Manifold Learning and Sparse Representation

Tao Hu
Department of Computer Science
Peking University
taohu@pku.edu.cn

1 Solution List

please click to jump to the corresponding solution.

- Solution 148

  ****

- Solution 190
- Solution 194
- Solution 195
- Solution 196
- Solution 197
- Solution 201
- Solution 229
- Solution 240

2 Linear Algebra

Problem Prove the following three duality relationships between norms:

- $\|\cdot\|_p^* = \|\cdot\|_q, where\ p^{-1} + q^{-1} = 1$
- $\|\cdot\|_2^* = \|\cdot\|_F$
- $\|\cdot\|_F^* = \|\cdot\|_F$

Solution
(a).

$$\|z\|_p^* = sup\{< z, x >: \|x\|_p \leq 1\}$$
$$\sum_{i=1}^n z_i x_i \leq \sum_{i=1}^n |x_i z_i| \leq \|zx\|_1 \leq \|z\|_q \|x\|_p \leq \|z\|_p$$

the above inequality uses Holder's inequality,and the fact that: $\|x\|_p \leq 1$
now ,we will try to find a clever x,which satisfies:$\sum_{i=1}^n z_i x_i = \|z\|_q, and \|x\|_p \leq 1$

$$\text{* firstly, let } x = sign(z) \cdot |z|^{q-1}$$
$$\text{then,} \sum_{i=1}^n z_i x_i = \sum_{i=1}^n z_i \cdot sign(z_i) \cdot |z_i|^{q-1} = \sum_{i=1}^n |z_i|^q = \|z\|_q^q$$

on the other hand,we can get:

$$\|x\|_p^p = \sum_{i=1}^n |x_i|^p = \sum_{i=1}^n |sign(z_i|_i|^{q-1})|^p = \sum_{i=1}^n |z_i|^{p(q-1)} = \sum_{i=1}^n |z_i|^q = \|z\|_q^q$$

let $y = \frac{x}{\|x\|_p}$

$$\sum_{i=1}^n z_i y_i = \sum_{i=1}^n z_i \frac{x_i}{\|x\|_p} = \frac{1}{\|x\|_p} \sum_{i=1}^n \sum_{i=1}^n x_i z_i$$

here :

$$\|x\|_p = (\|x\|_p^p)^{\frac{1}{p}} = (\|z\|_q^q)^{\frac{1}{p}} = \|z\|_q^{\frac{q}{p}}$$

therefore

$$\sum_{i=1}^{n} z_i y_i = \|z\|_q^{\frac{q}{p}} \sum_{i=1}^{n} x_i z_i$$

since $\sum_{i=1}^{n} x_i z_i = \|z\|_q^q$,we can fomulate $\sum_{i=1}^{n} z_i y_i$ into this:

$$\sum_{i=1}^{n} z_i y_i = \|z\|_q^{\frac{q}{p}} \cdot \|z\|_q^q = \|z\|_q^{\frac{-q+pq}{p}} = \|z\|_q^q$$

so ,we find a clever variant y=$\frac{x}{\|x\|_p}$ let $sum_{i=1}^{n} z_i y_i = \|z\|_q, and \|y\|_p \leq 1$ as desired,completing the proof.

(b).

$$\|z\|_2^* = sup\{< z, x >: \|x\|_2 \leq 1\}$$

for the nuclear norm,

$$\|x\|_2 = \sigma_1(x), \|z\|_* = \sum \sigma_i(z)$$

therefore,we seek to prove that:

$sup\{< z, x >: \sigma_1(x) \leq 1\} = sup\{tr(z^T x) : \sigma_1(x) \leq 1\}$ *is really equal to* $\sum_i \sigma_i(z)$?

* first prove that $sup\{< z, x >: \|x\|_2\} \geq \sum_i \sigma_i(z)$

$$letz = U\Sigma V^H = \sum_i \sigma_i U_i V_i^H$$

use SVD above,and define:

$$\hat{Q} = UV^H = UIV^H$$

all singular value of $\hat{Q}$ is 1,and $\sigma_1 \hat{Q} = 1$

$$< \hat{Q}, z >=< UV^H, UIV^H >= tr(VU^H \cdot U\Sigma V^H) = tr(V^H V U^H U \Sigma) = tr(\Sigma) = \sum_i \sigma_i$$

Note the use of the identity tr(ABC) = tr(CAB) ,this is always true when both multiplications are well-posed.

since the supremum cannot be smaller than this single instance,we have :

$$sup\{< z, x >\} \geq sup\{\hat{Q}, x\} = \sum_i \sigma_i$$

* secend,let's proof another direction:

$$\begin{aligned}
&sup\{< z, x >: \sigma_1(x) \leq 1\} \\
&= sup\{tr(x^H \cdot U\Sigma V^H) : \sigma_1(x) \leq 1\} \\
&= sup\{tr(V^H \cdot x^H U\Sigma) : \sigma_1(x) \leq 1\} \\
&= sup\{< UxV^H, \Sigma >: \sigma_1(x) \leq 1\} \\
&= sup\{\sum_i \sigma_i (UxV^H)_{ii} : \sigma_1(x) \leq\} \\
&= sup\{\sum_i \sigma_i U_i x V_i^H : \sigma_1(x) \leq 1\} \\
&\leq sup\{\sum_i^n \sigma_i \sigma_1(x)\} \\
&= \sum_{i=1}^{n} \sigma_i
\end{aligned}$$

the inequality above comes from the fact that $\|U_i\| = 1, \|V_i\| = 1$,and

$$U_i x V_i \leq sup\{U^H x V : \|U\|_2 = 1, \|V\|_2 = 1\} = \sigma_1(z)$$

therefore,we have:

$$sup\{< z, x >\} \leq \sum_i \sigma_i$$

we have the $\leq$ *and* $\geq$,so equality is confirmed.
(c).

