大数据技术与风险管理应用

欧阳夫

南开大学金融学院

二零一九年一月十一日

内容大纲

内容大纲

统计学习、大数据与风险管理

讲座目的

- 回顾风险管理中所涉及到的常用统计学习模型。
- 介绍常用的大数据方法在分析建模过程中的应用。

统计学习（Statistical Learning）

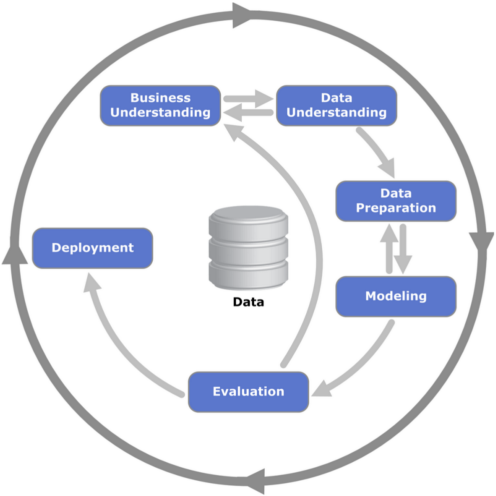
对大量数据进行探索和分析，深化对数据的理解，以便发现有意义的模式、规则，进行有效预测的过程。

- 有监督学习（预测模型），无监督学习（探索数据中有意义的模式）。

大数据时代

- 容易收集大量数据，结合统计学习方法，可改善企业运营管理。
- 数据复杂性（维度）增加，如何筛选有效信息成为挑战。
- 对大数据进行分析与利用的能力成为企业组织核心竞争力。

路径与应用



- 统计学习方法结合大数据在银行业的应用：盈利分析、欺诈发现、违约管理、信贷风险评估、客户细分与拓展等。
- 本次讲座主要涉及到其中的数据准备和建模步骤。

统计学习模型的基本框架和任务

统计学习模型的基本框架和任务

给定自变量 $X \in \mathbb{R}^p$，我们希望找到一个函数 $f(\cdot)$，使得 $f(X)$ 可以被用来预测因变量 $Y \in \mathbb{R}$ 的取值，并使预测的误差最小，即

$$f = \arg\inf_{g \in \mathcal{F}} E[(Y - g(X))^2]$$

我们知道这个问题的解是 $f(X) = E[Y|X]$，从而预测模型可以写为

$$Y = f(X) + e = E[Y|X] + e$$

其中 e 为预测的误差项。（有监督）统计学习的任务即是使用观测到的自变量数据 $(x_1, ..., x_N)$ 和因变量数据 $(y_1, ..., y_N)$ 得到函数 $f(x) = E[Y|X = x]$ 的一个估计 $\widehat{f}(x)$，并评估其预测误差。在这一过程中，我们通常需要对数据进行划分，并明确自变量与因变量的具体类型。

数据与变量

数据集划分

- 训练（training）数据：用于拟合模型
- 验证（validation）数据：用于评估备选模型，并进行模型选择
- 测试（test）数据：用于对模型的普适性进行评价

变量类型

- 名义变量：对观测进行分类的变量，取值没有含义，如性别。
- 定序变量：变量取值本身没有意义，但其排序有意义，如满意度评价。
- 定距变量：变量取值的差有意义，但比值没有意义，如考试成绩。
- 定比变量：变量的取值的排序和比值均有意义，如收入水平。
- 分类/离散变量 ={名义变量，定序变量}。
- 数值/连续变量 ={定距变量，定比变量}。

内容大纲

线性回归 (Linear Regression)

在线性回归中，因变量 y 与自变量 $x=(x_1,...,x_p)$ 的关系可以表述为

$$y=E[y|x]+e=f(x)+e$$

其中 $f(x)=\beta_1 x_1+\cdots+\beta_p x_p$，而则 e 是与 x 相互独立，均值为 0，方差为 σ^2 的扰动项。

- 适用问题：y 为（可正可负）连续变量，$(x_1,...,x_p)$ 为连续或名义变量。
- 目的：获得对模型参数 $\beta=(\beta_1,...,\beta_p)$ 及 σ^2 的一致估计 $(\widehat{\beta},\widehat{\sigma}^2)$，在此基础上对因变量进行预测 $\widehat{y}=\widehat{\beta}_1 x_1+\cdots+\widehat{\beta}_p x_p$，并评估预测误差 $E[(y-\widehat{y})^2]$。
- 应用举例：研究恐慌指数（VIX）与股票指数收益率的（负）相关关系。
- 线性回归的实现：R 语言 lm 命令。

