第十一章：特征选择与稀疏学习

主讲：连德富 特任教授 | 博士生导师

邮箱：liandefu@ustc.edu.cn

手机：13739227137

主页：http://staff.ustc.edu.cn/~liandefu

特征

- 特征
  - 描述物体的属性

- 特征的分类
  - 相关特征: 对**当前学习任务**有用的属性
  - 无关特征: 与**当前学习任务**无关的属性
  - 冗余特征: 其所包含信息能由其他特征推演出来

例子：西瓜的特征

西瓜的**特征**

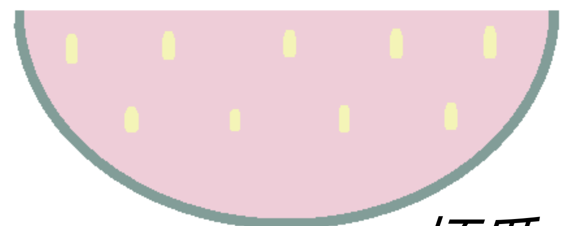
- 颜色
- 纹理
- 触感

→ **无关特征**

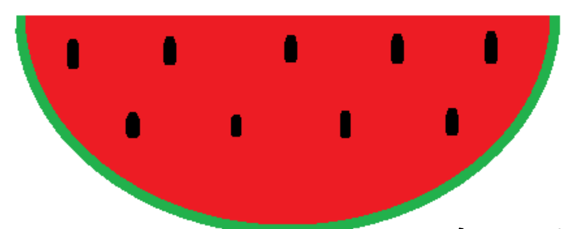
- 根蒂
- 声音

→ **相关特征**

当前任务：西瓜是否是好瓜



坏瓜



好瓜

特征选择

- 特征选择
  - 从给定的特征集合中选出**任务相关**特征子集
  - 必须确保不丢失重要特征
- 原因
  - 减轻维度灾难：在少量属性上构建模型
  - 降低学习难度：留下关键信息

