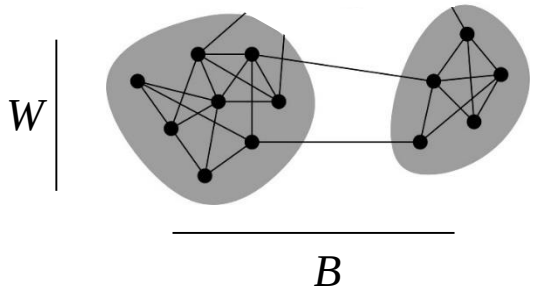
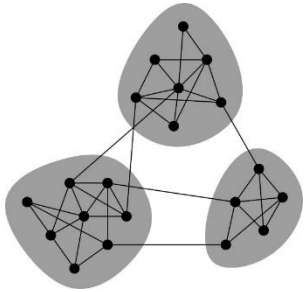
第二十二讲 聚类、分类

2024.6.3



K-均值聚类 (K -means clustering)

K-means （K -均值）聚类方法将对象/物体划分为K 类，使得类内距离极小。 K -均值聚类法是应用非常广泛的一种数据聚类方法（Lloyd，1957；MacQueen 1967）。



平方和分解

单因素方差分析中我们已知有如下平方和分解：

假设$\mathbf{x}_1,...,\mathbf{x}_n \in R^p, K \geq 2$,假设$C_1,...,C_K$是指标集$I=\{1,2,...,n\}$的一个划分(即$C_1,...,C_K$互斥，且$\bigcup C_i = I$). 则总平方和分解为(向量模平方形式的)组内平方和W和组间平方和B:

$$SS_T = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_k\|^2 + \sum_{k=1}^{K} \sum_{i \in C_k} \|\bar{\mathbf{x}}_k - \bar{\mathbf{x}}\|^2 \overset{\Delta}{=} W + B$$

组间平方和B相对于组合平方和W（比如BW^{-1}）越大，表明数据越有可能确实是聚集为K个类的。

K均值聚类问题

给定n个点$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \in R^p$，以及类的个数$K$.

目标：寻找最优的$I = \{1,2,...,n\}$的划分$C = \{C_1, ..., C_K\}$，以及各组的中心$\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}$使得组内平方和最小

$$\min_{C_k, \boldsymbol{\mu}_k, k=1,..,K} \sum_{k=1}^{K} \sum_{i \in C_k} \| \mathbf{x}_i - \boldsymbol{\mu}_k \|^2$$

划分：C_k's互斥且$\bigcup_{k=1}^{K} C_k = I$

K-means算法

K-means算法（Lloyd，贪心算法）

- 当划分给定时，第k个类的中心$\boldsymbol{\mu}_k$的最优估计为平均值

$$\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i, \quad |C_k| \text{为} C_k \text{中点的个数}$$

- 而当中心$\boldsymbol{\mu}_k, k = 1,..., K$给定时，$C_k$容易决定：

$$C_k = \{1 \le i \le n: \quad \| \mathbf{x}_i - \boldsymbol{\mu}_k \| \le \| \mathbf{x}_i - \boldsymbol{\mu}_s \|, \text{对所有} s \ne k \}$$

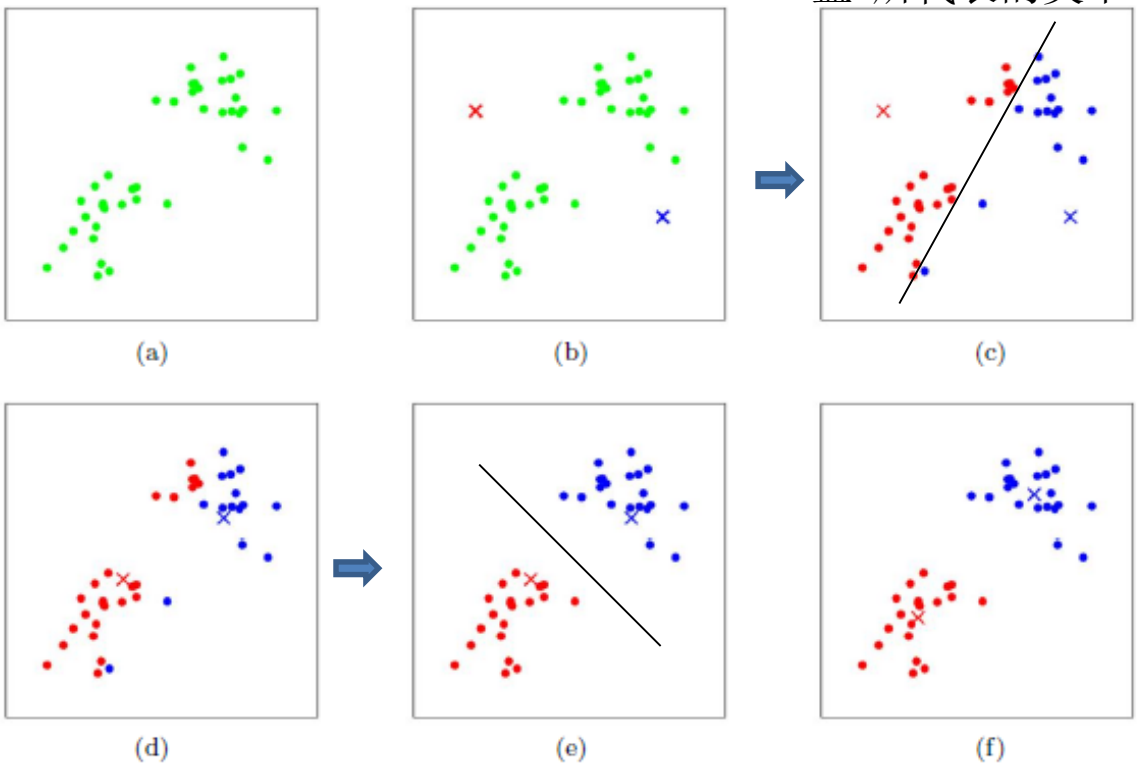
反复迭代上述两步，直至收敛。

K-means算法示意图

(a)数据

(b)任取两个初始中心(以红x 、蓝色x表示）

(c)红x 、蓝色x中垂线将数据分别划分到红、蓝x所代表的类中



(d) 计算红点和蓝点的均值，作为两类新的中心(红、蓝x)。

(e)红、蓝x)的中垂线将数据分别重新划分到新的红、蓝x所代表的类中

(f) 重新计算中心，再次分类，与上次相同。停止迭代。

注：

- 初始步可以指定初始中心，也可以指定初始的划分。
- K-means, K-medoid方法的算法思想几乎一样，都是中心、划分两个步骤的反复迭代。是原型*prototype*方法，以中心代表类，称为原型，
- K-means得到的解未必一定是全局最优解，但非常高效，其计算复杂度为$O(npKi)$,其中$i=$迭代次数，i通常较小$(i<10)$.
- 为了避免得到局部最优解，通常取多组不同起始点，多次进行聚类，从中选取目标函数最小者，*R*函数*kmeans*:

  $>kmeans(\text{x}，centers=k,nstart=25)$

  #25次,每次随机选取*x*的*k*行作为*k*类的中心初值。

高斯混合分布模型

高斯混合分布假设数据点的分布是若干高斯正态分布的混合，允许不同的类有不同的方差。 K-means聚类方法可看作是高斯混合分布模型的一种特殊情况（球对称情况）。

Gaussian mixture

记$f_k(\mathbf{x})$为正态分布$N_p(\boldsymbol{\mu}_k, \Sigma_k)$的密度函数，若随机向量$\mathbf{x}$的概率密度

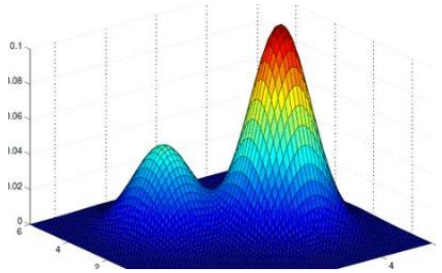
$$f(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}),\ p_k \geq 0,\ \ p_1 + \ldots + p_K = 1,$$

称为是一个高斯混合分布（Gaussian Mixture Model）。

记$\Theta_k = \{\boldsymbol{\mu}_k, \Sigma_k,\ p_k\}$为第$k$个类的所有参数。

为什么假设多个正态混合？

总体不一致，不是单个正态，而是概率分布具有多个峰，因此假设总体由多个正态子总体（类）混合而成。具体到每个样本，我们并不知道它来自于哪个总体。



混合高斯作为潜变量模型

假设$\mathbf{x}$有一个对应的类别标号G (潜变量，不可观测)，其分布为

$$P(G=k)=p_k,\ p_k\geq 0,\ 1\leq k\leq K,\ \sum_{k=1}^{K}p_k=1.$$

假设给定$G=k$时，$\mathbf{x}\sim N(\boldsymbol{\mu}_k,\Sigma_k)$,即$\mathbf{x}\,|\,G=k\sim f_k(\mathbf{x})$，则$\mathbf{x}$的边际分布：

$$f(\mathbf{x})=\sum_{k=1}^{K}P(G=k)P(\mathbf{x}\,|\,G=k)=\sum_{k=1}^{K}p_k f_k(\mathbf{x}).$$

混合高斯的EM算法

EM算法试图预测每个样本所属的类别$G_i, i=1,...,n$，其原理和步骤（E-step和M-step反复迭代）本质上与k-means算法相同：

E-step：在各类中心已知条件下，E-step将每个样本划分到其所属的类，这需要计算每个样本点属于每个类的条件概率/条件期望，然后将每个样本点划分到概率最大的一类中。

M-step：当所有样本所属类别已知的条件下，用极大似然方法估计每个类的中心（以及其它参数）。

EM算法细节: 条件期望E用于预测潜变量， 极大似然M用于估计参数

- M - step（Maximization - 极大似然）：估计类的中心

当知道每个样本所属类别时(即已知G_i，$i=1,...,n$), 求各类中心：

$$\hat{\boldsymbol{\mu}}_k = \sum_{G_i=k} \mathbf{x}_i / n_k = \sum_{i \in \text{class } k} \mathbf{x}_i / n_k$$ （第k类内样本的平均值）

以及$\hat{\Sigma}_k = \sum_{G_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}} / n_k, \hat{p}_k = \sum_{i=1}^{n} 1_{(G_i=k)} / n, k = 1,...,K.$

- E - step（Expectation - 条件期望）：预测G_i

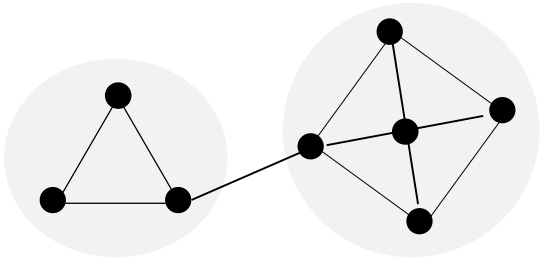
给定$\Theta_k, k=1,...,K$, 将每个数据点划分到概率最大的类中：

若$f_k(\mathbf{x}_i) > f_l(\mathbf{x}_i), \forall l \neq k$，则$G_i$预测为$\hat{G}_i = k,$

即$\hat{G}_i = \arg\max_k P(G_i = k \mid \hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k).$