$$\|\cdot\|_2 = \|\sigma\|_2$$

σ is the sigular value of matrix,we know $\|\sigma\|_2^* = \|\sigma\|_2$ from (a)

$$\|\cdot\|_F^* = \|\sigma\|_2^* = \|\sigma\|_2 = \|\cdot\|_F$$

Problem
try to proof the following deviations:

- $\frac{\partial(XY)}{\partial t} = \frac{\partial X}{\partial t} Y + \frac{\partial Y}{\partial t} X$, *where X, Y is matrix, t is scalar.*
- $\frac{\partial(a^T x)}{\partial x} = a$, *where a is matrix, x is vector.*
- $\frac{\partial(x^T A x)}{\partial x} = (A + A^T)x$, *where A is matrix, x is vector.*

Solution
(a).

$$\begin{aligned} M_{ij} &= \left(\sum_{k=1}^{p} [\frac{\partial x_{ij}}{\partial t} \cdot y_{kj} + x_{ik} \cdot \frac{\partial y_{jk}}{\partial t}]\right)_{ij} \\ &= \left(\sum_{k=1}^{p} \frac{\partial x_{ij}}{\partial t} \cdot y_{kj} + \sum_{k=1}^{p} x_{ik} \cdot \frac{\partial y_{jk}}{\partial t}\right)_{ij} \\ &= \begin{pmatrix} \frac{\partial X_{11}}{\partial t} & \cdots & \frac{\partial X_{1p}}{\partial t} \\ \vdots & . & \vdots \\ \frac{\partial X_{m1}}{\partial t} & \cdots & \frac{\partial X_{mp}}{\partial t} \end{pmatrix} \cdot Y + X \cdot \begin{pmatrix} \frac{\partial Y_{11}}{\partial t} & \cdots & \frac{\partial Y_{1n}}{\partial t} \\ \vdots & . & \vdots \\ \frac{\partial Y_{p1}}{\partial t} & \cdots & \frac{\partial Y_{pn}}{\partial t} \end{pmatrix} \\ &= \frac{\partial X}{\partial t} Y + \frac{\partial Y}{\partial t} X \end{aligned}$$

(b).
let notate a is m-by-n matrix,x is a n-by-1 vector.

$$a^T x = \begin{pmatrix} \sum_{i=1}^{n} a_{i1} x_i \\ \vdots \\ \sum_{i=1}^{n} a_{im} x_i \end{pmatrix}$$

$a^T x$ is a m-by-1 vector ,therefore$\frac{\partial(a^T x)}{\partial x}$ can be represented as:

$$\frac{\partial(a^T x)}{\partial x} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & \cdots & \cdots & a_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1n} & \cdots & \cdots & a_{nn} \end{pmatrix} = a$$

(c).
let notate x is a n-by-1 vector,and A is a n-by-n matrix.
and

$$x = (x_1, ..., x_n)^T.$$

$$A = \begin{pmatrix} a_{11} & \cdots & a_{nn} \\ \vdots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

$$x^T Ax = \begin{pmatrix} \sum_{i=1}^n a_{i1}x_i & \sum_{i=1}^n a_{i2}x_i & \sum_{i=1}^n a_{i3}x_i & \cdots \sum_{i=1}^n a_{in}x_i \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ v_n \end{pmatrix}$$

$$= \sum_{i=1}^n \sum_{j=1}^n (a_{ij} + a_{ji}) x_i x_j$$

therefore

$$\begin{aligned} \frac{\partial (x^T Ax)}{\partial x} &= [\frac{\partial (x^T Ax)}{\partial x_k}]_k = [\frac{\partial (\sum_{i=1}^n \sum_{j=1}^n (a_{ij}+a_{ji}) x_i x_j)}{\partial x_k}]_k \\ &= [\frac{\partial (\sum_{j=1}^n (a_{kj}+a_{jk}) x_j)}{\partial x_k}]_k \\ &= (A + A^T)x \end{aligned}$$

Problem

$$X \in R^{3*3}, A(X) = X_{11} + X_{12} - X_{31} + 2X_{33}, \textit{solve } A^*$$

Solution

operator A means: elementwise multiply and sum.which mapping $R^{3*3} \rightarrow R^1$,the correspondense matrix is:

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

so ,y is a scalar,and we can get A^* is a operator which mapping $R^1 \rightarrow R^{3*3}$,

$$A^*(x) = x \cdot \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

Problem

Use von Neumann inequality to prove that the solution to

$$\min_{\mathbf{Y}} tr(\mathbf{YKY}^T), \quad s.t. \quad \mathbf{YY}^T = \mathbf{I},$$

is the tailing eigenvectors of $\mathbf{K}$*. The solution to*

$$\max_{\mathbf{Y}} tr(\mathbf{YKY}^T), \quad s.t. \quad \mathbf{YY}^T = \mathbf{I},$$

is the leading eigenvectors of $\mathbf{K}$*. Try whether the above can be deduced by using Lagrange multiplier.*

Solution

TODO

3 Data Geometry

Problem

练习46. 用*Data for MATLAB hackers*（*http://www.cs.nyu.edu/~roweis/data.html*）的*UMist Faces*人脸数据来测试以上*1*、*3*（$p=1$）、*4*、*5*哪个距离相对来说最适合分类，判断准则为：类内样本平均距离/类间样本平均距离。

Solution

run demo.m in the root directory, we can get following result:

- cosine distance: 1.067270
- euclidean distance: 0.764861
- L1 distance: 0.712962
- angular distance: 0.742154

the judge principle is as follows:

$$score = \frac{1}{N}\sum_{i=1}^{N}\frac{ave(\sum_{m\in C_i}\sum_{n\in C_i} distance(m,n))}{ave(\sum_{m\in C_i}\sum_{n\notin C_i} distance(m,n))}$$

where N is the class number,here is 20;ave means the average of the distance sum;distance can be cosine,enclidean,L1,angular distance etc.

and the feature of each face is flattend by the image matrix of each face.

because the score of L1 distance is smallest,so we'd better choose L1 distance in this dataset.