例子：判断是否好瓜时的特征选择

西瓜的**特征**

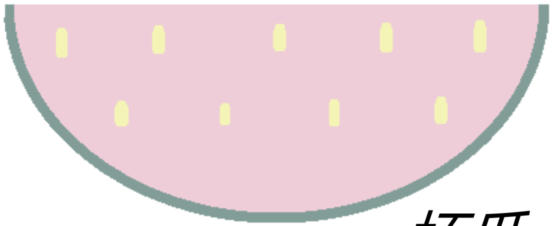
- 颜色
- 纹理
- 触感

→ **无关特征**

- 根蒂
- 声音

→ **相关特征**

当前任务：西瓜是否是好瓜



坏瓜

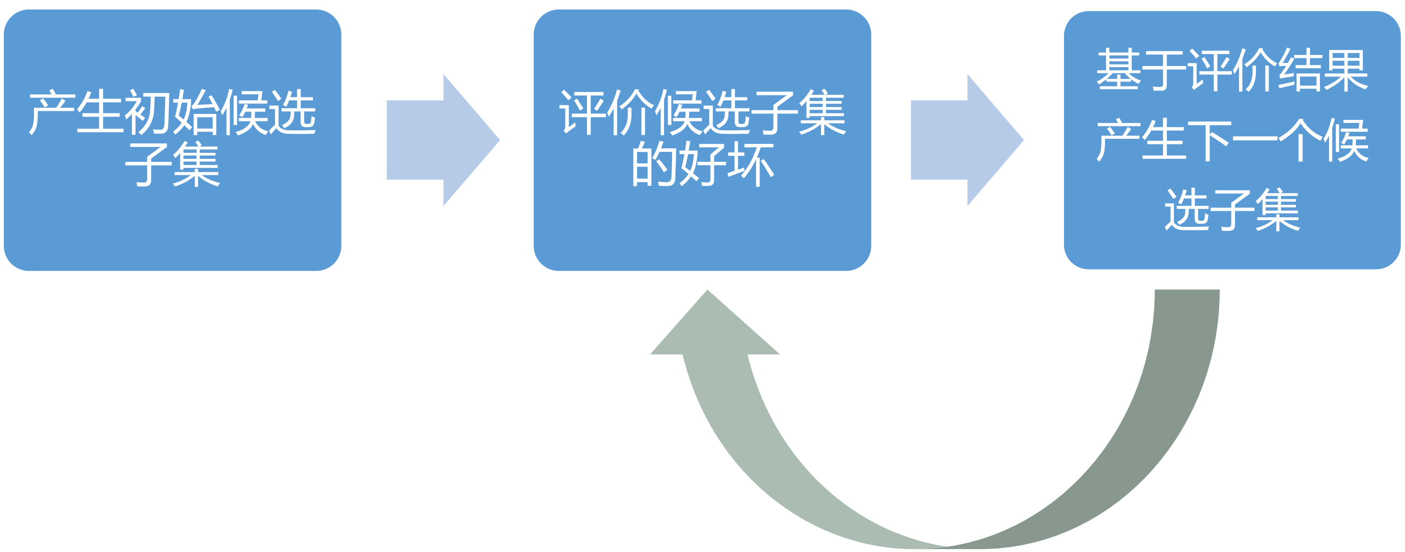
好瓜

特征选择：选择当前任务相关特征

特征选择的一般方法

- 遍历所有可能的子集
  - 计算上遭遇组合爆炸，**不可行**

- 可行方法



两个关键环节：子集搜索和子集评价

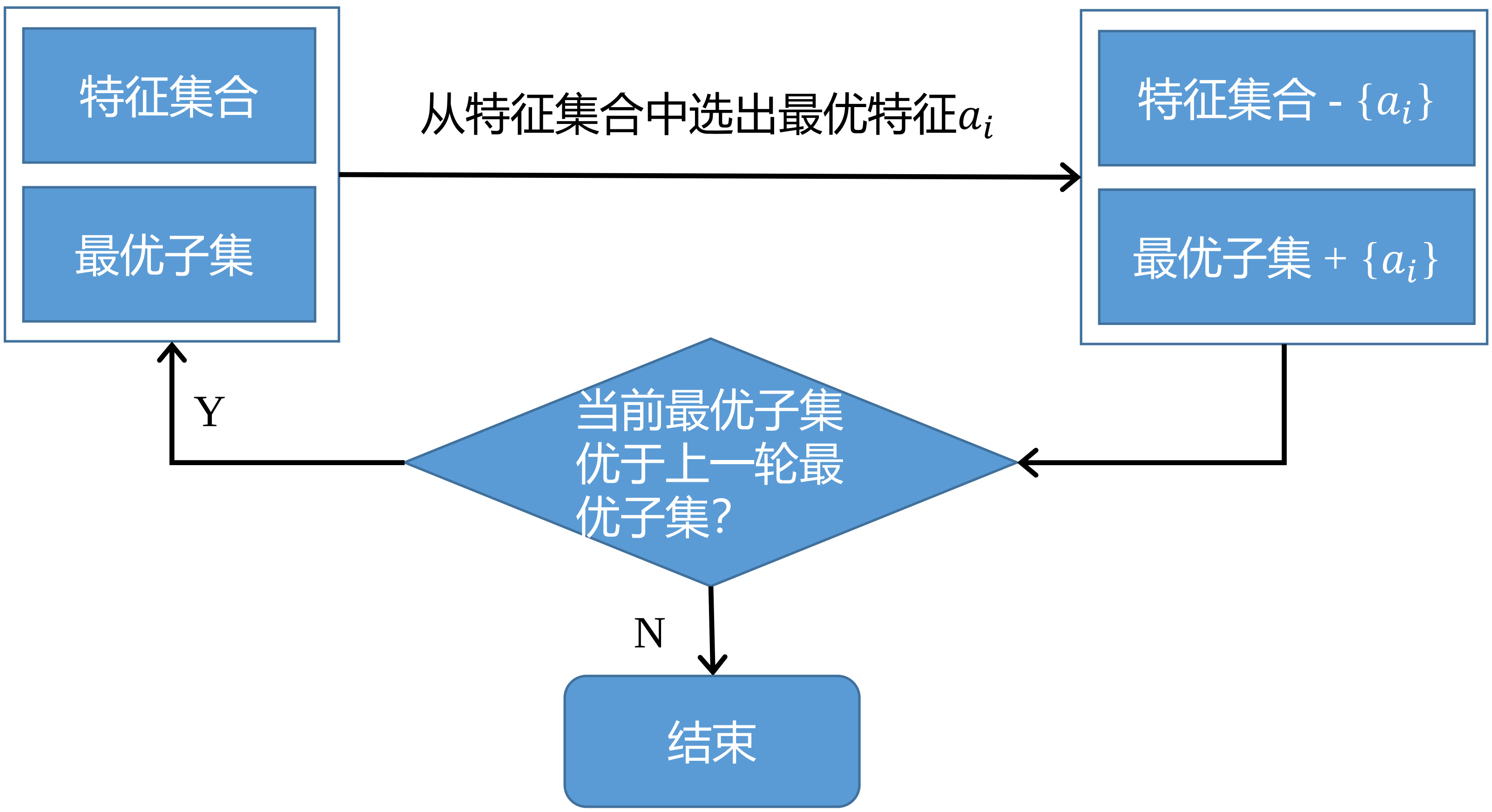
子集搜索

- 用贪心策略选择包含重要信息的特征子集

  - 前向搜索：逐渐增加相关特征
  - 后向搜索：从完整的特征集合开始，逐渐减少特征
  - 双向搜索：每一轮逐渐增加相关特征，同时减少无关特征

前向搜索

- 最优子集初始为空集，特征集合初始时包括所有给定特征



子集评价

- 特征子集确定了对数据集的一个划分
  - 每个划分区域对应着特征子集的某种取值

- 样本标记对应着对数据集的真实划分

通过估算这两个划分的差异，就能对特征子集进行评价；与样本标记对应的划分的差异越小，则说明当前特征子集越好

用信息熵进行子集评价

- 特征子集A确定了对数据集D的一个划分
  - A上的取值将数据集D分为V份，每一份用D^v表示
  - $\mathrm{Ent}(D^v)$表示D^v上的信息熵
- 样本标记Y对应着对数据集D的真实划分
  - $\mathrm{Ent}(D)$表示D上的信息熵

D上的信息熵定义为

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

第i类样本所占比例为p_i

特征子集A的信息增益为：

$$\mathrm{Gain}(A) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Ent}(D^v)$$

常见的特征选择方法

- 将特征子集搜索机制与子集评价机制相结合，即可得到特征选择方法

常见的特征选择方法大致分为如下三类：

- 过滤式
- 包裹式
- 嵌入式

过滤式选择

先用特征选择过程过滤原始数据，再用过滤后的特征来训练模型；特征选择过程与后续学习器无关

- Relief (Relevant Features) 方法 [Kira and Rendell, 1992]
  - 为每个初始特征赋予一个“**相关统计量**”，度量特征的重要性
  - 特征子集的重要性由子集中每个特征所对应的相关统计量之和决定
  - 设计一个阈值，然后选择比阈值大的相关统计量分量所对应的特征
  - 或者指定欲选取的特征个数，然后选择相关统计量分量最大的指定个数特征

如何确定相关统计量?

Relief方法中相关统计量的确定

- 猜对近邻（near-hit）： $\boldsymbol{x}_i$的同类样本中的最近邻$\boldsymbol{x}_{i,nh}$
- 猜错近邻（near-miss）： $\boldsymbol{x}_i$的异类样本中的最近邻$\boldsymbol{x}_{i,\mathrm{nm}}$
- 相关统计量对应于属性j的分量为

$$\delta^j = \sum_i -\mathrm{diff}\left(x_i^j, x_{i,nh}^j\right)^2 + \mathrm{diff}\left(x_i^j, x_{i,nm}^j\right)^2$$

若j为离散型，则$x_a^j = x_b^j$时$\mathrm{diff}\left(x_a^j, x_b^j\right) = 0$，否则为1；若$j$为连续型，则$\mathrm{diff}\left(x_a^j, x_b^j\right) = \left|x_a^j - x_b^j\right|$，注意$x_a^j$，$x_b^j$已规范化到[0,1]区间

- 相关统计量越大，属性j上，猜对近邻比猜错近邻越近，即属性j对区分对错越有用
- Relief方法的时间开销随采样次数以及原始特征数线性增长，运行效率很高

Relief方法的多类拓展

- Relief方法是为二分类问题设计的，其扩展变体Relief-F [Kononenko, 1994]能处理多分类问题
- 数据集中的样本来自$|\mathcal{Y}|$个类别，其中$\boldsymbol{x}_i$属于第k类
- 猜中近邻：第k类中$\boldsymbol{x}_i$的最近邻$\boldsymbol{x}_{i,nh}$
- 猜错近邻：第k类之外的每个类中找到一个$\boldsymbol{x}_i$的最近邻作为猜错近邻，记为$\boldsymbol{x}_{i,l,nm}(l = 1,2, \ldots, \quad ; l \neq k)$
- 相关统计量对应于属性的分量为

$$\delta^j = \sum_i \left(-\mathrm{diff}\left(x_i^j, x_{i,nh}^j\right)^2 + \sum_{l \neq k} p_l \times \mathrm{diff}\left(x_i^j, x_{i,l,nm}^j\right)^2 \right)$$

p_l为第l类样本在数据集D中所占的比例

包裹式选择

- 包裹式选择直接把最终将要使用的学习器的性能作为特征子集的评价准则
    - 包裹式特征选择的目的就是为给定学习器选择最有利于其性能、“量身定做”的特征子集
    - 包裹式选择方法直接针对给定学习器进行优化，因此从最终学习器性能来看，包裹式特征选择比过滤式特征选择更好
    - 包裹式特征选择过程中需多次训练学习器，计算开销通常比过滤式特征选择大得多

LVW包裹式特征选择方法

- LVW（Las Vegas Wrapper）[Liu and Setiono, 1996] 在拉斯维加斯方法框架下使用随机策略来进行子集搜索，并以最终分类器的误差作为特征子集评价准则

- 基本步骤
  - 在循环的每一轮随机产生一个特征子集
  - 在随机产生的特征子集上通过交叉验证推断当前特征子集的误差
  - 进行多次循环，在多个随机产生的特征子集中选择误差最小的特征子集作为最终解*

*若有运行时间限制，则该算法有可能给不出解

嵌入式选择

- 嵌入式特征选择是将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成，在学习器训练过程中自动地进行特征选择

- 考虑最简单的线性回归模型，以平方误差为损失函数，并引入L_2范数正则化项防止过拟合，则有

$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_2^2$$

岭回归 (ridge regression)
[Tikhonov and Arsenin, 1977]

- 将L_2范数替换为L_1范数，则有**LASSO** [Tibshirani, 1996]

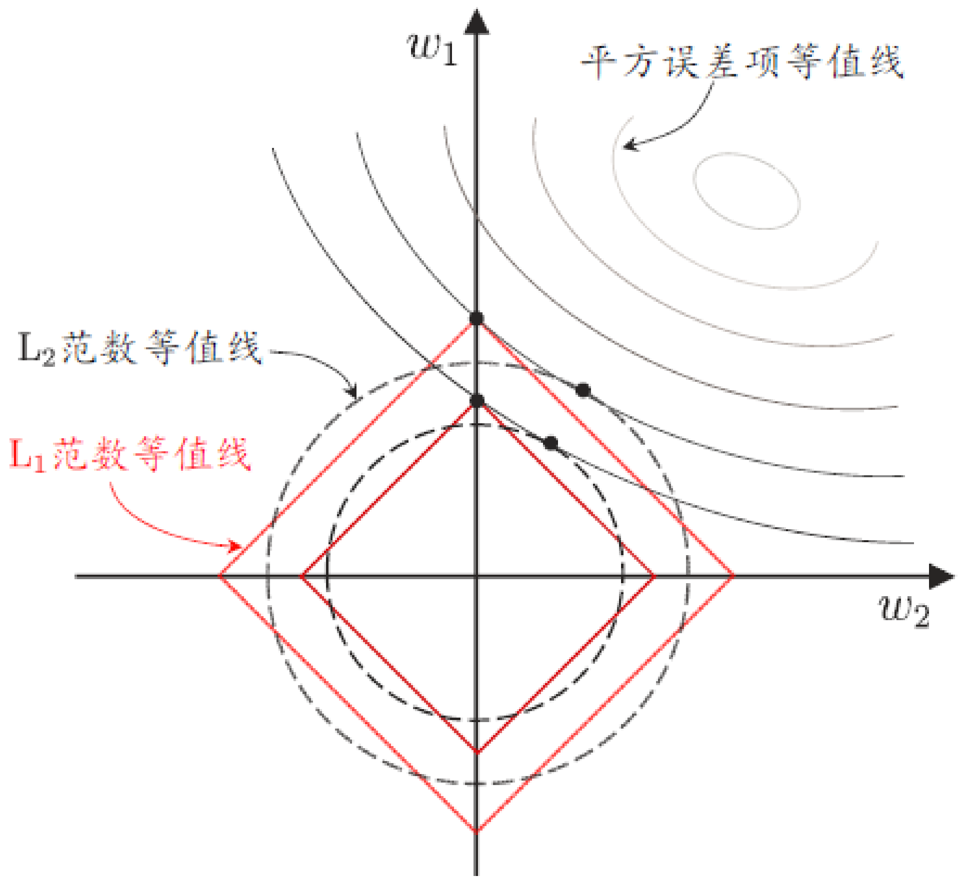
$$\min_{\boldsymbol{w}} \sum_{i=1}^{m} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|_1$$

