ (wlog X centered, ignore $\frac{1}{n}$)

$S^TX^TXS = U\Sigma U^T \cdot \lambda^2 I = U\lambda^2 \cdot \Sigma U^T$

Variance: $tr(\Sigma)$, $tr(\Sigma) = \sum_i \lambda_i = \sum_{i=1}^{5} \lambda_i / 0.7$

$\Sigma' = \lambda^2 \cdot \Sigma = \sum_i \lambda^2 \cdot \lambda_i = \lambda^2 \cdot \sum_{i=1}^{5} \lambda_i / 0.7$

$\Sigma' = \sum_i \lambda_i' = \sum_i \lambda^2 \lambda_i = \lambda^2 \sum_{i=1}^{5} \lambda_i / 0.7$

b) $Y_2^TY_2 = R^TX^TXR = \underbrace{R^TU}_{V}\Sigma\underbrace{U^TR}_{V^T} = V\Sigma V^T$

$UU^T = I_n$

c) $P = S \cdot \underbrace{D(1, -1, 1, \ldots)}_{\begin{pmatrix} 1 & & \\ & -1 & \\ & & \ddots \end{pmatrix} \Rightarrow \text{orthogonal}} = 5 \cdot D_{1,-1}$

$\begin{pmatrix} 1 & & \\ & 1 & \\ & & \ddots \end{pmatrix}$ ⇒ orthogonal

$Y_3^T Y_3 = S \cdot D_{1,1} X^T X D_{1,1} \cdot S =$

$= 2S \cdot \underbrace{D_{1,-1} U}_{V'} \Sigma \underbrace{U^T D_{1,-1}}_{V'^T} = 2S \cdot V' \Sigma V'^T$

d) $Q = \begin{pmatrix} 1 & & \\ & 2 & \\ & & 3 \\ 0 & & \ddots \end{pmatrix}$ Different scale in each original direction

⇒ can affect eigendecomposition in arbitrary ways

$Q^T X^T X Q =$

$= Q^T U \Sigma U^T Q$

ΣV not orthogonal!

e) $Y_5 = X + \vec{1}\vec{\mu}^T$

PCA: Eigendecomp of $(X - \vec{1}\vec{\mu}^T)^T (X - \vec{1}\vec{\mu}^T)$ (data mean)

Subtract mean ⇒ no effect

f) $Y_6 = X \cdot A$

rank(S) ⇒ $X \cdot A$ in S-dim space

⇒ $A^T X^T X A = V'' \Sigma' V''^T$

$\begin{pmatrix} \sigma'_1 & & & & \\ & \sigma'_2 & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix}$

⇒ 100% of variance

Problem 4: You are given $N = 4$ data points: $\{\boldsymbol{x}_i\}_{i=1}^4, \boldsymbol{x}_i \in \mathbb{R}^3$, represented with the matrix $\boldsymbol{X} \in \mathbb{R}^{4\times 3}$.

$$\boldsymbol{X} = \begin{bmatrix} 4 & 3 & 2 \\ 2 & 1 & -2 \\ 4 & -1 & 2 \\ -2 & 1 & 2 \end{bmatrix}$$

Hint: In this task the results of all (final and intermediate) computations happen to be integers.

a) Perform principal component analysis (PCA) of the data $\boldsymbol{X}$, i.e. find the principal components and their associated variances in the transformed coordinate system. Show your work.

b) Project the data to two dimensions, i.e. write down the transformed data matrix $\boldsymbol{Y} \in \mathbb{R}^{4\times 2}$ using the top-2 principal components you computed in (a). What fraction of variance of $\boldsymbol{X}$ is preserved by $\boldsymbol{Y}$?

c) Let $\boldsymbol{x}_5 \in \mathbb{R}^3$ be a new data point. Specify the vector $\boldsymbol{x}_5$ such that performing PCA on the data including the new data point $\{\boldsymbol{x}_i\}_{i=1}^5$ leads to exactly the same principal components as in (a).

In-class Exercises

Wednesday, 22 January 2020 14:21

$x_i \in \mathbb{R}^D \quad i = 1..N$

find $z_i \in \{1..K\}$ s.t. $z_i = z_j \Rightarrow x_i$ similar to x_j

K centroids $\mu_k \in \mathbb{R}^D$

$$J(Z, \mu) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \| x_i - \mu_k \|_2^2$$

Lloyd's algorithm

1. initialize μ_k
2. find $z = \min_z J(z, \mu)$
3. find $\mu = \min_\mu J(z, \mu)$
4. iterate 2+3 until convergence

Problem 6: Consider a modified version of the K-means objective, where we use L_1 distance instead.

$$\mathcal{J}(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||_1$$

This variation of the algorithm is called K-*medians*. Derive the Lloyd's algorithm for this model.

Step 2

$$z_{ik} = \begin{cases} 1 & \text{if } k = \arg\min_j \| x_i - \mu_j \|_1 \\ 0 & \text{else} \end{cases}$$

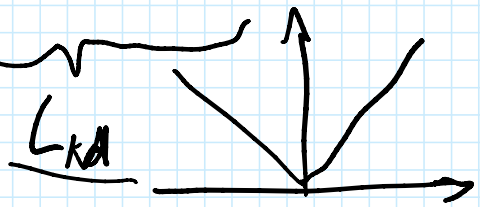
Step 3

$$\mu_k = \arg\min_\mu \sum_{i=1}^{N} z_{ik} \| x_i - \mu_k \|_1$$

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \sum_{d=1}^{D} | x_{id} - \mu_{kd} | = \sum_{k=1}^{K} \sum_{d=1}^{D} \sum_{i=1}^{N} z_{ik} | x_{id} - \mu_{kd} |$$

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \sum_{d=1}^{D} |x_{id} - \mu_{kd}| = \sum_{k=1}^{K} \sum_{d=1}^{D} \sum_{i=1}^{N} z_{ik} |x_{id} - \mu_{kd}|$$

L_{kd}



$$\frac{\partial L_{kd}}{\partial \mu_{kd}} = \sum_{i=1}^{N} z_{ik} \frac{\partial}{\partial \mu_{kd}} |x_{id} - \mu_{kd}| =$$

$$\frac{\partial |x_{id} - \mu_{kd}|}{\partial \mu_{kd}} = \begin{cases} -1 & \text{if } \mu_{kd} > x_{id} \\ 1 & \text{if } \mu_{kd} < x_{id} \\ 0 & \text{if } \mu_{kd} = x_{id} \end{cases}$$

$$= \underbrace{-\sum_{i=1}^{N} z_{ik} \, \mathbb{I}[\mu_{kd} > x_{id}]}_{A} + \underbrace{\sum_{i=1}^{N} z_{ik} \, \mathbb{I}[\mu_{kd} < x_{id}]}_{B} \doteq 0$$

$$A = B$$

$$\frac{\partial L_{kd}}{\partial \mu_{kd}} = 0 \iff \mu_{kd} = \text{median}\{x_{id} \mid z_{ik} = 1\}$$