谱聚类

基于相似度（距离/邻接/权重）有关的矩阵的特征向量进行聚类。



图表示

谱聚类通常在图(graph)或网络的框架中讨论或演示。给定n个物件的$n \times n$相似度矩阵$A = (a_{ij})$，$a_{ij} \geq 0$，我们将物件当作图的节点，若$a_{ij} > 0$，则称节点i, j是相邻的，即它们之间有一条边（连线），记作$i \sim j$，而相似度a_{ij}称为该条边的权重。$A = (a_{ij})$通常称为权重矩阵，当$a_{ij} = 0$或1时，称为邻接矩阵。

邻接矩阵

最重要的是相似度/权重为0-1的情形，

$$a_{ij} = \begin{cases} 1, & i \sim j \\ 0, & \text{无边} \end{cases}$$

此时$A = (a_{ij})$称为邻接矩阵(adjacency matrix)。

下面我们考虑的相似系数矩阵$A = (a_{ij})$一般是 元素非负的对称权重矩阵，但通常会以0-1邻接矩阵演示算法。

度数

第i个节点的度数: $d_i = \sum_{j=1}^{n} a_{ij}$

度数矩阵: $D = \mathrm{diag}(d_1, \ldots, d_n)$

记 $\mathbf{d} = (d_1, \ldots, d_n)^\top = A\mathbb{1}$

拉普拉斯矩阵

拉普拉斯矩阵(Laplacian):

$$L = D - A$$

作为线性变换，L与拉普拉斯算子有类似的作用，即拉普拉斯矩阵比较每个节点i与邻点的平均。说明如下：

仅考虑0-1邻接情形（其它类似）。假设节点i的属性/特征标签x_i，$\mathbf{x} = (x_1, \ldots, x_n)^\top$，令$\mathbf{y} = L\mathbf{x}$，令 $\bar{x}_i = \sum_{j:j\sim i} x_j / d_i$ 为与节点 i 邻接/相似的节点的平均，则

$$y_i = d_i x_i - \sum_{j:j\sim i} x_j = d_i(x_i - \bar{x}_i),$$

表示节点 i 与其相邻点的平均值的差异 。$L\mathbf{x} = 0$意味着每个节点与其相邻的点无差异。

拉普拉斯算子：$\Delta f = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2},$

$$\frac{\partial^2 f}{\partial x^2} \approx \frac{f'(x+h) - f'(x)}{h} \approx \frac{[f(x+h) - f(x)]/h - [f(x) - f(x-h)]/h}{h} = \frac{f(x+h) + f(x-h) - 2f(x)}{h^2}$$

热传导方程：$\frac{\partial T}{\partial t} = \kappa \Delta T$，某点的温度变化率与该点附近的温差成正比。

如果加权图的节点由K个互不连通的子集构成，则我们认为每个子集构成一个类(*cluster*)。如果图是全连通的，我们希望划分节点为K个cluster，不同cluster之间的相似性/连通性最小。

2-cluster mincut

假设节点分属若干不同的类，*cut*定义为相邻但不同类的边的相似系数/权重之和

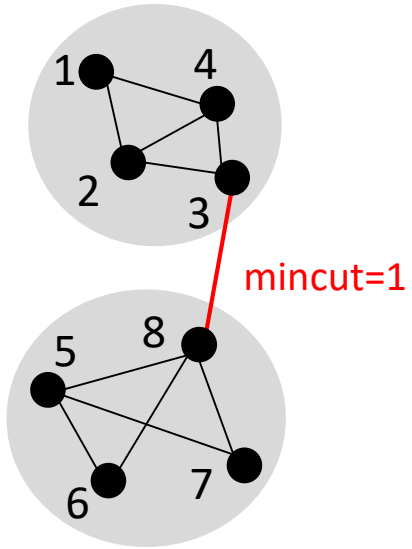
$$\text{cut} = \sum_{i,\ j} a_{ij} 1(i \sim j，\text{但} i, j \text{不同类})$$

最优划分使得cut数最小。

为什么称为*cut*？对于0－1图，即$a_{ij} = 0/1$的情形：

cut＝相邻但不同类的边的个数，

如果移除这些边，则图可以划分为若干互不连通的子集。右图中移除边2－3和3－4可得到不连通的(1,2,4)和(3,5,6,7,8)；移除边3－8，则得到不连通的(1,2,3,4)和（5,6,7,8）。后者是更好的划分，我们希望划分/聚类使得cut最小。



2-cluster 组合优化

下面给出mincut问题的数学表示，为此引入每个节点的类别标签，先考虑两类问题。记节点 i 的类别标签$x_i = 1$或0，则

$$cut = \sum a_{ij} 1_{(x_i \neq x_j)} = \sum a_{ij}(x_i - x_j)^2$$

若a_{ij}很大，则应该有 $x_i = x_j$

这是组合优化问题。与谱配列中一样，我们有

命题1. $\sum a_{ij}(x_i - x_j)^2 = 2\mathbf{x}^\top L\mathbf{x}$， $L = D - A$

验证：$\sum_{i,j} a_{ij}(x_i - x_j)^2 = 2\sum_j d_i x_i^2 - 2\sum_{i,j} a_{ij} x_i x_j = 2\mathbf{x}^\top (D - A)\,\mathbf{x} = 2\mathbf{x}^\top L\mathbf{x}$

放松$x_i = 1$或0的限制，对 $\mathbf{x} = (x_1, \ldots., x_n)^\top \in S^{n-1}$极小化

2-cluster 二次优化

定义$L = D - A$为拉普拉斯矩阵，mincut问题转化为：

$$\text{min}cut = \min \sum a_{ij}(x_i - x_j)^2 = 2 \min_{\mathbf{x}} \mathbf{x}^\top L\mathbf{x} \text{ s.t. } \mathbf{x} \in S^{n-1}, \ \mathbf{x} \perp \mathbb{1}$$

约束：为了解唯一及避免平凡解

得到的解$\mathbf{x}$是L的特征向量，含有cluster信息（参见命题2，3），将它们划分为两类即可 。

K类 mincut

考虑将n个节点划分为K个类：$C_1,\ldots,C_K$，以$V_{n\times(K-1)}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathsf{T}}$代表节点的类别属性，其中$\mathbf{x}_i^{\mathsf{T}}=(1_{(i\in C_1)},\ldots,1_{(i\in C_{K-1})})$为$C_1,\ldots,C_{K-1}$的示性变量

$$\text{cut}=\sum a_{ij}1_{(\mathbf{x}_i\neq\mathbf{x}_j)}=\frac{1}{2}\sum_{i,j}a_{ij}\|\mathbf{x}_i-\mathbf{x}_j\|^2=tr(V^{\mathsf{T}}LV)$$

其中V每一行仅有一个1，其余全为0.

现在放松V分量为0-1的要求，但为了唯一性，增加V各列模长为1，相互正交、且与**1**正交的限制，即$V^{\mathsf{T}}V=I_{K-1}$

拓展的两类 mincut

定义$L=D-A$为拉普拉斯矩阵，mincut问题转化为：

$$\min cut=\min\sum a_{ij}\|\mathbf{x}_i-\mathbf{x}_j\|^2\Leftrightarrow\min_{V\in R^{n\times(K-1)}}\text{tr}\,V^{\mathsf{T}}LV$$

s.t. $V^{\mathsf{T}}V=I_{K-1}$.

谱聚类对L的前K个最小特征根（包括0）的特征向量进行K-means聚类，原因参见下页命题2-3。

全连通加权图：对任何两个节点i, j,存在节点$x, y, ..., z$,使得$i \sim x \sim y \sim ... \sim z \sim j$,其中$a_{ix} > 0, a_{xy} > 0, ..., a_{zj} > 0$.

L的特征向量与连通性

命题2.（1）$L \geq 0$，最小特征根为0，$\mathbf{1}$是特征根0对应的特征向量；（2）图是全连通的（即对任何两个节点，都存在一个若干边组成的路径连结它们）当且仅当L的0特征根重数为1（即除了常数因子之外，$\mathbf{1}$是唯一特征向量）。

拉普拉斯矩阵L的最小非0特征根称为Fiedler数

证明：（1）由命题1，对任何$\mathbf{x}$，$\mathbf{x}^{\mathsf{T}} L\mathbf{x} = \frac{1}{2}\sum a_{ij}(x_i - x_j)^2 \geq 0$，所以$L \geq 0$.

因为$L\mathbf{1} = (D - A)\mathbf{1} = D\mathbf{1} - A\mathbf{1} = \mathbf{d} - \mathbf{d} = 0$,所以0是特征根，$\mathbf{1}$是特征向量。

（2）若$\mathbf{x}$是0特征根的特征向量，则$L\mathbf{x} = 0\mathbf{x} = \mathbf{0} \Rightarrow$

$$\mathbf{x}^{\mathsf{T}} L\mathbf{x} = \frac{1}{2}\sum_{i\sim j} a_{ij}(x_i - x_j)^2 = 0 \Rightarrow \text{当} a_{ij} > 0 \text{时，} x_i = x_j,$$

若图是全连通的，则$x_1 = ... = x_n$,即$\mathbf{x} = \mathbf{1}x_1$，唯一，所以0重数为1.

命题3. 拉普拉斯矩阵L的0特征根的重数K(L的零空间维数)等于图（包括加权图）的互不连通的全连通子图的个数，且零空间由 $\mathbf{1}_{G_1},...,\mathbf{1}_{G_K}$ 张成，其中$G_1,....,G_K$为连通子图的节点集合。

证明：不同的子图之间没有连结，所以可以重排L成为分块对角阵：

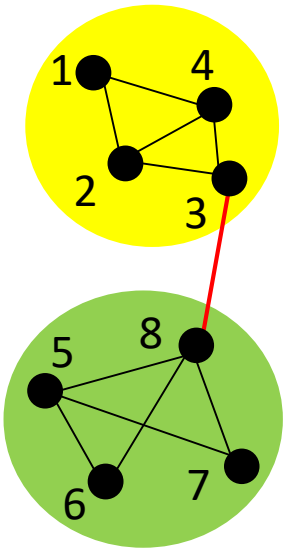
$$L=\begin{pmatrix} L_{G_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & L_{G_K} \end{pmatrix},$$

由命题2，在连通子图上特征向量$\mathbf{x}$的分量取值全部相同，

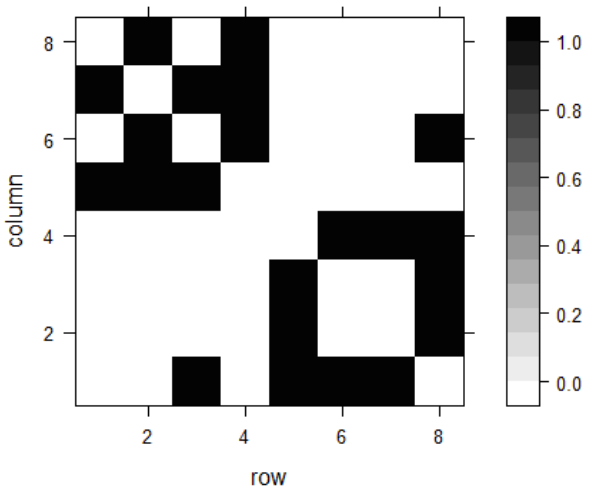
不同的全连通子集上取值可能不同，即$\mathbf{x}=c_1\mathbf{1}_{G_1}+...+c_K\mathbf{1}_{G_K}$.证毕。