$$\|\boldsymbol{w}\|_1 = \sum_{d} |w_d|$$

易获得稀疏解，是一种嵌入式特征选择方法

使用L_1范数正则化易获得稀疏解

等值线即取值相同的点的连线



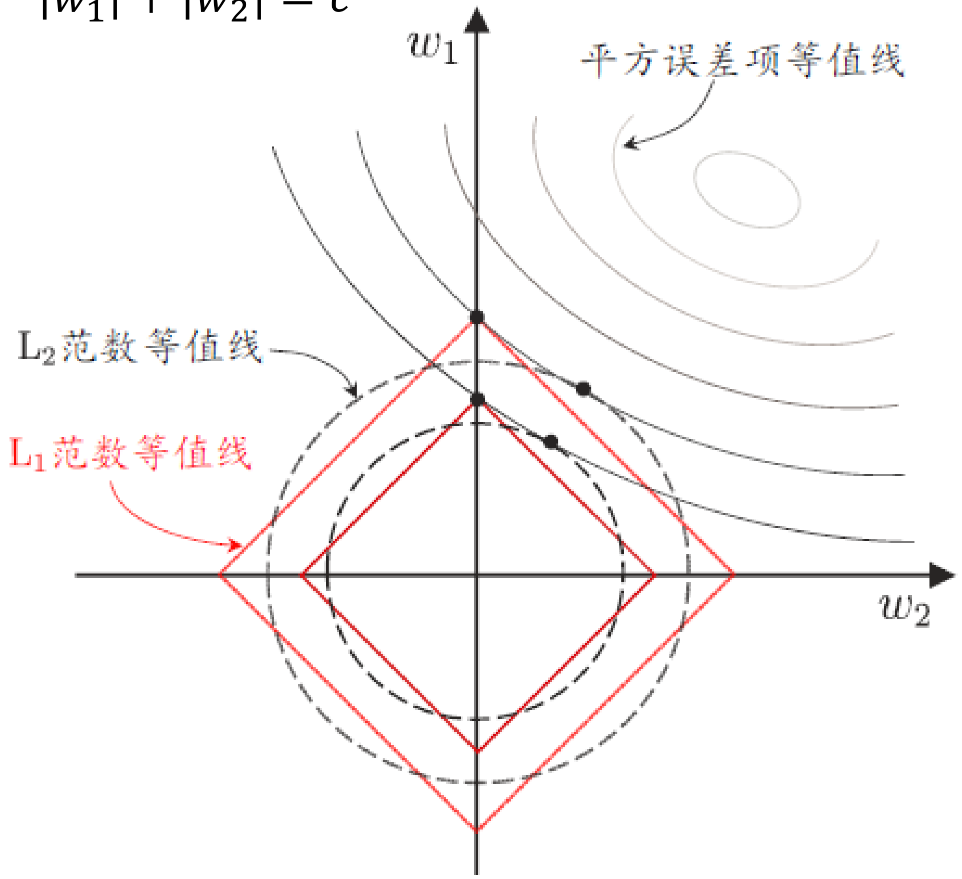
$$\|\boldsymbol{w}\|_1 = c$$

$$|w_1| + |w_2| = c$$

$$\begin{cases} w_1 + w_2 = c, \text{if } w_1 \geq 0, w_2 \geq 0 \\ w_1 - w_2 = c, \text{if } w_1 \geq 0, w_2 < 0 \\ -w_1 + w_2 = c, \text{if } w_1 < 0, w_2 \geq 0 \\ -w_1 - w_2 = c, \text{if } w_1 < 0, w_2 < 0 \end{cases}$$

使用L_1范数正则化易获得稀疏解

$|w_1| + |w_2| = c$



等值线即取值相同的点的连线

- 假设$\boldsymbol{x}$仅有两个属性，那么$\boldsymbol{w}$有两个分量w_1和w_2. 那么目标优化的解要在平方误差项与正则化项之间折中,即出现在图中平方误差项等值线与正则化等值线相交处.

从图中看出,采用L_1范数时交点常出现在坐标轴上,即产生w_1或者w_2为0的稀疏解

L_1正则化问题的求解(1)

- 可使用近端梯度下降(Proximal Gradient Descend，简称PGD) 求解

[Boyd and Vandenberghe, 2004]

- 考虑更一般的问题

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1$$

- 假设$f(\boldsymbol{w})$为凸函数，且$\nabla f(\boldsymbol{w})$满足L-Lipschitz条件， 即存在常数$L > 0$, 使得

$$\forall \boldsymbol{w}, \boldsymbol{w}' \quad \|\nabla f(\boldsymbol{w}') - \nabla f(\boldsymbol{w})\| \leq L\|\boldsymbol{w}' - \boldsymbol{w}\|$$

L_1正则化问题的求解(1)

- 假设$f(\boldsymbol{x})$为凸函数，且$\nabla f(\boldsymbol{x})$满足L-Lipschitz条件， 即存在常数$L>0$, 使得

$$\forall \boldsymbol{w},\boldsymbol{w}' \quad \|\nabla f(\boldsymbol{w}')-\nabla f(\boldsymbol{w})\| \le L\|\boldsymbol{w}'-\boldsymbol{w}\|$$

- 等价于$g(\boldsymbol{w})=\frac{L}{2}\|\boldsymbol{w}\|^2-f(\boldsymbol{w})$为凸函数

$$\boxed{f(\boldsymbol{w}') \ge f(\boldsymbol{w})+\nabla f(\boldsymbol{w})^{\top}(\boldsymbol{w}'-\boldsymbol{w})} + \boxed{f(\boldsymbol{w}) \ge f(\boldsymbol{w}')+\nabla f(\boldsymbol{w}')^{\top}(\boldsymbol{w}-\boldsymbol{w}')}$$

$$\Rightarrow \left(\nabla f(\boldsymbol{w})-\nabla f(\boldsymbol{w}')\right)^{\top}(\boldsymbol{w}-\boldsymbol{w}') \ge 0$$

$$\Rightarrow \left(\nabla f(\boldsymbol{w})-\nabla f(\boldsymbol{w}')\right)^{\top}(\boldsymbol{w}-\boldsymbol{w}') \le L\|\boldsymbol{w}-\boldsymbol{w}'\|^2$$

$$\left(\nabla g(\boldsymbol{w}')-\nabla g(\boldsymbol{w})\right)^{\top}(\boldsymbol{w}'-\boldsymbol{w}) = \left(L(\boldsymbol{w}'-\boldsymbol{w})-\left(\nabla f(\boldsymbol{w}')-\nabla f(\boldsymbol{w})\right)\right)(\boldsymbol{w}'-\boldsymbol{w})$$

$$\ge 0$$

L_1正则化问题的求解(1)

