

22

Tree

树

没有闭合回路的图



人类的历史，本质上是思想的历史。

Human history is, in essence, a history of ideas.

—— 赫伯特·乔治·威尔斯 (Herbert George Wells) | 英国小说家和历史学家 | 1866 ~ 1946



- networkx.all_pairs_lowest_common_ancestor() 寻找最近共同祖先
- networkx.draw_networkx_edge_labels() 绘制边标签
- networkx.draw_networkx_edges() 绘制图边
- networkx.draw_networkx_labels() 绘制节点标签
- networkx.draw_networkx_nodes() 绘制图节点
- networkx.minimum_spanning_tree() 计算最小生成树
- seaborn.clustermap() 绘制热图树形图
- seaborn.heatmap() 绘制热图

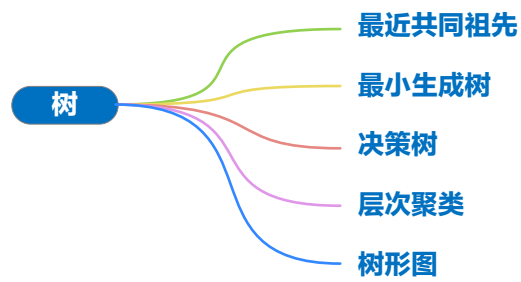
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



22.1 树

上一章提过，在图论中，树是一种特殊的无向图，它树是一个没有闭合回路的图，其中任意两个节点之间都有唯一的路径。树有如下性质。

- ▶ 连通性：一棵树是连通的，即任意两个节点之间都存在路径。一个树有 n 个节点时，它具有 $n - 1$ 条边。这确保了树的连接性。
- ▶ 无环性：树是无环的，不存在任何形式的回路或环。
- ▶ 唯一路径性：任意两个节点之间有唯一的简单路径。

图 1 所示的互联网上的路由网络是树形结构。这个例子来自 NetworkX，请大家自行学习：

https://networkx.org/documentation/stable/auto_examples/graphviz_layout/plot_lanl_routes.html

图 2 所示动物分类也是采用的树形结构。

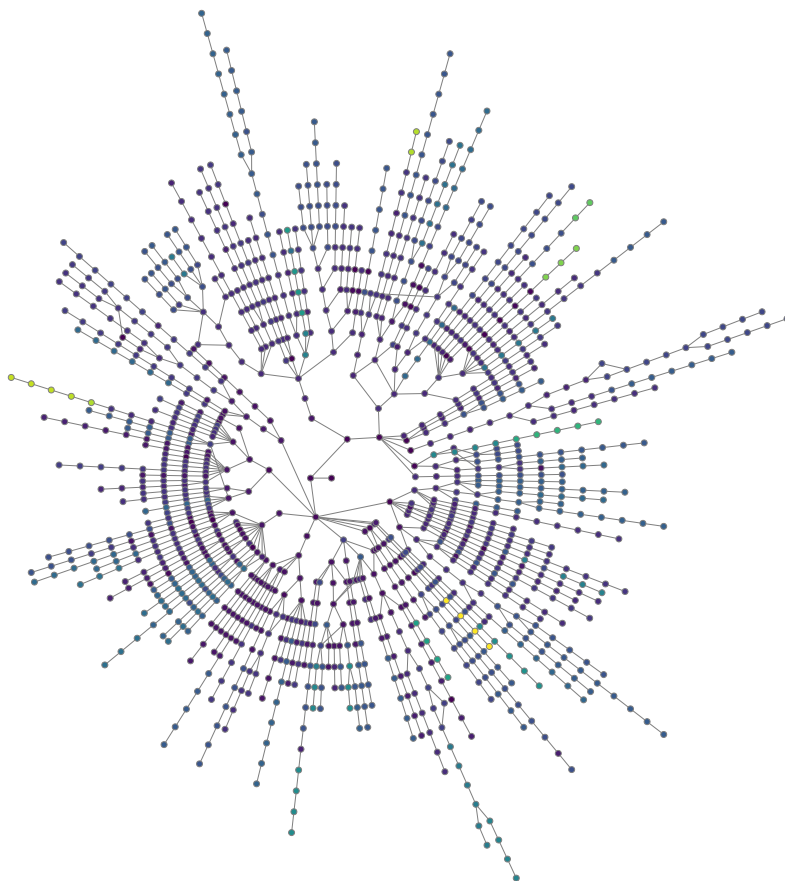


图 1. 可视化互联网上的 186 个站点到洛斯阿拉莫斯国家实验室的路由 LANL Routes 信息

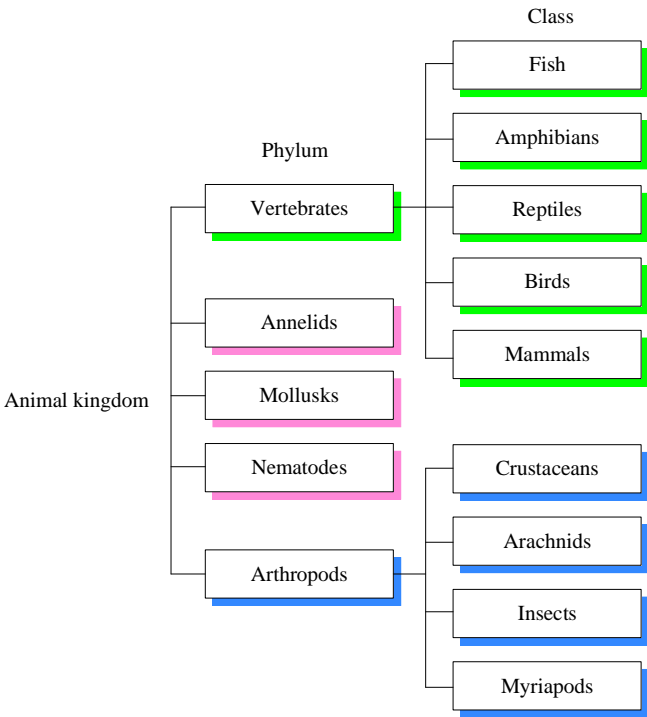


图 2. 动物分类

表 1 展示的数据列出了猫科动物的分类，从科 (family) 开始到亚科 (subfamily)、属 (genus)、亚种 (subspecies)，然后是常用名，最后一列是灭绝的危险等级。例如，猎豹 (cheetah) 被分类为 Felidae 科，Acinonychinae 亚科，Acinonyx 属，其学名为 Acinonyx jubatus。

图 3 用环形树状图可视化这些数据，这幅图可以帮助我们理解不同猫科动物之间的关系和它们的分类体系。图 4 用水平树状图展示表 1 数据。

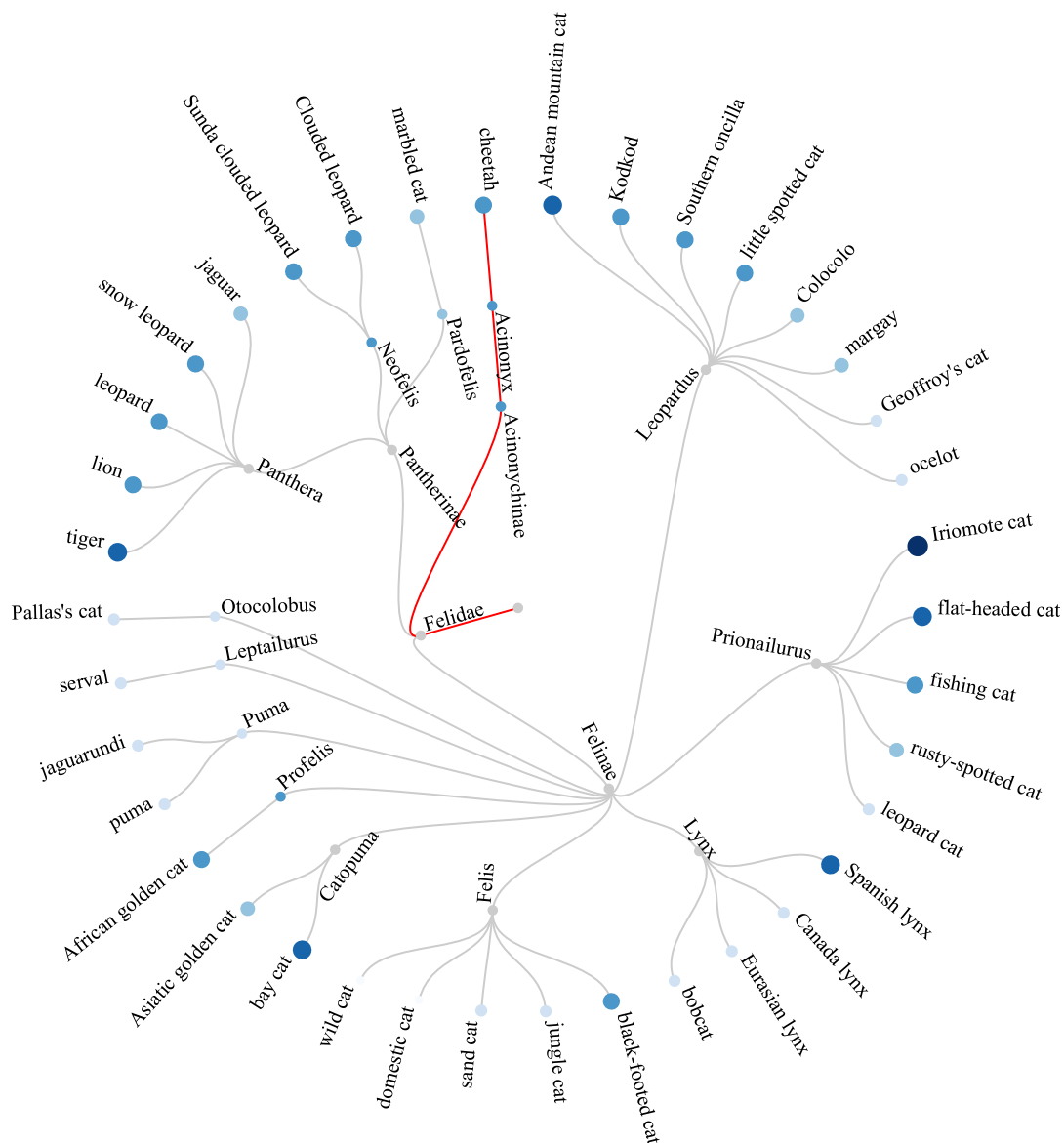
这两幅图和数据都来自 <https://www.rawgraphs.io/>，非常推荐大家尝试使用这个网站提供的可视化工具。