广义线性模型（Generalized Linear Models）

广义线性回归可以被看作是线性回归的扩展，适用于因变量是名义、计数或非负变量等情形。因变量 y 与自变量 $x = (x_1, ..., x_p)$ 的关系可以表述为

$$E[y|x] = g(\beta_1 x_1 + \cdots + \beta_p x_p)$$

其中 $g(\cdot)$ 为一一对应且连续可导的连接函数。

因变量取值为正数的模型

$$\log(y) = \beta_1 x_1 + \cdots + \beta_p x_p + e$$

- 获得对模型参数 β 的一致估计 $\widehat{\beta}$，在此基础上对因变量进行预测 $\log(\widehat{y})$，并评估预测误差。
- 应用举例：波动率（Volatility）预测，计算 Value at Risk（VaR）。

广义线性模型

Logit 与 Probit 模型

- 适用问题：y 为二值名义变量，如 0-1，因此 $E[y|x] = P(y = 1|x)$。
- 对 Logit 模型，$g(\cdot)$ 为 Logistic 分布的分布函数；对 Probit 模型，$g(\cdot)$ 为标准正态分布的分布函数。
- 目的：获得对模型参数 β 的一致估计 $\widehat{\beta}$，在此基础上对 $P(y = 1|x)$ 及 $\partial P(y = 1|x)/\partial x_j$，$j = 1, ..., p$ 进行预测，并评估预测误差。
- 应用举例：贷款违约率的预测与评估。
- Logit 与 Probit 模型均可以被扩展应用到 y 为 J（$J > 2$）值名义变量或定序变量的问题中。

广义线性模型

Poisson 模型

- 适用问题：y 为计数变量，如 0, 1, 2,...，代表事件发生的次数。
- $E[y|x] = \exp(\beta_1 x_1 + \cdots + \beta_p x_p)$。
- 目的：获得对模型参数 β 的一致估计 $\widehat{\beta}$，在此基础上对 $P(y=k|x)$ 及 $\partial P(y=k|x)/\partial x_j$，$j=1,...,p$ 进行预测，并评估预测误差。
- 应用举例：未/提前支付的信用分期付款次数的预测。

广义线性模型的实现：R 语言 glm 命令。

k 近邻法（k-NN，k-Nearest Neighbors）

k-NN 的基本想法可以理解为

$$f(x_0) = E[y|x = x_0] \approx E[y|x \in N_k(x_0)] \approx \sum_{x_i \in N_k(x_0)} y_i/k$$

其中 x_0 为 p 维空间中一个给定的向量，$N_k(x_0)$ 为数据中与 x_0 距离最小的 k 个观测的集合。

- 适用问题：任意类型的 y 和 x。
- 应用举例：线性与广义线性模型中的例子均适用。
- 目的：获得对 $f(x)$ 的估计，并评估估计/预测误差。
- 优点：不依赖对于 $f(x)$ 的函数形式的假设。
- 缺点：高维数据应用中的维数灾难（curse of dimensionality）问题。
- k-NN 的实现：R 语言 knnreg 命令（caret 程序包）。

判别分析（Discriminant Analysis）

- 适用问题：y 为名义变量（取值为 1-J），$(x_1, ..., x_p)$ 为连续变量。
- 目的：预测事件 $y = j$ 发生的条件概率，$P(y = j|x)$。
- 原理：使用贝叶斯定理对观测进行分类。

$$P(y = j|x) = \frac{P(y = j)f(x|y = j)}{\sum_{l=1}^{J} P(y = l)f(x|y = l)}$$

  其中 $P(y = j)$ 为事件 $y = j$ 发生的概率，$f(x|y = j)$ 为给定 $y = j$ 下的 x 的联合密度函数。

- $P(y = j)$ 可以通过样本均值估计。假设 $f(x|y = j)$ 为均值为 μ_j，协方差矩阵为 Σ_j 的联合正态分布，则 μ_j 与 Σ_j 均可通过样本矩估计。
- 判别分析的实现：R 语言 lda 及 qda 命令（MASS 程序包）。
- 相关方法：朴素贝叶斯方法 (Naive Bayes)。

