题目：K-means Clustering and Principal Component Analysis

姓名:张胤民 学号:201694069

一、 实现功能简介

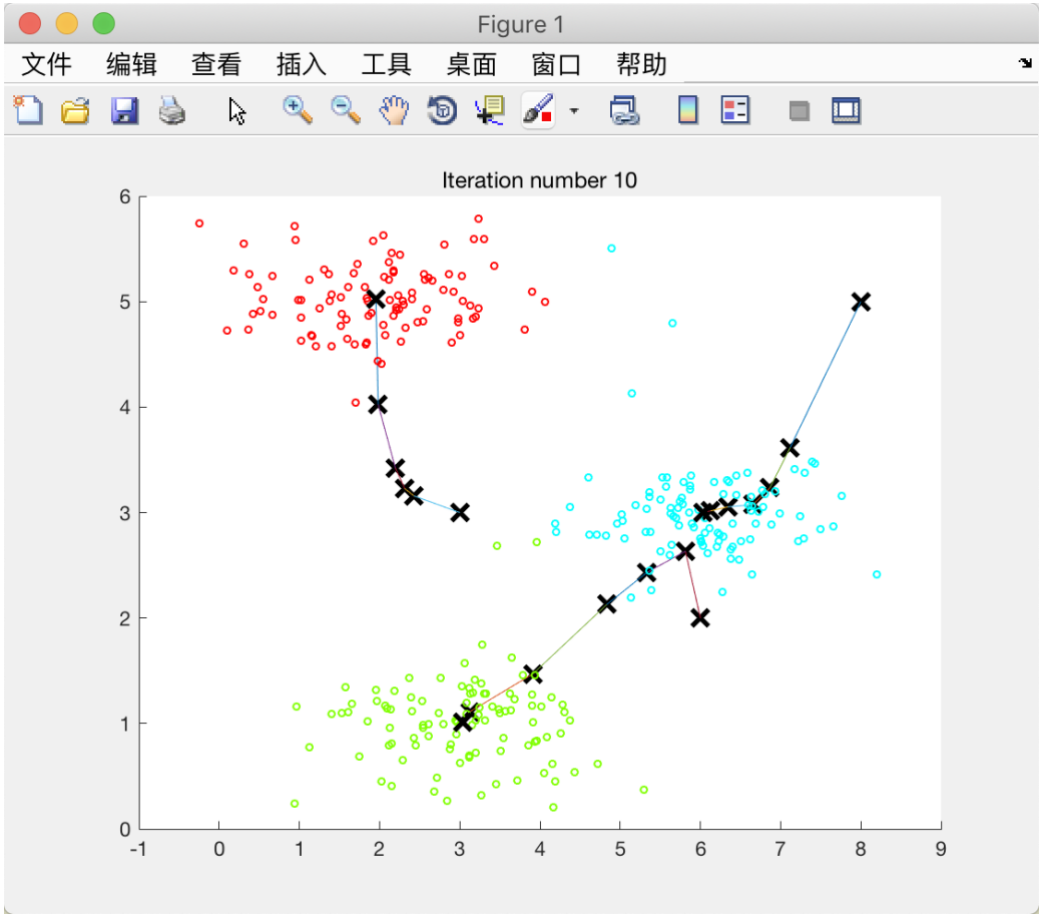
主要功能:

1. 实现 **K-means** 聚类算法.
2. 使用**主成分分析(PCA)**来确定数据的低维表示.