表 1. 猫科动物分类，部分数据；数据来自 <https://www.rawgraphs.io/>

Family	Subfamily	Genus	Subspecies	Name	Risk of Extinction
Felidae	Acinonychinae	Acinonyx	Acinonyx jubatus	cheetah	4
Felidae	Felinae	Catopuma	Catopuma badia	bay cat	5
Felidae	Felinae	Catopuma	Catopuma temminckii	Asiatic golden cat	3
Felidae	Felinae	Felis	Felis catus	domestic cat	1
Felidae	Felinae	Felis	Felis chaus	jungle cat	2
Felidae	Felinae	Leopardus	Leopardus colocolo	Colocolo	3
Felidae	Felinae	Leopardus	Leopardus geoffroyi	Geoffroy's cat	2
Felidae	Felinae	Leptailurus	Leptailurus serval	serval	2
Felidae	Felinae	Lynx	Lynx canadensis	Canada lynx	2
Felidae	Felinae	Lynx	Lynx lynx	Eurasian lynx	2
Felidae	Felinae	Lynx	Lynx pardinus	Spanish lynx	5
Felidae	Felinae	Lynx	Lynx rufus	bobcat	2
Felidae	Felinae	Otocolobus	Otocolobus manul	Pallas's cat	2
Felidae	Felinae	Prionailurus	Prionailurus bengalensis	leopard cat	2
Felidae	Felinae	Profelis	Profelis aurata	African golden cat	4
Felidae	Felinae	Puma	Puma concolor	puma	2
Felidae	Felinae	Puma	Puma yagouaroundi	jaguarundi	2
Felidae	Pantherinae	Neofelis	Neofelis diardi	Sunda clouded leopard	4
Felidae	Pantherinae	Neofelis	Neofelis nebulosa	Clouded leopard	4
Felidae	Pantherinae	Panthera	Panthera leo	lion	4

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。
代码及 PDF 文件下载：<https://github.com/Visualize-ML>
本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

Felidae	Pantherinae	Panthera	Panthera onca	jaguar	3
Felidae	Pantherinae	Pardofelis	Pardofelis marmorata	marbled cat	3

图 3. 环形树形图，来源：<https://www.rawgraphs.io/>

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

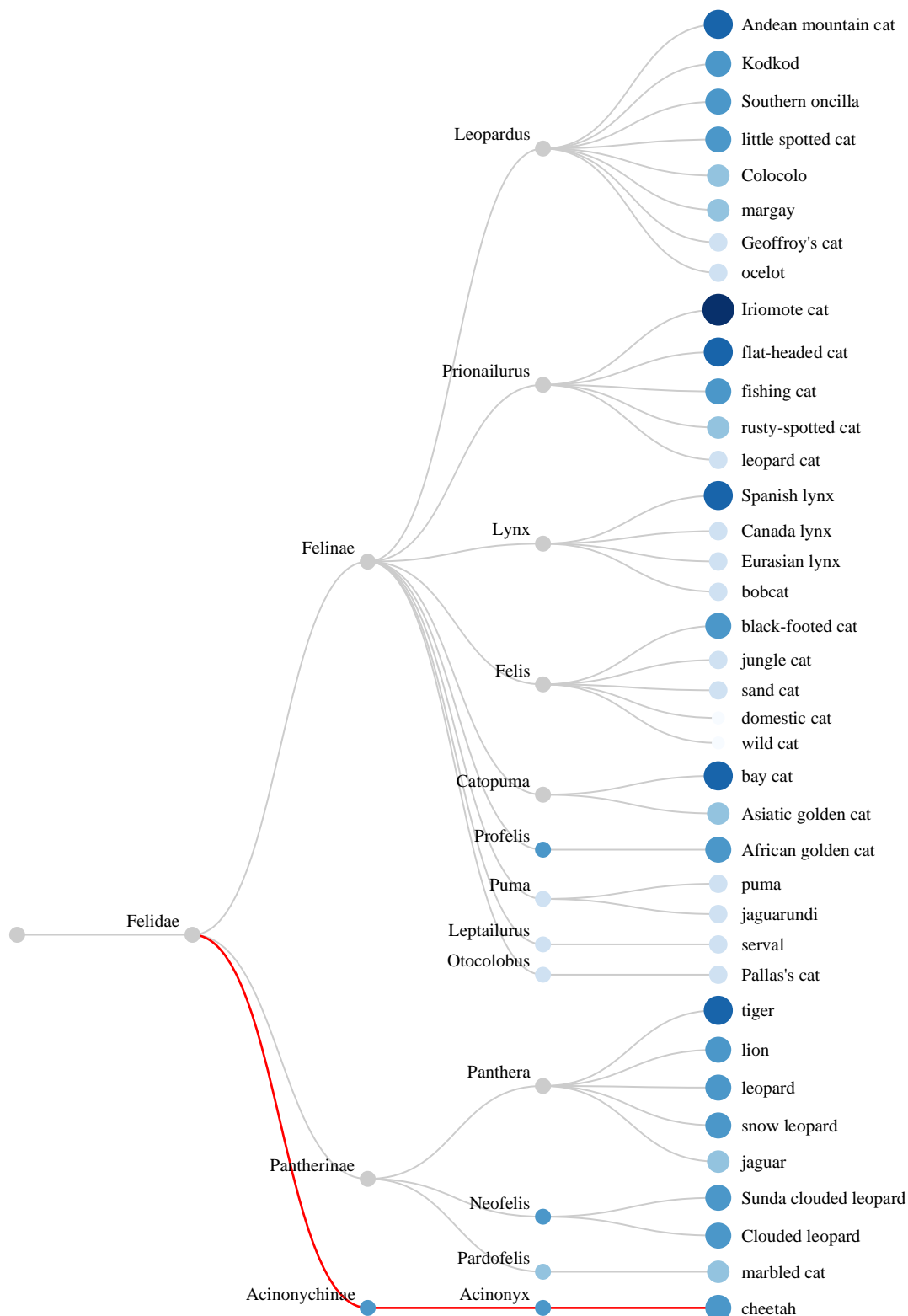
图 4. 水平树形图, 来源 <https://www.rawgraphs.io/>

图 5 所示的太阳爆炸图也可以看做是一种树形图。

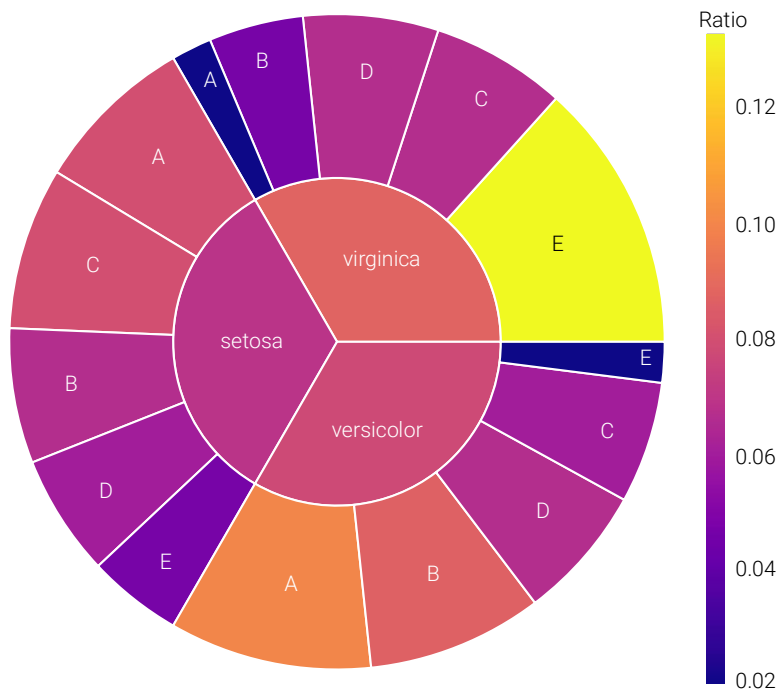


图 5. 太阳爆炸图本质上也是树状图，图片来自《编程不难》

表 2 所示为 26 个英文字母的摩尔斯电码，图 6 所示为根据电码规则绘制的树图。这个示例也是来自 NetworkX，请大家自行学习。

https://networkx.org/documentation/stable/auto_examples/graph/plot_morse_trie.html

表 2. 英文字母的摩尔斯电码

字母	摩尔斯电码	字母	摩尔斯电码
A	. -	N	- .
B	- . . .	O	---
C	- . - .	P	. - - .
D	- . .	Q	- - . -
E	.	R	. - .
F	. . - .	S	. . .
G	- - .	T	-
H	U	. . -
I	. .	V	. . . -
J	. - - -	W	. - -
K	- . -	X	- . . -
L	. - . .	Y	- . - -
M	- -	Z	- - . .