神经网络（Neural Networks）

神经网络包含了众多的模型与方法，在统计与人工智能领域有广泛的应用。

这里我们介绍最简单（常用）的单层感知器（single layer perceptron）模型。

神经网络建模

目的：寻找自变量 $X=(X_1,...,X_p)$ 的函数 $f(X)=(f_1(X),...,f_k(X))$，来预测因变量 $Y=(Y_1,...,Y_k)$ 的值。

- 利用（线性）组合函数与激活函数 $\psi(\cdot)$ 构建神经元

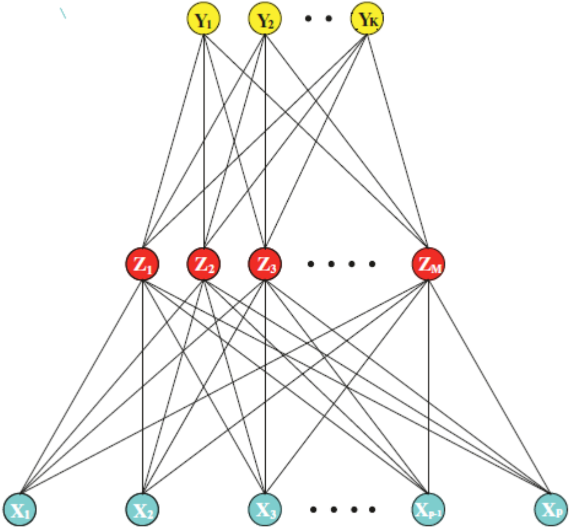
$$Z_m=\psi(\alpha_{0m}+\alpha_m^T X),\ m=1,...,M$$

- 利用神经元 $Z=(Z_1,...,Z_M)$ 的线性组合生成输出向量

$$T_k=\beta_{0k}+\beta_k^T Z,\ k=1,...,K$$

- 通过输出函数对 T 进行转换生成预测 $f_k(X)=g_k(T)$，$k=1,...,K$。

神经网络



神经网络

- 神经网络中包含 Z 的层成为隐藏层。理论上，使用更多的隐藏层有可能用更少的神经元及参数达到理想的预测效果。
- 神经网络的实现：R 语言 nnet 程序包。
- 应用举例：线性与广义线性模型中的例子均适用。
- 优点：神经网络作为通用近似器（universal approximator），在给予足够的数据与神经元数量（M）的条件下，理论上可以很好地近似任何连续函数 $f(X)$。
- 缺点：模型复杂度高，对模型的解释比较困难，通常只用于预测。

内容大纲

Lasso 和 Elastic Net 方法

在大数据时代，我们往往需要处理高维数据，即拥有大量信息（变量）的数据。在线性回归或广义线性回归中使用过多的变量会导致模型过度拟合，增加产生共线性的可能，模型估计的结果也更加难以解释。

这里我们简要介绍 Lasso 和 Elastic Net（EN）两种规则化（regularization）回归方法。它们能够在回归的过程中“自动”进行变量选择。

考虑线性回归模型。假设我们可以获得 N 个观测 $\{(x_i, y_i)\}_{i=1}^N$，其中 $x_i = (x_{i1}, ..., x_{ip})$ 与 y_i 分别是第个 i 观测的自变量与因变量。

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i$$

通常我们可以使用最小二乘法（OLS）来估计模型的参数 $\beta = (\beta_1, ..., \beta_p)$:

$$\widehat{\beta}^{OLS} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

Lasso 和 Elastic Net 方法

Lasso 和 EN 回归方法通过对模型的复杂程度加以限制（惩罚）来控制模型自变量的数量，即使某些系数的估计值为 0。

Lasso：对于 $t \geq 0$，

$$\widehat{\beta}^{Lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} \quad s.t. \ \sum_{j=1}^{p} |\beta_j| \leq t$$

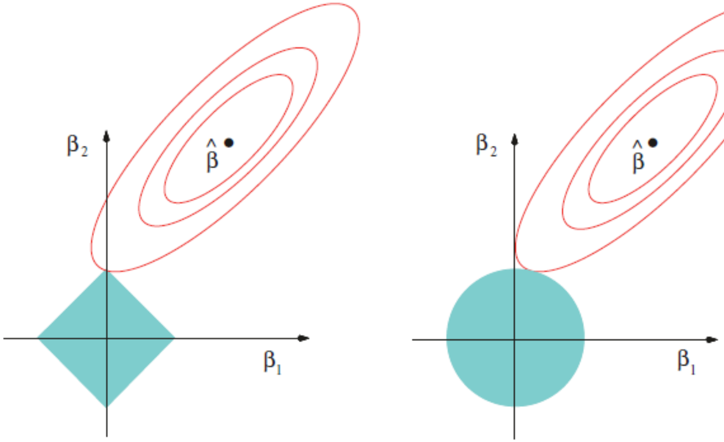
Elastic Net：对于 $0 \leq \alpha \leq 1$, $t \geq 0$，

$$\widehat{\beta}^{EN} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right\} \quad s.t. \ \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \frac{\beta_j^2}{2} \leq t$$