Problem

练习48. *1.* 证明：$\frac{1}{2}\sum_{i,j}W_{ij}\|\mathbf{f}_i-\mathbf{f}_j\|^2 = tr(\mathbf{F}^T\mathbf{L}_W\mathbf{F})$，其中$\mathbf{F}=(\mathbf{f}_1,\cdots,\mathbf{f}_n)$.

2. 请推导$\frac{1}{2}\sum_{i,j}W_{ij}(f_i-f_j)^2$在约束$\sum_i W_{ii}f_i^2=1$下的*Laplacian*矩阵。

Solution

(a).

first,notice that f_i is a scalar.

$$\begin{aligned}&\frac{1}{2}\sum_{i,j}W_{ij}\|f_i-f_y\|^2\\ =&\frac{1}{2}\sum_{ij}^{k}W_{ij}f_i^2+\frac{1}{2}\sum_{ij}^{k}W_{ij}f_j^2-\sum_{ij}^{k}W_{ij}f_if_j\end{aligned}$$

because W is symmetic matrix:$W_{ij}=W_{ji}$

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j} W_{ij}\|f_i - f_y\|^2 \\
=\ &\frac{1}{2}\sum_{ij}^{k} W_{ij}f_i^2 + \frac{1}{2}\sum_{ij}^{k} W_{ij}f_j^2 - \sum_{ij}^{k} W_{ji}f_i f_j \\
=\ &\sum_{ij}^{k} W_{ij}f_i^2 - \sum_{ij}^{k} W_{ji}f_i f_j \\
=\ &f^T D f - f^T W f = f^T(D-W)f \\
=\ &f^T L_W f
\end{aligned}$$

$f^T L_W f$ is a scalar,so we have:

$$f^T L_W f = tr(f^T L_W f)$$

therefore

$$\frac{1}{2}\sum_{i,j} W_{ij}\|f_i - f_y\|^2 = tr(f^T L_W f)$$

(b).
from (a) we can know:

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j} W_{ij}\|f_i - f_y\|^2 \\
=\ &\sum_{ij}^{k} W_{ij}f_i^2 - \sum_{ij}^{k} W_{ji}f_i f_j \\
=\ &\sum_{i}^{k} f_i^2[w_{ii} + \sum_{j=1,j\neq i}^{k} w_{ij}] - \sum_{ij}^{k} W_{ji}f_i f_j
\end{aligned}$$

from the constraint

$$\sum_i W_{ii}f_i^2 = 1$$

we know:

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j} W_{ij}\|f_i - f_y\|^2 \\
=\ &1 + \sum_{j=1,j\neq i}^{k} w_{ij}f_i^2 - \sum_{ij}^{k} W_{ji}f_i f_j \\
=\ &f^T D^- f - f^T W f = f^T(D^- - W)f
\end{aligned}$$

finally, the Laplacian matrix is $D^- - W$
where $D^- = diag(\{d_i^-\}), d_i^- = \sum_{j\neq i} W_{ij}$, W is weight matrix

Problem

于是可以定义目标函数：

$$\frac{1}{2}\sum_{i,j=1}^{k} \pi_i p_{ij}\left(\frac{g_i}{\sqrt{\pi_i}} - \frac{g_j}{\sqrt{\pi_j}}\right)^2, \quad s.t. \quad \|\mathbf{g}\| = 1. \tag{3.34}$$

它可以写成矩阵形式：

$$\mathbf{g}^T(\mathbf{I} - \mathbf{\Theta})\mathbf{g}, \quad s.t. \quad \|\mathbf{g}\| = 1, \tag{3.35}$$

其中$\mathbf{\Theta} = \frac{1}{2}(\mathbf{\Pi}^{\frac{1}{2}}\mathbf{P}\mathbf{\Pi}^{-\frac{1}{2}} + \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{P}^T\mathbf{\Pi}^{\frac{1}{2}})$，$\mathbf{\Pi} = \mathrm{diag}(\pi)$。

Solution

Problem

练习51. 用*Data for MATLAB hackers*（*http://www.cs.nyu.edu/~roweis/data.html*）的*UMist Faces*两个人的人脸数据，计算出图像间的欧氏距离，再用*Gauss*核$\exp(-d/\sigma)$转化为相似度矩阵$\mathbf{W}$，其中σ取作类内距离标准差（方差开根号）的平均值，最后利用*NCut*进行聚类，汇报一下准确率。

Solution

we write two types of PCA: pca.m and normalized_pca.m

we set $\sigma = \frac{1}{2}(\sigma_1 + \sigma_2)$,where σ_i is the standard variance of distance matrix(N-by-N).

and we get the similarity matrix using gaussian kernel: $exp(-d/\sigma)$.

from the deduction of the textbook,we should optimize:

$$min_y\ Ncut(x) = \frac{y^T L_w y}{y^T D_w y}$$

the minimum can be reached by the second smallest eigenvector of $\tilde{L_w}$,from the program I wrote,we can calculate y.

if y_i >0,we think $x_i = 1$,o.w we think $x_i = -1$

I wrote a program named "ncut_binary_classfier.m", the my accuracy is 5

4 Linear Dimensionality Reduction

Problem

练习57. 用*Data for MATLAB hackers*（*http://www.cs.nyu.edu/~roweis/data.html*）的*UMist Faces*一个人的人脸数据做主成分分析和标准化变量的主成分分析，分别显示平均脸和前*6*个主成分。

Solution

notation:

n: data numbers

m:data dimensions

d: finally reduced data dimensions

U:m-by-d,transfer matrix.