- 假设$f(\boldsymbol{x})$为凸函数，且$\nabla f(\boldsymbol{x})$满足L-Lipschitz条件， 即存在常数$L>0$, 使得

$$\forall \boldsymbol{w}, \boldsymbol{w}' \quad \|\nabla f(\boldsymbol{w}') - \nabla f(\boldsymbol{w})\| \leq L\|\boldsymbol{w}' - \boldsymbol{w}\|$$

- 等价于$g(\boldsymbol{w}) = \frac{L}{2}\|\boldsymbol{w}\|^2 - f(\boldsymbol{w})$为凸函数
- 也等价于$f(\boldsymbol{w}') \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})(\boldsymbol{w}' - \boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2$

$$g(\boldsymbol{w}') \geq g(\boldsymbol{w}) + \nabla g(\boldsymbol{w})(\boldsymbol{w}' - \boldsymbol{w})$$

$$\frac{L}{2}\|\boldsymbol{w}'\|^2 - f(\boldsymbol{w}') \geq \frac{L}{2}\|\boldsymbol{w}\|^2 - f(\boldsymbol{w}) + \left(L\boldsymbol{w} - \nabla f(\boldsymbol{w})\right)^\top (\boldsymbol{w}' - \boldsymbol{w})$$

$$f(\boldsymbol{w}') \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\top (\boldsymbol{w}' - \boldsymbol{w}) - L\boldsymbol{w}(\boldsymbol{w}' - \boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{w}'\|^2 - \frac{L}{2}\|\boldsymbol{w}\|^2$$

$$f(\boldsymbol{w}') \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})(\boldsymbol{w}' - \boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2$$

L_1正则化问题的求解(1)

- 假设$f(\boldsymbol{x})$为凸函数，且$\nabla f(\boldsymbol{x})$满足L-Lipschitz条件， 即存在常数$L > 0$, 使得

$$\forall \boldsymbol{w}, \boldsymbol{w}' \quad \|\nabla f(\boldsymbol{w}') - \nabla f(\boldsymbol{w})\| \leq L\|\boldsymbol{w}' - \boldsymbol{w}\|$$

- 等价于$g(\boldsymbol{w}) = \frac{L}{2}\|\boldsymbol{w}\|^2 - f(\boldsymbol{w})$为凸函数
- 也等价于$f(\boldsymbol{w}') \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})(\boldsymbol{w}' - \boldsymbol{w}) + \frac{L}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2$

$$= \frac{L}{2}\left\|\boldsymbol{w}' - \left(\boldsymbol{w} - \frac{1}{L}\nabla f(\boldsymbol{w})\right)\right\|_2^2 + \text{const}$$

$$f(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1 \leq \frac{L}{2}\left\|\boldsymbol{w} - \left(\boldsymbol{w}^k - \frac{1}{L}\nabla f\left(\boldsymbol{w}^k\right)\right)\right\|_2^2 + \lambda\|\boldsymbol{w}\|_1 + \text{const}$$

L_1正则化问题的求解(1)

- $f(\boldsymbol{w})+\lambda\|\boldsymbol{w}\|_1 \le \frac{L}{2}\left\|\boldsymbol{w}-\left(\boldsymbol{w}^k-\frac{1}{L}\nabla f(\boldsymbol{w}^k)\right)\right\|_2^2+\lambda\|\boldsymbol{w}\|_1+\text{const}$

$$\boldsymbol{w}^{k+1}=\underset{\boldsymbol{w}}{\operatorname{argmin}}\frac{L}{2}\left\|\boldsymbol{w}-\boxed{\left(\boldsymbol{w}^k-\frac{1}{L}\nabla f(\boldsymbol{w}^k)\right)}\right\|_2^2+\lambda\|\boldsymbol{w}\|_1$$

$$\boldsymbol{w}^{k+1}=\underset{\boldsymbol{w}}{\operatorname{argmin}}\frac{L}{2}\|\boldsymbol{w}-\boxed{\boldsymbol{z}^k}\|_2^2+\lambda\|\boldsymbol{w}\|_1$$

$$[\boldsymbol{w}^{k+1}]_i=\arg\min_{w_i}\frac{L}{2}\left(w_i-z_i^k\right)^2+\lambda|w_i|$$

$$\frac{L}{2}\left(w_i-z_i^k\right)^2+\lambda|w_i|$$

$$=\begin{cases}\frac{L}{2}\left(w_i-z_i^k+\frac{\lambda}{L}\right)^2+const, \text{if } w_i\ge 0\\ \frac{L}{2}\left(w_i-z_i^k-\frac{\lambda}{L}\right)^2+const, \text{if } w_i<0\end{cases}$$

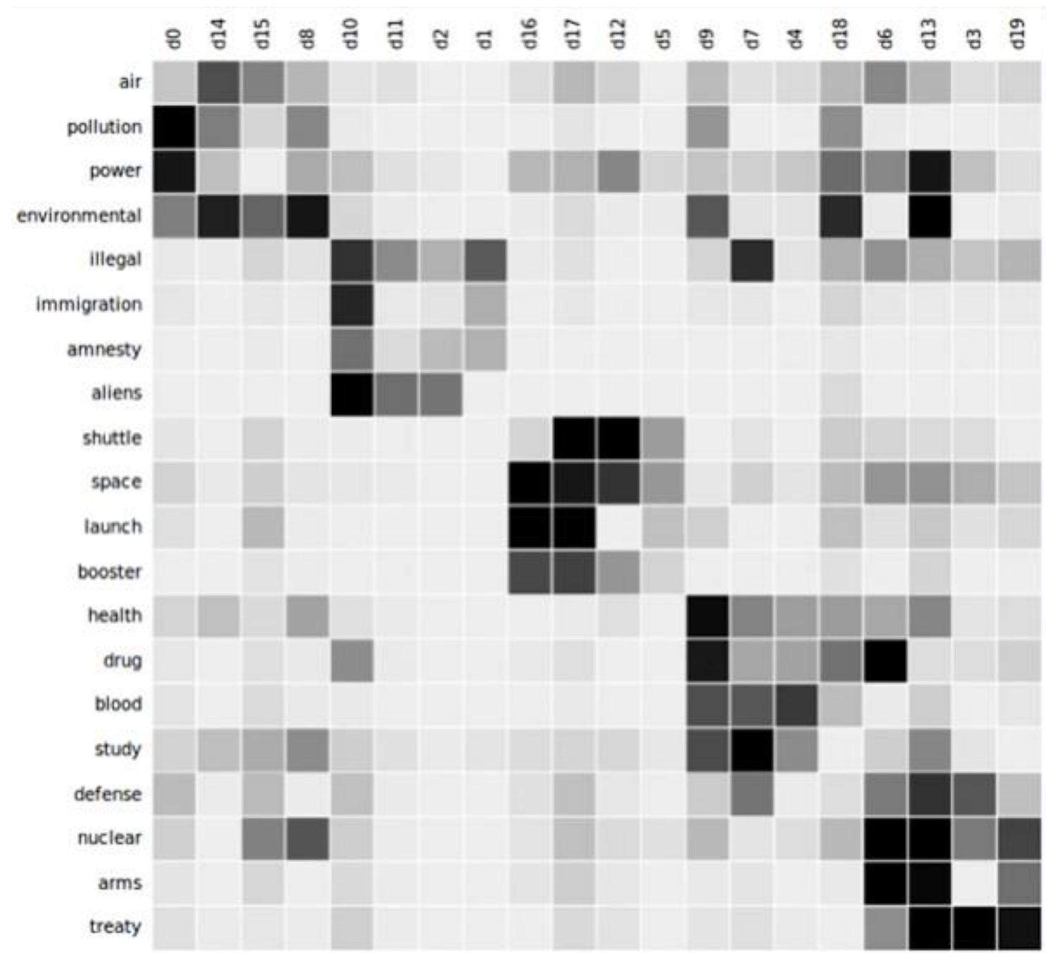
$$w_i=\begin{cases}z_i^k-\frac{\lambda}{L}, \text{if } \frac{\lambda}{L}<z_i^k\\ 0, \text{if } \frac{\lambda}{L}\ge\left|z_i^k\right|\\ z_i^k+\frac{\lambda}{L}, \text{if} z_i^k<-\frac{\lambda}{L}\end{cases}$$