EN 可以被看作是 Lasso 和岭回归（ridge regression）的加权平均，更适合分析含有较多相关性较高的自变量的问题。

Lasso 和 Elastic Net 回归方法

Lasso vs. OLS



Lasso 和 Elastic Net 方法

应用举例 1：研究股票指数收益率的影响因素

- 因变量：沪深 300 指数收益率。
- 自变量：标普指数、日经指数、富时指数、PMI、汇率、利率、国际油价……滞后 L 期的值。

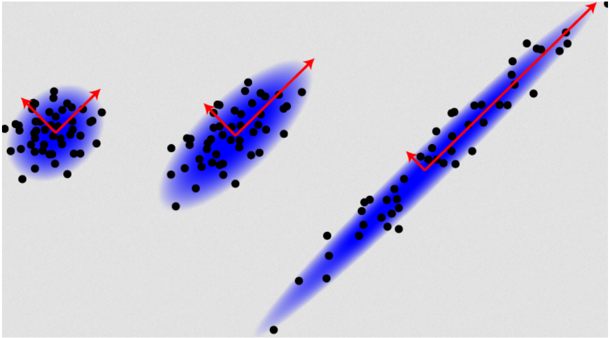
应用举例 2：预测流感发病率

- 因变量：流感发病率。
- 自变量：与流感相关的 p 个关键词在 Google 上的搜索次数。

- Lasso 和 EN 回归的基本想法可以被推广应用到（部分）广义线性模型，如 Park and Hastie (2007)，Friedman et al (2008)。
- Lasso 和 Elastic Net 回归方法的实现：R 语言 glmnet 程序包。

主成分分析（PCA，Principal Component Analysis）

PCA 的主要目的是构造输入变量的少数线性组合，尽可能解释数据的变异性（方差）。这些线性组合被称为主成分，它们形成的降维数据可以用于进一步的分析。可以通过下面的示意图来对 PCA 的工作原理加以理解：



主成分分析

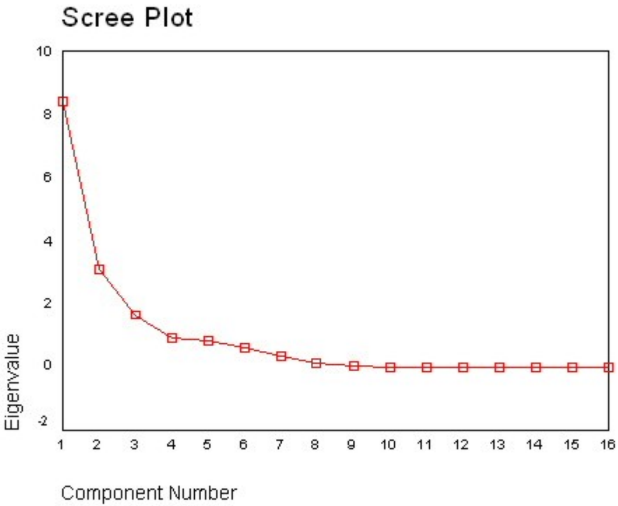
主成分个数的选择

- Kaiser 准则，即保留解释输入变量总方差的比例大于平均解释比例的主成分。
- $(X_1, ..., X_p)$ 总方差被前 q 个主成分解释的比例（$\geq 85\%$）。
- 保留的主成分的可解释性。
- 崖底碎石图（scree plot）绘出的特征值与其排序的关系（选择拐点出现之前一点）。

PCA 的实现

R 语言 princomp 命令。进行 PCA 之前通常需要标准化数据。

主成分分析


Scree Plot
10
8
6
4
2
0
-2
Eigenvalue
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Component Number

探索性因子分析（FA，Factor Analysis）

FA 的基本想法是，每一个输入变量的变异性都可以归结于少数潜在的公共因子和一个与公共因子无关的特殊因子。它的主要目的在于通过寻找少量公共因子来实现数据降维，并使用降维数据进行进一步分析。