$\bar{x}$: the mean of data x

$\bar{\sigma}$: the standard variance of data x.

dimension reduce procedure:

$$y = U^T(x - \bar{x})/\bar{\sigma}$$

primary component recovery procedure:

$$x = \bar{\sigma}Uy + \bar{x}$$

the normal component face is :


face-1.bmp


face-2.bmp


face-3.bmp


face-4.bmp


face-5.bmp


face-6.bmp

the normalized component face is :


normalized_face-1.bmp


normalized_face-2.bmp


normalized_face-3.bmp


normalized_face-4.bmp


normalized_face-5.bmp


normalized_face-6.bmp

Problem

练习76. *Prove that the solution to*

$$\begin{aligned} \min \quad & \|\mathbf{X}-\mathbf{A}\|_F^2 \\ \mathbf{X} &\succcurlyeq \mathbf{0}, \\ \mathit{rank}(\mathbf{X}) &\leq r \end{aligned}$$

is $\mathbf{X}=\mathbf{U}_r \max(\mathbf{\Lambda}_r, 0)(\mathbf{U}^r)^T$*, where* $\mathbf{A}$ *is a symmetric matrix and* $\mathbf{U}_r$ *and* $\mathbf{\Lambda}_r$ *consist of the first* r *eigenvectors and eigenvalues of* $\mathbf{A}$*, respectively.*

Solution

$$A = U\begin{pmatrix} A_r & 0 \\ 0 & A_{n-r} \end{pmatrix}U^T$$

其中U为可逆矩阵，对应着一系列基本变换。因此可以对X进行一系列基本变换得到:

$$X = U\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}U^T \text{ 其中} X_{11} \text{为r*r，} X_{22} \text{为(n-r)*(n-r)}$$

又由于$rank(X) \leq r$并且X为整个优化问题的变量,所以不妨令$X_{12}, X_{21}, X_{22} = 0, X_{11}$为r*r对角矩阵，这样的X依然是满足优化条件的。

$$\begin{aligned} & argmin_X \|X-A\|_F^2 \\ =\ & argmin_X \|U(X_{11}-A_r)U^T\|_F \\ =\ & argmin_X\ tr(U^TU(X_{11}-A_r)) \\ =\ & argmin_X\ tr(X_{11}-A_r) \end{aligned}$$

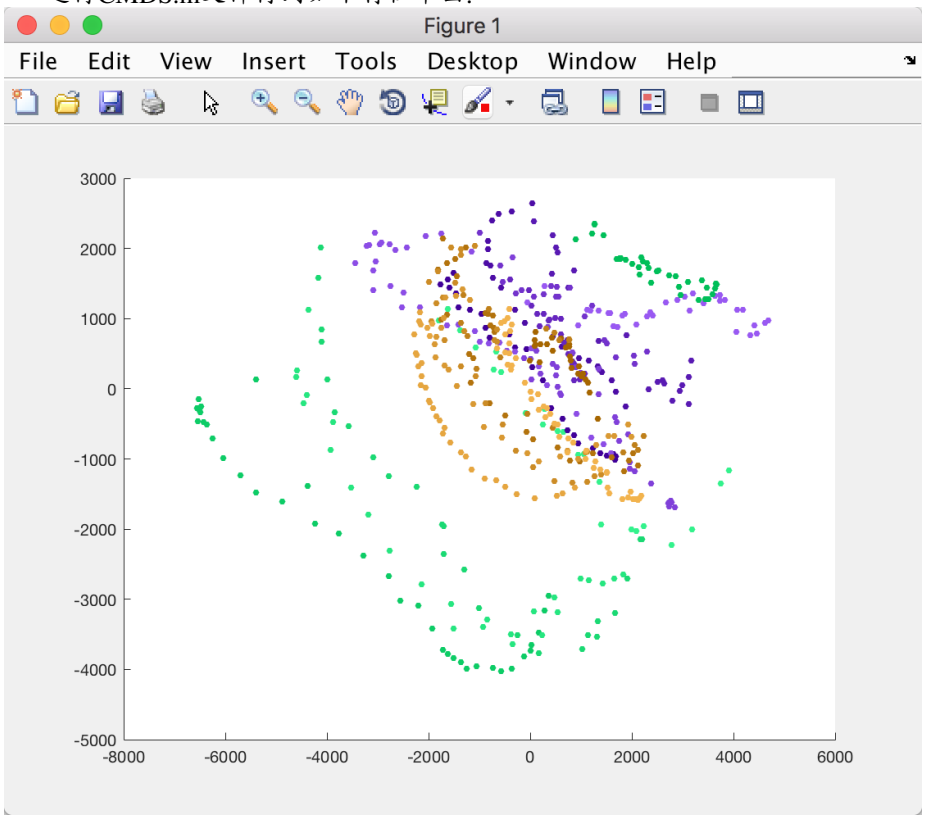
要使上述值最小，并且有X正定，所以可以得知:

$$X = U_r max(Diag_r, 0)(U^r)^T$$

Problem

练习77. *Choose one distance for UMist Faces in Data for MATLAB hackers (http://www.cs.nyu.edu/~roweis/data.html), do CMDS with $k = 2$, and draw the points on a plane. Also report what distance you have used.*

CMDS按照以下步骤执行:
1,构建centering Gram matrix.
2,对G^c进行谱分解.
3,选择合适的d($d \leq r$),求解最终的Y.
运行CMDS.m文件得到如下特征平面:



Problem

练习93. *Apply random projection to UMist Faces in Data for MATLAB hackers (http://www.cs.nyu.edu/~roweis/data.html), use the same recognition algorithm as eigenfaces (reorthogonalized* $\mathbf{R}$ *replacing* $\mathbf{F}^T$ *therein), and report recognition rates.*

Solution

Random Projection有以下步骤:
1,切分train,test数据集，这里train 取475个，test取100个。

2,根据type-1,2,或者type-3构造R矩阵，这里我用了type-3。
3,特征降维,$z = R(Z - \bar{x})$。
4,应用K近邻法预测test数据集中的label。
执行rp.m文件求得准确率为86%.

