HW1 Solution

18340013 陈琮昊

3.2

a）聚类的目的是为了使得同类的样本尽可能相似，即同类样本距离尽可能近，等价于同类样本距离"类中心"样本尽可能近，即同类样本紧凑、聚成团。直观来看需要去寻找能代表类i的"表示"(样本)，构造出各样本的类标记μ_i，得到式(3.5)的最优化形式。

b)

i.

对于固定的 μ_i,有$J(\gamma,\mu)=\sum_{i=1}^{K}\sum_{j=1}^{M}\gamma_{ij}||x_j-\mu_i||^2$

对于每个j, 下列式子中正好有一个是非零的：

$$\gamma_{1j}||x_j-\mu_1||^2,\gamma_{2j}||x_j-\mu_2||^2,\ldots,\gamma_{Kj}||x_j-\mu_K||^2$$

取 $\gamma_{cj}=1(\text{其中}\,c=\arg\min_i||x_j-\mu_i||^2)$, 也就是说，以最小距离为指标将$x_j$分配到组$c$

$$||x_j-\mu_c||^2$$

ii.

对于固定的 γ, 对于类i修改其代表μ_i为:

$$\mu_i=\frac{\sum_{j=1}^{M}\gamma_{ij}x_j}{\sum_{j=1}^{M}\gamma_{ij}}$$

即将当前类的平均向量作为代表值。

iii.回到步骤i迭代。

c） 将M个样本划分为k类，至多有M^k种划分方式。这是一个很大但是有限的数。算法每次迭代的过程中，总是基于`old clustering`进行更新，如果`new clustering`不同于`old clustering`，说明`new clustering`比`old clustering`有更小的类间距离。又由于算法在有限域中迭代，且环的大小只能为1(否则出现一个`clustering`比自身的类间距离小)，所以算法必定在有限步内终止。

4.6

a)

$E[(y-f(x;D))^2] = E_D[(F(x)-f(x;D)+\epsilon)^2]$
$= E_D[(F(x)-f(x;D))^2 + \epsilon^2 + 2\epsilon(F(x)-f(x;D))]$
$= E_D[(F(x)-f(x;D))^2] + E[\epsilon^2] + 2E_D[\epsilon(F(x)-f(x;D))]$
因为ϵ与其他都独立，故下面两式成立：
$E[\epsilon^2] = (E[\epsilon])^2 + Var(\epsilon) = \sigma^2$
$E_D[\epsilon(F(x)-f(x;D))] = E_D[F(x)-f(x;D)]E[\epsilon] = 0$
代入：$E[(y-f(x;D))^2] = E_D[(F(x)-f(x;D))^2] + \sigma^2$
$= (E_D[F(x)-f(x;D)])^2 + Var(F(x)-f(x;D)) + \sigma^2$
因为 $E_D[F(x)-f(x;D)] = F(x) - E_D[f(x;D)]$
且 $Var(F(x)-f(x;D)) = Var(f(x;D)) = E_D[(f(x;D)-E_D[f(x;D)])^2]$
所以 $E[(y-f(x;D))^2] = (F(x)-E_D[f(x;D)])^2 + E_D[(f(x;D)-E_D[f(x;D)])^2] + \sigma^2$
$= bias^2 + variance + noise$

即$bias^2 = (F(x)-E_D[f(x;D)])^2$，$variance = E_D[(f(x;D)-E_D[f(x;D)])^2]$，$noise = \sigma^2$

b)

$E[f] = E[\frac{1}{k}\sum_{i=1}^{k} y_{nn}(i)] = E[\frac{1}{k}\sum_{i=1}^{k} F(x_{nn}(i)) + \epsilon]$
$= E[\frac{1}{k}\sum_{i=1}^{k} F(x_{nn}(i))] + E[\epsilon] = \frac{1}{k}E[\sum_{i=1}^{k} F(x_{nn}(i))] = \frac{1}{k}\sum_{i=1}^{k} F(x_{nn}(i))$

c)

$E[(y-f(x;D))^2] = (F(x)-E[f])^2 + E[(f-E[f])^2] + \sigma^2$
$= (F - \frac{1}{k}\sum_{i=1}^{k} F(x_{nn}(i)))^2 + E[\frac{1}{k}\sum_{i=1}^{k}(y_{nn}(i) - F(x_{nn}(i)))^2] + \sigma^2$
$= (F - \frac{1}{k}\sum_{i=1}^{k} F(x_{nn}(i)))^2 + E[\frac{1}{k}\sum_{i=1}^{k}(\epsilon_{nn}(i))^2] + \sigma^2$