如果一个图由不连通的K个子图构成，那么0特征根的所有K个特征向量包含所有子图信息。 实际问题中的图一般是全连通的（特征根0的重数为1）， L的$K-1$个最小非0特征根对应的特征向量可用来聚类。

例1. 下图的邻接矩阵如下



$$A=\begin{array}{c|cccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 4 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 7 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 8 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \end{array}$$



拉普拉斯矩阵L的最小4个特征根为: 0, 0.3, 2, 2.4，对应的特征向量为

特征根	0	0.3	2	2.4
1	-0.35	-0.45	0	0.69
2	-0.35	-0.37	0	-0.12
3	-0.35	-0.18	0	-0.65
4	-0.35	-0.37	0	-0.12
5	-0.35	0.37	0	0.12
6	-0.35	0.37	-0.71	0.12
7	-0.35	0.37	0.71	0.12
8	-0.35	0.26	0	-0.17

第二列特征向量说明

节点1-4为一类，

节点5-8为一类。

谱聚类算法

目标：给定n个物体的相似度矩阵A，将n个物体聚为K类，K已知。

- 假设$A_{n\times n}$是n个个体/物体的相似度矩阵（如果是0-1矩阵，可认为是普通图的邻接矩阵，否则认为是加权图的权重矩阵）。
- 计算度数矩阵$D = diag(\mathbf{d})$，其中$\mathbf{d} = (d_1,...,d_n)^{\mathsf{T}} = A\mathbf{1}$，拉普拉斯$L = D - A$
- 计算L的谱分解，取其前K个最小特征根的对应的特征向量

$$V = (\mathbf{v}_1,...,\mathbf{v}_K)$$

包括0特征根(可能重复)及其它最小非0特征根的特征向量.

- 记V各行为$\mathbf{y}_1,...,\mathbf{y}_n \in R^K$,即$V = (\mathbf{y}_1,...,\mathbf{y}_n)^{\mathsf{T}}$

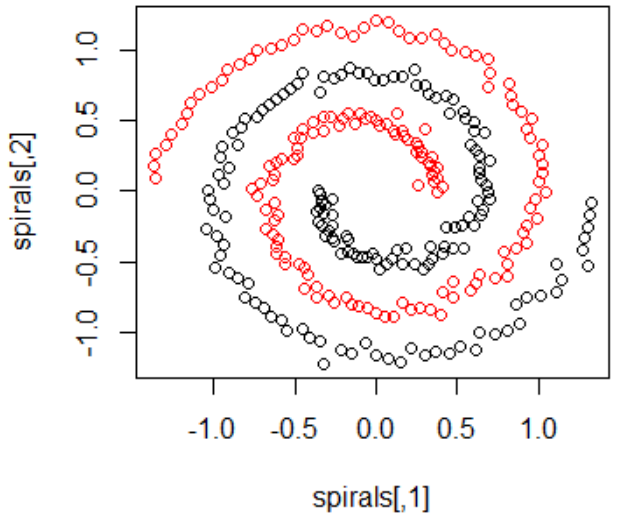
对$\mathbf{y}_1,...,\mathbf{y}_n \in R^K$进行$K$-means聚类，将$n$个节点聚成$K$类。

如果问题给定的是距离矩阵(d_{ij})，则可以通过某种变换转化为相似度矩阵，通常用Gaussian kernel：$a_{ij} = \exp(-d_{ij}^2 / 2\sigma^2), i \neq j.$

谱聚类R函数：
specc （package:kernlab)；cluster_fast_greedy （package:igraph）；fastClustering（package:sClust）…

例1. kernlab中的数据集spirals是二维平面上的两类螺旋状数据（300×2），我们使用谱聚类方法进行聚类，点之间的距离用欧氏距离，点之间的相似度（权重）用高斯核计算。基于拉普拉斯矩阵的两个最小特征根对应的特征向量进行k-means聚类，聚类结果如下图两种颜色。

```
library(kernlab)
data(spirals)
sc <- specc(spirals, centers=2,kernel="rbfdot") # 高斯核
plot(spirals, col=sc)
```

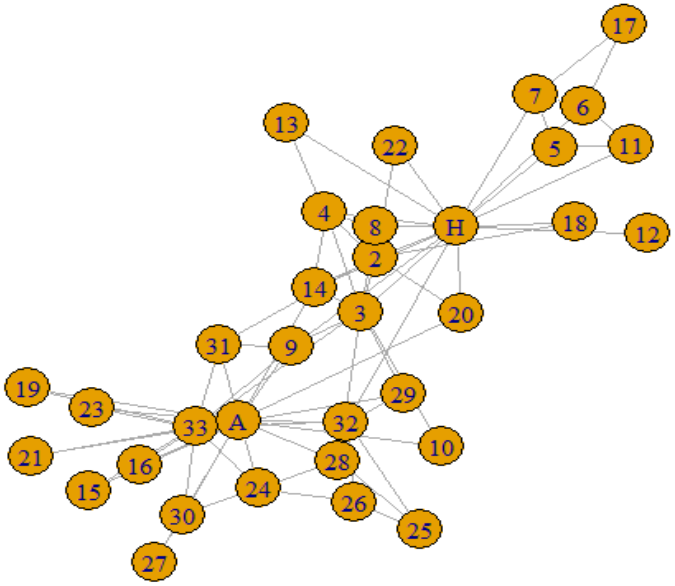


因为有300个样本的原始数据，我们不需要图或加权图进行可视化。图的概念在本例没有任何作用，

例2 （karate club network） 大学跆拳道俱乐部会员交往数据，邻接矩阵元素表达交往密集程度（下图），其中A, H分别是经理和教练，其它32个节点为会员。在收集本数据期间，该俱乐部分裂为两家俱乐部，起因是A,H的矛盾。

邻接矩阵

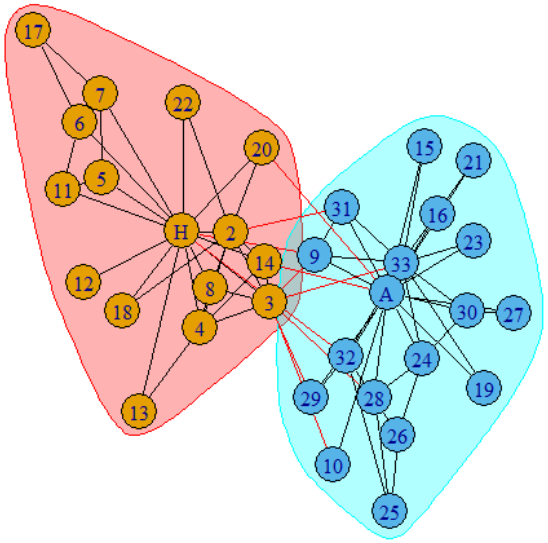
Mr Hi	.	1	1	1	1	1	1	1	1	.	1	1	1	1	.	.	.	1	.	1	.	1	.	.	.	.	.	.	.	.	1	.	.
Actor 2	1	.	1	1	.	.	.	1	.	.	.	.	.	1	.	.	.	1	.	1	.	1	.	.	.	.	.	.	.	1	.	.	.	
Actor 3	1	1	.	1	.	.	.	1	1	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	1	.
Actor 4	1	1	1	.	.	.	.	1	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Actor 5	1	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Actor 6	1	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Actor 7	1	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Actor 8	1	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	
Actor 9	1	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	1		
Actor 10	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1		
Actor 11	1	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
Actor 12	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
Actor 13	1	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
Actor 14	1	1	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1		
Actor 15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1		
Actor 16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1		
Actor 17	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
Actor 18	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
Actor 19	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1		
Actor 20	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1		
Actor 21	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1		
Actor 22	1	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.		
Actor 23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1		
Actor 24	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	1	.	.	1	1	
Actor 25	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	.	.	.	1	.	.	
Actor 26	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	1	.	.	
Actor 27	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1	
Actor 28	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	1	
Actor 29	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1	
Actor 30	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	.	.	1	1	
Actor 31	.	1	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	
Actor 32	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	1	.	.	.	1	1		
Actor 33	.	.	1	.	.	.	.	.	1	.	.	.	.	.	1	1	.	.	1	.	1	.	1	1	.	.	.	.	.	1	1	1	.	1
John A	.	.	.	.	.	.	.	.	1	1	.	.	.	1	1	1	.	.	1	1	1	.	1	1	.	.	1	1	1	1	1	1	1	.



```
#igraph 中的数据集 karate具有igraph特殊格式，需要用as_adj函数转化为（提取）一般的邻接矩阵
library(igraph)
library(igraphdata)
data(karate)
A=as_adj(karate) #提取karate中的邻接矩阵
A=as.matrix(A)
faction = vertex_attr(karate)$Faction #提取faction：真实的派别标号
library(kernlab)
A=as.kernelMatrix(A)
clust =specc(A,centers=2) # 聚类
table(faction,clust) # 比较聚类结果与真实类别
```

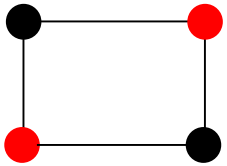
```
        clust
faction  1  2
      1 18  0
      2  1 15
```

只有一个人（Actor 3）分错了类。



附录：谱分析

偶数长度的环（二染色），聚类会出现困难。



略过P21-P31

矩阵的特征根称为谱（spectrum），基于特征根、特征向量的分析方法称为谱方法。前述谱聚类方法基于拉普拉斯矩阵$L = D - A$ 的谱。这里我们进一步探讨为什么拉普拉斯矩阵的小特征根的特征向量包含了数据聚簇信息。为了简单，只考虑0-1图。

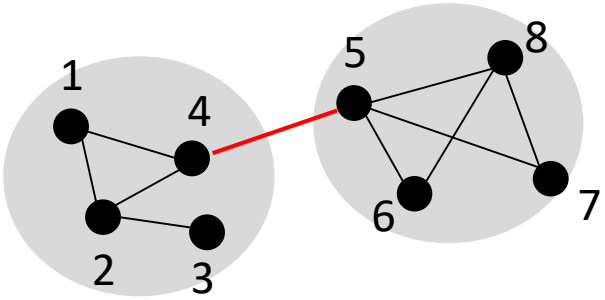
考虑 Laplacian $L = D - A$的最小特征根0的特征向量$\mathbf{x}$

$$0 = L\mathbf{x} = (D - A)\mathbf{x} \Leftrightarrow D\mathbf{x} = A\mathbf{x} \Leftrightarrow D^{-1}A\,\mathbf{x} = \mathbf{x}$$

所以 $\mathbf{x}$是$B = D^{-1}A$的特征根1的特征向量（我们将看到1是B的最大特征根）。这说明为了考察L的最小特征根和特征向量，我们可以考虑$B = D^{-1}A$的最大特征根及其特征向量。

$B = D^{-1}A$ 称为归一化/标准化的邻接矩阵，每行总和为1，是概率分布。

例A1 将下图聚为两类



B的特征根： 1.00 0.86 0.23 -0.73 -0.67 -0.50 -0.19 0.00

L的特征根： 5.29 4.00 4.00 3.26 2.00 1.15 0.30 0.00

B的最大的两个特征向量

```
[1,] -0.35 -0.38
[2,] -0.35 -0.43
[3,] -0.35 -0.50
[4,] -0.35 -0.23
[5,] -0.35  0.22
[6,] -0.35  0.33
[7,] -0.35  0.33
[8,] -0.35  0.34
```

