



机器学习（研究生）

特征选择、稀疏学习与 半监督学习

Feature Selection, Sparse Learning & Semi-Supervised Learning

赵凯 上海大学

<https://kaizhao.net/teaching/shu-ml-grad>

本讲内容索引

01 子集搜索与评价

相关、无关与冗余特征 · 子集搜索 · 子集评价 · 三类方法

02 过滤式与包裹式选择

Relief · Relief-F · LVW

03 嵌入式选择与 L_1 正则化

岭回归与 LASSO · 为什么 L_1 得到稀疏解 · 近端梯度下降

04 稀疏表示与字典学习

稀疏表示 · 字典学习的目标 · 交替优化

05 未标记样本

监督信息的类型 · 主动学习与半监督学习 · 聚类假设与流形假设

06 生成式方法

高斯混合模型 · 有标记与未标记的对数似然 · EM 求解

07 半监督支持向量机

低密度分隔 · TSVM 的目标 · 伪标记与局部搜索

08 半监督聚类

必连与勿连约束 · 约束 k 均值 · 约束种子 k 均值

09 小结

本讲要点 · 中英文术语对照 · 课后作业

属性太多，标记太少

第五讲用降维应对高维；本讲从另外两个方向处理“数据不够好用”

属性太多

有经验的人只需看看根蒂、听听敲声就知道是否好瓜：许多属性与任务无关

⇒ **特征选择**：从原有特征中选出有用的子集；
稀疏学习：让模型只用少数特征或少数“字典元素”

标记太少

瓜农只指认了几个好瓜与坏瓜，地里还有大量没有标记的瓜

⇒ **半监督学习**：不依赖外界交互，同时利用少量有标记样本与大量未标记样本

问题 哪些特征有用？怎样判断一个特征子集的好坏？未标记样本没有标记，又怎能帮助学习？

记号 周志华《机器学习》第 11 章 11.1 节 p.247；第 13 章 13.1 节 p.293。

本讲内容

Part 1 特征选择与稀疏学习

选出有用的特征，学出稀疏的表示

子集搜索与评价

过滤式与包裹式选择

嵌入式选择与 L_1 正则化

稀疏表示与字典学习

Part 2 半监督学习

利用未标记样本

未标记样本

生成式方法

半监督支持向量机

半监督聚类

符号约定

沿用前几讲的记号；新增与特征选择、半监督学习有关的记号

数据	$D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^d$, x_{ij} 为样本 $\mathbf{x}_i$ 在第 j 个属性上的取值
特征子集	候选特征集合 $\{a_1, \dots, a_d\}$; 特征子集记为 A
稀疏性	$\ \mathbf{w}\ _1 = \sum_j w_j $, $\ \mathbf{w}\ _2^2 = \sum_j w_j^2$; $\mathbf{w}$ 的大多数分量为 0 称为稀疏
半监督	有标记样本 $D_l = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_l, y_l)\}$; 未标记样本 $D_u = \{\mathbf{x}_{l+1}, \dots, \mathbf{x}_{l+u}\}$, $l \ll u$, $l + u = m$

记号 周志华《机器学习》第 11 章用上标 x_i^j 表示属性，本课程统一写作 x_{ij} （第一个下标为样本，第二个为属性）。

01

子集搜索与评价

Subset Search & Evaluation

相关、无关与冗余特征 · 子集搜索 · 子集评价 · 三类方法

特征选择

从给定的特征集合中选出相关特征子集的过程，称为特征选择（feature selection）

相关特征	对当前学习任务有用
无关特征	对当前任务没什么用
冗余特征	信息可由其他特征推演出来，如“底面积”之于“底面长”“底面宽”

