MACHINE LEARNING

机器学习

Feature Selection

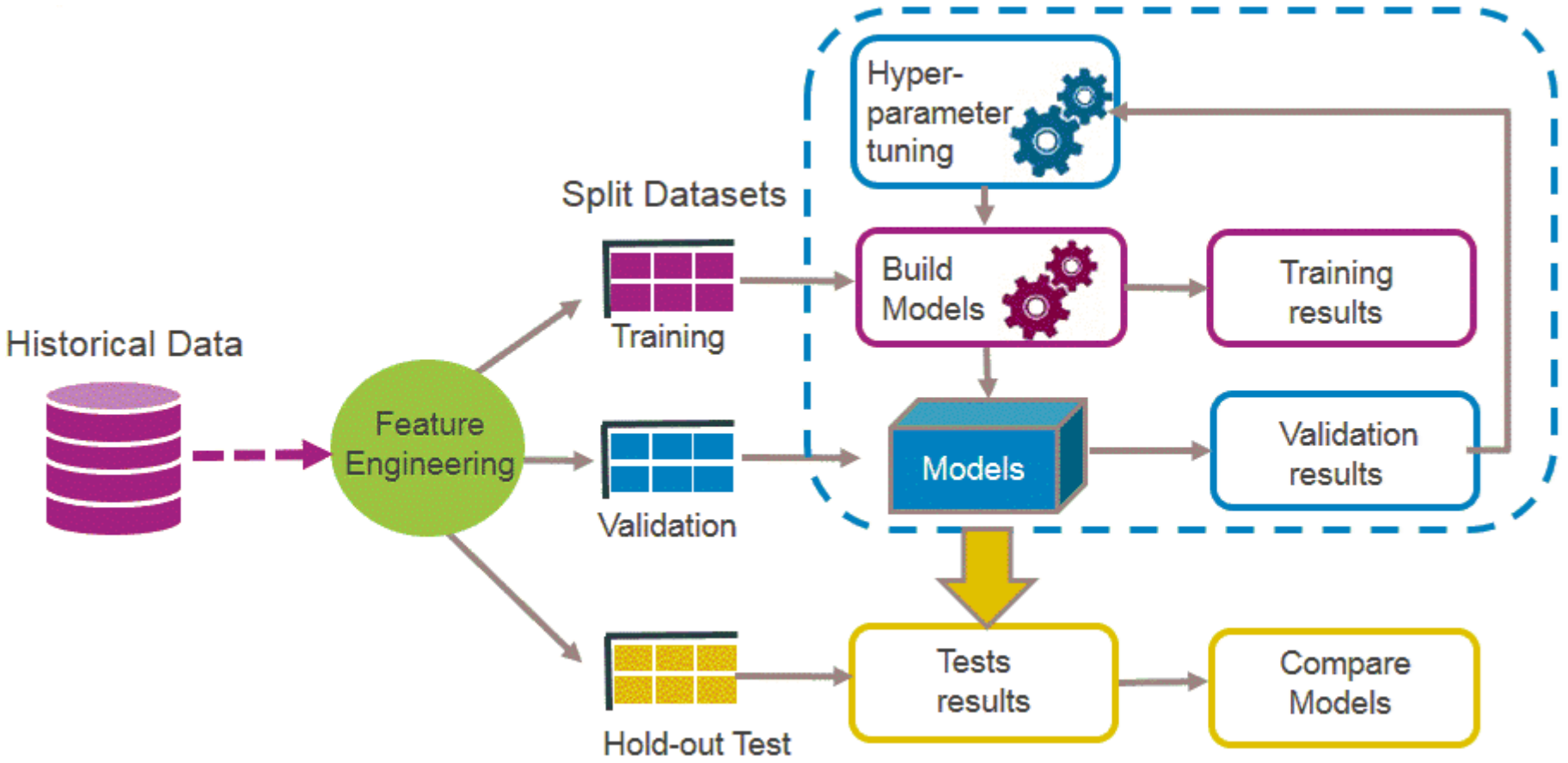
特征工程

参考：
《机器学习》

Machine Learning Course

涵义

特征工程指的是这样一个过程：将数据转换为能更好的（特征）表示，从而提高后续机器学习的性能。



范围

- 特征构建
- 特征探索
- 特征增强
- 特征衍生
- 特征选择
- 特征转换
- 特征学习

特征选择

- 定义

不同于前面提到的特征抽取方法（例如PCA算法）通过某种数学变换将原始高维属性空间转变为一个低维“子空间 subspace”，特征选择从给定的特征集合中选取出相关特征子集，使构造出来的模型更好。

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \xrightarrow{\text{特征选择}} \begin{bmatrix} x_1 \\ x_1 \\ \\ x_{d'} \end{bmatrix}$$

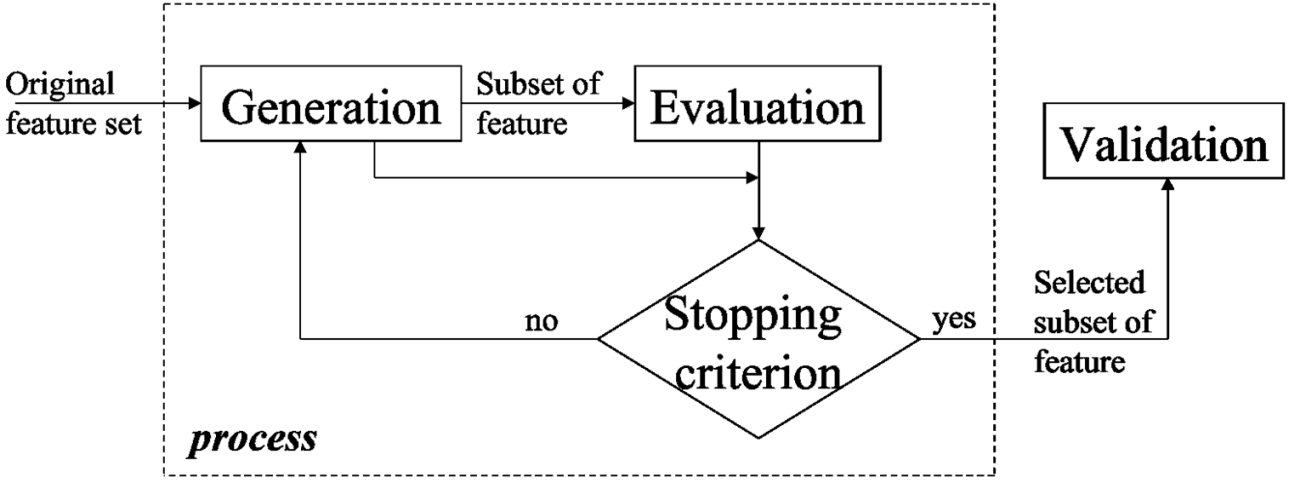
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \xrightarrow{\text{特征抽取}} \begin{bmatrix} y_1 \\ y_2 \\ \\ y_{d'} \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \right)$$