L的最小的两个特征向量

```
[1,]  0.31 -0.35
[2,]  0.37 -0.35
[3,]  0.53 -0.35
[4,]  0.16 -0.35
[5,] -0.26 -0.35
[6,] -0.37 -0.35
[7,] -0.37 -0.35
[8,] -0.37 -0.35
```

这两个向量都把1-4，5-7聚集成两类

B作为线性变换

假设$\mathbf{x}=(x_1,...,x_n)^{\mathsf{T}}\in R^n, x_i$代表节点的某种属性，比如类别、次序、信息、能量等等。线性变换$\mathbf{y}=B\mathbf{x}$将属性$\mathbf{x}$在互相连结的节点之间重新分配（流动）：

$$\mathbf{y}=B\mathbf{x}=D^{-1}A\mathbf{x}=(\bar{x}_1,\bar{x}_2,..\bar{x}_n)^{\mathsf{T}},\quad \bar{x}_i=\frac{1}{d_i}\sum_{j:\,j\sim i}x_j$$

即 $y_i=\bar{x}_i=$与i邻接的节点属性的平均

幂次迭代

一般情况下，反复B-变换（幂次迭代）

$$\mathbf{x}^{(k)}=B\mathbf{x}^{(k-1)},\quad k=1,2,...,\text{初值}\mathbf{x}^{(0)}$$

会使连结紧密的节点的属性变得趋同，不连通的节点互不影响。

因此B的最大特征根的特征向量通常（但有例外）会有如下特点：

相同或接近相同的分量对应的节点属于同一类

例如，假设B的最大特征根的特征向量具有形式(a,a,a,b,b,b,b)

那么很有可能节点1,2,3为一类，节点4,5,6,7为另外一类。

幂次迭代不收敛情形

幂次迭代也会出现不收敛的情况，例如只有两个邻接节点时，

$$B = A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, B\mathbf{x} = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}, B^2\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, B^3\mathbf{x} = B\mathbf{x} = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}, \ldots$$

即幂次变换在$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} x_2 \\ x_1 \end{pmatrix}$之间震荡，不收敛。右下图也不收敛，

一般地，如果图具有某种对称性（若B的特征根关于0对称，等价地如果可以对节点染两种颜色，使得相邻点颜色不同），那么幂次迭代不收敛。

幂次迭代的收敛性

对于一般方阵C(未必对称),幂次迭代 在一定条件下收敛到绝对值最大的特征根对应的特征向量。

命题4. 假设$n\times n$矩阵C是可相似对角化的 $C=P\Lambda P^{-1}$,其中$\Lambda=diag(\lambda_1,...,\lambda_n)$为特征根对角阵，$P=(\mathbf{v}_1,...,\mathbf{v}_n)$的列为特征向量。对任意$\mathbf{x}\in R^n$,记$\mathbf{c}=P^{-1}\mathbf{x}=(c_1,...,c_n)^{\mathsf{T}}$。

(1) 假设 $|\lambda_1|>|\lambda_2|\geq\cdots\geq|\lambda_n|\geq 0$, $C^k\mathbf{x}/\lambda_1^k\to\mathbf{v}_1c_1, k\to\infty$,
(即极限是特征向量$\mathbf{v}_1c_1$，除非$c_1=0$，即$\mathbf{x}$恰好与$\mathbf{v}_1$无关).

（2）假设 $|\lambda_1|=...=|\lambda_m|>|\lambda_{m+1}|\geq\cdots\geq|\lambda_n|\geq 0$,且$\lambda_1,...,\lambda_m$同号（重根），则$C^k\mathbf{x}/\lambda_1^k\to\mathbf{v}_1c_1+...+\mathbf{v}_mc_m, k\to\infty.$

（3）若(2)中的$\lambda_1,...,\lambda_m$不全同号，则不收敛。

注: (3)说明，如果矩阵C的最大特征根和最小特征根符号相反，绝对值相同，则幂次迭代不收敛。

证明:(1) 因为$\Lambda^k/\lambda_1^k = diag(1,\lambda_2^k/\lambda_1^k,\lambda_3^k/\lambda_1^k....) \to diag(1,0,...,0), k \to \infty,$

所以$C^k\mathbf{x}/\lambda_1^k = P\Lambda^k\mathbf{c}/\lambda_1^k \to (\mathbf{v}_1,...,\mathbf{v}_n)\begin{pmatrix}1 & & \\ & 0 & \\ & & 0\end{pmatrix}\begin{pmatrix}c_1\\ \vdots \\ c_n\end{pmatrix} \to \mathbf{v}_1c_1。$

（2）$\Lambda^k/\lambda_1^k = diag(1^k,...,1^k,\lambda_{m+1}^k/\lambda_{m+1}^k....) \to diag(1,...,1,0,...,0),$

$C^k\mathbf{x}/\lambda_1^k = P\Lambda^k\mathbf{c}/\lambda_1^k \to (\mathbf{v}_1,...,\mathbf{v}_n)\begin{pmatrix}1 & & \\ & 1 & \\ & & 0\end{pmatrix}\begin{pmatrix}c_1\\ \vdots \\ c_n\end{pmatrix} \to \mathbf{v}_1c_1+...+\mathbf{v}_mc_m$

（3）若$\lambda_1,...,\lambda_m$绝对值相同，但不全同号，比如λ_1与其它符号相反，则Λ^k/λ_1^k的第一个对角元$(-1)^k$不收敛。

当$k \to \infty, C^k\mathbf{x}/\lambda_1^k \Rightarrow \pm\mathbf{v}_1c_1+...+\mathbf{v}_mc_m$

标准化邻接矩阵的性质

命题5. (1)$B=D^{-1}A$的所有特征根λ为实数，$|\lambda|\le 1$.

(2)B最大特征根为$\lambda_1=1$，$\mathbf{1}_n$是对应的特征向量.

(3)$\lambda_1=1$的重数为1当且仅当图全连通的.

(4) $\lambda_1=1$的重数为k当且仅当图由k个互不连通的连通子图$G_1,...,G_k$构成，则$\lambda_1=1$的特征空间的基为$\mathbf{v}_1,...,\mathbf{v}_k, v_{ij}=1_{(i\in G_j)}, j=1,...,k; i=1,...,n.$

(5)（幂次迭代不收敛情形）假设图是全连通的，则下述陈述等价：

(*a*) B有特征根-1（幂次迭代不收敛）

(*b*) 节点可由两种颜色染色，使得邻接节点颜色不同(这称为二分图)，

(*c*) 没有奇数长度的环，

(*d*) B的所有特征根(谱)关于0对称。

证明：(1)$D^{-1}A$与$D^{-1/2}AD^{-1/2}$有相同的特征根，后者是对称矩阵，

所以$B=D^{-1}A$的特征根都是实数。假设λ为$B=D^{-1}A$的特征根，

对应的特征向量为$\mathbf{x}$，$B\mathbf{x}=\mathbf{x}\lambda$，设$|x_k|=\max_{1\le i\le n}|x_i|>0$，由$\lambda x_k=\bar{x}_k=\sum_{j:j\sim k}x_j/d_k$

$|\lambda x_k|=|\sum_{j:j\sim k}x_j/d_k|\le\sum_{j:j\sim k}|x_j|/d_k\le\sum_{j:j\sim k}|x_k|/d_k=|x_k|$，所以$|\lambda|\le 1$.

(2)因为$B\mathbf{1}_n = D^{-1}A\mathbf{1}_n = \mathbf{1}_n$，所以$\lambda_1 = 1$是特征根，$\mathbf{1}_n$是对应的特征向量。

(3) 假设图是全连通的，我们只需证明除了一个常数倍数之外，$\mathbf{1}_n$是$\lambda_1 = 1$的唯一特征向量。设$\mathbf{x}$是B的对应于$\lambda_1 = 1$的特征向量，

$$B\mathbf{x} = \mathbf{x} \Leftrightarrow D^{-1}A\mathbf{x} = \mathbf{x} \Leftrightarrow A\mathbf{x} = D\mathbf{x} \Leftrightarrow (D-A)\mathbf{x} = 0 \Rightarrow 0 = \mathbf{x}^{\mathsf{T}}(D-A)\mathbf{x}$$

$$= \mathbf{x}^{\mathsf{T}}L\mathbf{x} \overset{\text{引理}}{=} \frac{1}{2}\sum_{i\sim j}(x_i - x_j)^2 \Rightarrow x_1 = \ldots = x_n \Rightarrow \mathbf{x} \propto \mathbf{1}_n.$$

(4) 假设图由k个互斥的连通子图$G_1,\ldots,G_k$构成，重排邻接矩阵

$$A = \begin{pmatrix} A_1 & & \\ & \ddots & \\ & & A_k \end{pmatrix}, \text{ 记} \mathbf{v}_1 = \begin{pmatrix} \mathbf{1}_{G_1} \\ \vdots \\ \mathbf{0} \end{pmatrix}, \ldots, \mathbf{v}_k = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{1}_{G_k} \end{pmatrix}.$$

设$\mathbf{x}$是B的对应于$\lambda_1 = 1$的特征向量，$B\mathbf{x} = \mathbf{x} \Leftrightarrow (D-A)\mathbf{x} = \sum_{i\sim j}(x_i - x_j)^2/2 = 0$

$\Rightarrow$ 对每个子图 $\sum_{i\sim j, i, j\in Gs}(x_i - x_j)^2 = 0$,即每个子图上的节点对应的$x's$全部相等。

$\Rightarrow$ $\mathbf{x}$是$\mathbf{v}_1,\ldots,\mathbf{v}_k$的线性组合.

由（3），除了一个常数倍数之外，$\mathbf{1}_{G_s}$ 是A_s对应于特征根1的唯一特征向量，从而$\mathbf{v}_1,...,\mathbf{v}_k$ 都是 A的对应于特征根1的特征向量,故$\lambda_1 = 1$的 重数为k.