稀疏表示与字典学习

- 将数据集考虑成一个矩阵，每行对应一个样本，每列对应一个特征

- 矩阵中有很多零元素，且非整行整列出现

- 稀疏表达的优势：
  - 文本数据线性可分
  - 存储高效



能否将稠密表示的数据集转化为“稀疏表示”，
使其享受稀疏表达的优势？

稀疏表示与字典学习

- 一般的学习任务中，并没有字典可用，需学习出这样一个字典

- 为普通稠密表达的样本找到合适的字典，将样本转化为稀疏表示，这一过程称为字典学习

- 给定数据集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\},\boldsymbol{x}_i\in\mathbb{R}^d$
- 最简单的字典学习的优化形式为

$$\min_{\boldsymbol{B},\boldsymbol{\alpha}_i}\sum_{i=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{B}\boldsymbol{\alpha}_i\|_2^2+\lambda\sum_{i=1}^{m}\|\boldsymbol{\alpha}_i\|_1$$

采用交替迭代优化进行求解

$\boldsymbol{B}\in\mathbb{R}^{d\times k}$为字典矩阵、$\boldsymbol{\alpha}_i\in\mathbb{R}^k$为样本的稀疏表示

k称为字典的词汇量，通常由用户指定

字典学习的解法(1)

- 固定字典**B**，参考LASSO的方法求解

$$\min_{\boldsymbol{\alpha}_i} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{m} \|\boldsymbol{\alpha}_i\|_1$$

- 以$\boldsymbol{\alpha}_i$为初值更新字典$\boldsymbol{B}$

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m] \in \mathbb{R}^{d\times m}$$

$$\min_{\boldsymbol{B}} L = \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{\alpha}_i\|_2^2 \quad = \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{A}\|_F^2$$

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_m] \in \mathbb{R}^{k\times m}$$

$$\nabla_{\boldsymbol{B}} L = -2(\boldsymbol{X} - \boldsymbol{B}\boldsymbol{A})\boldsymbol{A}^\top = 0 \quad \Rightarrow \quad \boldsymbol{B} = \boldsymbol{X}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}$$

字典学习的解法(1)

- 当k很大的时候，求逆计算代价高，可以通过逐列更新

$$L = \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{A}\|_F^2 = \left\| \boldsymbol{X} - \sum_{i=1}^{k} \boldsymbol{b}_i \boldsymbol{a}_{(i)}^\top \right\|$$

$\boldsymbol{b}_i$表示字典$\boldsymbol{B}$的第i列

$\boldsymbol{a}_{(i)}$表示矩阵$\boldsymbol{A}$的第i行

$$= \left\| \boxed{\left(\boldsymbol{X} - \sum_{j \neq i} \boldsymbol{b}_i \boldsymbol{a}_{(i)}^\top \right)} - \boldsymbol{b}_i \boldsymbol{a}_{(i)}^\top \right\|_F^2$$

$$= \left\| \boxed{\boldsymbol{E_i}} - \boldsymbol{b}_i \boldsymbol{a}_{(i)}^\top \right\|_F^2$$

$$\nabla_{\boldsymbol{b}_i} L = -2\left(\boldsymbol{E}_i - \boldsymbol{b}_i \boldsymbol{a}_{(i)}^\top\right)\boldsymbol{a}_{(i)} = 0 \Rightarrow \boldsymbol{b}_i = \frac{\boldsymbol{E}_i \boldsymbol{a}_{(i)}}{\boldsymbol{a}_{(i)}^\top \boldsymbol{a}_{(i)}}$$

反复迭代以获得最优解

压缩感知

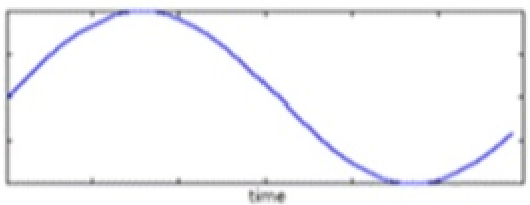
- 能否利用部分数据恢复全部数据?

- 数据传输中，能否利用接收到的压缩、丢包后的数字信号，精确重构出原信号?

- 压缩感知 (compressive sensing) [Cándes et al., 2006, Donoho, 2006] 为解决此类问题提供了新的思路.

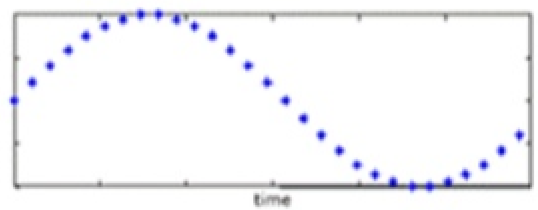
针对信号采样的技术，它通过一些手段，实现了“压缩的采样”，准确说是**在采样过程中完成了数据压缩的过程。**