特征选择

- 目的

提升后续任务效果
降低后续任务的开销
对数据能有更进一步地认识

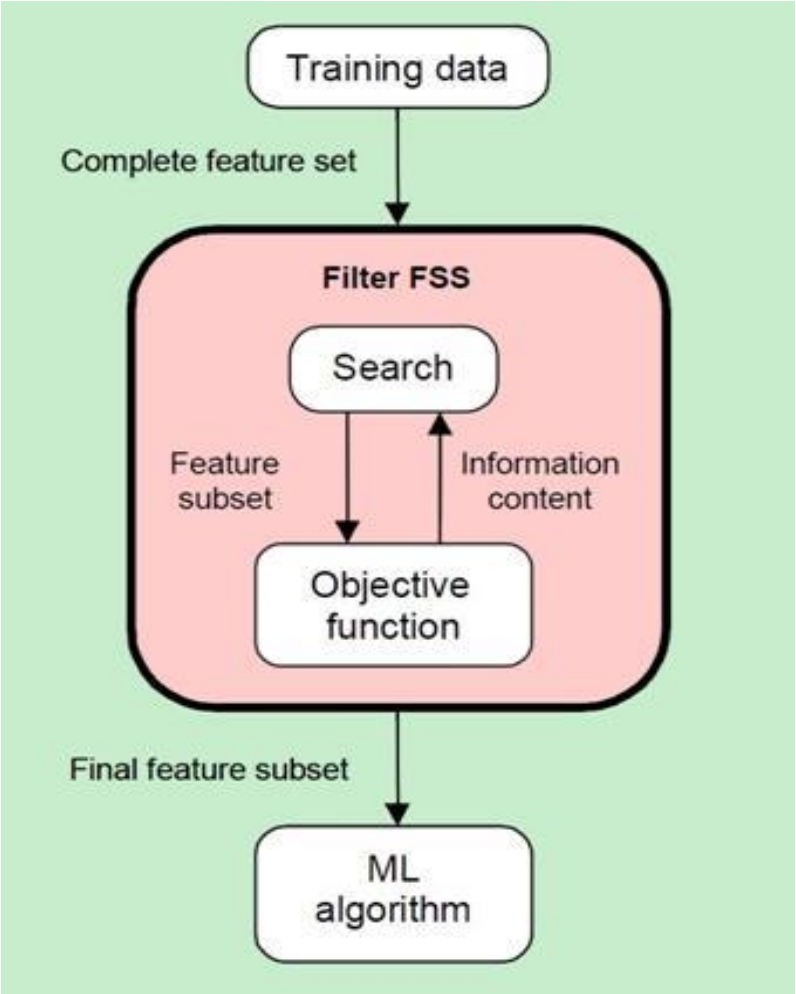
- 框架



过滤式特征选择

- 特点

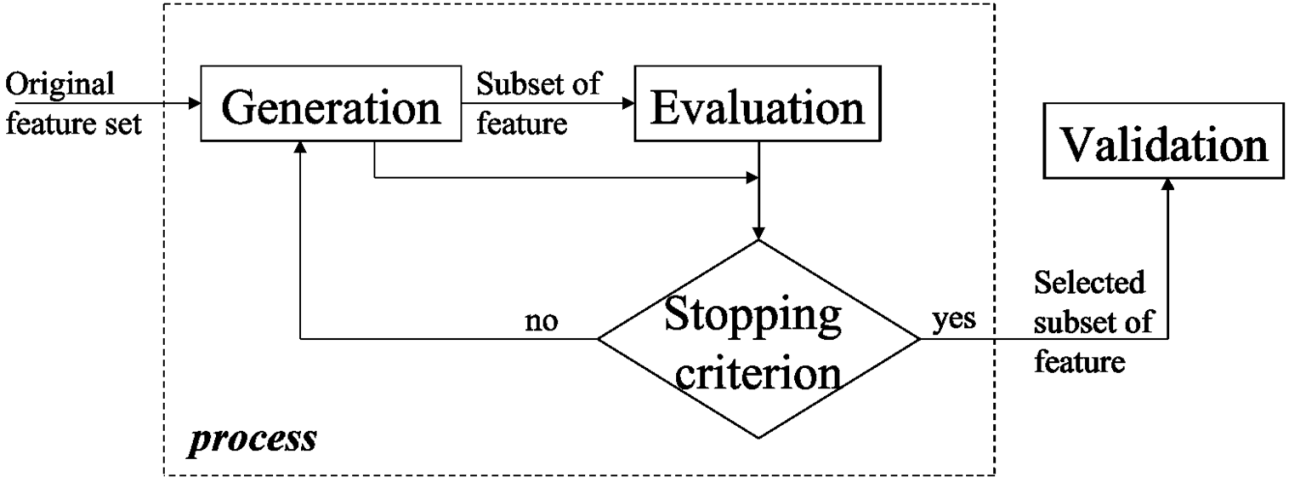
过滤式方法先对数据集进行特征选择，然后再训练学习器，特征选择过程与后续学习器无关.这相当于先用特征选择过程对初始特征进行“过滤”，再用过滤后的特征来训练模型



过滤式特征选择

- 基于单变量评价的过滤式特征选择

前提：为每个特征设定一个分数，这个分数可以是其区分数据的能力等
框架实现：
特征子集的评估 Evaluation = 特征子集中每个特征的分数相加
特征子集的生成 Generation = 每次挑选最高的特征加入到当前的特征子集中
停止策略 Stopping criterion = 特征个数或欲加入子集的特征的分数小于某个阈值



过滤式特征选择

- 基于单变量评价的过滤式特征选择

单变量评价标准示例： Relief (Relevant Features)

给定训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

对每个示例 x_i

“猜中近邻”(near-hit) $= \boldsymbol{x}_i$ 的同类样本 中最近邻 $\boldsymbol{x}_{i,\mathrm{nh}}$