(5)$(a) \Leftrightarrow (b)$

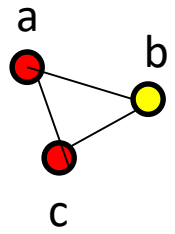
设$\mathbf{x} \neq \mathbf{0}$是$B$的对应于$\lambda = -1$的特征向量，$B\mathbf{x} = -\mathbf{x} \Leftrightarrow D^{-1}A\mathbf{x} = -\mathbf{x}$

$$\Leftrightarrow (D+A)\mathbf{x} = 0 \Rightarrow 0 = \mathbf{x}^{\mathsf{T}}(D+A)\mathbf{x} = \frac{1}{2}\sum_{i\sim j}(x_i + x_j)^2$$

$\Rightarrow$ 若$i \sim j$，则$x_i = -x_j$,所以相邻的节点或者符号相反,或者都为0。

因为图是全连通的，任何两个节点之间都有一条路径连结它们，所以所有x_i绝对值相同且非0，相邻的节点符号相反。

$(c) \Leftrightarrow (b)$如果存在奇数环(下图),那么不可能用两种颜色染色使得相邻节点的颜色不同。



$(a) \Leftrightarrow (d)$

当-1是B的特征根时，我们可以将节点重新排列，使得前面m个都是同一种颜色（互不连结），后面$n-m$个是另外一种颜色（互不连结），但两种颜色间有连结。邻接矩阵如下

$$A=\begin{pmatrix} 0 & C \\ C^{\mathsf{T}} & 0 \end{pmatrix} \Rightarrow B=D^{-1}A=\begin{pmatrix} 0 & B_1 \\ B_2 & 0 \end{pmatrix}$$

假设λ是B的一个特征根，对应的特征向量为$\mathbf{x}=\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$，记$\mathbf{y}=\begin{pmatrix} \mathbf{x}_1 \\ -\mathbf{x}_2 \end{pmatrix}$

由$B\mathbf{x}=\mathbf{x}\lambda$，即$\begin{pmatrix} 0 & B_1 \\ B_2 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}=\begin{pmatrix} B_1\mathbf{x}_2 \\ B_2\mathbf{x}_1 \end{pmatrix}=\begin{pmatrix} \lambda\mathbf{x}_1 \\ \lambda\mathbf{x}_2 \end{pmatrix}$。令$\mathbf{y}=\begin{pmatrix} \mathbf{x}_1 \\ -\mathbf{x}_2 \end{pmatrix}$，则

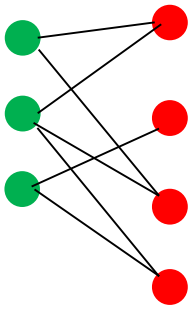
$$B\mathbf{y}=\begin{pmatrix} 0 & B_1 \\ B_2 & 0 \end{pmatrix}\begin{pmatrix} \mathbf{x}_1 \\ -\mathbf{x}_2 \end{pmatrix}=\begin{pmatrix} -B_1\mathbf{x}_2 \\ B_2\mathbf{x}_1 \end{pmatrix}=\begin{pmatrix} -\lambda\mathbf{x}_1 \\ \lambda\mathbf{x}_2 \end{pmatrix}=-\lambda\mathbf{y}$$

所以$-\lambda$也是B的特征根，对应的特征向量为$\mathbf{y}=\begin{pmatrix} \mathbf{x}_1 \\ -\mathbf{x}_2 \end{pmatrix}$。

所以B的所有特征根关于0对称。证毕。

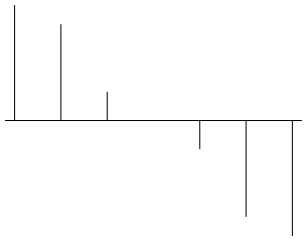
关于二分图

命题5(5)刻画了二分图的性质，总结如下

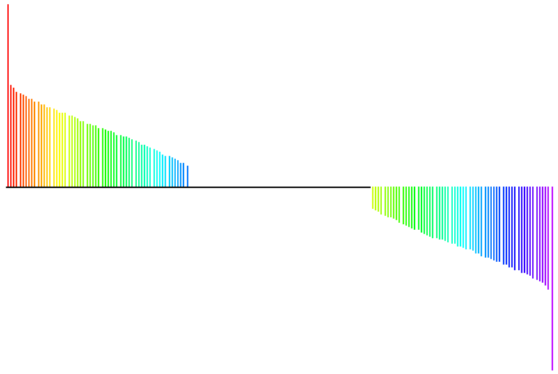


二分图(bipartite)
⇔ 两种着色，相邻的点不同色（左图）
⇔ $\lambda_{\min}(B) = -1$
⇔ B的特征根关于0对称（右图）

谱图(B的特征根)



Spectrum of B
(ordered eigenvalues)



Spectrum：an ordered arrangement by a particular characteristic (such as frequency or energy)

分类：logistic回归和神经网络

吴恩达(Andrew Ng)认为机器学习有6个核心算法:

- Linear Regression: Straight & Narrow
- Logistic Regression: Follow the Curve
- Gradient Descent: It's All Downhill
- Neural Networks: Find the Function
- Decision Trees: From Root to Leaves
- K-Means Clustering: Group Think

https://www.deeplearning.ai/the-batch/issue-146/

1. 分类和预测介绍

分类和预测对未知的随机变量进行"估计"，当待预测随机变量是类别的时候，称为分类（classification）或判别（discriminant）问题。当待预测随机变量是连续变量的时候，称为回归预测（prediction）问题。

数据：$(y_i, \mathbf{x}_i), i = 1, \ldots, n,$
求解$f(\mathbf{x}, \theta) = E(y|\mathbf{x})$

y相应或类别，

$\mathbf{x}$自变量或特征

训练

当y是类别时，$p = f(\mathbf{x}, \theta)$是概率

$\hat{\theta}$，拟合曲线$y = f(\mathbf{x}, \hat{\theta})$

回归: $\hat{\theta}$=argmin$\sum(y_i - f(\mathbf{x}_i, \theta))^2$

分类: $\hat{\theta}$=argmin$(-\sum[y_i \log p_i + (1 - y_i)\log(1 - p_i)])$

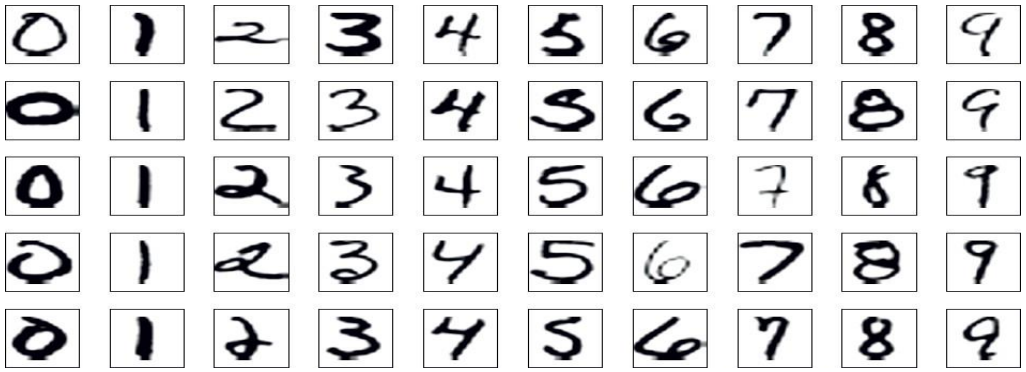
$p_i = f(\mathbf{x}_i, \theta)$

判别
预测

$\hat{y}_0 = f(\mathbf{x}_0, \hat{\theta})$

测试数据集（不参与训练）评价效果

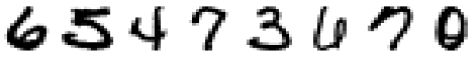
例1 (手写体识别). 50 个数字0-9的手写体样本如下。每个手写数字是16x16像素图像，每个像素点的黑白强度在0-255之间，将像素强度矩阵拉直成长度为196的向量$\mathbf{x}$。



数据： $\mathbf{x}_1, \ldots, \mathbf{x}_{50}$：50个196x1的手写数字的像素向量，
$y_1, \ldots, y_{50}$：每个手写体代表的数字/类别，类别取值0-9。

训练： 求解判别/预测准则。

预测： 判别新的手写体（下图）分别是什么数字（即判别/预测它们所属的类别）。



两个学派

统计学分为两大流派：频率学派（Fisher学派，古典）和贝叶斯学派，随着人工智能的发展，贝叶斯学派越来越被重视。

对于判别分析，Fisher的方法称为Fisher线性判别分析(LDA: linear discriminant analysis)，贝叶斯方法可得到类似的线性判别以及二次判别或其它非线性判别.

Ronald Fisher （1890-1962）英国统计学家.

$$SS_T = SS_W + SS_B$$

Thomas Bayes （1701-1761），英国统计学家、哲学家。发现了贝叶斯公式（先验-后验）：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

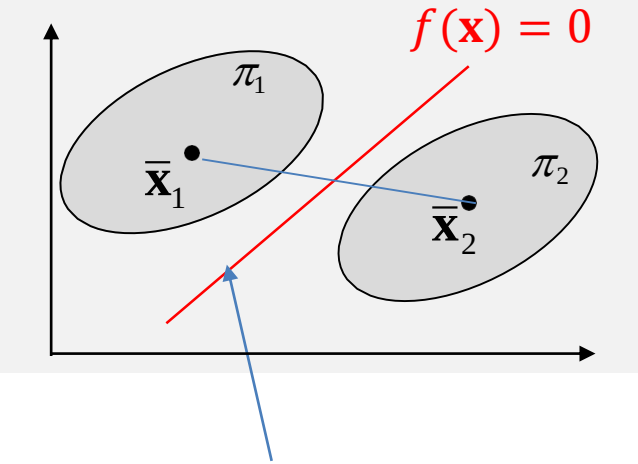
背景：Fisher线性判别(LDA)

两类Fisher线性判别

训练数据：　第一类π_1：$\mathbf{x}_{11},\mathbf{x}_{12},\ldots,\mathbf{x}_{1n_1}\in R^p$，　样本均值和方差$\bar{\mathbf{x}}_1, S_1$;

　　第一类π_2：$\mathbf{x}_{21},\mathbf{x}_{22},\ldots,\mathbf{x}_{2n_2}\in R^p$，　样本均值和方差$\bar{\mathbf{x}}_2, S_2$.

假设两类训练数据等方差，估计为：$S=\dfrac{1}{n_1+n_2-2}\left((n_1-1)S_1+(n_2-1)S_2\right)$



Fisher 判别：
若$f(\mathbf{x})>0$，判定$\mathbf{x}\in\pi_1$;
否则判定$\mathbf{x}\in\pi_2$

$$f(\mathbf{x})=(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\top}S^{-1}\left(\mathbf{x}-\frac{\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2}{2}\right)=0$$

$\Leftrightarrow$ 距离两类中心的等(马氏)距线

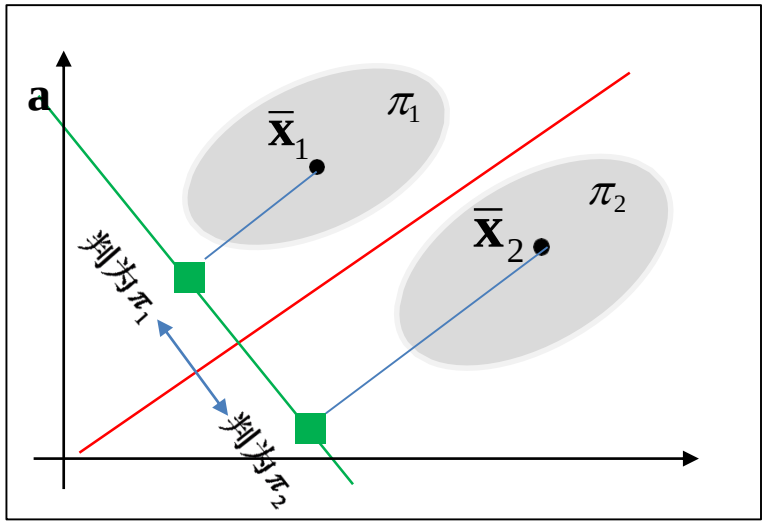
$$(\mathbf{x}-\bar{\mathbf{x}}_1)^{\top}S^{-1}(\mathbf{x}-\bar{\mathbf{x}}_1)=(\mathbf{x}-\bar{\mathbf{x}}_2)^{\top}S^{-1}(\mathbf{x}-\bar{\mathbf{x}}_2)$$

*Fisher*观点：

寻找方向**a**，使得两类数据区分度最大

$$\max \frac{(\mathbf{a}^{\mathsf{T}}\bar{\mathbf{x}}_1 - \mathbf{a}^{\mathsf{T}}\bar{\mathbf{x}}_2)^2}{\mathbf{a}^{\mathsf{T}} S \mathbf{a}} \Rightarrow \mathbf{a}_{opt} = S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

判别函数$f(\mathbf{x}) = \mathbf{a}_{opt}{}^{\mathsf{T}}\mathbf{x} + b$ (经过中心$(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)/2$).



