1. 聚类分析（Cluster Analysis）

定义：聚类分析是一种无监督学习方法，旨在将数据集中的对象根据特征相似性划分成若干个组（簇）。同一簇内的对象相似度较高，簇与簇之间相似度较低。

常见聚类算法：

1. **K-means聚类**
   该算法通过迭代地将数据点分配到K个簇中，使得每个簇内的点之间的距离最小化，而簇之间的距离最大化。

   算法步骤：

   - 随机选择K个初始质心（每个簇的中心点）。
   - 将每个数据点分配到距离其最近的质心的簇中。
   - 计算每个簇的新质心（每个簇的平均值）。
   - 重复以上步骤直到质心不再变化或达到预设的迭代次数。

   公式：

   1. 计算每个数据点 x_i 到簇质心 c_k 的距离：

   $$d(x_i, c_k) = \|x_i - c_k\|^2$$

   2. 更新质心：

   $$c_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

   其中，C_k 表示簇 k 中的所有数据点，$|C_k|$ 是簇的大小。

2. **层次聚类（Hierarchical Clustering）**

   - **凝聚型（自下而上）**：每个数据点最初被视为一个独立的簇，逐步合并相似的簇，直到所有数据点都在一个簇中。
   - **分裂型（自上而下）**：将整个数据集视为一个簇，逐步分裂成更小的簇。

聚合规则：

- **单链接**：簇之间的距离为最短的点对之间的距离。
- **完全链接**：簇之间的距离为最远的点对之间的距离。
- **平均链接**：簇之间的距离为簇内所有点对的平均距离。

应用：客户细分、市场分析等。

---

2. 判别分析（Discriminant Analysis）

定义：判别分析是一种用于分类的统计方法，它试图通过找到不同类别之间的区分边界来进行分类。其目标是最大化类别之间的差异，并最小化同一类别内部的差异。

1. **线性判别分析（LDA）**
   LDA是最常用的判别分析方法，它假设每个类别的样本都服从高斯分布，并且所有类别的协方差矩阵相同。

   步骤：

   - **计算类内散度矩阵** S_W 和类间散度矩阵 S_B。

     $$S_W = \sum_{i=1}^{k} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^T$$

     其中，μ_i 是类 C_i 的均值，x_j 是类 C_i 中的样本点。

     $$S_B = \sum_{i=1}^{k} n_i (\mu_i - \mu)(\mu_i - \mu)^T$$

     其中，n_i 是第 i 类的样本数，μ 是所有样本的总体均值。

   - **计算判别函数**：LDA通过最大化类间散度与类内散度的比值来构造判别函数。

     $$w = S_W^{-1}(\mu_1 - \mu_2)$$

- **分类决策**：通过判别函数将样本归类。

2. **二次判别分析（QDA）**
   QDA与LDA的区别在于，QDA假设每个类别的协方差矩阵不同。其判别函数为：

$$g_i(x) = \ln |\Sigma_i| - \frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) + \ln P(C_i)$$

   其中，Σ_i 是第 i 类的协方差矩阵，$P(C_i)$ 是类别 C_i 的先验概率。

---

3. 主成分分析（PCA）

定义：主成分分析（PCA）是一种降维技术，旨在通过线性变换将数据映射到新的坐标系统，使得最大方差方向上的投影成为数据的新特征。

步骤：

1. **标准化数据**：将每个特征减去均值并除以标准差，使得每个特征的均值为0，方差为1。
2. **计算协方差矩阵**：计算标准化数据的协方差矩阵 Σ：

$$\Sigma = \frac{1}{n} X^T X$$

   其中 X 是标准化后的数据矩阵。
3. **特征值分解**：求解协方差矩阵的特征值和特征向量，特征向量代表数据的主成分方向，特征值表示主成分的方差。
4. **选择主成分**：根据特征值的大小选择前K个主成分。K通常根据累积方差解释比例来选择。
5. **降维**：将原始数据投影到前K个特征向量构成的空间上。

公式：

- 协方差矩阵 Σ 的特征值分解：

$$\Sigma v_i = \lambda_i v_i$$

其中 λ_i 是特征值，v_i 是特征向量。

---

4. 因子分析（Factor Analysis）

定义：因子分析是一种探索数据中潜在结构的降维方法，它试图用少数几个因子来解释数据集中的方差。

模型公式：

$$X = \Lambda F + \epsilon$$

其中，X 是观测数据，Λ 是因子载荷矩阵，F 是因子得分，ϵ 是误差项。

步骤：

1. **选择因子数**：可以使用碎石图、最大似然估计等方法来确定因子的个数。
2. **提取因子**：可以使用主成分分析（PCA）或最大似然法来提取因子。
3. **旋转因子**：为了解释的方便性，通常会对因子进行旋转（如正交旋转或斜交旋转）。