“猜错近邻”(near-miss) $= x_i$ 的异类样本中寻找其最近邻 $x_{i,\mathrm{nm}}$

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

若属性 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

若属性 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

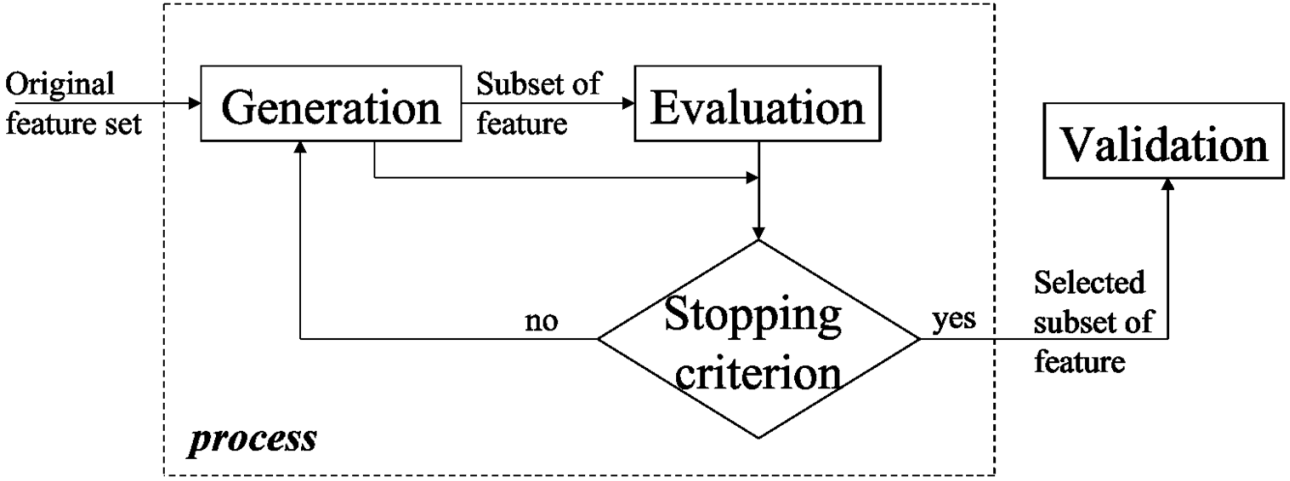
其他单变量评价标准：

1. 相关系数
2. 假设检验

过滤式特征选择

- 基于多变量评价的过滤式特征选择

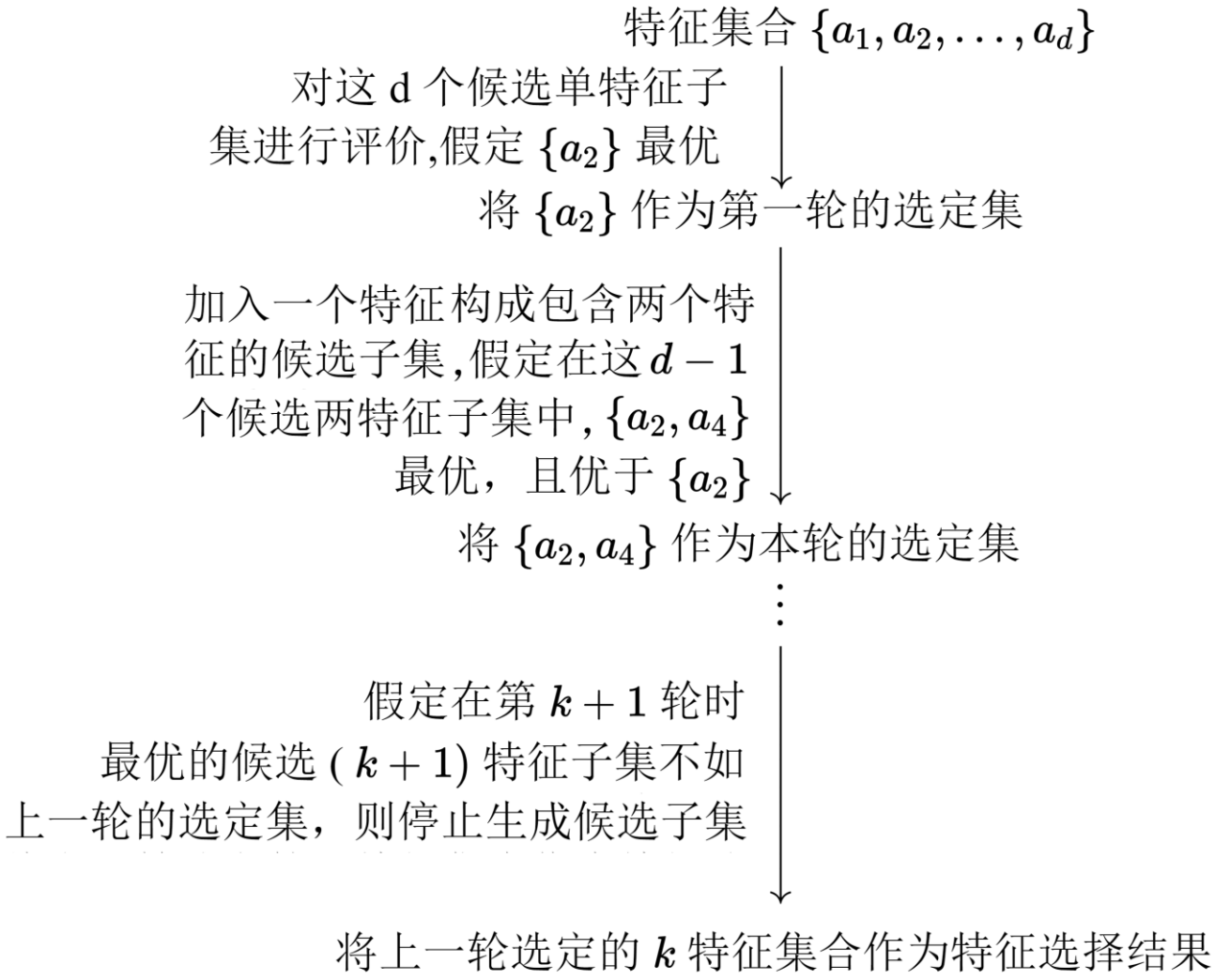
前提：拥有为一个特征子集设定分数的方法
框架实现：
特征子集的评估 Evaluation＝多变量评价方法的输出
特征子集的生成 Generation＝ 多用启发式搜索
停止策略 Stopping criterion＝特征个数或其他