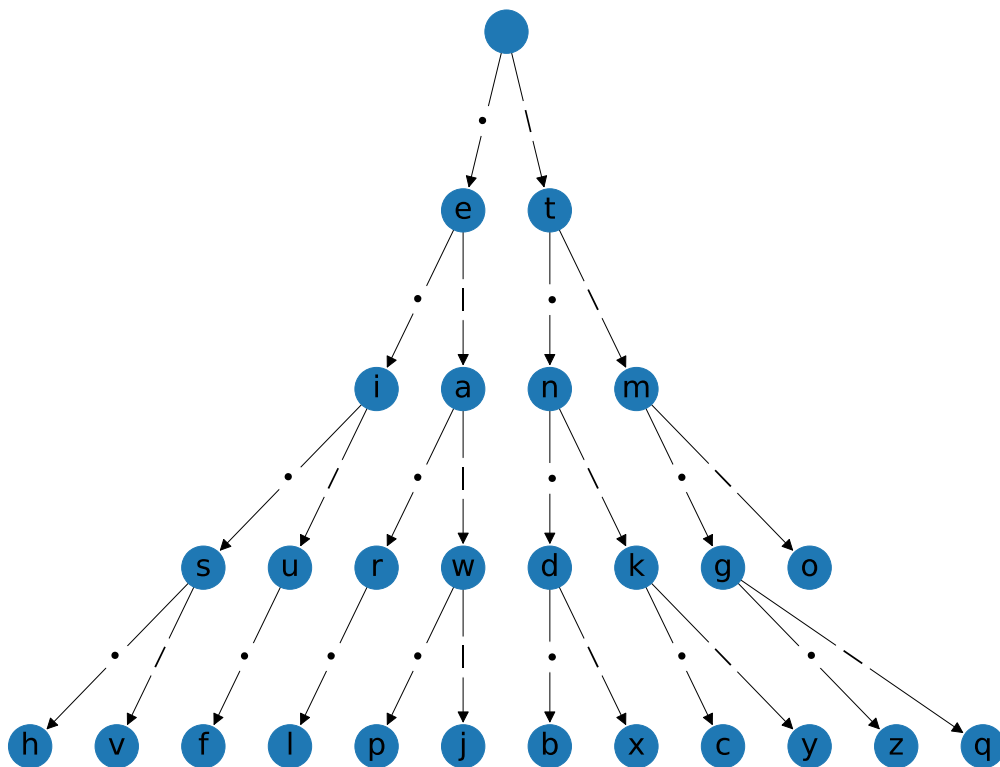


图 6. 26 个英文字母摩斯码构造的树图

在机器学习算法中，树有很多应用案例。

- ▶ 树可以用于搜索算法，解决最短路径问题。
- ▶ 决策树是一种机器学习模型，它使用树结构来表示决策规则。决策树在分类和回归问题中都有广泛的应用。
- ▶ 层次聚类算法使用树结构来表示数据点之间的相似性关系。这种树形结构有助于理解数据的层次性结构，并可视化聚类结果。
- ▶ 随机森林是一种集成学习方法，它包括多个决策树，并通过投票或平均来提高预测性能。树的集成有助于减少过拟合，提高模型的鲁棒性。
- ▶ 在神经网络中，树状结构被用于表示网络的分层结构。这种分层结构有助于提取输入数据的层次性特征。

总结来说，树是一种基本的数据结构，具有一些重要的性质和用途；本章就专门聊聊树这种图。

22.2 最近共同祖先

最近共同祖先 (Lowest Common Ancestor, LCA) 是指在一个树状结构中，两个节点最低的共同祖先节点。在树中，每个节点都有一个父节点 (除了根节点)，而根节点是没有父节点的节点。

考虑一个树状结构，例如家谱 (认祖归宗) 或计算机科学中的树数据结构，每个节点代表一个个体或对象，而边表示父子关系。如图 7 所示，给定树中的两个节点 (a 和 b)，它们的最低共同祖先是指在树中

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

向上移动，直到找到两个节点的最小的公共祖先节点 c 。请大家自己找到图 7 树中节点 d 、 e 的共同祖先。

在计算机科学中，可以在一个文件系统的目录结构中使用 LCA 算法来确定两个文件的共同祖先目录。

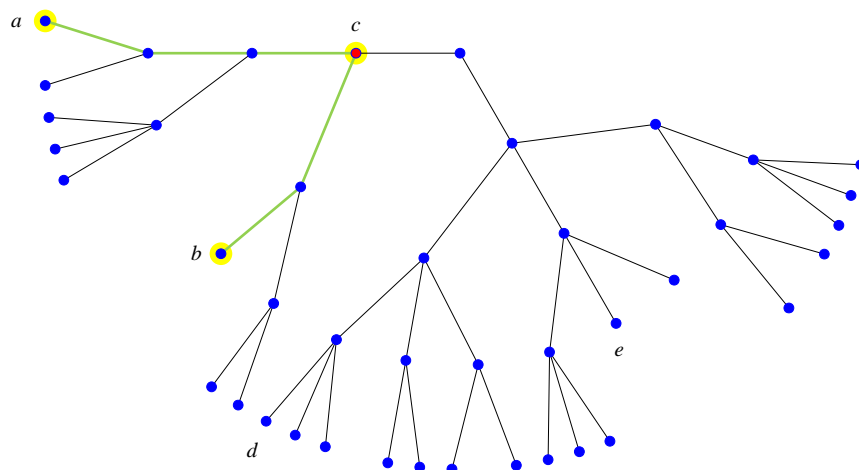


图 7. 最近共同祖先

下面，让我们看看 NetworkX 给出的一个示例。

代码中利用 `networkx.all_pairs_lowest_common_ancestor(G, ((1, 3), (4, 9), (13, 10)))` 找到：

- ▶ 节点 1 和 3 的 LCA 为节点 7；
- ▶ 节点 4 和 9 的 LCA 为节点 6；
- ▶ 节点 10 和 13 的 LCA 为节点 11。

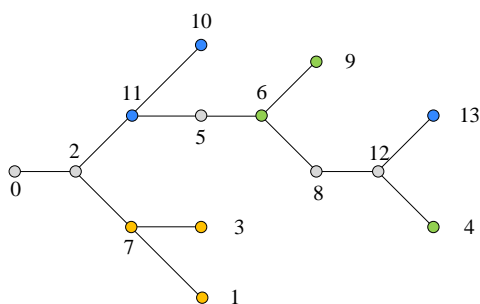


图 8. NetworkX 中最近共同祖先示例

请大家自行学习如下示例。

https://networkx.org/documentation/stable/auto_examples/algorithms/plot_lca.html

22.3 最小生成树

在图论中，**最小生成树** (Minimum Spanning Tree, MST) 是一个连通无向图中的一棵生成树。生成树是一个无环的连通子图，它包含图中的所有节点；但是只包含足够的边，使得这棵树是连通的且权重之和最小。

简单来说，对有 n 个节点的图遍历，遍历后的子图包含原图中所有的点且保持图连通，最后结构一定是一个具有 $n - 1$ 条边的树，这个子图叫**生成树** (spanning tree)。