因子模型

$$
\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1q}F_q + e_1 \\
X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2q}F_q + e_2 \\
&\vdots \\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pq}F_q + e_p
\end{aligned}
$$

其中 $(F_1, ..., F_q)$ 为 q 维（$q < p$）公共因子，$(e_1, ..., e_p)$ 为特殊因子，$L = \{l_{ij}\}$ 称为载荷矩阵。

探索性因子分析

因子模型的估计

- 载荷矩阵的估计：主成分法、极大似然法。
- 对 $(F_1, ..., F_q)$ 的估计可形成降维数据，用于进一步分析。
- R 语言 factanal 命令。通常在进行 FA 之前需要标准化数据。

选取公共因子个数 q

- Kaiser 准则。
- $(X_1, ..., X_p)$ 总方差被 q 个公共因子解释的比例（$\geq 85\%$）。
- 保留的公共因子的可解释性。
- 假设检验（极大似然法）。
- 崖底碎石图（拐点出现之前一点）。

多维标度分析 (MDS, Multidimensional Scaling)

MDS 是一种在低位空间展示高维数据的可视化方法，也常用于对高维数据进行降维，各观测在低维空间所对应的坐标被用来进行进一步的分析。

应用举例：银行倒闭分析

Mar-Molinero and Serrano (2011) 使用 66 家西班牙银行的 9 个财务比率来分析影响银行倒闭的财务因素：

1. 该研究首先通过 MDS 将数据的维数由 9 降为 6，并发现降维后的第一个维度是预测银行倒闭的有效指标；
2. 以每个财务比率为因变量，以降维后的前两个维度的指标作为自变量进行回归，发现第一个维度与营利性密切相关，而第二个维度受流动性影响较大；
3. 说明营利性指标可以有效预测银行是否倒闭；
4. 降维后的数据可以用于进一步建立预测银行倒闭的统计模型。

多维标度分析

给定一组 p 维空间的观测 $x_1, \ldots, x_N \in \mathbb{R}^p$，MDS 旨在找到一组 k $(k < p)$ 维空间中的观测 $z_1, \ldots, z_N \in \mathbb{R}^k$，使得两个空间中的观测之间的距离或相似度 “大致匹配”。

度量 (metric) MDS 寻找一组 $(z_1^*, \ldots, z_N^*)$ 最小化基于距离测度的应力方程 (stress function)

$$S_M(z_1, \ldots, z_N) = \sum_{i \neq i'} (d_{ii'} - \|z_i - z_{i'}\|)^2$$

其中 $d_{ii'}$ 为 x_i 与 $x_{i'}$ 之间的距离。

非度量 (non-metric) MDS 则侧重于距离的排序，其应力方程为

$$S_{NM}(z_1, \ldots, z_N) = \frac{\sum_{i \neq i'} (\phi(d_{ii'}) - \|z_i - z_{i'}\|)^2}{\sum_{i \neq i'} \|z_i - z_{i'}\|^2}$$

其中 $\phi(\cdot)$ 为（未知）单调函数。

多维标度分析

实践中使用非度量 MDS 较多，因为其对与数据中存在的异常值更加稳健。

选取维数 k

1. 应力函数的值足够小，达到一定的拟合优度（≤ 0.05 很好，≤ 0.025 非常好）。
2. 绘制崖底碎石图，选择拐点。

MDS 的实现

R 语言 cmdscale 命令。

内容大纲

关联规则分析 (Association Rules)