5 NonLinear Dimensionality Reduction

Problem

课后作业98. *Prove that the function in* (5.14) *is indeed a kernel.*

Solution
Problem

课后作业99. *Use Data for MATLAB hackers* （`http://www.cs.nyu.edu/~roweis/data.html`） *and the polynomial kernel with* $c = 1$ *and* $r = 1, \cdots, 5$ *to repeat the experiment in Chapter PCA and see whether using KPCA can improve the recognition rate and which* r *is the best choice.*

Solution
Problem

课后作业103. *Deduce the solution to* (5.28) *by the method of Lagrange multiplier and compare with the above solution. Using the data in Exercise 104 to check whether they are identical.*

Solution
Problem

课后作业104. *Use one person's face images in Data for MATLAB hackers* （`http://www.cs.nyu.edu/~roweis/data.html`） *as the data for dimensionality reduction with LLE, where the target dimension is 2, and see whether it is better than CMDS and ISOMAP, e.g., whether the face images are arranged better in some order. Different neighborhood sizes and distance measures are encouraged to be tried.*

Solution

Problem

课后作业105. *Use one person's face images in Data for MATLAB hackers* （*http://www.cs.nyu.edu/~roweis/data.html*） *as the data for dimensionality reduction with LTSA, where the target dimension is 2, and see whether it is better than CMDS, ISOMAP, and LLE, e.g., whether the face images are arranged better in some order. Different neighborhood sizes and distance measures are encouraged to be tried.*

Use the above data to check whether the method described in section 5.4.2 is the same as that in section 5.4.3.3, i.e., $\mathbf{Y} = \mathbf{T}$*?*

Solution
Problem

课后作业107. *Use one person's face images in Data for MATLAB hackers* （*http://www.cs.nyu.edu/~roweis/data.html*） *as the data for dimensionality reduction with LE, where the target dimension is 2, and see whether it is better than CMDS, ISOMAP, LLE, and LTSA, e.g., whether the face images are arranged better in some order. Different neighborhood sizes and distance measures are encouraged to be tried.*

Solution
Problem

课后作业109. *Use one person's face images in Data for MATLAB hackers* （*http://www.cs.nyu.edu/~roweis/data.html*） *as the data for dimensionality reduction with Graph-Laplacian type Dmaps, where the target dimension is 2, and see whether it is better than CMDS, ISOMAP, LLE, LTSA, and LE, e.g., whether the face images are arranged better in some order. Different neighborhood sizes and distance measures are encouraged to be tried.*

Solution

Problem

课后作业110. *Use one person's face images in Data for MATLAB hackers（**http://www.cs.nyu.edu/~roweis/data.html**）as the data for dimensionality reduction with MVU, where the target dimension is 2, and see whether it is better than CMDS, ISOMAP, LLE, LTSA, LE, and Dmaps, e.g., whether the face images are arranged better in some order. Different neighborhood sizes and distance measures are encouraged to be tried.*

Solution

6 Convex Analysis

课后作业147.

1. 证明下列函数是凸的：
   - $f_1(x_1,\cdots,x_n)=\begin{cases}-(x_1x_2\cdots x_n)^{1/n}, & \text{if } x_1>0,\cdots,x_n>0,\\ +\infty, & \text{otherwise.}\end{cases}$
   - $f_2(\mathbf{x})=\ln(e^{x_1}+\cdots+e^{x_n})$.
   - $f_3(\mathbf{x})=\|\mathbf{x}\|_2^p$, 其中$p\geq 1$.
   - $f_4(\mathbf{x})=e^{\beta\mathbf{x}^T\mathbf{A}\mathbf{x}}$，其中$\mathbf{A}$为对称半正定矩阵，$\beta>0$.
2. 证明：如果$\alpha_i\geq 0$, $x_i>0$, $\sum\limits_{i=1}^{n}\alpha_i=1$, 那么
   $$x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_n^{\alpha_n}\leq\alpha_1x_1+\alpha_2x_2+\cdots+\alpha_nx_n,$$
   且等号成立当且仅当$x_1=x_2=\cdots=x_n$.
3. 证明定理*123-2.*
4. 证明：方向导数$f'(x;y)$关于y是凸函数。
5. 求向量*2*-范数的次梯度。
6. 证明定理*137-1.*
7. 设$A\succ 0$, 证明：$\frac{1}{2}x^TAx+b^Tx+\frac{1}{2}(y-b)^TA^{-1}(y-b)\geq\langle x,y\rangle$.
8. 求矩阵*(2,0)*范数：$\|(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n)\|_{2,0}=\#\{j|\|\mathbf{x}_j\|_2\neq 0\}$在$C=\{\mathbf{X}|\|\mathbf{X}\|_{2,1}\leq 1\}$上的凸包络。

1. (1).$\frac{\partial^2 f_1}{\partial x_i} \geq 0$

every x_i is equivalent to some extent. so the function is convex.
(2,3)
just acquire the second-order derivation of f,so the function 2, 3 are both convex as 1.
(4). $\frac{\partial^2 f_4(x)}{\partial x} = e^{2\beta x^T Ax}\beta^2 A^2$
since A is semi-positive. so the function is convex.
2
take log function in both side,make use of the convexity of ln(x).
3.
TODO
4.
acquire the second-order derivation can get proof.
5.
TODO
6.
utilize the definition of convex set and convex function can easily get proofed.
(7).
$\frac{1}{2}x^T Ax + b^T x + c$
the conjugate function is:
$f*(y) = \frac{1}{2}(y-b)A^{-1}(y-b) - c$
then utilize the fenchel's inequality:
$f(x) + f*(y) \geq x^T y \forall x, y$

课后作业148.