如图9上图所示，这幅图有9个节点，图中每条边都有自己的权重。我们可以很容易找到一个树(图9下图)，连通所有的节点；这棵树就是所谓生成树，有8条边。

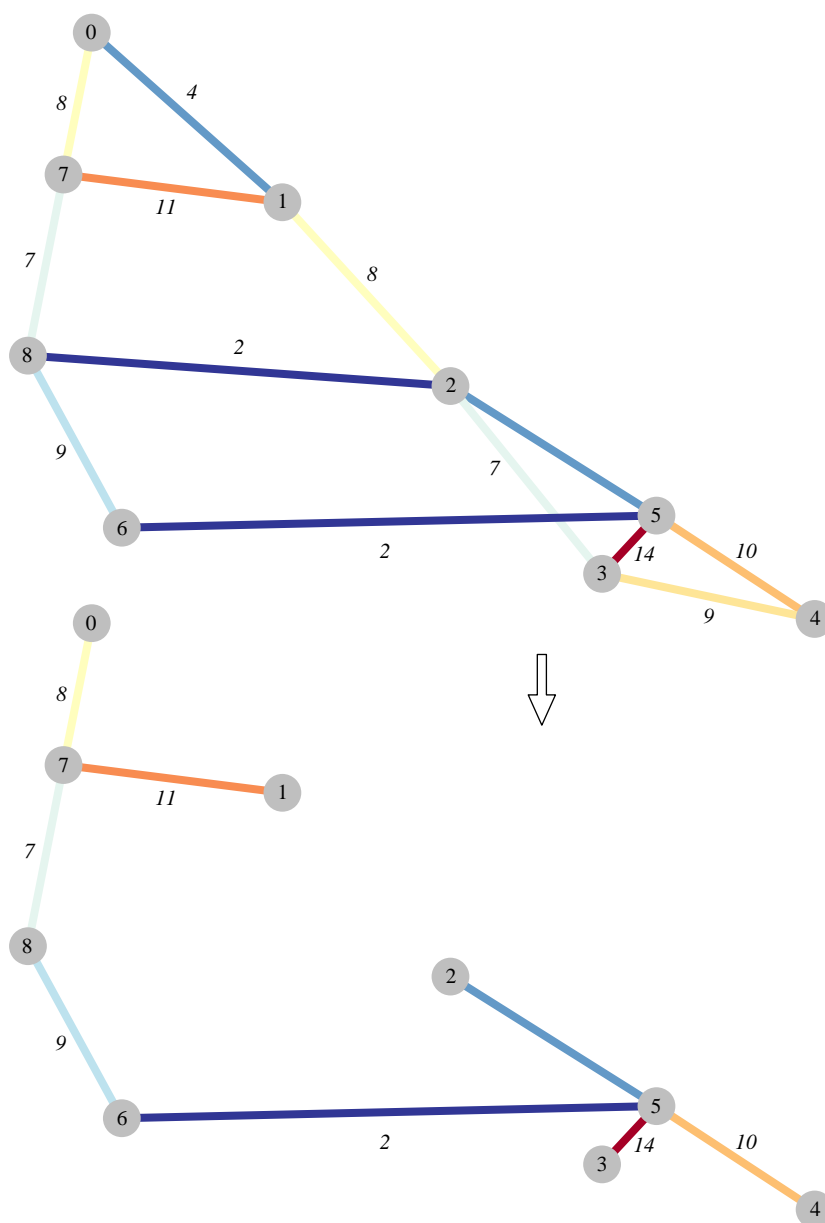


图 9. 生成树

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

“最小”生成树就是在所有可能的生成树中，边的权重之和最小的那棵树。图 10 所示为找到的最小生成树。Bk6_Ch22_01.ipynb 完成本例，这段代码参考 NetworkX 官方示例。

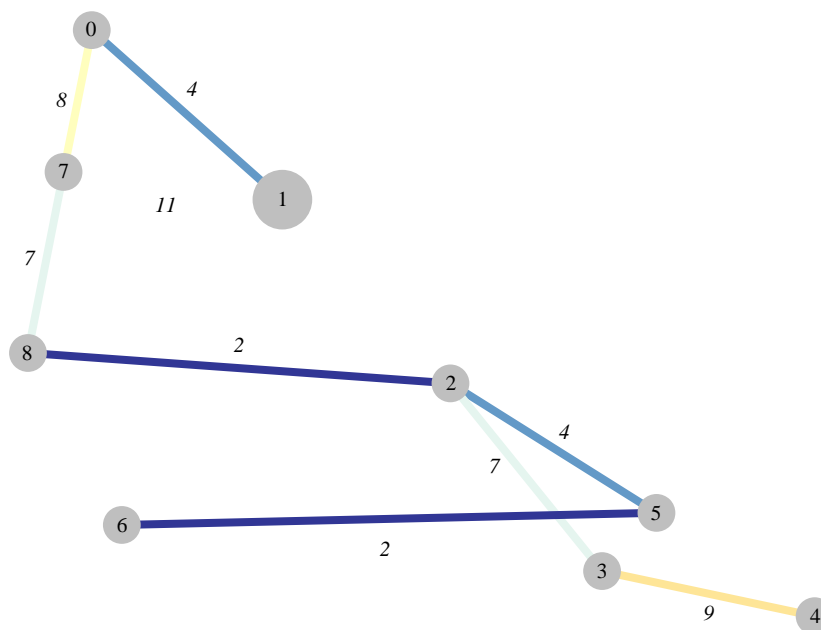


图 10. 最小生成树

最小生成树的应用非常广泛。比如，在设计通信网络时，连接各个节点的成本可能不同。通过找到最小生成树，可以以最小的总成本连接所有节点，确保网络的高效性和经济性。再如，在电路设计中，节点可以表示电路中的元件，边的权重可以表示连接这些元件的成本或电阻。找到最小生成树可以帮助设计出成本最低或电阻最小的电路。

还有，在城市规划中，道路或铁路的建设成本不同。通过最小生成树算法，可以找到以最小的总成本连接城市中各个区域的交通网络。在电力网络设计中，连接不同发电站和消费站的输电线路的成本可能不同。最小生成树可以用于确定最经济的电力传输网络。

22.4 决策树：分类算法

决策树 (decision tree) 是机器学习中常用的分类算法。如图 11 所示，决策树树形结构主要由**结点** (node) 和**子树** (branch) 构成。结点又分为**根结点** (root node)、**内部结点** (internal node) 和**叶结点** (leaf node)。每一个根节点和内部结点一般都是二叉树，向下构造**左子树** (left branch) 和**右子树** (right branch)，构造子树的过程也是将结点数据划分为两个子集的过程。

以包含两个特征的样本点 $x = (x_1, x_2)$ 的分类过程为例。图 12 展示了决策树的第一步划分，首先判断第一个特征 x_1 。当样本数据中 $x_1 \geq a$ 时， x 被划分到右子树；而当样本数据 $x_1 < a$ 时， x 被划分到左子树。经过第一步二叉树划分，原始数据被划分为 A 和 B 两个区域。

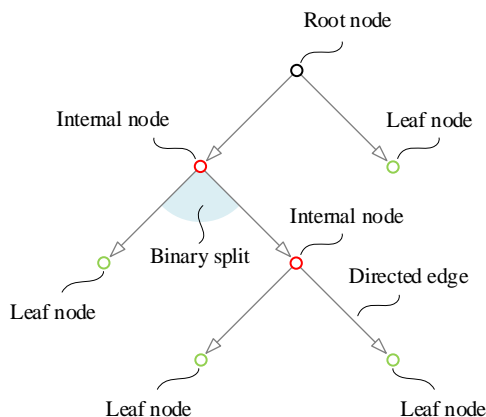


图 11. 决策树树形结构

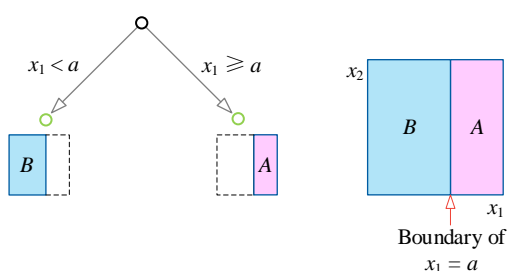


图 12. 决策树第一步划分

接下来，图 13 展示了决策树的第二步划分，为图 12 左子树内部结点衍生出一个新的二叉树，对第二个特征 x_2 进行判断。当样本数据中 $x_2 \geq b$ 时， x 被划分到右子树，而当样本数据中 $x_2 < b$ 时， x 被划分到左子树。经过第二步二叉树划分，原本的 B 数据区域被划分为 C 和 D 两个部分。

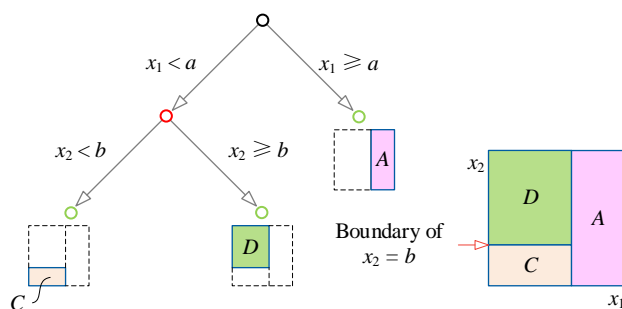


图 13. 决策树第二步划分

图 14 展示了决策树的第三步划分，为图 13 右子树内部结点衍生出又一个新的二叉树。此时，再回到第一个特征 x_1 来进行判断。当样本数据中 $x_1 \geq c$ 时， x 被划分到右子树；样本数据中 $x_1 < c$ ， x 被划分到左子树。经过第三步二叉树划分，原本的 D 数据区域被划分为 E 和 F 两个区域。