模拟信号采样

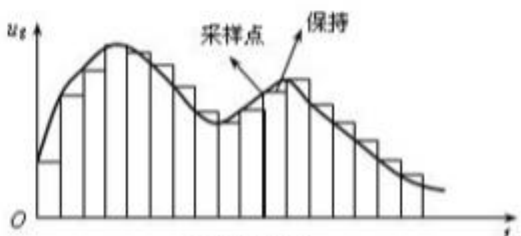


模拟信号

采样



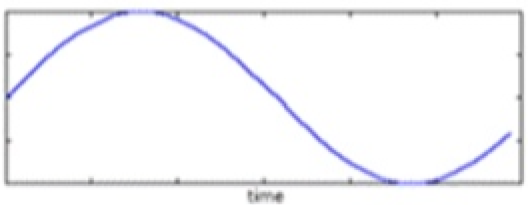
数字信号



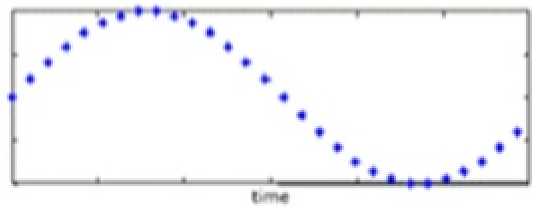
采样波形图

用多大的采样频率，才能完全恢复模拟信号？

模拟信号采样

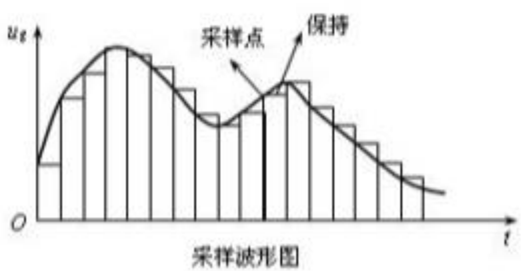


采样



模拟信号

数字信号



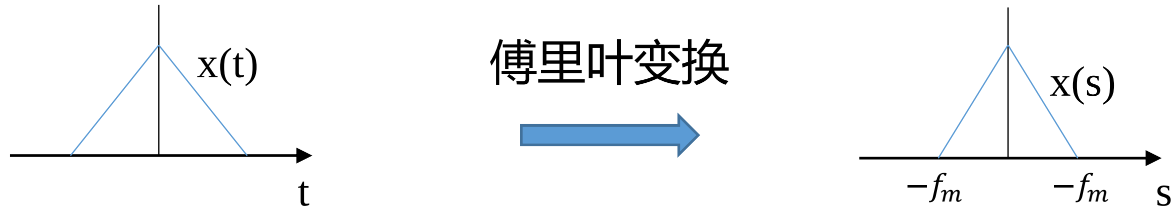
奈奎斯特采样定理

为了不失真地恢复模拟信号，采样频率应该**大于模拟信号频谱中最高频率的2倍**。

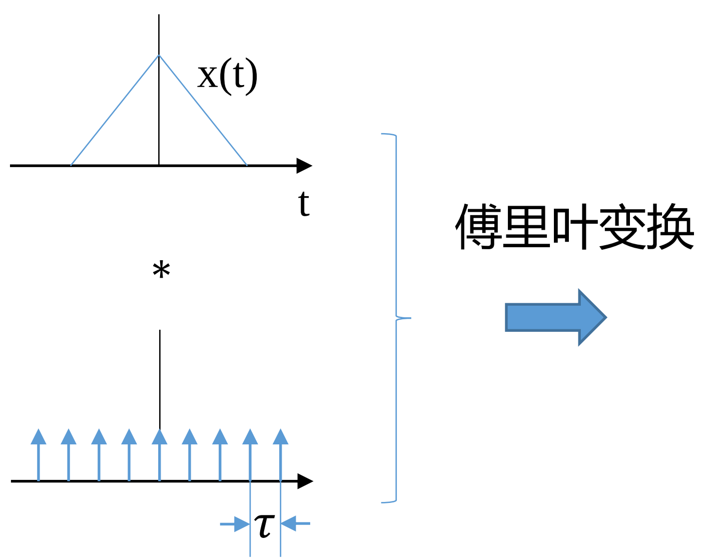
奈奎斯特
1889-1976

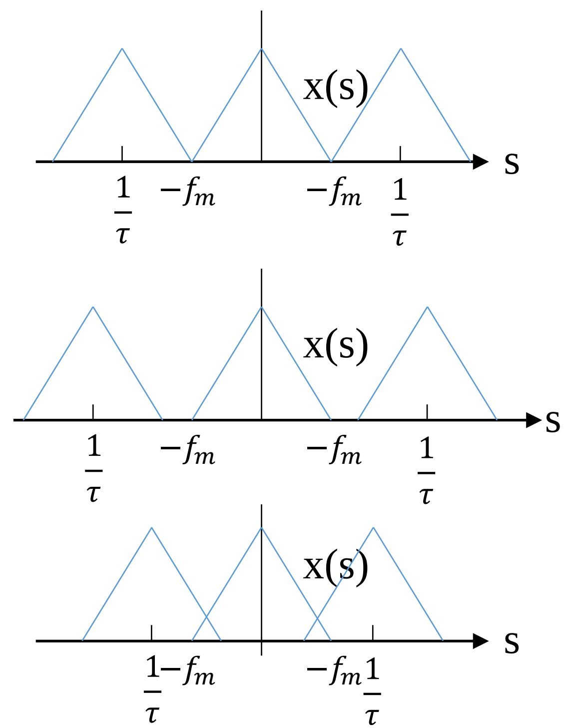
模拟信号采样

时域　　　　　　　　　　　　　　频域



原始信号





混叠

时域以τ为间隔进行采样，频域会以1/τ为周期发生周期延拓

压缩感知

- 2004年，陶哲轩等人证明，如果信号是稀疏的，那么可以由远低于采样定理要求的采样点重建恢复，并与2007年正式提出压缩感知这个概念

陶哲轩

Emmanuel Candes

David Donoho

压缩感知

- 采样频率应该大于模拟信号频谱中最高频率的2倍

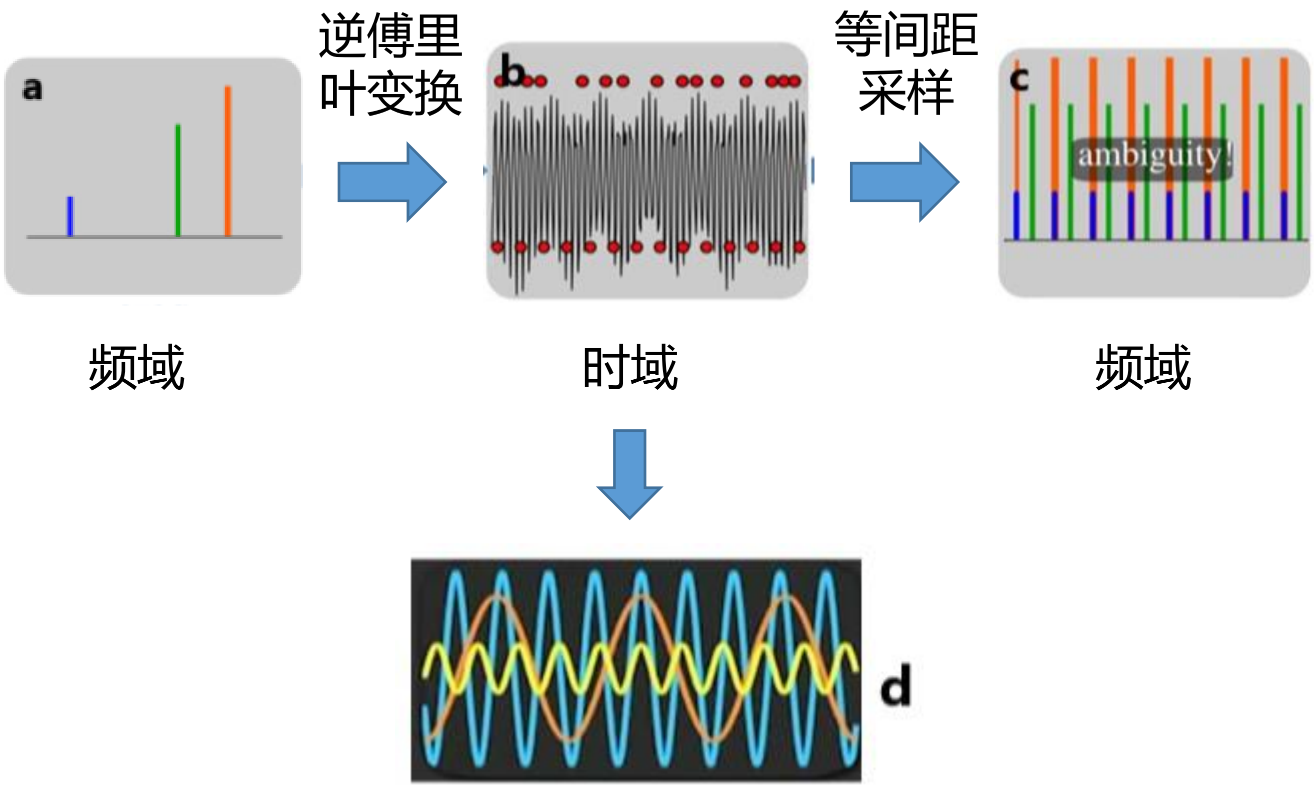
  奈奎斯特采样定理

- 给定采样频率采样，意味着等间距采样

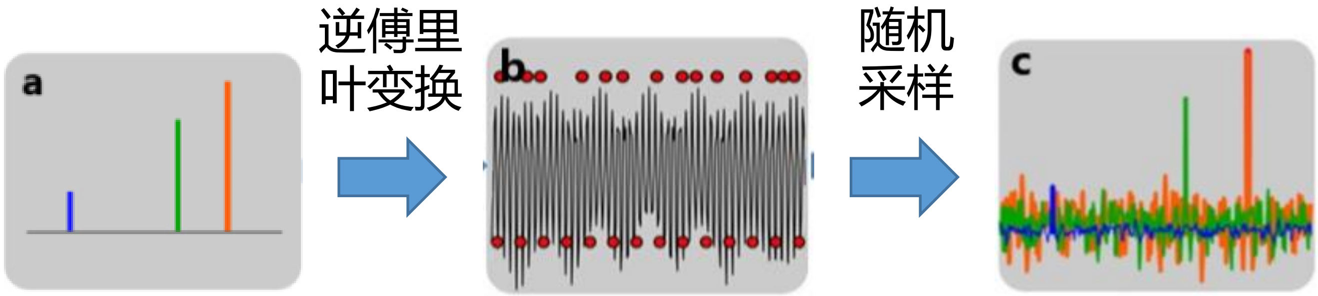
- 等间距采样，频域将以以1/τ为周期延拓，采样频率低的时候会一起混叠

- 如果是不等间距采样或者随机采样呢？

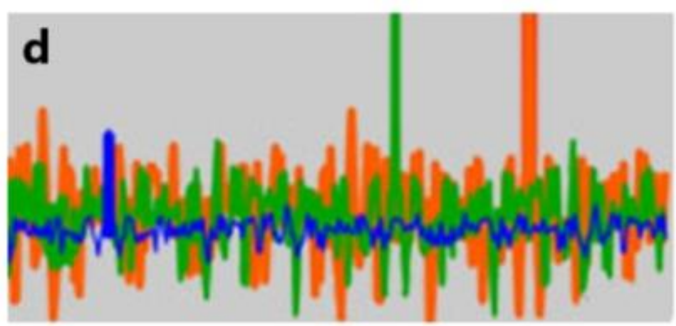
压缩感知



压缩感知



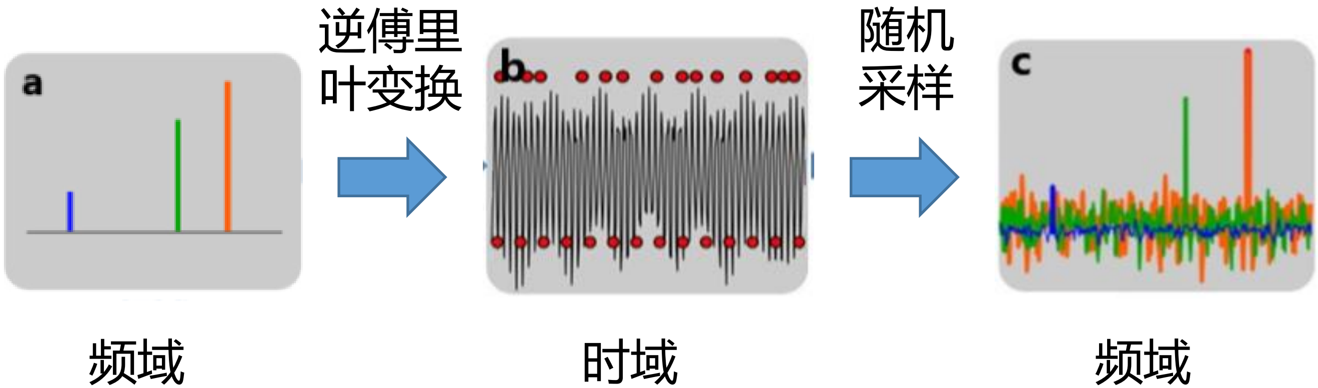
频域　　　　时域　　　　频域



频域就不再是以固定周期进行延拓了，而是会产生大量不相关（incoherent）的干扰值

随机采样使得频谱不再是整齐地搬移，而是一小部分一小部分胡乱地搬移

压缩感知—前提条件



信号在频域稀疏

随机采样

稀疏性（sparsity）

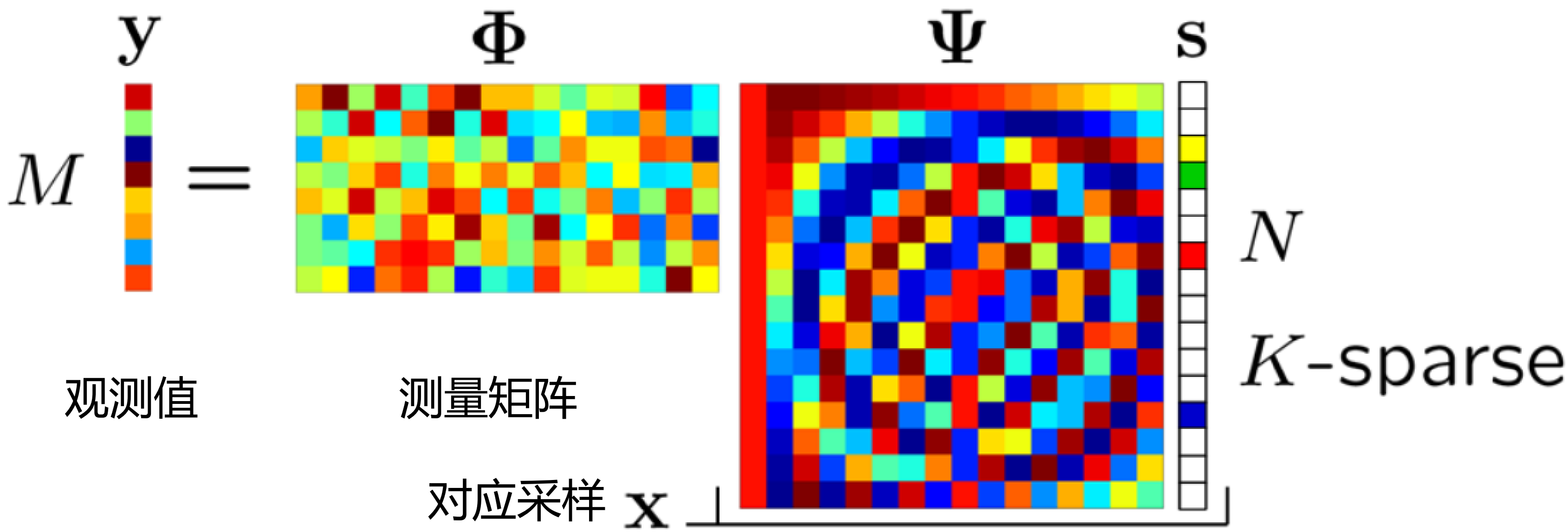
不相关性（incoherence）

压缩感知

一般的自然信号x本身并不是稀疏的，需要在某种稀疏基上进行稀疏表示

对x在基Ψ上进行稀疏表示，$x = \Psi s$

基、字典　　编码



观测值　　测量矩阵

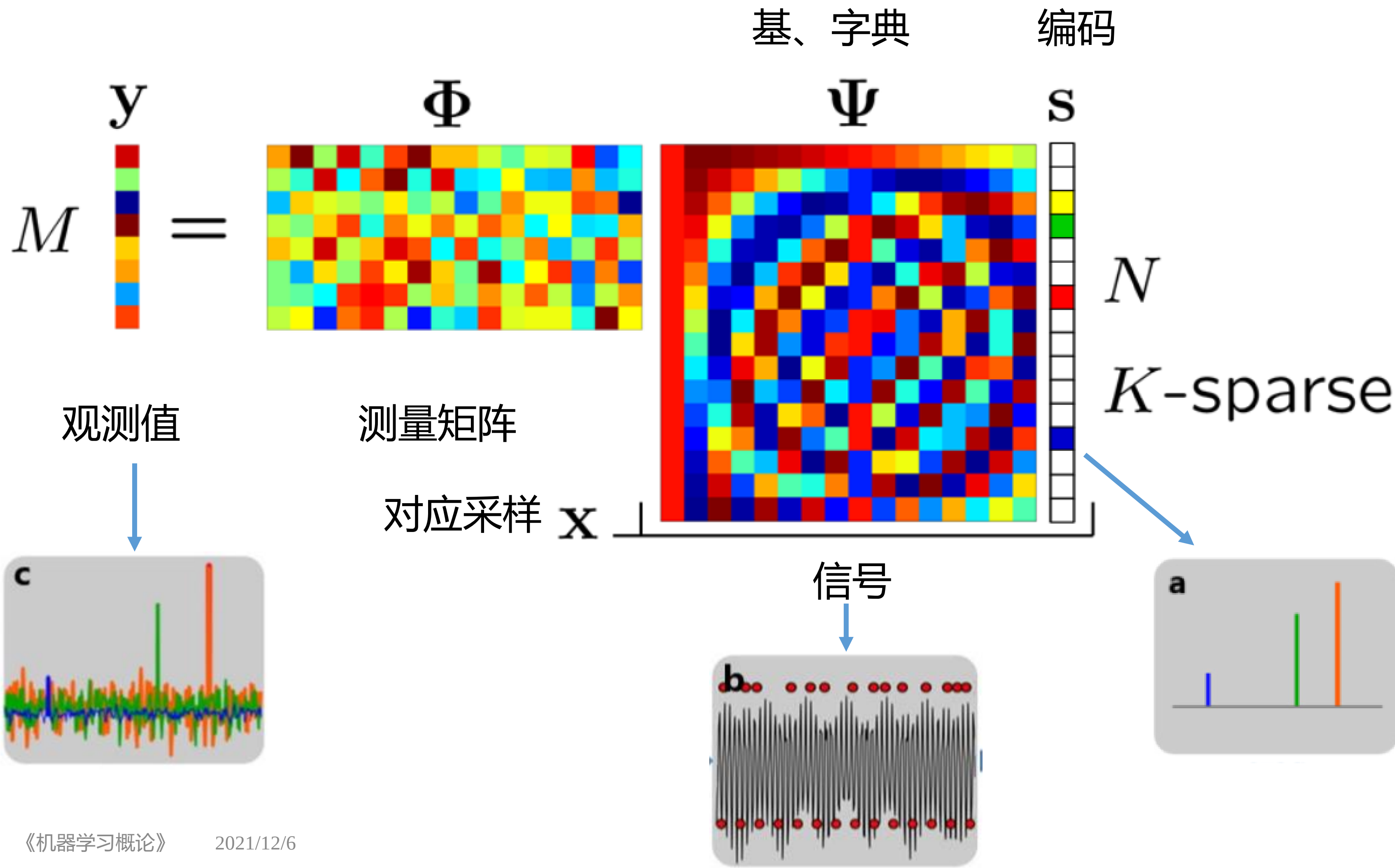
对应采样

信号

测量矩阵将高维信号投影到低纬空间

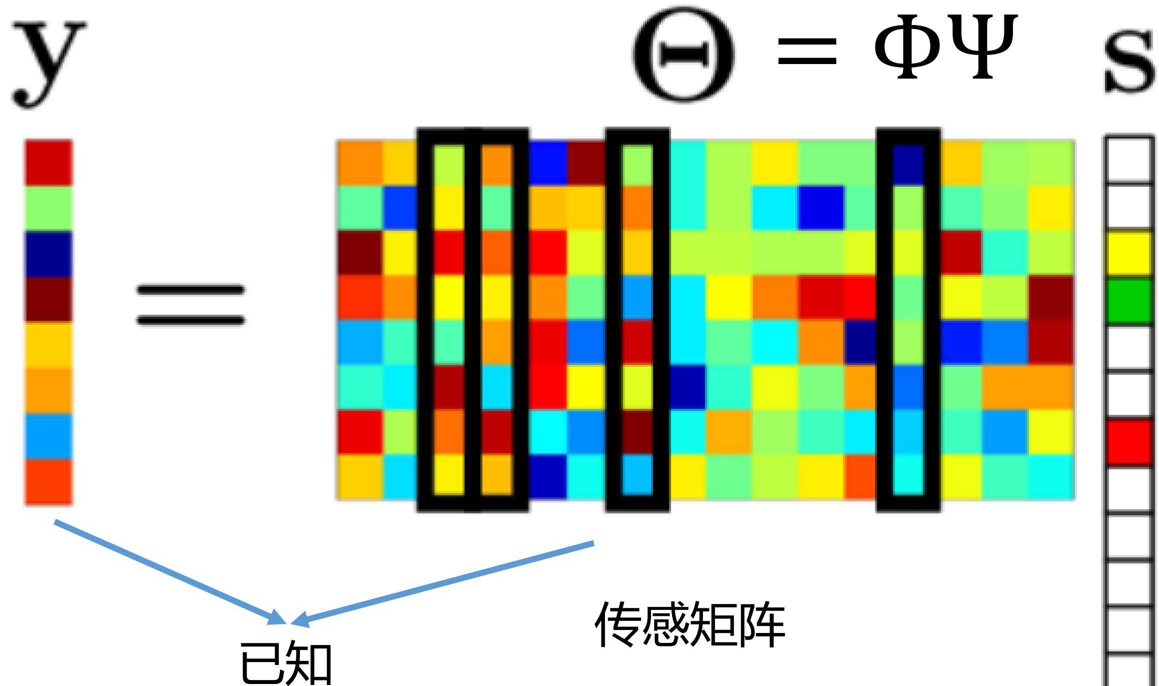