过滤式特征选择

- 基于多变量评价的过滤式特征选择

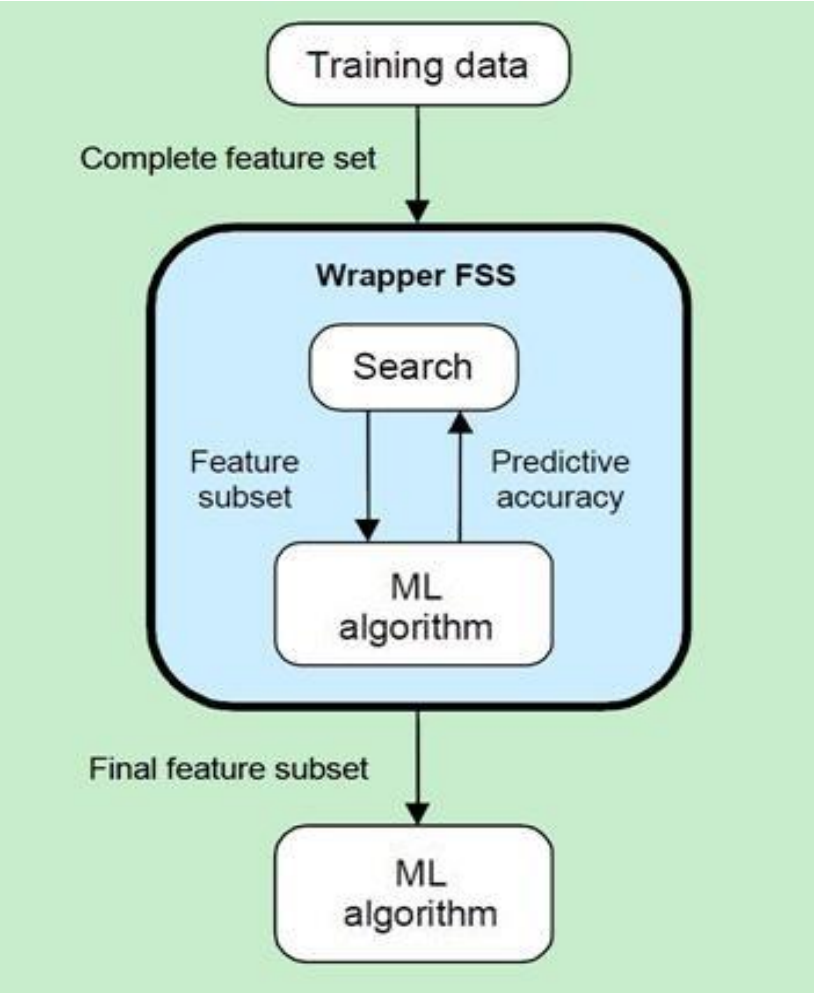
示例：基于序列前向选择 (SFS , Sequential Forward Selection)



包裹式特征选择

- 特点

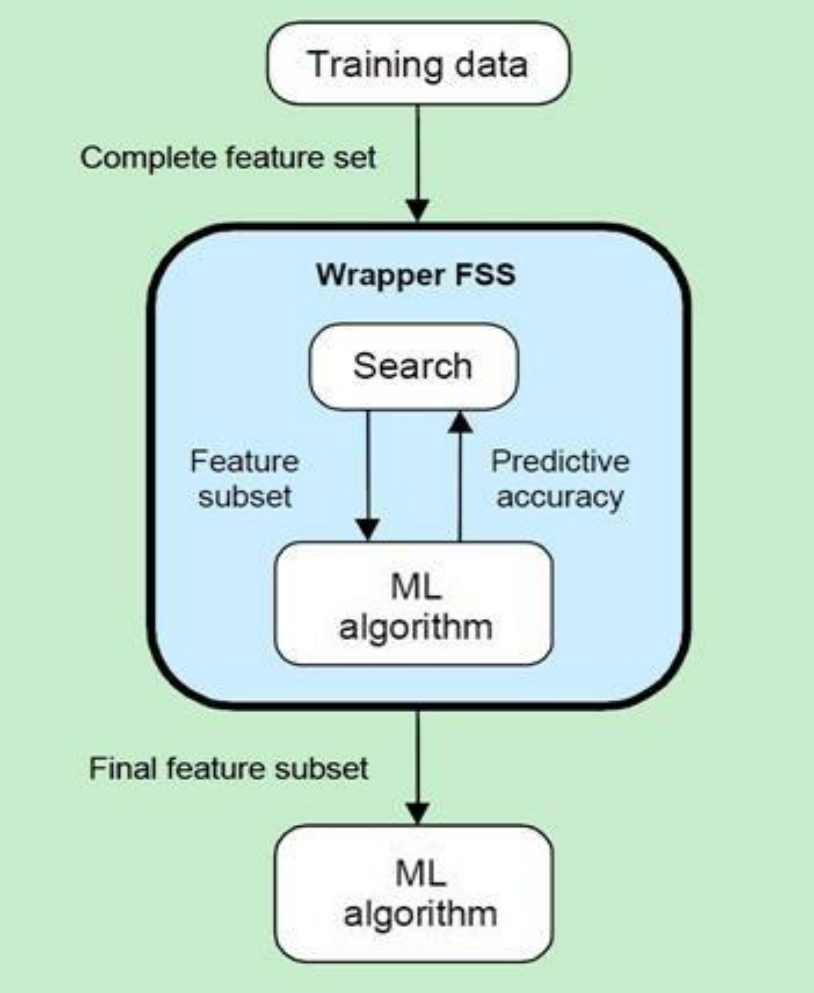
与过滤式特征选择不考虑后续学习器不同，包裹式特征选择直接把最终将要使用的学习器的性能作为特征子集的评价准则.换言之，包裹式特征选择的目的就是为给定学习器选择最有利于其性能、“量身定做”的特征子集.



包裹式特征选择

- 实现

类似于基于多变量评价的过滤式特征选择。但特征子集的评估 Evaluation将改为：基于某个特征子集学习后续任务的模型，即后续算法学习的模型的性能作为特征子集的评价准则