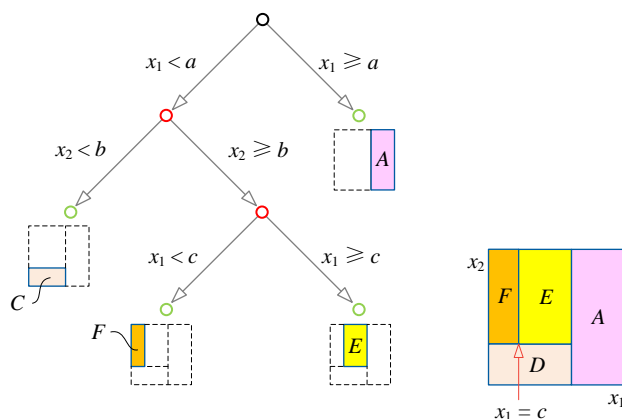


图 14. 决策树第三步划分

下面看个实例。图 15 给出的样本数据有两个特征：个人收入 (x_1) 和信用评分 (x_2)；样本数据有两个分类：优质贷款 (C_1) 和劣质贷款 (C_2)。

根据图 15 数据，可以直观判断，当个人收入和信用评分两者越高，则贷款质量越高，越不容易出现劣质贷款。下面介绍借助决策树分类方法获得判断好坏贷款的决策边界。

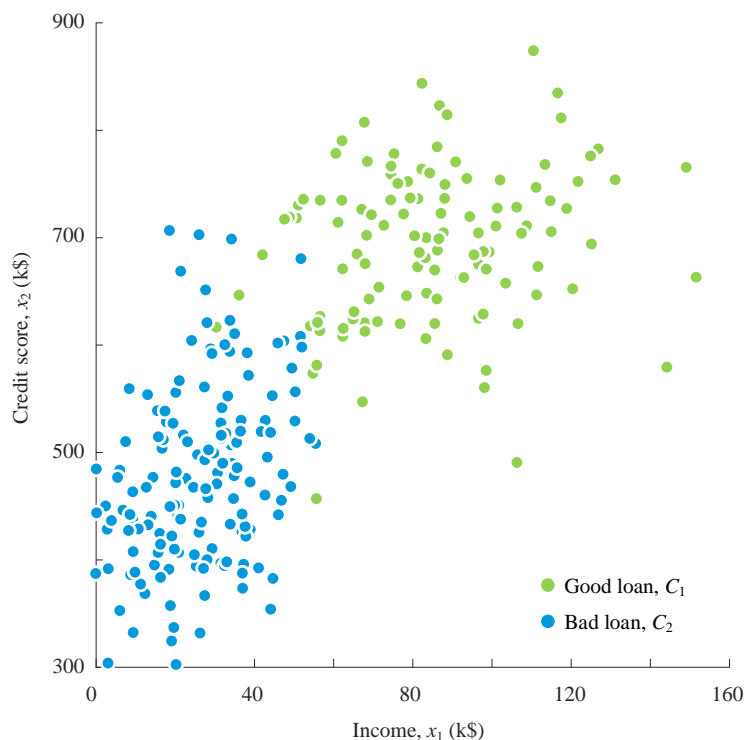


图 15. 根据个人收入和信用评分判断好坏贷款

图 16 至图 19 分别展示了整个分类过程中各步的具体划分。图 20 将图 16 到图 19 集中在一起，展示了整个决策树。

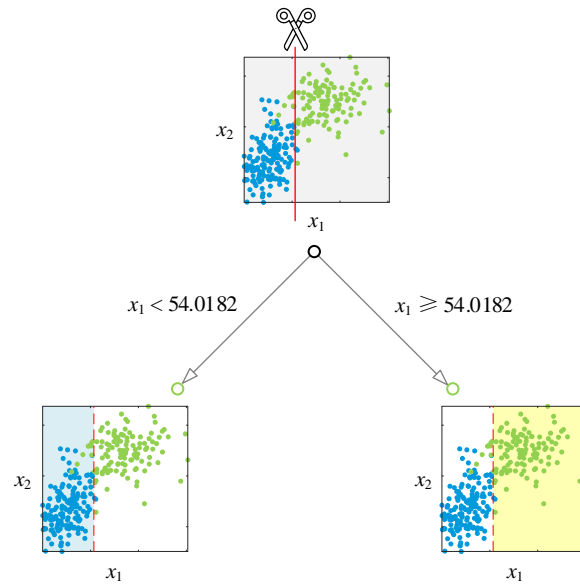


图 16. 好坏贷款数据分类，决策树第一步划分 (沿 x_1)

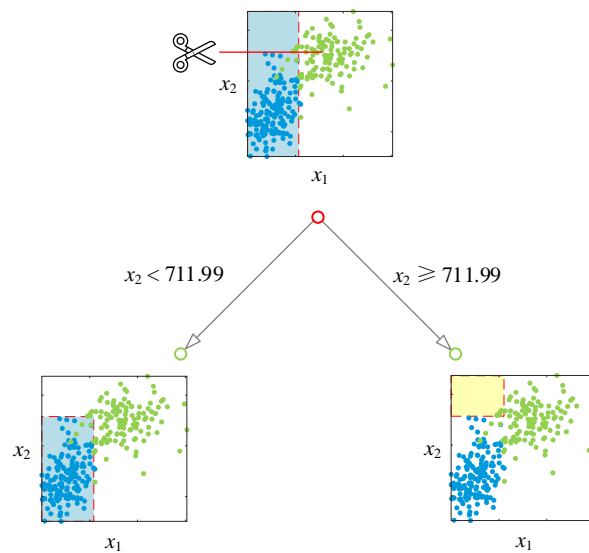


图 17. 好坏贷款数据分类，决策树第二步划分 (沿 x_2)

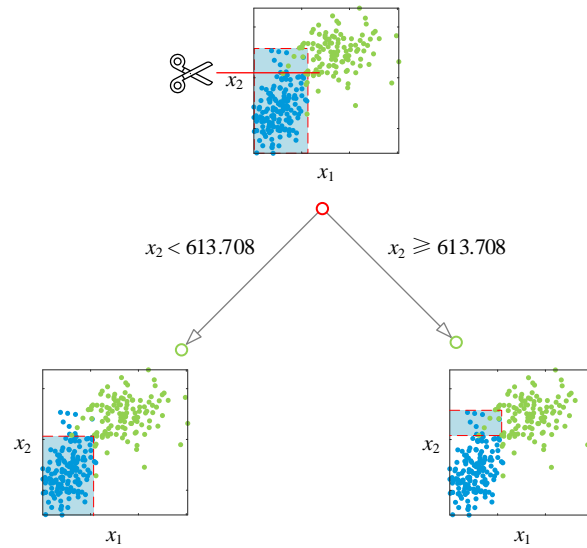


图 18. 好坏贷款数据分类，决策树第三步划分 (沿 x_2)

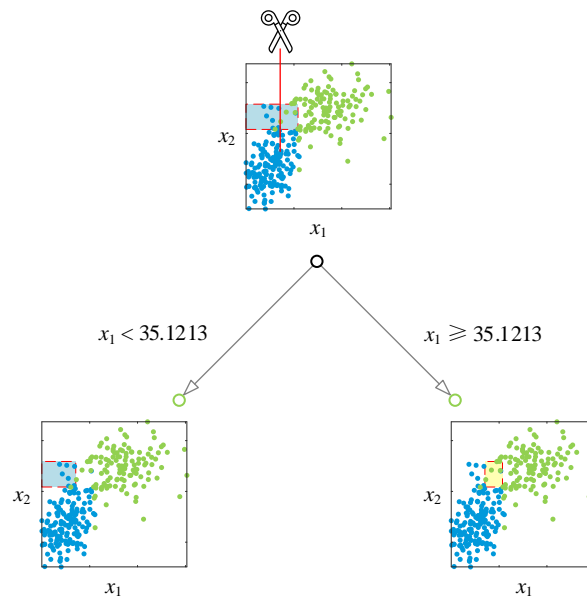


图 19. 好坏贷款数据分类，决策树第四步划分 (沿 x_1)