K类Fisher线性判别

K个类，中心为$\bar{\mathbf{x}}_j \in R^p, j = 1,\ldots,K$,假设各类数据的方差相同，所有数据的样本方差为$S$,对任何$\mathbf{x} \in R^p$,定义马氏度量

$$M_j(\mathbf{x}) = (\mathbf{x} - \bar{\mathbf{x}}_j)^{\mathsf{T}} S^{-1} (\mathbf{x} - \bar{\mathbf{x}}_j).$$

*Fisher*线性判别准则：若$M_k(\mathbf{x}) < M_j(\mathbf{x}), j \neq k$, 则判定$\mathbf{x} \in \pi_k$.

```
> library(MASS)
> mylda = lda(train.data, class)   #训练
> predict(mylda, test.data)        #预测
```

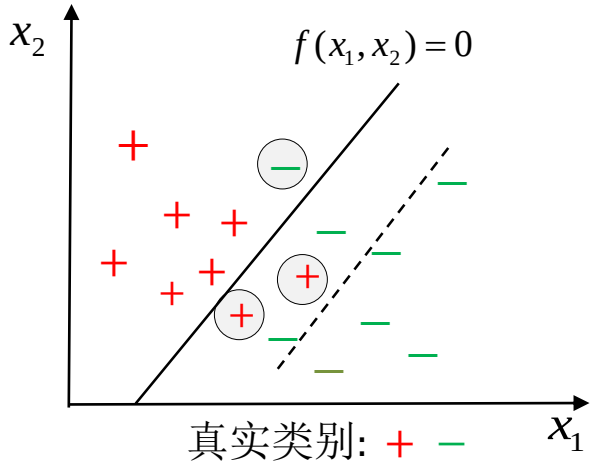
评价分类效果

评价判别效果

Fisher 线性判别(LDA)函数

$$f(x_1, x_2) = ax_2 + bx_1 + c$$

若$f(x_1, x_2) > 0$，判别为阳性 +；
否则，判为阴性 − .



Fisher判别(实线):
Recall(+) = 7/8，Precision (+) = 6/7
虚线判别:
Recall(+) = 8/8， Precision (+) = 8/11

Fisher判别将2个+ 错判为−；1个−错判为+。如果正确判定+是重要的（比如疾病），那么应该下调判别阈值，比如：$f(x_1, x_2) > -1$时，判为+，这可以增加正确判别+的概率：

$$\text{Recall}(+) = P\ (\text{判为} + | \text{真实为} +),$$

极端地，若判别阈值为负无穷大，则所有点判为+，recall=1，但此时所有判别为+的对象中有大约一半是将−误判为+，精度只有0.5

$$\text{Precision}(+) = \frac{\text{正确判别为+的个数}}{\text{所有判别为+的个数}}$$

若关注重点是 - ，同样定义Recall(-),Precision(-). 无论如何，评估分类效果需综合考虑准确度和精确度，比如两者的几何平均F1。

判别效果度量

测试数据与训练数据格式完全相同（含真实的类别标号），但不参与训练。用训练得到的方法预测测试数据的类别标号，与真实类别标号比较考察效果。

两个类：＋positive阳性; – negative阴性，
四种结果：TP: true positive； FP: false positive； FN: false negative TN: true negative

		真正的类别 +	真正的类别 –
判定的类别	+	TP	FP
	–	FN	TN

常用的分类正确率如下（前4个2 × 2的边际概率）：

准则	定义	解释
召回率recall(+), 灵敏度Sensitivity	TP/(TP+FN)	真阳性被判对的比例
精确度Precision (+)	TP/(TP+FP)	判为阳性的判别中正确的比例
准确度 accuracy	(TP+TN)/(TP+FP+TN+FN)	所有判别中正确判别的比例
F1 score=2/(1/recall+1/precision)	2TP/(2TP+FP+FN)	灵敏度和精确度的几何平均

召回率(–)，特异度Specificity	**TN/(FP+TN)**	**真阴性被判对的比例**
精确度 Precision(–)	TN/(FN+TN)	判为阴性的判别中判对的比例

2. Logistic回归/贝叶斯分类

- Logistic 回归研究二值（二类）响应变量与自变量(feature)的关系，多项式回归对多类问题判别分类。
- 贝叶斯方法基本与logistic/多项回归相同（假设正态）。
- 主流机器学习一般不采用Fisher's LDA，而是采用logistic。

Logistic回归模型

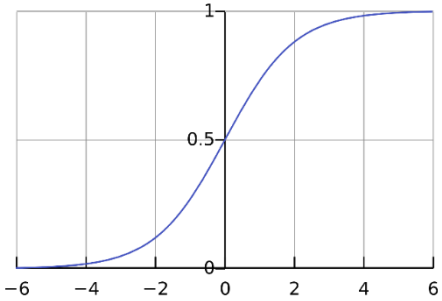
数据$(\mathbf{x}, y)$：$y=0$或1，特征$\mathbf{x}$。假设概率(回归函数)$p=\mathrm{P}(y=1|\mathbf{x})=E(y|\mathbf{x})$具有形式

$$p(\boldsymbol{\theta},\mathbf{x})=\frac{\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})}{1+\exp(\alpha+\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})},$$

其中$\boldsymbol{\theta}=(\alpha,\boldsymbol{\beta})$为参数,$\boldsymbol{\beta}=\log(p/(1-p))=\log(\text{odds ratio})$.

$$sigmoid\text{函数}:\sigma(x)=\frac{e^x}{1+e^x}$$

连续化的示性函数



训练/拟合

数据$(y_i, \mathbf{x}_i), i = 1,\ldots,n$, 其中$y_i = 0$ or $1, \mathbf{x}_i \in R^p$为特征(*feature*).

训练 : 极小化交叉熵损失(≈ 极大似然), 估计未知参数$\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$。

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{\theta}, \mathbf{x}_i)^{y_i} (1 - p(\boldsymbol{\theta}, \mathbf{x}_i))^{1-y_i} = \prod_{i=1}^{n} \left(\frac{\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i)} \right)^{y_i} \left(1 - \frac{\exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i)}{1 + \exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i)} \right)^{1-y_i}$$

交叉熵损失

$$loss = -\log L(\hat{\boldsymbol{\theta}}) = -\sum_{i=1}^{n} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)], \quad \hat{p}_i = p(\hat{\boldsymbol{\theta}}, \mathbf{x}_i)$$

logistic≈ 线性判别

对于类别y_0未知的新的特征$\mathbf{x}_0$, 若$p(\boldsymbol{\theta}, \mathbf{x}_0) > C$（阈值), 则预测$\hat{y}_0 = 1$.

$$p(\boldsymbol{\theta}, \mathbf{x}_0) > C \Leftrightarrow \alpha + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x} > \log(C/(1 - C))$$

logistic预测实际上是线性判别， 与*Fisher*线性判别(*LDA* : Fisher's linear discriminant analysis)形式上基本相同， 但容许可调阈值，以满足具体的预测要求（Fisher LDA固定阈值）。

多项回归模型

$K \geq 2$个类别时，类别标号$y = 1,2,...,K$。多项回归(multinomial regression)假设

$$P(y=k \mid \mathbf{x}) = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{x})}{\sum_{j=1}^{K}\exp(\alpha_j + \boldsymbol{\beta}_j^{\mathsf{T}}\mathbf{x})}, \quad k = 1,...,K$$

其中$\alpha_k, \boldsymbol{\beta}_k, k = 1,...,K$是未知参数，$\alpha_1 = 0$，$\boldsymbol{\beta}_1 = 0.$

- 多项回归基本上等价于多类线性判别方法（但求解方法不同）。
- 多项回归可认为是来自于多组正态问题。具体如下：

性质：假设各类数据服从方差相等的正态：$\mathbf{x} \mid y = k \sim N_p(\boldsymbol{\mu}_k, \Sigma)$，则由贝叶斯公式，$P(y=k \mid \mathbf{x})$一定具有多项式回归的形式。

证明：由*Bayes*公式，假设各类先验概率为$p_k = P(y=k)$，则后验概率

$$P(y=k \mid \mathbf{x}) = \frac{p_k f_{N_p(\boldsymbol{\mu}_k,\Sigma)}(\mathbf{x})}{\sum_j p_j f_{N_p(\boldsymbol{\mu}_j,\Sigma)}(\mathbf{x})} = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k^{\mathsf{T}}\mathbf{x})}{\sum_{j=1}^{K}\exp(\alpha_j + \boldsymbol{\beta}_j^{\mathsf{T}}\mathbf{x})}$$

其中$\boldsymbol{\beta}_k = \Sigma^{-1}\boldsymbol{\mu}_k, \alpha_k = \log(p_k) - \frac{1}{2}\boldsymbol{\mu}_k^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_k.$

二次判别

若各类数据的方差不相等，假设正态$N(\boldsymbol{\mu}_k, \Sigma_k)$，则由贝叶斯公式知后验概率具有形式

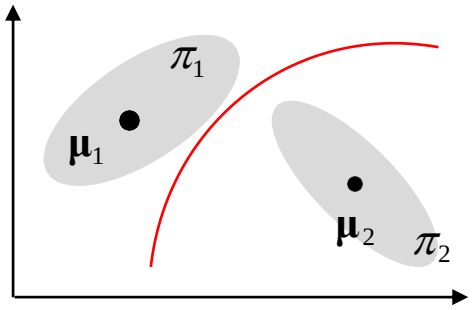
$$P(y=k \mid \mathbf{x}) = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k{}^{\mathsf{T}}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathsf{T}}\Sigma_k^{-1}\mathbf{x})}{\sum_{j=1}^{K}\exp(\alpha_j + \boldsymbol{\beta}_j{}^{\mathsf{T}}\mathbf{x} - \frac{1}{2}\mathbf{x}^{\mathsf{T}}\Sigma_k^{-1}\mathbf{x})},$$

此时判别准则可以取为：

对任何新的特征$\mathbf{x}_0$,对所有k计算上述后验概率，

若$P(y_0=k \mid \mathbf{x}_0) > P(y_0=i \mid \mathbf{x}_0), \text{all } i \neq k,$则判定$\hat{y}_0 = k.$

注意$P(y=k \mid \mathbf{x}) > P(y=i \mid \mathbf{x}) \Leftrightarrow \mathbf{x}^{\mathsf{T}}Q\mathbf{x} + \mathbf{b}^{\mathsf{T}}\mathbf{x} + a > 0,$称为二次判别.