压缩感知问题就是在已知测量值y和测量矩阵Φ的基础上，求解欠定方程组$y = \Phi x$得到原信号x

压缩感知



基、字典
编码
$\mathbf{y}$
Φ
Ψ
$\mathbf{s}$
M
$=$
N
K-sparse
观测值
测量矩阵
对应采样
$\mathbf{x}$
信号
c
b
a

压缩感知



$\boldsymbol{y}$、$\boldsymbol{\Theta}$ ⇨ $\boldsymbol{s}$

若$M = N$，则可以轻松解出$\boldsymbol{s}$

若$M < N$，则根据$\boldsymbol{\Theta}$的RIP特性重构

RIP特性—k-RIP

- 存在某个$\epsilon > 0$，对任意含有k个非零元素的向量v,

$$满足\ 1-\epsilon \le \frac{\|\boldsymbol{\Theta v}\|_2}{\|\boldsymbol{v}\|_2} \le 1+\epsilon,$$

矩阵$\boldsymbol{\Theta}$必须维持向量$\boldsymbol{v}$的长度只发生小量变化

- Baraniuk证明：RIP的等价条件是测量矩阵和稀疏基的不相关性

不相关性

$$\Phi \longleftrightarrow \Psi$$

陶哲轩和Candès又证明:

独立同分布的高斯随机测量矩阵可以成为普适的压缩感知测量矩阵

压缩感知的优化目标和解法

- 若$\mathbf{A}$满足k限定等距性，则可通过下面的优化问题近乎完美地从$\boldsymbol{y}$中恢复出稀疏信号$\boldsymbol{s}$，进而恢复出$\boldsymbol{x}$：

$$\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_0 \;\; s.t.\, \boldsymbol{y} = \boldsymbol{\Theta}\boldsymbol{s}$$

- L_0范数最小化是NP难问题。不过，L1范数最小化在一定条件下与L0范数最小化问题共解 [Cándes et al., 2006]：将上式转化为共解的L_1范数最小化问题

$$\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_1 \;\; s.t.\, \boldsymbol{y} = \boldsymbol{\Theta}\boldsymbol{s}$$

- 转化为LASSO的等价形式再通过近端梯度下降法求解，即使用“基寻踪去噪”（Basis Pursuit De-Noising）[Chen et al., 1998]

矩阵补全

客户对书籍的喜好程度的评分

	《笑傲江湖》	《万历十五年》	《人间词话》	《云海玉弓缘》	《人类的故事》
赵大	5	?	?	3	2
钱二	?	5	3	?	5
孙三	5	3	?	?	?
李四	3	?	5	4	?