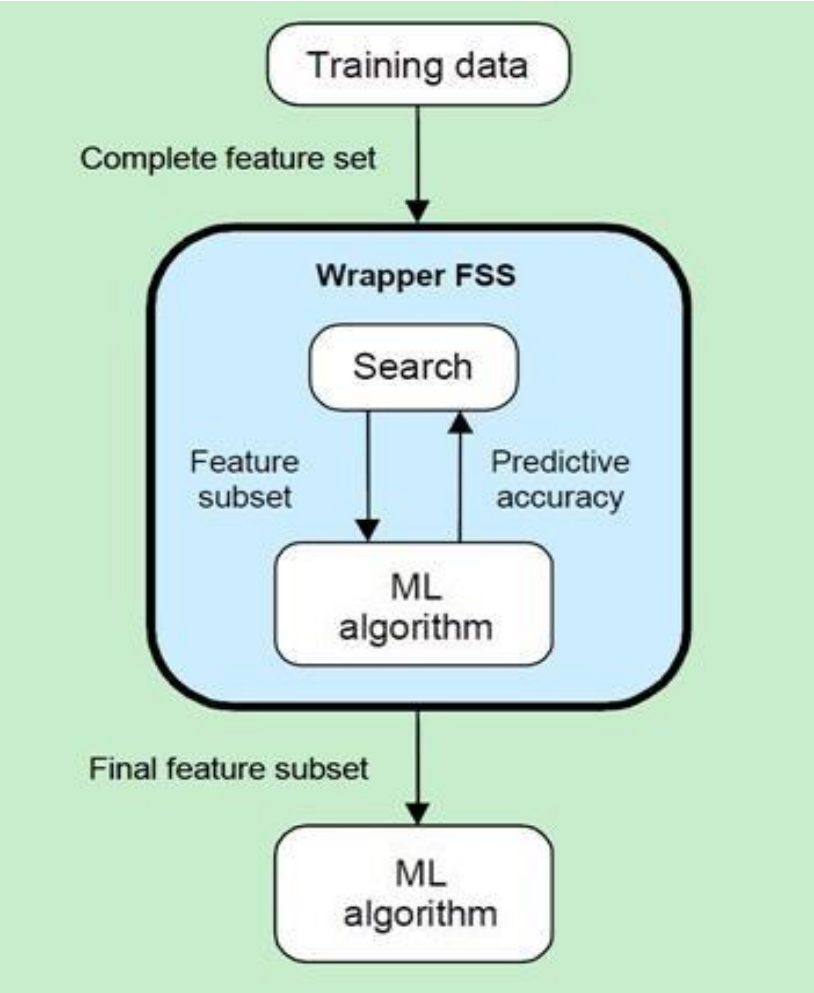


包裹式特征选择

- 实现

特征子集的评估 Evaluation：基于某个特征子集学习后续任务的模型，此模型的性能作为特征子集的评价准则

特征子集的生成 Generation：一般用启发式搜索或随机搜索



包裹式特征选择

- 示例

Las Vegas Wrapper (LVW)

给定: 学习算法$\mathfrak{L}$; 数据集 D; 特征集 A; 停止条件控制参数T

求: 最优子集 A^*

$E=\infty; d=|A|; A^*=A;$

$t=0$

while $t<T$ do 随机产生特征子集 A';

随机产生特征子集 A';

$d'=|A'|$

$E'=\text{CrossValidation}\left(\mathfrak{L}\left(D^{A'}\right)\right);$

if $(E'<E)\vee((E'=E)\wedge(d'<d))$ then

$t=0$

$E=E'$

$d=d'$

$A^*=A'$

else

$t=t+1$

end if

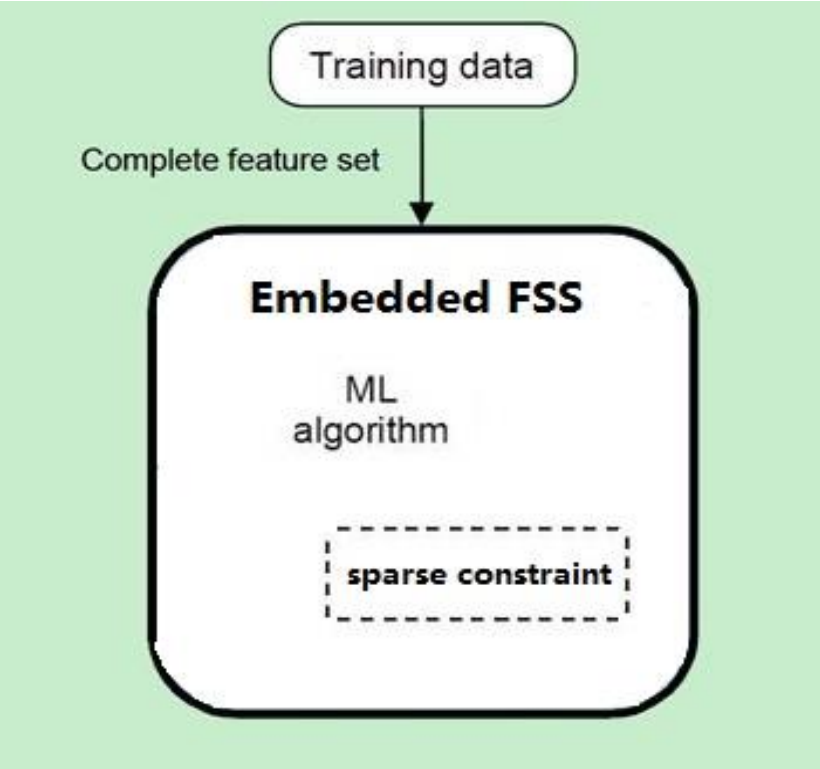
end while

嵌入式特征选择

- 特点

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，即在学习器训练过程中自动地进行了特征选择.