附1： 支持向量机（SVM: support vector machine)

SVM模型试图将空间中的不同类的点尽可能地用比较宽的间隔(margin，下图方框)分开：间隔边缘的点称为支持向量，间隔的宽度即两类支持向量之间的距离，SVM最大化间隔的宽度。

对任一判别分类器H, 各类中离H最近的点$\mathbf{x}$称为支持向量。支持向量$\mathbf{x}$到判别分类线H之间的垂直距离

$$d(\mathbf{x}, H)$$

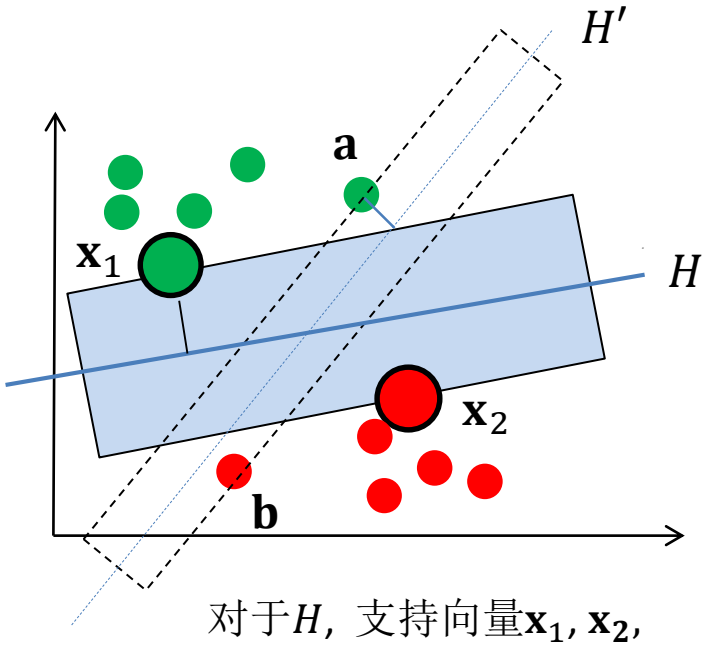
越大越好。

$$SVM: \ H_{svm} = \max_{H} \min_{\mathbf{x}} d(\mathbf{x}, H)$$

SVM将原始数据映射到更高维度，比如

$$\mathbf{x} = (x_1, x_2) \rightarrow (x_1, x_2, x_1 x_2),$$

再应SVM，得到非线性分类器。



对于H, 支持向量$\mathbf{x}_1, \mathbf{x}_2$,

对于H', 支持向量$\mathbf{a}, \mathbf{b}$

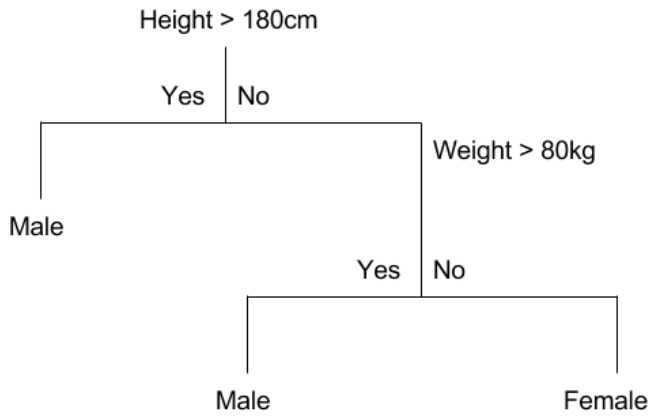
$$d(\mathbf{x}_1, H) > d(\boldsymbol{a}, H')$$

```
> library( kernlab )
> sv <-ksvm(Country ~Freshwater+Marine, data = salmon, type = “C-svc”)
> ploy(sv)
```

附2：决策分类树、随机森林

前述判别分类方法不容易解释，有时也不实用。决策分类树(decision tree)二分随机向量的各个分量，方法直观，决策准则如下：

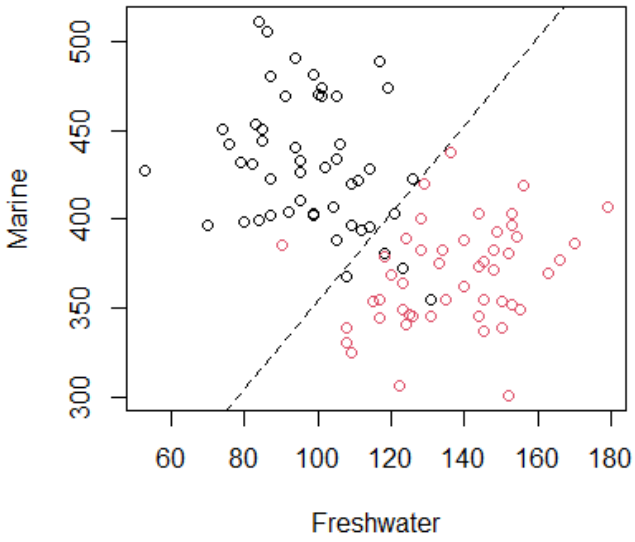
if condition1 *and* condition2 *and* condition3 *then* outcome



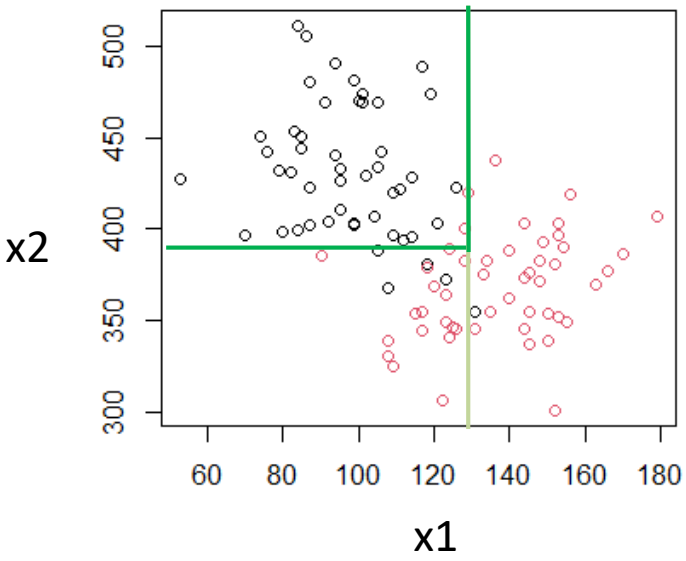
分类树

二分树概述：树的结构从根部节点（判别力最强的变量x1）开始，以大于或小于某个阈值C_k递归划分每个变量，节点左侧分枝准则成立(yes)， 右侧不成立(no)，每一个节点处的数字 n1/n2 表示n1 个第1类，n2个第2类。在终止节点处投票，若n1>n2，则判为第1类。
极小化错分率求解阈值C_k。

线性判别



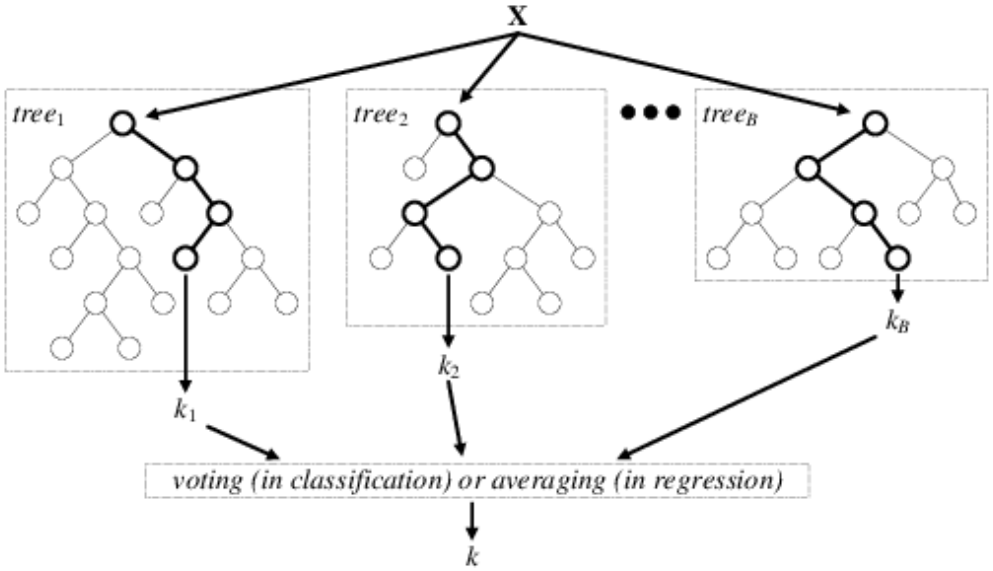
分类树



If (x1 >130) and (x2<390) red

随机森林

随机森林方法抽取部分变量，形成分类或回归树，并对每个样本分类。如此反复多次（形成森林），投票决定每个样本点的类别。该方法提高了CART的预测效果。



R packge： ： rpart，C50， randomForest

```
> mytree= rpart(class~variables , method="class")
> prune(mytree, cp ) #cp ： complexity parameter
> myfit = randomForest(class~features,data ntree=1000)
> predict(myfit, newdata)
```

3. 深度神经网络 (DNN, Deep Neural Network)

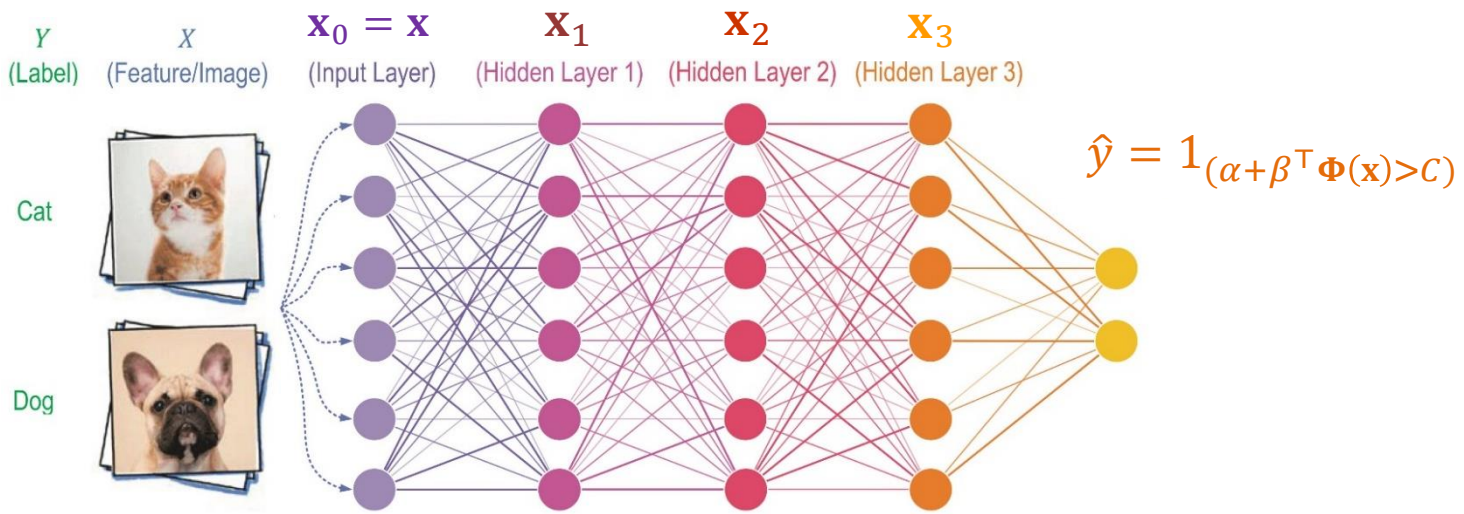
神经网络(DNN，deep neural network) 模仿神经系统特别是大脑的认知方式，将输入$\mathbf{x}$经过h次递归复合变换（非线性）生成具有判别能力的数据

$$\mathbf{x}_h = f^h\left(f^{h-1}(\cdots f^1(\mathbf{x}))\right),\ h\text{:深度}$$

基于$\mathbf{x}_h$应用logistic/线性判别(二分类为例)，DNN模型如下

$$P(y=1|\mathbf{x}) = \frac{1}{1+\exp[-\alpha-\boldsymbol{\beta}^\top \mathbf{x}_h]}$$
$$= \frac{1}{1+\exp\left[-\alpha-\boldsymbol{\beta}^\top f^h\left(f^{h-1}(\cdots f^1(\mathbf{x}))\right)\right]}$$