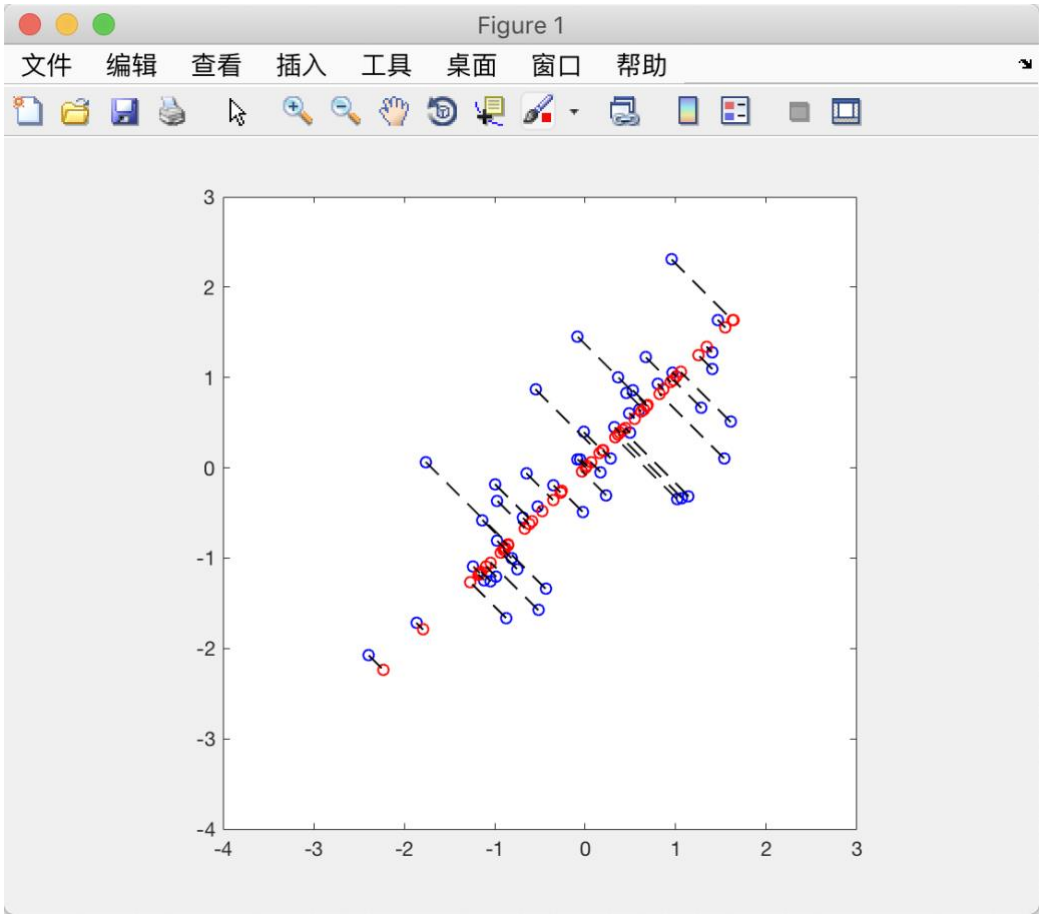
二、 具体编写代码及结果展示以及代码功能描述

1. 结果展示

- K-means:



- PCA



2. 代码实现

1) K-means 算法

a) computeCentroids.m

代码：

```
for i = 1:K
    idx_index = find(idx == i);
    ck = size(idx_index,1);
    centroids(i,:) = sum(X(idx_index, :),1) ./ ck;
end
```

实现功能:计算聚类中心平均值

$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} x^{(i)}$$

b) findClosestCentroids.m

代码：

```
for i = 1:size(X,1)
    idx(i) = 1;
    for j = 1:K
        if (sum((X(i,:) - centroids(j,:)) .^ 2) < sum((X(i,:) -
centroids(idx(i),:)) .^ 2))
            idx(i) = j;
        end
    end
end
```

实现功能:寻找最近聚类中心

$$c^{(i)} := j \quad \text{that minimizes} \quad ||x^{(i)} - \mu_j||^2,$$

c) featureNormalize.m

代码：

```
mu = mean(X);
X_norm = bsxfun(@minus, X, mu);
sigma = std(X_norm);
X_norm = bsxfun(@rdivide, X_norm, sigma);
```

实现功能:数据标准化

d) kMeansInitCentroids.m

代码：

```
randidx = randperm(size(X,1));
centroids = X(randidx(1:K),:);
```

实现功能:随机初始化

e) runkMeans.m

代码：

```
% Set default value for plot progress
if ~exist('plot_progress', 'var') || isempty(plot_progress)
    plot_progress = false;
end
% Plot the data if we are plotting progress
if plot_progress
    figure;
```

```
    hold on;
end

% Initialize values
[m n] = size(X);
K = size(initial_centroids, 1);
centroids = initial_centroids;
previous_centroids = centroids;
idx = zeros(m, 1);

% Run K-Means
for i=1:max_iters

    % Output progress
    fprintf('K-Means iteration %d/%d...\n', i, max_iters);
    if exist('OCTAVE_VERSION')
        fflush(stdout);
    end

    % For each example in X, assign it to the closest centroid
    idx = findClosestCentroids(X, centroids);

    % Optionally, plot progress here
    if plot_progress
        plotProgresskMeans(X, centroids, previous_centroids, idx, K,
i);
        previous_centroids = centroids;
        fprintf('Press enter to continue.\n');
        pause;
    end

    % Given the memberships, compute new centroids
    centroids = computeCentroids(X, idx, K);
end

% Hold off if we are plotting progress
if plot_progress
    hold off;
end
```

实现功能:运行 K-means 算法

2) PCA 算法

a) pca.m

代码：

```
sigma = X' * X / m;
[U,S,V] = svd(sigma);
```

实现功能:非奇异值分解(Singular Value Decomposition)

三、 小结（包括通过本内容的认识以及其他）

1. 认识

1) K-means

a) 目的:把 n 个点(可以是样本的一次观察或一个实例)划分到 k 个聚类中，使得每个点都属于离他最近的均值(此即聚类中心)对应的聚类,以之作为聚类的标准.

b) 特点: 这个问题在计算上是 NP 困难的,不过存在高效的启发式算法.一般情况下,都使用效率比较高的启发式算法,它们能够快速收敛于一个局部最优解.这些算法通常类似于通过迭代优化方法处理高斯混合分布的最大期望算法(EM 算法).而且,它们都使用聚类中心来为数据建模；然而 k-平均聚类倾向于在可比较的空间范围内寻找聚类,期望-最大化技术却允许聚类有不同的形状.

--Form wiki

2) PCA

a) 在多元统计分析中，主成分分析（英语：Principal components analysis，PCA）是一种分析、简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是，这也不是一定的，要视具体应用而

定。由于主成分分析依赖所给数据，所以数据的准确性对分析结果影响很大。

2. 心得

1) K-means

K-Means 的主要优点有：

- 原理比较简单，实现也是很容易，收敛速度快.
- 聚类效果较优.
- 主要需要调参的参数仅仅是簇数 k.

K-Means 的主要缺点有：

- K 值的选取不好把握.
- 如果各隐含类别的数据不平衡，比如各隐含类别的数据量严重失衡，或者各隐含类别的方差不同，则聚类效果不佳.
- 采用迭代方法，得到的结果只是局部最优.
- 对噪音和异常点比较的敏感.

2) PCA

- PCA 降维舍弃这部分信息之后能使样本的采样密度增大.
- 当数据受噪声影响时,最小的特征值所对应的特征向量往往和噪声有关,舍弃它们在一定程度上有去噪的效果.
- PCA 另一个重要特性是能将数据变换为元素之间彼此不相关的表示,可以消除数据中未知变化因素.可以用于数据白化.