- 能否将表中已经通过读者评价得到的数据当作部分信号，基于压缩感知的思想恢复出完整信号从而进行书籍推荐呢？从题材、作者、装帧等角度看（相似题材的书籍有相似的读者），表中反映的信号是稀疏的，能通过类似压缩感知的思想加以处理。

“矩阵补全”技术解决此类问题

矩阵补全的优化问题和解法

- 矩阵补全（matrix completion）技术的优化形式为

$$\min_{\boldsymbol{X}} \operatorname{rank}(\boldsymbol{X})$$
$$\mathrm{s.t.}\ X_{ij} = A_{ij}, (i,j) \in \Omega$$

约束表明，恢复出的矩阵中$\mathbf{X}_{ij}$应当与已观测到的对应元素相同

- **X**: 需要恢复的稀疏信号
- rank(**X**): **X**的秩
- **A**: 已观测信号
- Ω: **A**中已观测到的元素的位置下标的集合

- NP难问题. 将rank(**X**)转化为其凸包“核范数”（nuclear norm）

$$\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_*$$
$$\mathrm{s.t.}\ X_{ij} = A_{ij}, (i,j) \in \Omega$$

$$\|\boldsymbol{X}\|_* = \sum_{j=1}^{\min\{m,n\}} \sigma_j(\boldsymbol{X})$$

$\sigma_j(\boldsymbol{X})$ 为**X**的奇异值

$\|\boldsymbol{X}\|_*$矩阵的核范数为矩阵的奇异值之和

- 凸优化问题，通过半正定规划求解（SDP，Semi-Definite Programming）

满足一定条件时，只需观察到$O(mr\log^2 m)$个元素就能完美恢复**A** [Recht, 2011]

作业

- 11.5
- 11.7

阅读材料

- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.
- Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492-1496.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), 484-489.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. (2017). Mastering the game of go without human knowledge. *nature*, *550*(7676), 354-359.
- Zhu, J., Wen, C., Zhu, J., Zhang, H., & Wang, X. (2020). A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*.