示例包括决策树、L1正则化项等等

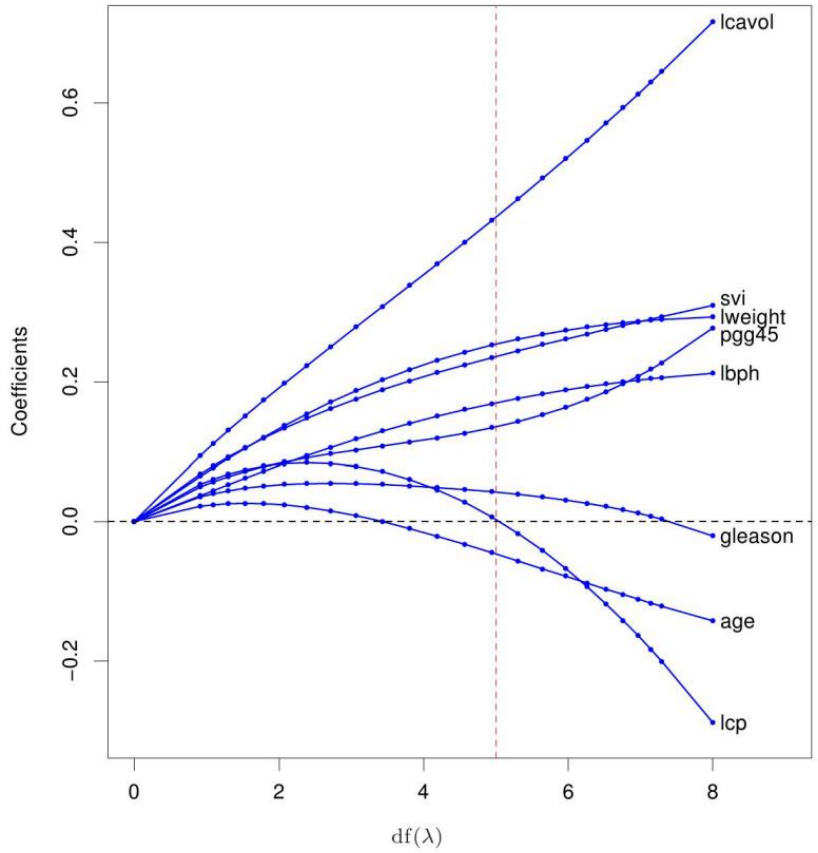


嵌入式特征选择

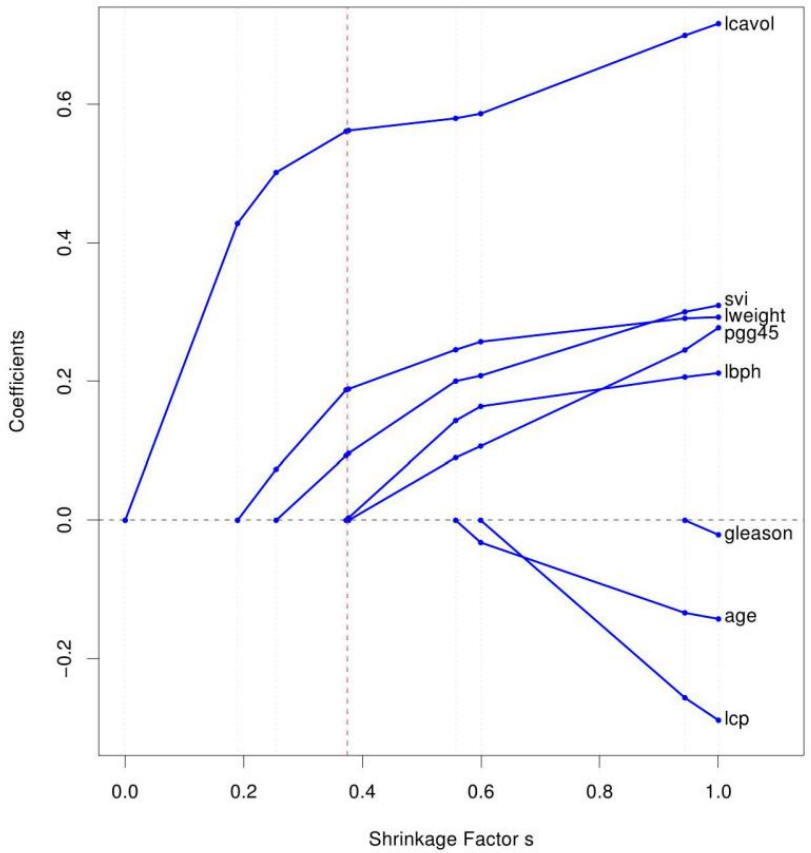
- LASSO

$$\min_{\boldsymbol{w},b} \| \boldsymbol{y} - \mathbf{X}\boldsymbol{w} \|_2^2 + \lambda \| \boldsymbol{w} \|_1$$

$$\| \boldsymbol{w} \|_1 = \sum_{j=1}^{d} |w_j|$$



(a) $||w||_2$ penalty



(b) $||w||_1$ penalty

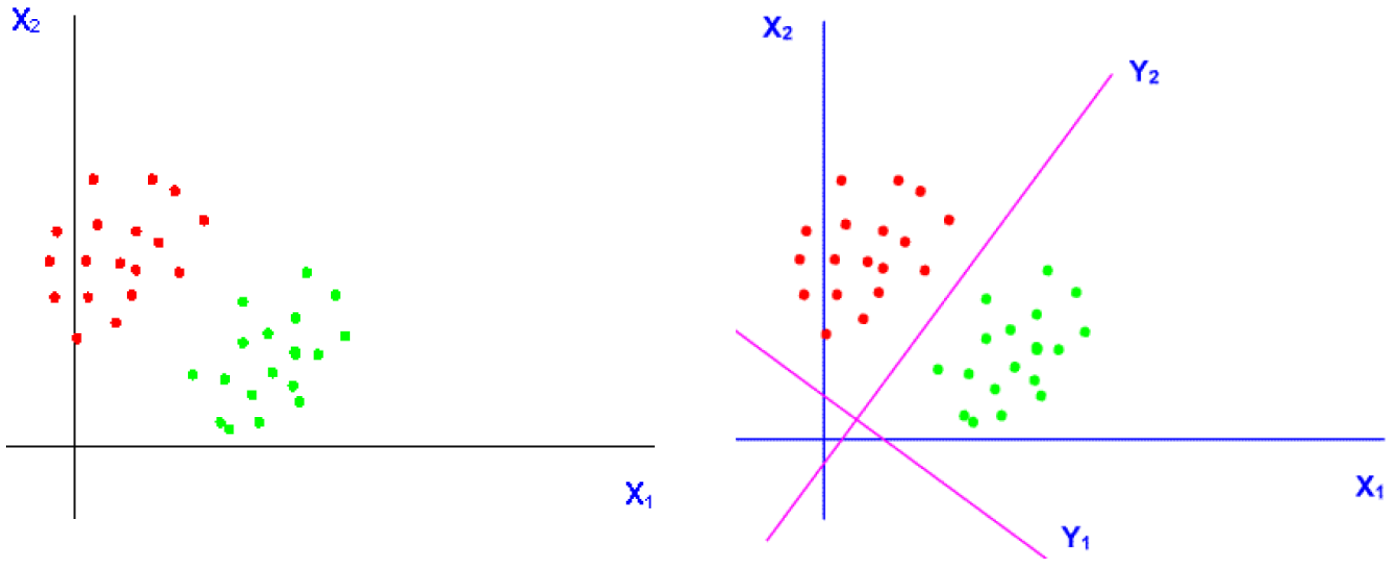
特征转换

- 动机

人们观测或收集到的数据样本虽是高维的，但与学习任务密切相关的也许仅是某个低维分布，即高维空间中的一个低维“嵌入” embedding

- 想法

通过某种数学变换将原始高维属性空间转变为一个低维“子空间”(subspace),在这个低维嵌入子空间中更容易进行学习.