1. Prove the last equalities in (6.14) *and* (6.15).

2. Find the envelope function and proximal mapping of $f(\mathbf{x}) = \|\mathbf{x}\|_2$.

3. Prove that if $f(\mathbf{x}) = g(\lambda\mathbf{x} + \mathbf{a})$, *where* $\lambda \neq 0$, *then* $P_c f(\mathbf{x}) = \frac{1}{\lambda}\left[P_c(\lambda^2 g)(\lambda\mathbf{x} + \mathbf{a}) - \mathbf{a}\right]$.

4. Prove that if $f(\mathbf{x}) = \lambda g(\mathbf{x}/\lambda)$, *where* $\lambda > 0$, *then* $P_c f(\mathbf{x}) = \lambda P_c(\lambda^{-1} g)(\mathbf{x}/\lambda)$.

5. Let $f(\mathbf{X})$ *be the indicator function on the set* S_+ *of psd matrices. Compute its proximal mapping.*

6. Use Moreau decomposition to prove that $\mathbf{x} = P_L(\mathbf{x}) + P_{L^\perp}(\mathbf{x})$, *where* L *is a subspace and* $L^\perp$ *is its orthogonal complement.*

3,scaling and translation of argument: with $\lambda \neq 0$:
$f(x) = g(\lambda x + a)$, $prox_f(x) = \frac{1}{\lambda}(prox_{\lambda^2 g}(\lambda x + a) - a)$

let $prox_f(x) = argmin_y f(y) + \frac{1}{2}||x-y||_2^2 = m$
we have:
$argmin_y g(\lambda y + a) + \frac{1}{2}||x-y||_2^2 = m$
$\lambda g^{'}(\lambda m + a) + m - x = 0$
on the other side:
$prox_{\lambda^2 g}(\lambda x + a) = argmin_y g(y) + \frac{1}{2\lambda^2}||y - \lambda x - a||_2^2$
derivation:
$g^{'}(y) + \frac{y-\lambda x-a}{\lambda^2}|_{y=\lambda m+a} = 0$
proof done.

4,scalar manipulation: with $\lambda > 0$:
$f(x) = \lambda g(\frac{x}{\lambda}), prox_f(x) = \lambda prox_{\lambda^{-1}g}(\frac{x}{\lambda})$

$f(x) = \lambda g(\frac{x}{\lambda})$
let $prox_f(x) = argmin_y f(y) + \frac{1}{2}||x-y||_2^2 = m$
$argmin_y \lambda g(\frac{y}{\lambda}) + \frac{1}{2}||x-y||_2^2 = m$
$g^{'}(\frac{y}{\lambda}) + y - x|_{y=m} = 0$
$g^{'}(\frac{m}{\lambda}) + m - x = 0$
on the other side:
$prox_{\lambda^{-1}g}(\frac{x}{\lambda}) = argmin_y g(y) + \frac{\lambda}{2}||y - \frac{x}{\lambda}||_2^2$
$g^{'}(y) + \lambda(y - \frac{x}{\lambda})|_{y=\frac{m}{\lambda}} = 0$
$prox_{\lambda^{-1}g}(\frac{x}{\lambda}) = \frac{m}{\lambda}$
proof done.

7 First-Order Sparse Representation

Problem

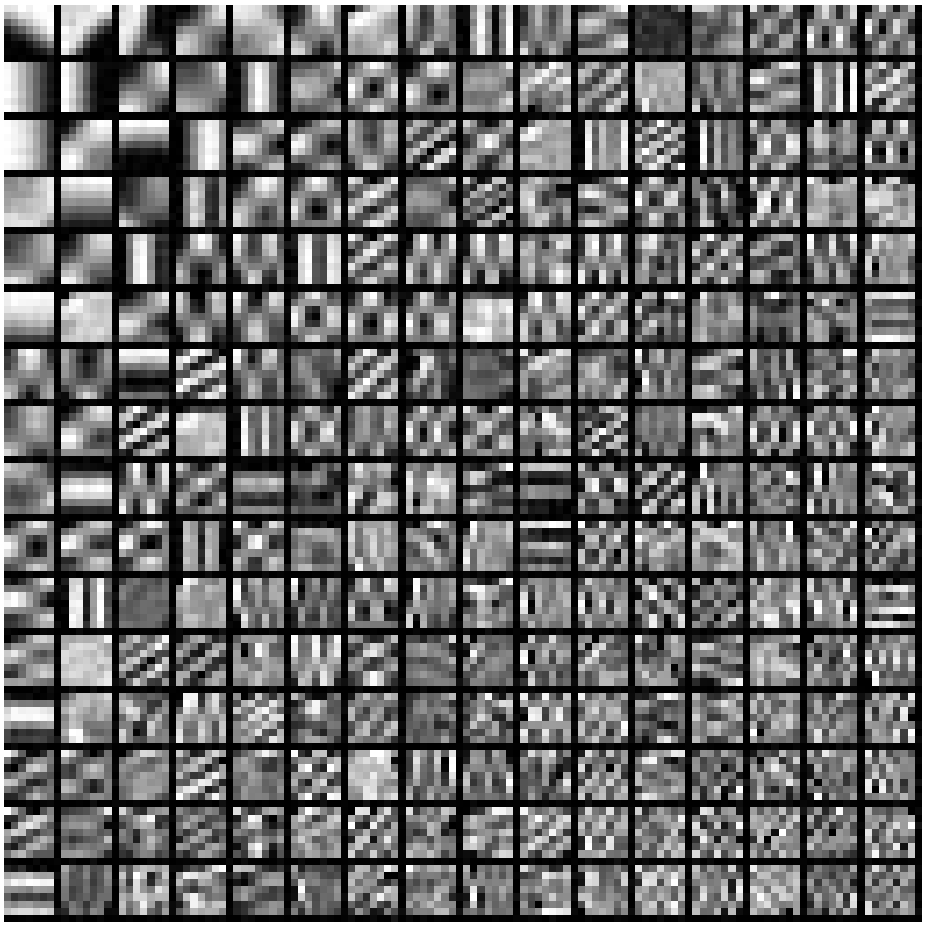
课后作业190. *Use image Lena and K-SVD to learn a 256-atom dictionary for* 8×8 *patches. Order the dictionary from "low frequency" to "high frequency" and rearrange in 2D in the zigzag manner.*