d）方差项是$E[\frac{1}{k}\sum_{i=1}^{k}(\epsilon_{nn}(i))^2]$，它随着$k$变大而变小。

e）偏置的平方项是$(F - \frac{1}{k}\sum_{i=1}^{k} F(x_{nn}(i)))^2$，它随着$k$变大而变大。

4.9

a）记C^T矩阵的元素记为c_{ij}^T。总代价为:$\sum_{j=1}^{K}\sum_{i=1}^{K} c_{ij} * a_{ij}$=$\sum_{j=1}^{K}\sum_{i=1}^{K} c_{ji}^T * a_{ij}$,由矩阵乘法的定义可知，该值为$C^T A$矩阵上所有对角线元素之和，即tr($C^T A$)。

b）我更喜欢规范化的混淆矩阵，因为和为1看起来比较方便，能够更直观的解释一些问题。

5.1

a）XX^T 的特征值为σ_i^2 ($i = 1, 2, \ldots\ldots m$),特征向量为U的每一列。

b）$X^T X$ 的特征值为σ_i^2 ($i = 1, 2, \ldots\ldots n$),特征向量为V的每一列（V^T的每一行）。

c）相等，证明：

假设$XX^T b = \lambda b$，λ为XX^T的特征值，b为XX^T的特征向量。
两边同时左乘X^T得：$X^T XX^T b = \lambda X^T b$,
令$X^T b = b'$,则有$X^T Xb' = \lambda b'$,故$X^T X$的特征值也为λ,但二者的特征向量不同。

d）X的奇异值为XX^T （$X^T X$）的特征值的算术平方根。

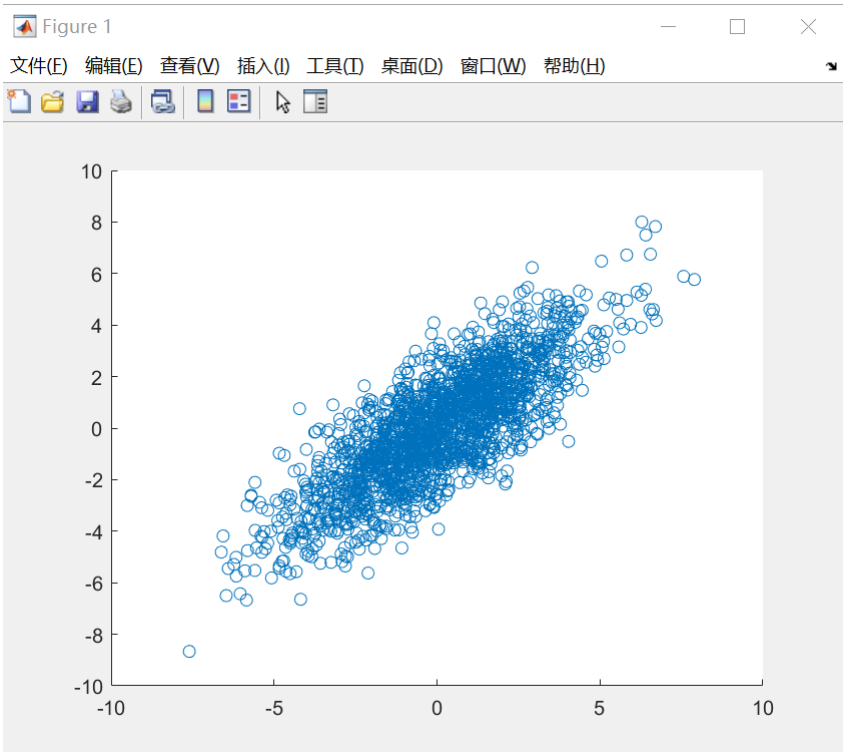
e）可以计算XX^T的特征值。因为XX^T 与$X^T X$的特征值相等，而XX^T是一个2×2的矩阵，显然要简单许多。

5.3

（a）代码如下：

```
clear
x=randn(2000,2)*[2 1;1 2];
figure(1);
scatter(x(:,1),x(:,2));
xlim([-10,10]);
ylim([-10,10]);
```