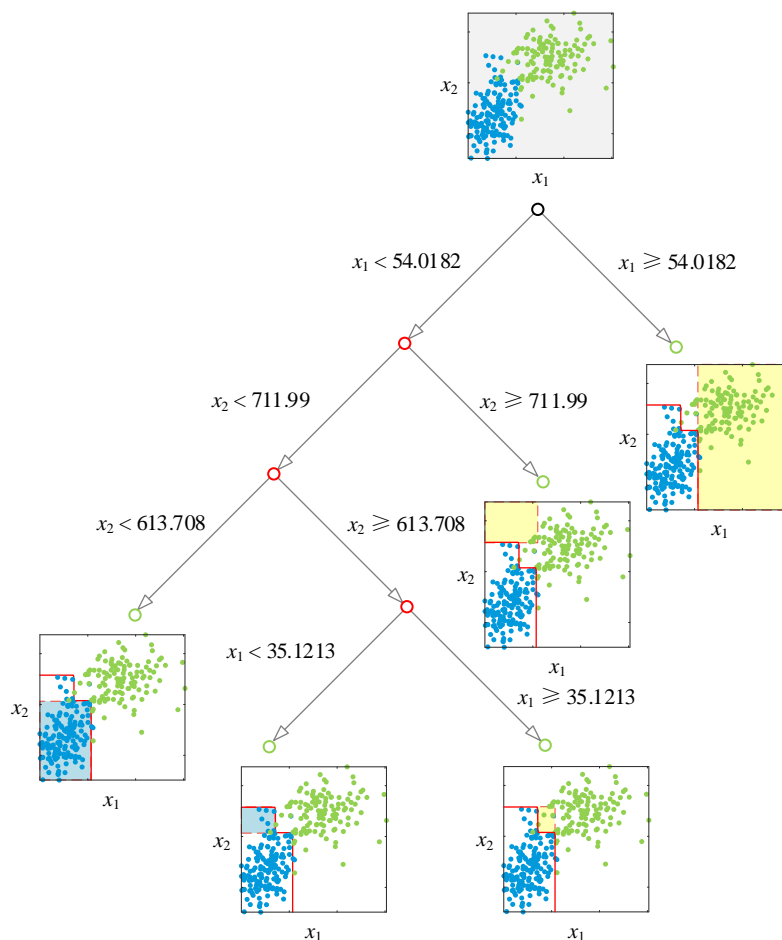


图 20. 好坏贷款数据分类，整个决策树

决策树分类算法有自己独特的优势。决策树的每个节点可以生长成一颗二叉树，这种基于某一特征的二分法很容易解释。

那么问题来了，如何在决策树的每一步中选择哪个特征进行判断，比如本例中 x_1 或 x_2 ？对于 x_1 或 x_2 ，如何找到最佳位置划分呢？这是鸢尾花书《机器学习》一册要回答的问题。

22.5 层次聚类

层次聚类 (hierarchical clustering) 算法是一种聚类算法。层次聚类依据数据之间的距离远近，或者亲密度大小，将样本数据划分为簇。层次聚类可以通过**自下而上** (agglomerative) 合并，或者**自上而下** (divisive) 分割来构造分层结构聚类。

图 21 所示为根据鸢尾花样本数据前两个特征——花萼长度和宽度——获得的层次聚类**树形图** (dendrogram)。

⚠ 注意，层次聚类算法为**非归纳聚类** (non-inductive clustering)。

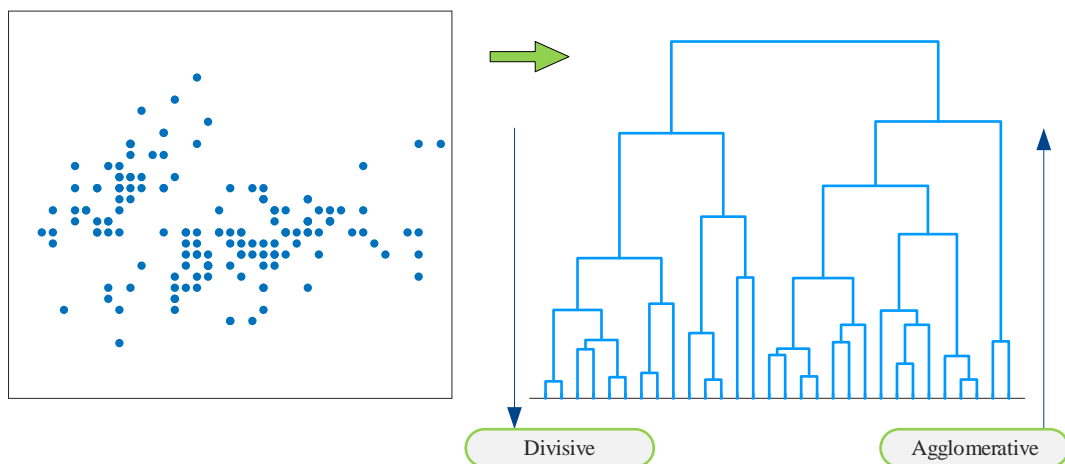


图 21. 区分“自上而下”和“自下而上”层次聚类

这一节采用图 22 样本数据讲解自下而上层次聚类。首先计算样本数据两两欧氏距离。图 23 展示图 22 数据两两距离的方阵构成的热图。请注意图 23 中用不同颜色圆圈 \circ 标记欧式距离，下文构造树形图将会用到这些结果。

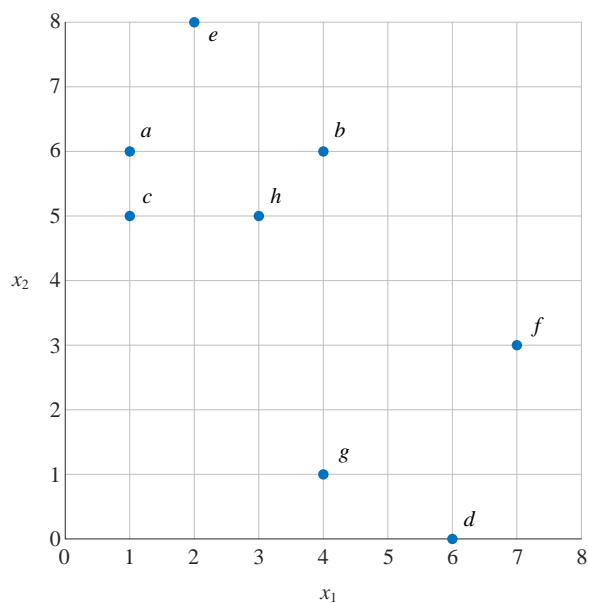


图 22. 样本数据

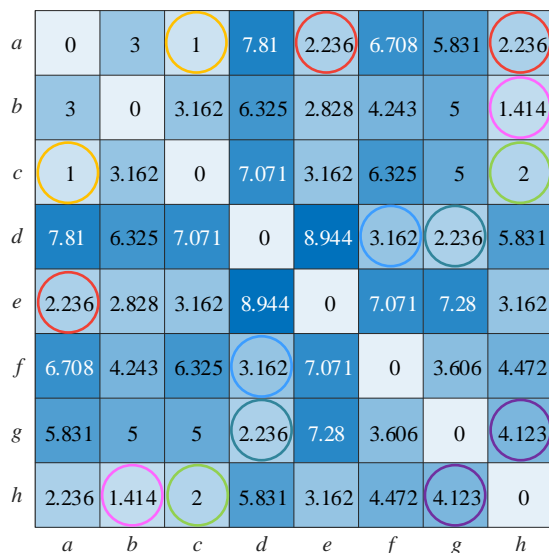


图 23. 8 个样本数据两两距离构成的方阵热图

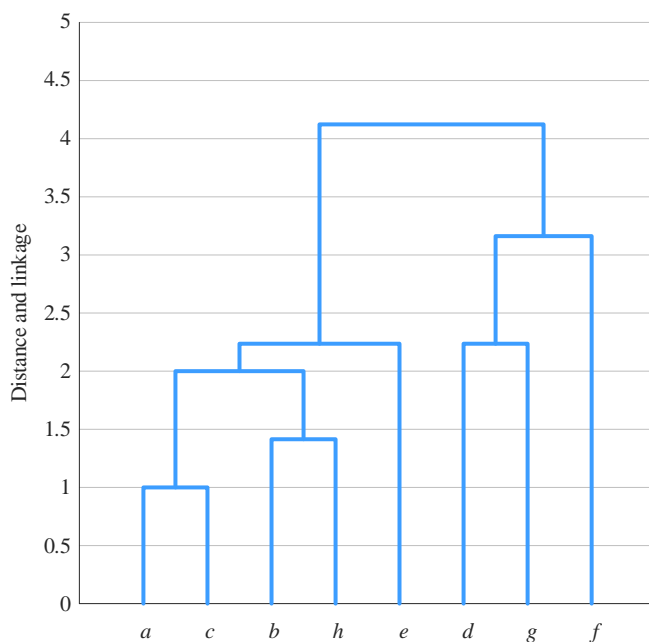


图 24. 数据树形图

图 24 展示了图 22 样本数据的树形图。树形图横轴对应样本数据编号，纵轴对应数据点间距离和簇间距离。

通过观察图 23，容易发现点 a 和 c 的欧式距离为 1，为两两距离中最短距离；点 a 和 c 可以构成最底层 C_1 簇，如图 25 所示。图 23 中，点 b 和 h 的欧式距离为 1.414，为两两距离中第二短；如图 26 所示，点 b 和 h 构成 C_2 簇。