Solution

atom: 256
patch size: 8*8
patch number: 65000
KSVD iteration times: 10

Y: 256*65000
dictionary A:256*64
data X: 64*65000

all image pixels are subtracted by the mean of pixels. the code is in `https://github.com/dongzhuoyao/sparserepresentation.git`
the final dictionary rearrange in 2D in the zigzag manner is :



---

for Basis Sorting Algorithm[2], n is data number, m is base amount.(we need to sort the m bases.) S is 1-m vector just responsible for the number recording. I^i is 1-m vector recording the indice. and $U^{(x_i)}$ is a set of base vector.

KSVD procedure:

Algorithm 7 The K-SVD dictionary-learning algorithm.

Initialize: $k = 0$, build $\mathbf{A}_{(0)} \in \mathbb{R}^{n \times m}$, either by using random entries, or using m randomly chosen examples, normalize the columns of $\mathbf{A}_{(0)}$.

while If the change in $\|\mathbf{Y} - \mathbf{A}_{(k)}\mathbf{X}_{(k)}\|_F^2$ is not small enough **do**

Sparse Coding: Use a pursuit algorithm to approximate the solution of

$$\hat{\mathbf{x}}_i = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{y}_i - \mathbf{A}_{(k-1)}\mathbf{x}\|^2, \quad s.t. \quad \|\mathbf{x}\|_0 \leq k_0,$$

obtaining sparse representations $\hat{\mathbf{x}}_i$ for $1 \leq i \leq M$. These form the matrix $\mathbf{X}_{(k)}$.

K-SVD Dictionary-Update: Use the following procedure to update the columns of the dictionary and obtain $\mathbf{A}_{(k)}$:

Repeat for $j_0 = 1, 2, \cdots, m$,

1. Define the group of examples that use the atom $\mathbf{a}_{j_0}$,

$$\Omega_{j_0} = \{i | 1 \leq i \leq M, \mathbf{X}_{(k)}[j_0, j] \neq 0\}.$$

2. Compute the residual matrix

$$\mathbf{E}_{j_0} = \mathbf{Y} - \sum_{j \neq j_0} \mathbf{a}_j \mathbf{x}_j^T,$$

where $\mathbf{x}_j^T$ are the j'th rows in the matrix $\mathbf{X}_{(k)}$.

3. Restrict $\mathbf{E}_{j_0}$ by choosing only the columns corresponding to Ω_{j_0}, and obtain $\mathbf{E}_{j_0}^R$.

4. Apply SVD on $\mathbf{E}_{j_0}^R = \mathbf{U}\Sigma\mathbf{V}^T$. Update the dictionary atom $\mathbf{a}_{j_0} = \mathbf{u}_1$, and the representations by $\mathbf{x}_{j_0}^R = \Sigma(1,1) \cdot \mathbf{v}_1$.

5. **Update k:** Increase k by 1.

end while

Ensure: Output $\mathbf{A}_{(k)}$.

Bases Sorting procedure:

Algorithm 8 Bases Sorting (BS) algorithm

Input: training data matrix $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, dictionary $U = (\mathbf{u}_1, \cdots, \mathbf{u}_m)$, and parameter $\epsilon > 0$.

Initialize the accumulation buffer: $\mathbf{s} = (0, \cdots, 0)$.

for $i = 1$ to n **do**

Calculate the sparsest representation of $\mathbf{x}_i$ by solving

$$\mathbf{v}_i = \arg\min_{\mathbf{v}} \|\mathbf{v}\|_1, \quad s.t. \quad \|\mathbf{x}_i - U\mathbf{v}\|_2^2 < \epsilon. \tag{7.86}$$

Identify the support of sparse representation vector $\mathbf{v}_i$ and the set of active bases:

$$\begin{aligned} I^{(i)} &= \{j \mid v_{ij} \neq 0,\ j = 1, \cdots, m\}, \\ U^{(\mathbf{x}_i)} &= \{\mathbf{u}_j | j \in I^{(i)}\}. \end{aligned} \tag{7.87}$$

Compute the magnitude vector $\mathbf{a}_i$ of active coefficients, i.e., $a_{ij} = |v_{ij}|$.

Sort the activated bases in $U^{(\mathbf{x}_i)}$ according to the descending order of the magnitude vector $\mathbf{a}_i$:

$$\pi(U^{(\mathbf{x}_i)}) = \text{sort}(I^{(i)}, \mathbf{a}_i). \tag{7.88}$$

Update the accumulation buffer: $s_j = s_j + 1$ for all $j \in I^{(i)}$.

end for

Compute the averaged order by

$$\pi(U) = \sum_{i=1}^{n} \pi(U^{(\mathbf{x}_i)}) \ ./\mathbf{s}, \tag{7.89}$$

where ./ is element-wise division.

Sort U in an ascending order of $\pi(U)$.