特征转换

- 线性变换

"线性代数的本质 - 03 - 矩阵与线性变换"

给定 d 维空间中的样本$\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) \in \mathbb{R}^{d \times m}$
希望将它们变换为d'维空间的样本$\boldsymbol{Z} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m) \in \mathbb{R}^{d' \times m}$

一般而言，$d' < d$，常用的变换有线性变换，记$\mathbf{W} \in \mathbb{R}^{d \times d'}$为变换矩阵，包含$d'$个$d$维基向量：$\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{d'}\}$，其对每个样本起到的线性变换作用可以写成$\boldsymbol{z}_i = \boldsymbol{W}^T \boldsymbol{x}_i$，即对整个数据集作了线性变换 $\boldsymbol{Z} = \boldsymbol{W}^T \boldsymbol{X}$

- 正交线性变换

若对任意两个基向量都满足$\boldsymbol{w}_k \perp \boldsymbol{w}_j$,则新坐标系是一个正交坐标系，此时的线性换称为正交线性变换

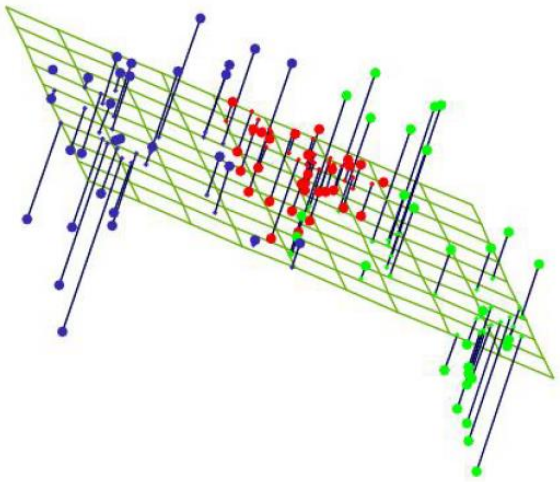
主成分分析

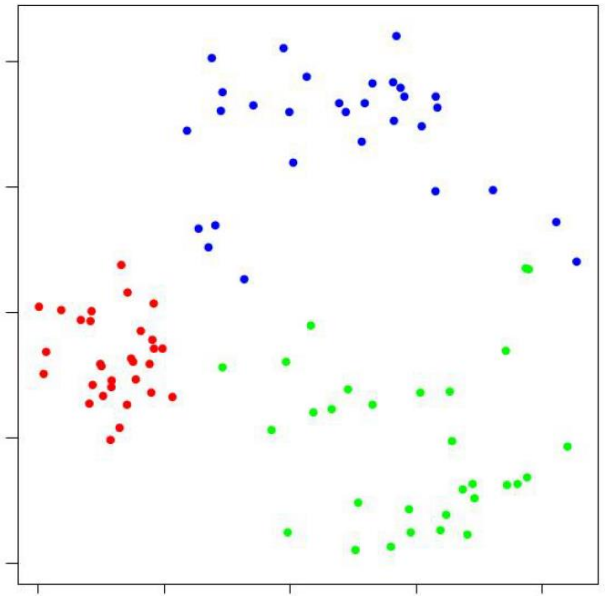
- 动机

对于正交属性空间中的样本点，若要用一个超平面(直线的高维推广)对所有样本进行恰当的表达，则这样的超平面，那么它大概应具有这样的性质：

>最近重构性：样本点到这个超平面的距离都足够近

> 最大可分性：样本点在这个超平面上的投影能尽可能分开



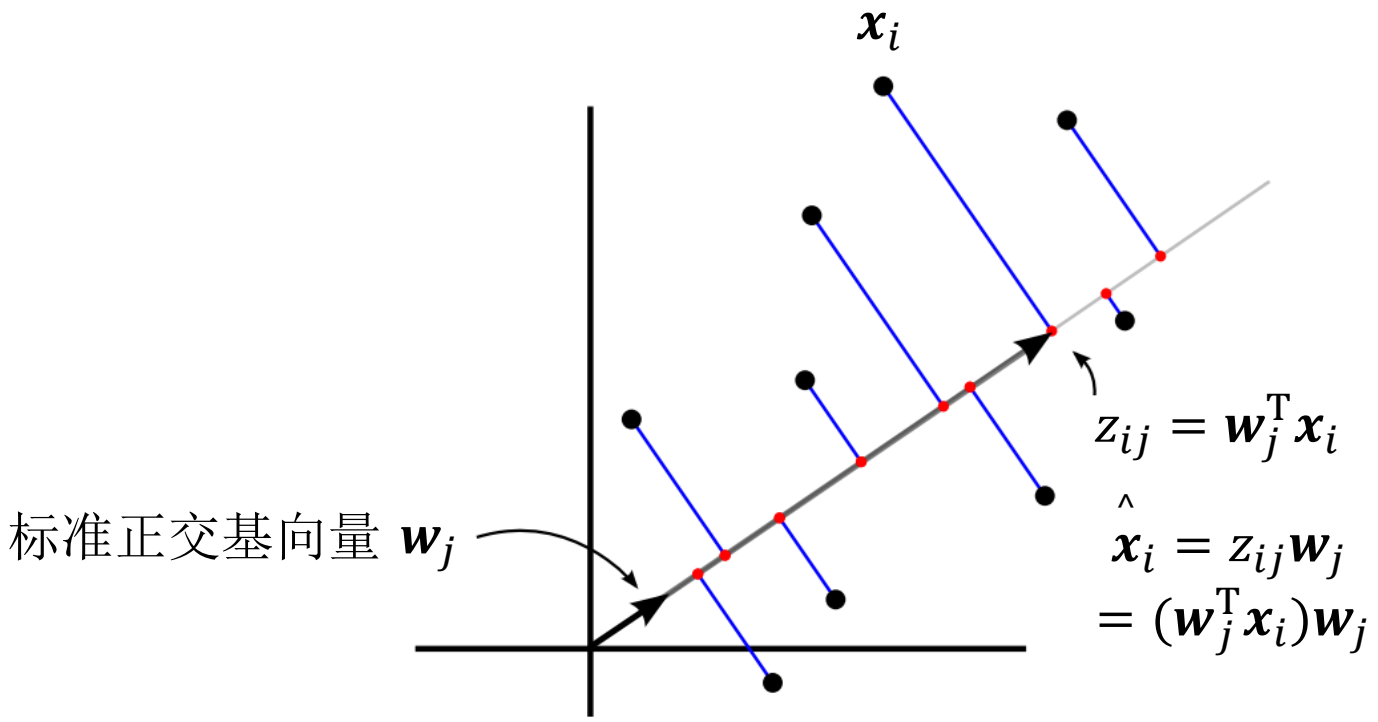


主成分分析

- 最近重构性

假定 $\sum_i \boldsymbol{x}_i = \mathbf{0}$

$$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j$$



主成分分析

- 最近重构性

> 重构错误

$$\sum_{i=1}^{m} \| \sum_{j=1}^{d'} z_{ij}\boldsymbol{w}_j - \boldsymbol{x}_i \|_2^2 = \sum_{i=1}^{m} \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_i - 2\sum_{i=1}^{m} \boldsymbol{z}_i^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i + \text{const}$$

$$\propto -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}(\sum_{i=1}^{m} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}})\mathbf{W})$$