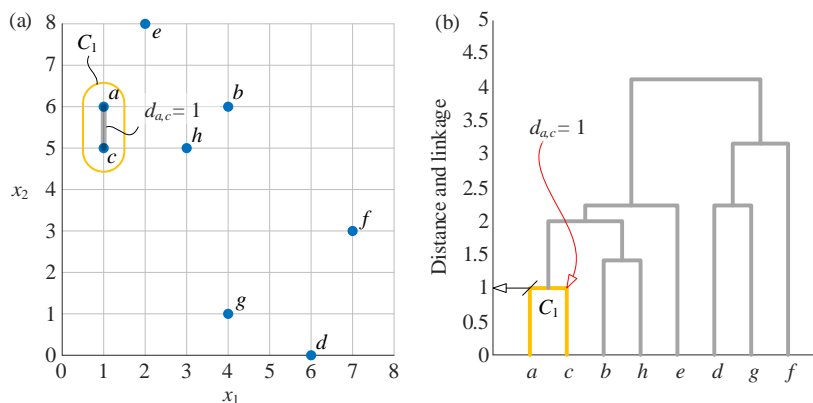


图 25. 构建树形图，第一步

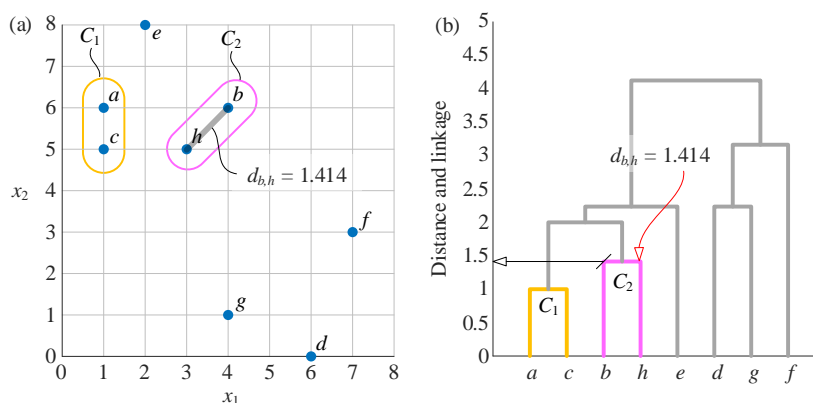


图 26. 构建树形图，第二步

下一步计算两个簇之间的**距离值** (linkage distance, linkage)，这里用 l 表示。图 27 展示的常用的 4 种簇间距离。本例中采用的是图 27 (a) 所示的**最近点距离** (single linkage 或 nearest neighbor)。这种距离指的是两个簇样本数据两两距离最近值。《机器学习》会介绍图 27 所有的簇间距离度量。

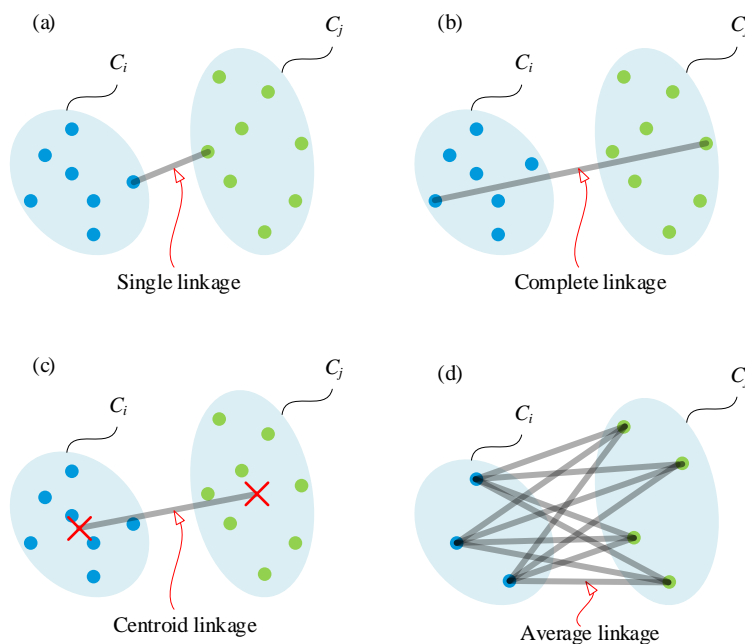


图 27. 簇间距离四种定义

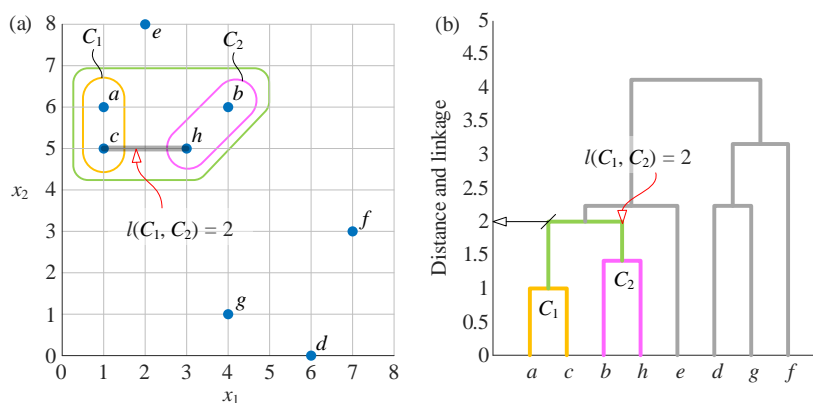


图 28. 构建树形图，第三步

观察图 28 可以发现， C_1 和 C_2 簇间最近点距离为点 c 和 h 之间距离，即 $l(C_1, C_2) = 2$ ； C_1 和 C_2 簇构成 C_3 。

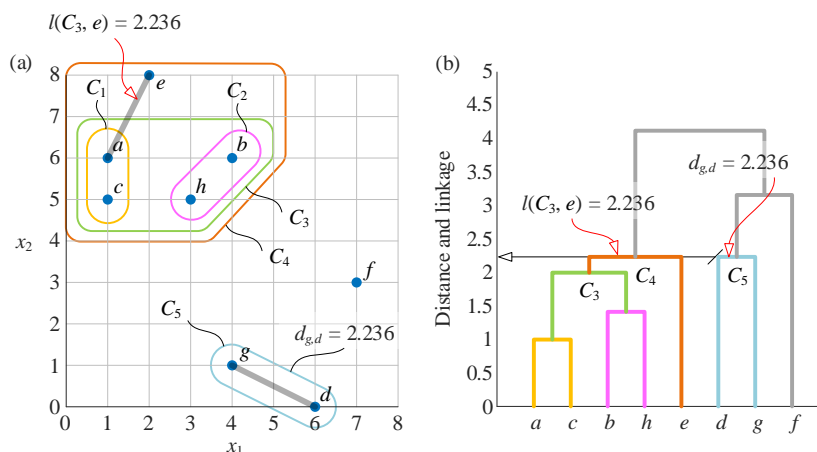


图 29. 构建树形图，第四步

如图 29 所示， e 点被视作簇， C_3 和 e 簇间最近点距离为 $l(C_3, e) = 2.236$ ；同样距离的还有，点 d 和 g 之间的欧式距离 $d_{d,g} = 2.236$ ；点 d 和 g 构成簇 C_5 。

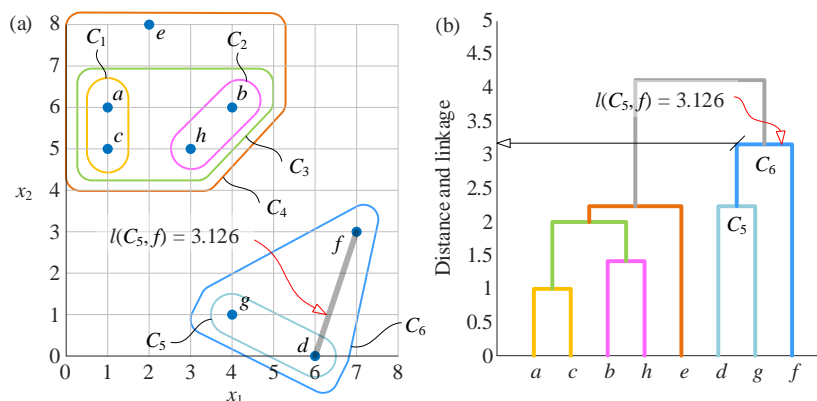
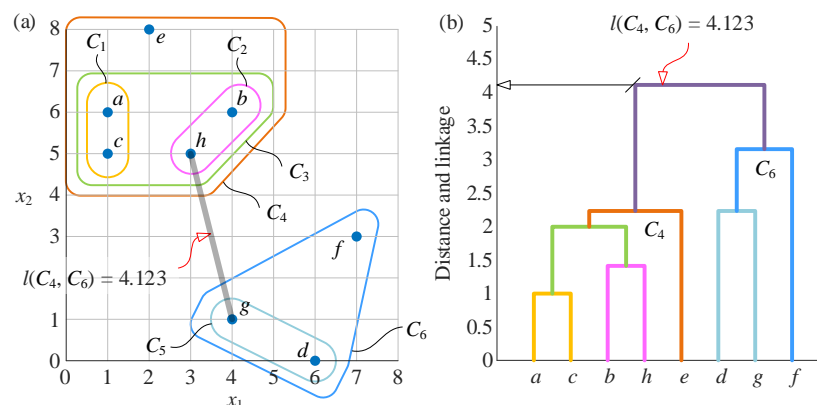


图 30. 构建树形图，第五步

然后，如图 30 所示，点 f 被视作簇，点 f 和 C_5 簇间最近点距离为 $l(C_5, f) = 2$ ； C_5 和 f 簇构成 C_3 。最后，如图 31 所示，簇 C_4 和 C_6 包含所有样本数据，两者簇间最近点距离为点 h 到 g 距离，即 $l(C_4, C_6) = d_{h,g} = 4.123$ 。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 31. 构建树形图，第六步