应用：用于心理学、市场研究等领域，通过发现潜在的隐含变量（因子）来解释多个观察变量之间的关系。

1. K-means聚类（K-means Clustering）

matlab 复制代码

```matlab
% 生成示例数据
data = [randn(100,2)+2; randn(100,2)-2; randn(100,2)+[5 5]];
% 执行K-means聚类，假设分为3类
k = 3;
[idx, C] = kmeans(data, k);

% 绘制聚类结果
figure;
gscatter(data(:,1), data(:,2), idx);
hold on;
plot(C(:,1), C(:,2), 'kx', 'MarkerSize', 12); % 聚类中心
title('K-means Clustering');
xlabel('Feature 1');
ylabel('Feature 2');
legend('Cluster 1', 'Cluster 2', 'Cluster 3', 'Centroids');
```

解释：

- `data` 是输入的二维数据集。
- `kmeans` 函数执行K-means聚类，`idx` 是数据点的簇标签，`C` 是簇中心。
- 绘图显示了数据点的分类和每个簇的中心。

---

2. 线性判别分析（LDA，Linear Discriminant Analysis）

matlab 复制代码

```matlab
% 生成示例数据
n = 100;
```

```
data = [randn(n,2)+2; randn(n,2)-2]; % 两个类别的二维数据
labels = [ones(n,1); 2*ones(n,1)]; % 类别标签

% 执行LDA
lda = fitcdiscr(data, labels);

% 查看LDA模型
disp(lda);

% 绘制LDA分类结果
figure;
gscatter(data(:,1), data(:,2), labels);
hold on;
h = ezplot(@(x,y)lda.Coeffs(1,2).Const + lda.Coeffs(1,2).Linear(1)*x + lda.Coeffs(1,2).Lin
set(h, 'LineWidth', 2);
title('LDA Classification');
xlabel('Feature 1');
ylabel('Feature 2');
legend('Class 1', 'Class 2', 'Decision Boundary');
```

复制代码

解释:

- `fitcdiscr` 函数用于执行线性判别分析，返回的 `lda` 对象包含分类模型。
- 使用 `ezplot` 函数绘制LDA的决策边界。

---

3. 主成分分析（PCA，Principal Component Analysis）

matlab

复制代码

```
% 生成示例数据
data = randn(100, 3);
```

```
% 标准化数据（每个特征减去均值并除以标准差）
data = (data - mean(data)) ./ std(data);

% 执行PCA
[coeff, score, latent] = pca(data);

% 绘制前两个主成分
figure;
scatter(score(:,1), score(:,2));
title('PCA: First Two Principal Components');
xlabel('Principal Component 1');
ylabel('Principal Component 2');
```

复制代码

解释：

- 使用 `pca` 函数进行主成分分析，`coeff` 是主成分载荷矩阵，`score` 是降维后的数据，`latent` 是每个主成分的方差。
- 绘制了前两个主成分的投影结果。

---

4. 因子分析（Factor Analysis）

matlab

复制代码

```
% 生成示例数据
data = randn(100, 5); % 5维数据

% 执行因子分析
[lambda, psi, T, stats] = factoran(data, 2); % 假设提取2个因子

% 查看因子载荷矩阵
disp('Factor Loadings:');
disp(lambda);
```

```
% 绘制因子得分图
figure;
scatter(T(:,1), T(:,2));
title('Factor Analysis: Factor Scores');
xlabel('Factor 1');
ylabel('Factor 2');
```

复制代码

解释：

- 使用 `factoran` 函数进行因子分析，`lambda` 是因子载荷矩阵，`T` 是因子得分。
- 绘制因子得分的二维投影。

---

5. 层次聚类（Hierarchical Clustering）

matlab

复制代码

```
% 生成示例数据
data = [randn(100,2)+2; randn(100,2)-2; randn(100,2)+[5 5]];

% 计算数据的欧氏距离
distances = pdist(data);

% 执行层次聚类
tree = linkage(distances, 'ward'); % 使用ward链接方法

% 绘制树状图
figure;
dendrogram(tree);
title('Hierarchical Clustering Dendrogram');
xlabel('Sample Index');
ylabel('Distance');
```

解释：

- `pdist` 函数计算样本间的距离，`linkage` 函数进行层次聚类，`ward` 是常用的最小方差链接方法。
- 使用 `dendrogram` 绘制树状图，展示聚类的层次结构。