8 Structured Sparsity

Group Lasso[1]

Problem

课后作业194. *Prove that the overlapped group Lasso* $\Omega^{\mathcal{G}}_{overlap}(\mathbf{w})$ *is a norm.*

Solution

positive homogeneity and positive definiteness hold trivially. we show the triangular inequality. Consider $w, w^{'} \in R^p$; let $(v_g)_{g\in G} and (v_g^{'})_{g\in G}$ be respectively optimal decompositions of w and $w^{'}$ so that$\Omega^G_{overlap}(w) = \sum_g \|v_g\|$ and $\Omega^G_{overlap}(w^{'}) = \sum_g \|v_g^{'}\|$.since $(v_g + v_g^{'})_{g\in G}$ is a priori non-optimal decomposition of $w + w^{'}$,we clearly have $\Omega^G_{overlap}(w + w^{'}) \leq \sum_{g\in G} \|v_g + v_g^{'}\| \leq \sum_g (\|v_g\| + \|v_g^{'}\|) = \Omega^G_{overlap}(w) + \Omega^G_{overlap}(w^{'})$

Problem

课后作业195. *Prove that*

$$\Omega^{\mathcal{G}}_{overlap}(\mathbf{w}) = \sup\{\boldsymbol{\alpha}^T\mathbf{w} | \boldsymbol{\alpha} \in \mathbb{R}^k, \|\boldsymbol{\alpha}_g\| \leq 1, \forall g \in \mathcal{G}\}.$$

Solution

the overlap lasso is :

$$\Omega^{G}_{overlap}(w) = inf_{v \in v_g, \sum_{g\in G}, v_g = w} \sum_{g\in G} \|v_g\|$$

Let us introduce slack variables $t = (t_g)_{g\in G} \in R^G$ and rewrite the optimization problem as follows:

$$min_{t\in R^G, v\in v_G} \sum_{g\in G} v_g = w \quad and \quad \forall g \in G, \|v_g\| \leq t_g$$

we can form a Lagrangian for this problem with the dual variables $\alpha \in R^p$ for the constraint $\sum_{g\in G} v_g = w$,and $(\beta, \gamma) \in V_G X R^G$ with $\|\beta_g\| \leq \gamma_g$ for the conic constraints $\|v_g\| \leq t_g$ and get:

$$L = \sum_{g\in G} t_g + \alpha^T(w - \sum_{g\in G} v_g) - \sum_{g\in G}(\beta_g^T v_g + \gamma_g t_g)$$

the minimum of L with respect to the primal variables t and v is non trivial only if $\gamma_g = 1$ and $\alpha_g = \beta_g$ for any $g \in G$.Therefore, we get the dual function:

$$min_{t,v} L = \begin{cases} \alpha^T w & if \quad \gamma_g = 1 \quad and \quad \alpha_g = -\beta_g \quad for \quad all \quad g \in G \\ -\infty & otherwise \end{cases} \tag{1}$$

by strong duality(since e.g. Slater's condition is fulfilled),the optimal value $\Omega^G_{overlap}(w)$ of the primal is equal to the maximum of the dual problem. maximizing this dual function over $\gamma_g = 1, \|\beta_g\| \leq \gamma_g$ and $\alpha_g = -\beta_g$ is equivalent to maximizing $a^T w$ over the vectors $\alpha \in R^p$ such that $\|\alpha_g\| \leq 1 \forall g \in G$.which proves the problem.

课后作业196. *Consider 2D DCT transform. We know that for image patches, their high frequencies are more likely to be zeros than low frequencies are. So if we want to recover an image patch, we want the high frequencies to be zeros* **before** *the low frequencies. Then how to design a group sparsity regularizer on such a prior? Please refer to Figure 7.8(b). The entries on the i-th anti-diagonal are called of frequency i. Then this prior means that if $i < j$ and entries of frequency i are zeros, then entries of frequency j are also zeros.*

课后作业197. *Explain why the unit balls of $\Omega^{\mathcal{G}}_{group}(\cdot)$ and $\Omega^{\mathcal{G}}_{overlap}(\cdot)$ for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ are as the left and the right figures in Figure 8.10, respectively.*

9 Second Order Sparsity: Linear Models

命题201. *The function $\|\mathbf{D}\,\mathrm{Diag}(\mathbf{w})\|_*$ is a norm w.r.t. the vector $\mathbf{w}$, provided that none of the columns of $\mathbf{D}$ is equal to $\mathbf{0}$.*

课后作业202. *Prove Proposition 201.*

定理228 ([84])**.** *Let $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD of data matrix $\mathbf{D}$. The optimal solution to*

$$\min_{\mathbf{A},\mathbf{Z}} \|\mathbf{Z}\|_* + \frac{\alpha}{2}\|\mathbf{D}-\mathbf{A}\|_F^2, \quad s.t. \quad \mathbf{A} = \mathbf{A}\mathbf{Z}, \tag{10.68}$$

is given by $\mathbf{A}^* = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^T$ and $\mathbf{Z}^* = \mathbf{V}_1\mathbf{V}_1^T$, where $\mathbf{\Sigma}_1$, $\mathbf{U}_1$, and $\mathbf{V}_1$ correspond to the top $r = \mathrm{argmin}_k \left(k + \frac{\alpha}{2}\sum_{i>k}\sigma_k^2\right)$ singular values and singular vectors of $\mathbf{D}$, respectively.

课后作业229. *Prove Theorem 228.*

命题238. *If f satisfies the EBD conditions (1), (2), and (3) on Ω, then also on $\Omega_1 \subseteq \Omega$, where $\Omega_1 \neq \emptyset$.*

命题239. *Let $\{f_i\}_{i=1}^m$ be a set of functions. If f_i satisfies the EBD conditions (1), (2), and (3) on Ω_i, $\forall i$, and $\cap_{i=1}^m \Omega_i \neq \emptyset$, then $\sum_{i=1}^m \lambda_i f_i$ also satisfies the EBD conditions on $\cap_{i=1}^m \Omega_i$, where $\lambda_i > 0$, $\forall i$.*

课后作业240. *Prove Propositions 238 and 239.*

10 Second Order Sparsity: Nonlinear Models

References

[1] Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

[2] Chun-Guang Li, Zhouchen Lin, and Jun Guo. Bases sorting: Generalizing the concept of frequency for over-complete dictionaries. *Neurocomputing*, 115:192–200, 2013.