关联规则分析是一种无监督统计学习方法，它可以从大量数据中挖掘变量之间有意义的相关关系。

应用举例

假设我们可以观测到 N 个个人的 p 条信息：

负债比，职业，性别，年龄，收入，婚姻状况，子女数，教育程度，租房/买房/与其他家人同住，⋯⋯

每一个信息可以被视为一个变量。我们可以通过关联规则分析来研究各变量之间的相关关系，尤其是发现那些与负债比有关的相关关系。

关联规则分析的实现

- Apriori 算法是使用最广泛的关联规则分析的基础算法。
- R 语言 arules 程序包。

关联规则的基本概念

- 首先，我们可以将每个变量的每种取值定义为一个项 (item)，令 $\mathcal{T} = \{i_1, ..., i_m\}$ 表示所有项的集合，$\mathcal{T}$ 的子集称为项集。
- 关联规则的形式为 $A \Rightarrow B$，其中 A 和 B 是两个项集，满足 $A \cap B = \emptyset$，A 称为关联规则的前项集，B 称为关联规则后项集。
- 任意项集 X 的支持度 $Supp(X)$ 定义为数据集中包含的所有项的比例，$\approx X$ 的概率。
- 关联规则 $A \Rightarrow B$ 的支持度 $Supp(A \Rightarrow B)$ 定义为 $Supp(A \cup B)$，即数据中同时包含 A 和 B 所有项的比例，$\approx A \cup B$ 的联合概率。
- 关联规则的置信度 $A \Rightarrow B$ 定义为 $Supp(A \cup B)/Supp(A)$, $\approx B$ 给定 A 的条件概率。
- 在进行关联规则分析时，需要设定最小支持度阈值 ($minsupp$) 和最小置信度阈值 ($minconf$)，支持度不小于 $minsupp$，且置信度不小于 $minconf$ 的关联规则被称为强关联规则。

提升度及关联规则分析的扩展

在实践中，只关注支持度和置信度往往是不够的，还需要考察提升值 $(lift)$，原因在于下列情况可能出现

$$Supp(A \cup B)/Supp(A) < Supp(B)$$

关联规则的提升值定义为 $(Supp(A \cup B)/Supp(A))/Supp(B)$。如果提升值大于 1，则和是正关联的，反之，则为负关联。只有当支持度和置信度都不小于相应阈值，且提升值大于 1 的关联规则才是有意义的。

关联规则分析的扩展

- 序列关联规则
- 多阈值关联规则
- 带因变量的关联规则

聚类分析 (Cluster Analysis)

聚类分析是一种无监督数据挖掘方法，有着非常广泛的应用。聚类分析的主要功能是将观测到的样本，依据其属性进行分类（clusters），使得被归入同一类的样本在属性上，较之属于其他类的样本，更加相似。聚类分析也常常被用来对数据进行统计性描述，以判断样本是否来自特征相异的总体，从而为后续的统计建模提供依据。

应用举例: 共同基金的分类

金融市场上的交易的共同基金可以根据其投资的板块，投资风格，和资本值大小等进行简单的分类。但是我们可以使用聚类分析方法，对其进行进一步细分:

例如，我们可以依据有关共同基金收益与风险相关的测度，如回报率，贝塔值，阿尔法值，夏普比率，标准差等，对共同基金加以细分。

常用的聚类方法为 K 均值聚类法 (K-means) 和分层聚类法 (hierarchical)。我们这里主要介绍前者 (R 语言 kmenas 命令)。

K 均值聚类算法

1. 初始化 K 个类别的中心 $c_1, ..., c_K$；
2. 在每次循环中，对给定 N 个观测（属性向量）$x_1, ..., x_N$：
   (1) 按照下面的规则重新分配观测

$$C(i) = \arg\min_{1\leq l\leq K} d(x_i, c_l), i = 1, ..., N$$

   其中 $d(\cdot,\cdot)$ 为相似度/距离的度量，$C(i)$ 表示观测 i 所属于的类别的编号。
   (2) 重新计算类别中心

$$c_l = \arg\min_c \sum_{i\in C(i)} d(x_i, c), l = 1, ..., K$$

3. 持续循环，直到所有类别中心的变化足够小。

K 均值聚类法是最小化类别内距离 $(\sum_{i=1}^N d(x_i, c_{C(i)}))$ 的算法，优点在于计算量小，适合处理大数据，而其缺点在于通常只能找到局部最优解。在实践中，可以通过设定不同初始中心，多次聚类，选取目标函数最小的结果。

K 均值聚类算法

聚类分析的基础是有关相似性的定义。而本质上，相似性可以被理解为属性向量的距离或相关性，所以，相似性的定义取决于属性向量中变量的类型。

- 名义变量：对称与非对称
- 定序或定距变量
- 定比变量
- 各类变量的混合

在聚类分析前，常常要对 (连续) 数据进行标准化，以消除变量方差对相似度的度量的影响。

确定类别个数 K

- 伪 F 统计量 (pseudo F statistic)：查看不同 K 相应的伪 F 统计量，选其中较大者。
- 数据相关领域内的经验和知识。

参考文献

- Friedman, J., Hastie, T. and Tibshirani, R. (2016): “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, 2nd Edition, Springer.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013): “An Introduction to Statistical Learning: with Applications in R”, 1st Edition, Springer.
- 张俊妮 (2018): “数据挖掘与应用：以 SAS 和 R 为工具”，第二版，北京大学出版社。
- Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018): “Foundations of Machine Learning”, 2nd Edition, The MIT Press.