$\mathbf{x}$	$\mathbf{x} \longrightarrow \mathbf{x}_h = \boldsymbol{\Phi}(\mathbf{x})$	$\boldsymbol{\Phi}(\mathbf{x}) \longrightarrow y$
embedding	非线性变换	logistic



其中隐藏层的递归变换通常取仿射Ax+b及正步(ReLU)激活运算 $[\ \ ]_+$

$$\mathbf{x}_k = f^k(\mathbf{x}_{k-1}) = [A_k\mathbf{x}_{k-1} + \mathbf{b}_k]_+,\ \ k = 1, \dots, h$$

最终得到

$$\mathbf{x}_h = f^h\left(f^{h-1}(\cdots f^1(\mathbf{x}))\right)$$

将$\mathbf{x}_h$视作feature数据，构造$\mathbf{x}_h$的线性函数（ $\mathbf{x}$ 的分段线性函数）

$$\Phi(\mathbf{x}) = \alpha + \boldsymbol{\beta}^\top\mathbf{x}_h$$

DNN模型

$$P(y = 1|\mathbf{x}\ \ ) = \frac{\exp[\Phi(\mathbf{x})]}{1+\exp[\Phi(\mathbf{x})]} = \frac{1}{1+\exp\left[-\alpha-\boldsymbol{\beta}^\top f^h\left(f^{h-1}(\cdots f^1(\mathbf{x}))\right)\right]}$$

DNN大量使用偏置(大的q）以及多次复合，以分段线性函数逼近任何函数$f(\mathbf{x}) = E(y|\mathbf{x})$

$\mathbf{x}$维数p，隐藏层维数 q

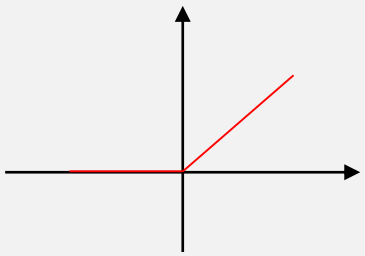
$$\mathbf{x}_k = f^k(\mathbf{x}_{k-1}) = [A_k\mathbf{x}_{k-1} + \mathbf{b}_k]_+$$

A_k：weight，$\mathbf{b}_k$：bias 偏置

ReLU激活函数（右图）

$$ReLU(x) = \begin{cases} x. x \geq 0 \\ 0, x < 0 \end{cases}$$

ReLU



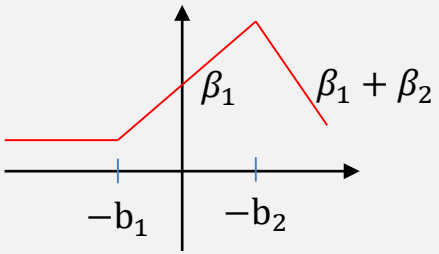
示例（ReLU激活⇔分段线性）：

假设$p = 1, q = 2, h = 1$

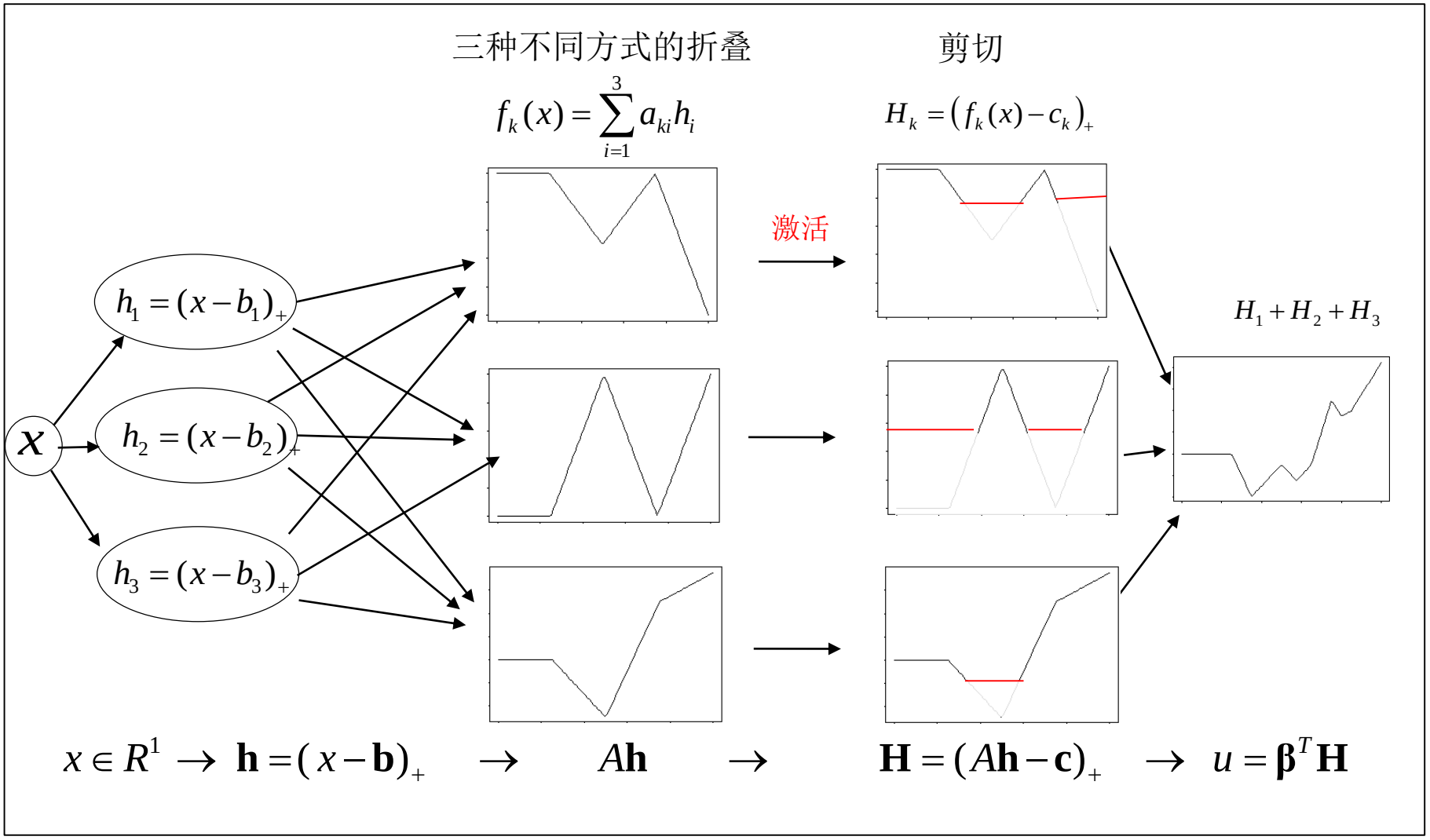
$$x \to \mathbf{x}_1 = [A_1 x + \mathbf{b}_1]_+ = \begin{bmatrix} (a_1 x + \mathrm{b}_1)_+ \\ (a_2 x + \mathrm{b}_2)_+ \end{bmatrix}$$

$$\Phi(\mathbf{x}) = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_1\alpha + \beta_1(x + \mathrm{b}_1)_+ + \beta_2(x + \mathrm{b}_2)_+$$

这是两段线性函数(不妨设a_1=1，a_2=1)，该函数在$-\mathrm{b}_1$，$-\mathrm{b}_2$处分段，斜率分别为β_1、$\beta_1 + \beta_2$。



DNN输入$p=1, q=3$, 层数$h=2$



梯度下降法求解参数

$$p(\mathbf{x},\Theta)=P(y=1\mid x)=\frac{\exp(\alpha+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_h)}{1+\exp(\alpha+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_h)}.$$

极小化交叉熵误差(−log似然)

$$J(\Theta)=-\sum_{i=1}^{n} y_i \log(p(\mathbf{x}_i,\Theta))+(1-y_i)\log(1-p(\mathbf{x}_i,\Theta)),$$

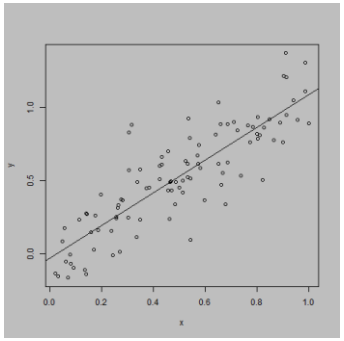
梯度下降法（GD，gradient descent）迭代求解最优Θ：

$$\Theta^{(k)}=\Theta^{(k-1)}-\eta\frac{\partial J}{\partial\Theta}\Big|_{\Theta=\Theta^{(k-1)}},\quad \eta:\text{learning rate}$$

梯度下降法阻止过拟合

- 大模型：多参数使得模型复杂到足以描述任何复杂函数。
- 梯度下降+规则化：梯度下降抑制拟合函数过于复杂。

例1. 100个参数的单层NN模型，拟合得到近似直线，没有出现过拟合。



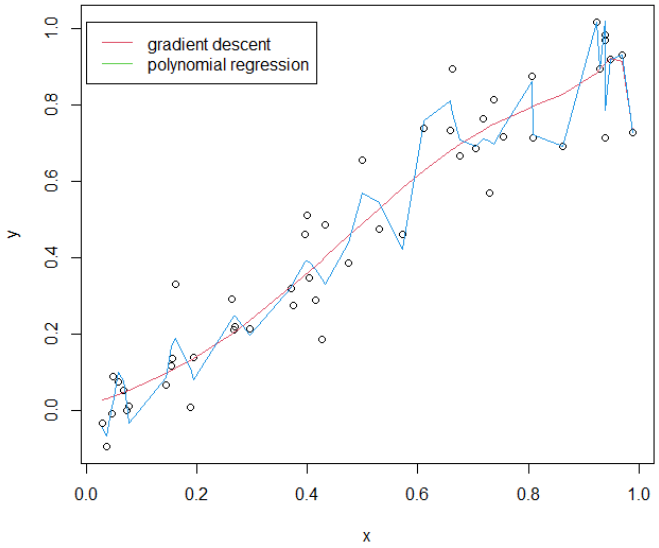
例2. 对多项式回归模型（可看作是单层ANN）

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

$k = 50, n = 100$

分别用LS算法和梯度下降法求解参数β_0，…, β_k

LS算法会出现过拟合（蓝，误差达到全局最小），而梯度下降法得到平滑估计（红，局部最优误差较大）



参考书籍/文献:

- G. Strang (2019) Linear algebra and learning from data.
- Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016) Deep Learning, the MIT press (中文版，2017）
- Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science. 28;313(5786):504-7

R package:
浅层NN： nnet, neuralnet,
深度DNN: keras, tensorflow