通过在特定层次切割树形图，可以得到相应的簇划分结果。比如在簇间最近点距离 3.126 和 4.123 之间切割树形图，可以得到 2 个聚类簇，具体如图 32(a) 所示。根据图 32(b) 可知，在簇间最近点距离 3.126 和 2.236 之间切割树形图，可以得到 3 个聚类簇。

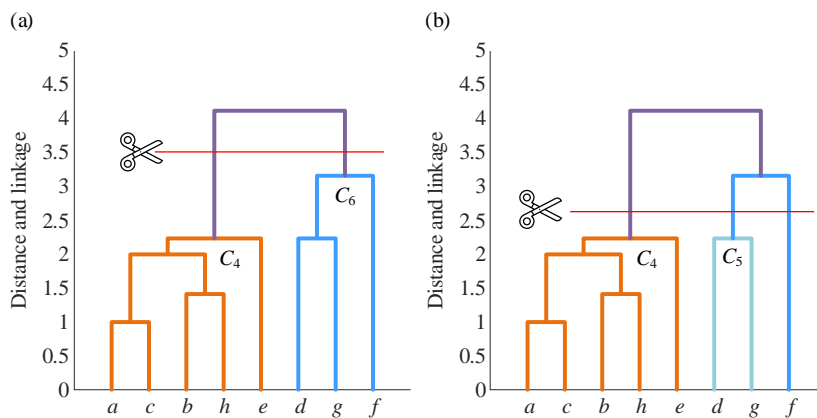


图 32. 在不同层次切割树形图获得 2 个和 3 个聚类簇

22.6 树形图：聚类算法

这一节介绍如何用**树形图** (dendrogram) 完成**聚类** (clustering)。树形图依托上一节介绍的**层次聚类**算法；简单总结一下，层次聚类算法依据数据之间的距离远近，或者亲近度大小，将样本数据划分为簇。

下载 12 只股票历史股价，初值归一走势如图 33 所示。计算日对数回报率，然后估算相关系数矩阵，如图 34 热图所示。相关系数相当于亲近度，相关系数越高，说明股票涨跌趋势越相似。利用树形图，我们可以清楚看到各种股票之间的关联。

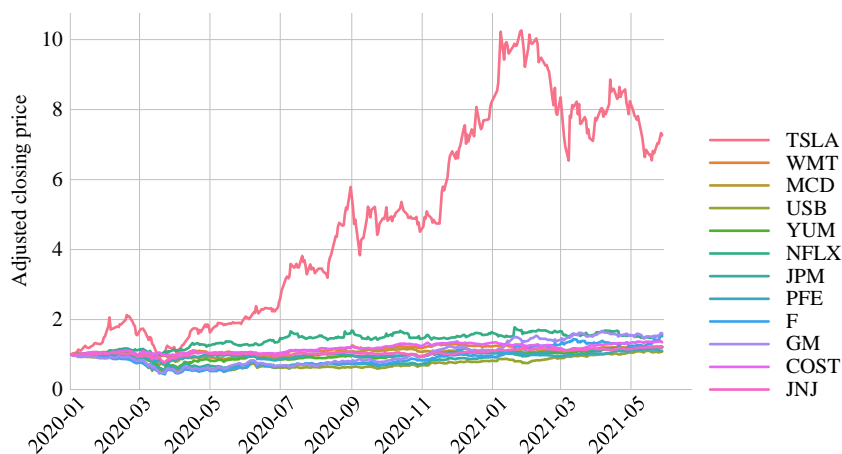


图 33. 12 只股票股价水平，初始股价归一化

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

PFE 和 JNJ 同属医疗，WMT 和 COST 同属零售，F 和 GM 同属汽车，USB 和 JPM 同属金融，MCD 和 YUM 同属餐饮；因此，它们之间相关性高并不足为奇。但是，本应该离汽车更近的 TSLA，却展现出和 NFLX 更高的相似性。

图 35 给出的树形图，直观地表达样本数据之间的距离/亲密度关系。树形图纵坐标高度表达不同数据之间的距离。

USB 和 JPM 之间相关性系数最高，因此 USB 和 JPM 距离最近，所以在树形图中首先将这两个节点相连，形成一个新的节点。然后，MCD 和 YUM 形成一个节点，F 和 GM 形成一个节点 ... 依据这种方式，树形自下而上不断聚拢。

图 35 树形图将股票按照相似度重新排列顺序。图 35 热图发生有意思的变化，热图中出现一个个色彩相近“方块”。每一个“方块”实际上代表着一类相似的数据点。因此，树形图很好揭示股票之间的相似性关系，这便是**聚类** (clustering) 算法的一种思路。

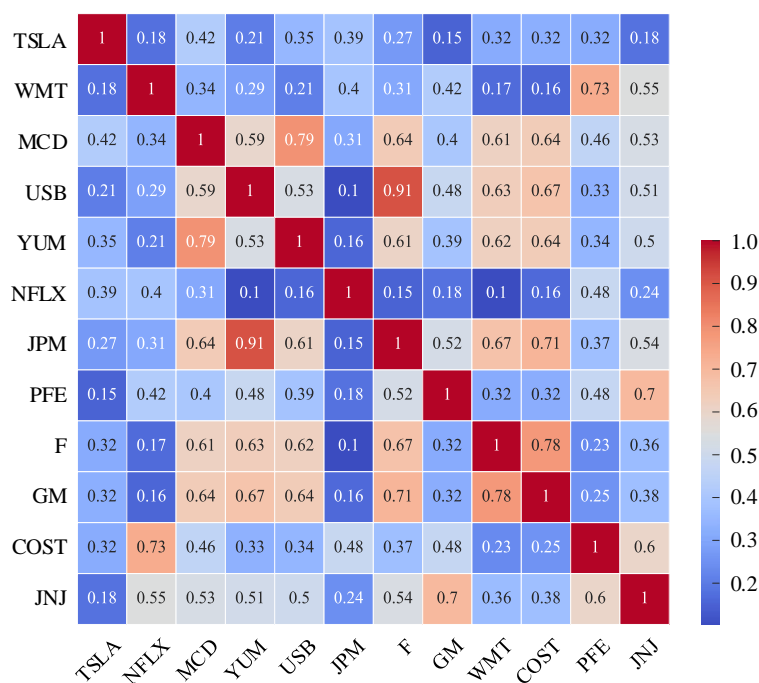


图 34. 12 只股票相关性热图

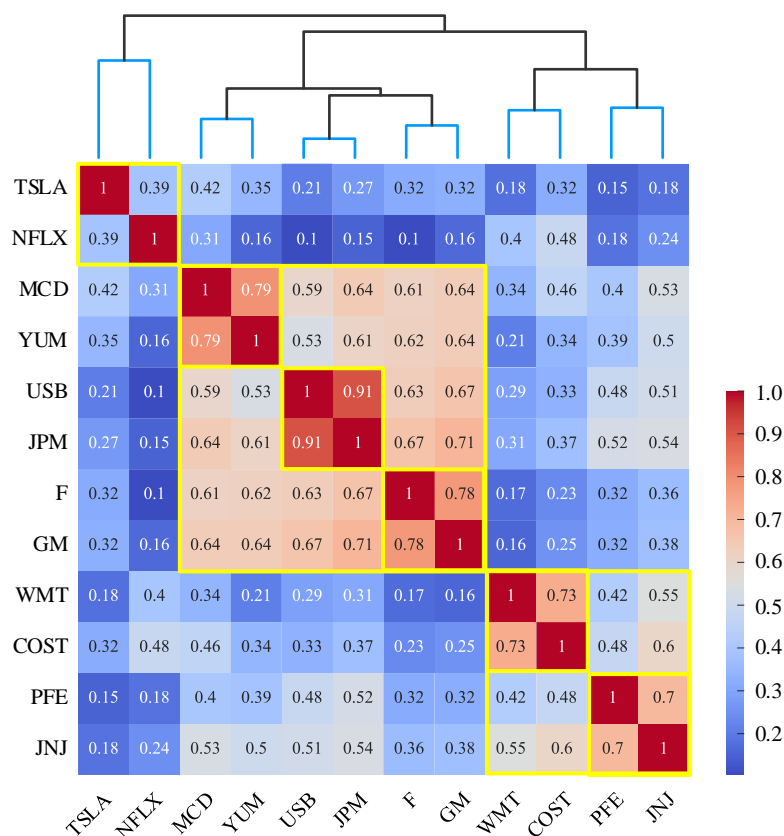


图 35. 根据树形图重组相关性热图



代码 Bk6_Ch22_02.ipynb 绘制图 33、图 34 和图 35。



本章介绍了一种特殊的图——树。请大家务必记住树的几个特点。本章后文还聊了聊几种和树有关的算法，最近共同祖先、最小生成树、决策树、树形图。

决策树是一种常用的分类算法，树形图则依托层次聚类算法，《机器学习》将专门介绍这两种算法。