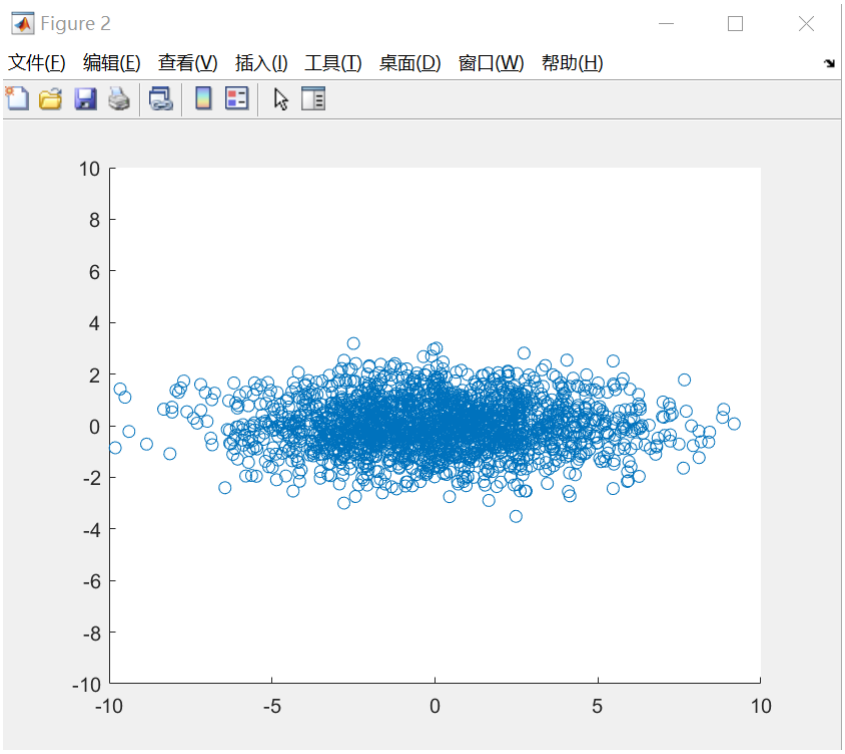
结果如图：



（b）代码如下：

```
%%接上面
[row,col]=size(x);
c=cov(x); %求矩阵x的协方差矩阵
[F,V]=eigs(c); %V为6个最大特征值对角阵，F的列向量为对应特征向量
meanx=mean(x); %求矩阵每列的平均值
temp=repmat(meanx,row,1); %堆叠矩阵
s=(x-temp)*F;
pca=s(:,1:2); %取第1、2列
figure(2)
scatter(pca(:,1),pca(:,2));
xlim([-10,10]);
ylim([-10,10]);
```

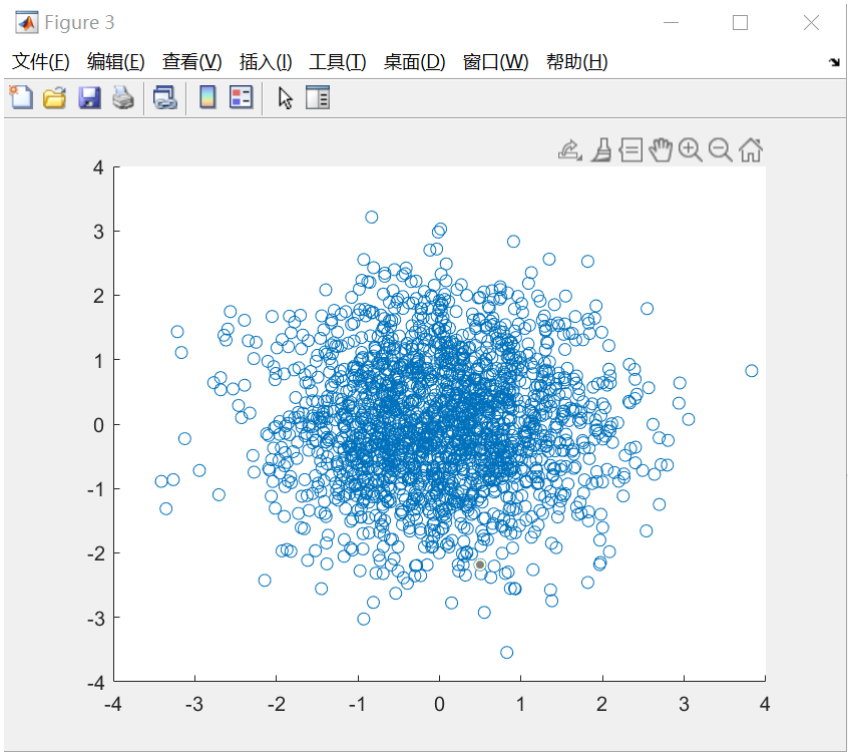
结果如图：



(c) 代码如下:

```
%%接上面
v=V^(-0.5);
whitenings=(x-temp)*(F*v);
whitening=whitenings(:,1:2);
figure(3)
scatter(whitening(:,1),whitening(:,2));
xlim([-4,4]);
ylim([-4,4]);
```

结果如图:



（d）因为`PCA`的过程的核心在于`“将代表原始数据的矩阵X的每一行减去该行的均值”`以及`“用前k个特征向量构成的矩阵乘原始矩阵得到降维后的矩阵”`。第一个步骤的减法就是一个平移的过程，而最后一步的矩阵乘法就是一个旋转的过程，因此若不进行降维（即在最后一步用所有特征向量构成的矩阵乘原始矩阵），`PCA`其实就是数据的平移后再旋转。如果不考虑降维这一作用的话，`PCA`的这一操作可以让数据变得更加分散。