> 最小重构错误

$$\min_{\mathbf{W}} -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$

$$\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$$

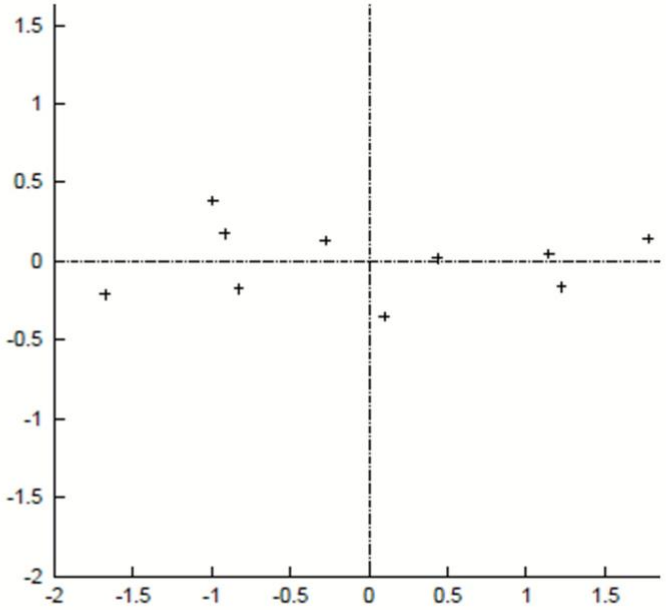
主成分分析

- 最大可
> 方差的意
在信号处理

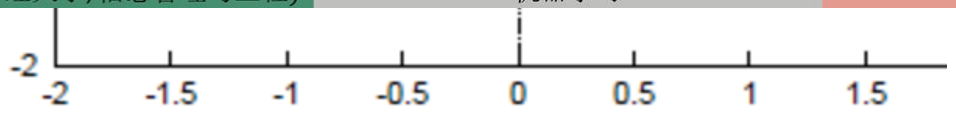
主成分分析

暂停 00:00:00 选择区域 音频 录制指针

- 最大可分性
> 方差的意义
在信号处理中认为信号具有较大的方差，噪声有较小的方差



涂文婷(上海财经大学, 信息管理与工程) 机器学习

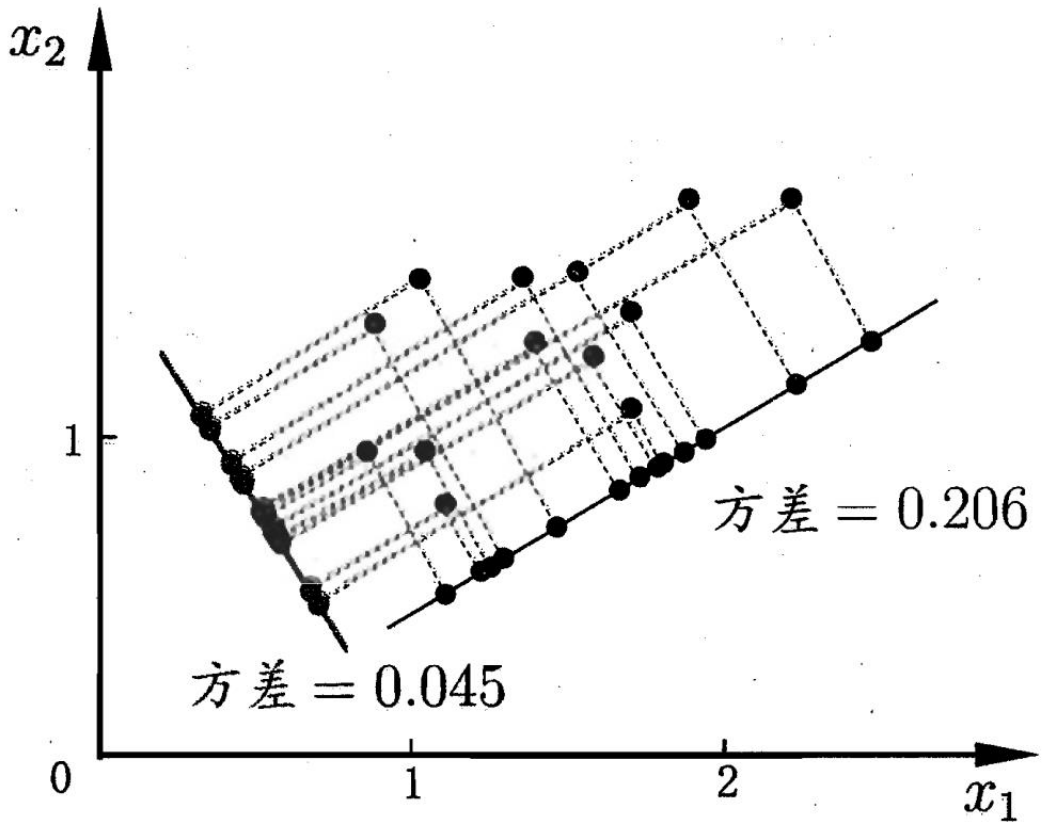


主成分分析

- 最大可分性

> 最大方差思想

使所有样本的投影尽可能分开对应于最大化投影点的方差



主成分分析

- 最大可分性

> 最大方差角度

投影后样本点的方差是 $\sum_i \mathbf{W}^{\mathrm{T}}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\mathbf{W}$

得优化任务为

$$\max_{\mathbf{W}} \quad \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W})$$
$$\text{s.t.} \quad \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$$

主成分分析

- 解

$$\min_{\mathbf{W}} -\mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}) \quad \text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I} \qquad \text{或} \qquad \max_{\mathbf{W}} \mathrm{tr}(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W}) \quad \text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I}$$

使用拉格朗日乘子法可得$\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{W} = \lambda\mathbf{W}$，即只需对协方差矩阵$\mathbf{X}\mathbf{X}^{\mathrm{T}}$进行特征值分解，将最大的$d'$个特征值对应的特征向量构成$\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_{d'}\}$即可

- 算法

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$
低维空间维数d'. 过程：
1: 对所有样本进行中心化: $\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$
2: 计算样本的协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$;
3: 对协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 做特征值分解;
4: 取最大的 d' 个特征值所对应的特征向量 $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'}$
输出： 投影矩阵 $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$