为什么要做特征选择

1. 减轻维数灾难：只在部分特征上建模（与第五讲的降维动机相似）
2. 去除无关特征，降低学习任务的难度

要点 “无关”是相对当前任务而言的；特征选择须确保**不丢失重要特征**。

记号 冗余特征若恰好对应完成任务所需的“中间概念”，则是有益的；本讲暂不涉及冗余特征。周志华《机器学习》第 11 章 11.1 节，pp.247–248。

子集搜索

遍历所有子集会遭遇组合爆炸：产生候选子集 → 评价 → 产生下一个候选子集

前向搜索	从空集开始，每轮加入一个使选定集最优的特征，直到不再变好
后向搜索	从完整特征集开始，每轮去掉一个无关特征
双向搜索	每轮既加入相关特征、又去掉无关特征

前向搜索的例子：

第 1 轮： $\{a_2\}$ 最优



第 2 轮： $\{a_2, a_4\}$ 最优



第 $k + 1$ 轮不如
第 k 轮：停止

要点 这些策略都是贪心的：只考虑使本轮选定集最优；贪心搜索不保证找到全局最优的子集。

记号 周志华《机器学习》第 11 章 11.1 节，p.248；子集搜索亦称子集“生成与搜索”。

子集评价

特征子集 A 把数据集 D 划分为 V 个子集 $D^1, \dots, D^V$ (每个子集中样本在 A 上取值相同)

$$\text{Gain}(A) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v), \quad \text{Ent}(D) = - \sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

- 信息增益越大，特征子集 A 包含的有助于分类的信息越多（第三讲决策树的划分准则）
- 一般地： A 确定的划分与标记确定的真实划分差异越小， A 越好

类别	特征选择与学习器训练的关系	代表方法
过滤式 (filter)	先选特征，再训练；与学习器无关	Relief
包裹式 (wrapper)	以最终学习器的性能作为评价准则	LVW
嵌入式 (embedding)	在训练过程中自动完成选择	L_1 正则化

记号 p_k : 第 k 类样本所占的比例；搜索与评价结合即得特征选择方法，如前向搜索 + 信息熵即为决策树。周志华《机器学习》第 11 章 11.1 节，式 (11.1)–(11.2)，pp.248–249。

02

过滤式与包裹式选择

Filter & Wrapper Methods

Relief · Relief-F · LVW

过滤式选择：Relief

Relief (Relevant Features): 用一个“相关统计量”向量度量各特征的重要性

对每个样本 x_i : 在同类样本中找最近邻 $x_{i,nh}$ (猜中近邻, near-hit), 在异类样本中找最近邻 $x_{i,nm}$ (猜错近邻, near-miss)

$$\delta_j = \sum_i \left[-\text{diff}(x_{ij}, x_{i,nh,j})^2 + \text{diff}(x_{ij}, x_{i,nm,j})^2 \right]$$

离散属性 $\text{diff} = 0$ (取值相同) 或 1

连续属性 $\text{diff} = |x_{aj} - x_{bj}|$, 取值已规范化到 $[0, 1]$

- 与猜中近邻的距离小于与猜错近邻的距离: 属性 j 有益, δ_j 增大
- 选 δ_j 大于阈值 τ 的特征, 或最大的 k 个

要点 只需训练一次、不训练学习器: 时间开销随采样次数与特征数线性增长 (另需近邻搜索), 是运行效率很高的过滤式方法。

记号 Kira & Rendell (1992)。 i 取遍所用的采样样本; Relief 只需在数据集的采样上估计相关统计量。周志华《机器学习》第 11 章 11.2 节, 式 (11.3), pp.249–250。

Relief-F：多分类

Relief 为二分类设计；Relief-F 在每个异类中各找一个猜错近邻

若 x_i 属于第 y_i 类：在该类中找猜中近邻 $x_{i,nh}$ ；在每个第 $l \neq y_i$ 类中找猜错近邻 $x_{i,l,nm}$

$$\delta_j = \sum_i \left[-\text{diff}(x_{ij}, x_{i,nh,j})^2 + \sum_{l \neq y_i} p_l \text{diff}(x_{ij}, x_{i,l,nm,j})^2 \right]$$

- p_l : 第 l 类样本在 D 中所占的比例，按类别比例加权
- 对不同样本得到的估计结果平均，即得各属性的相关统计量分量

问题 过滤式方法不考虑后续学习器。能否直接用最终学习器的性能来选特征？

记号 Kononenko (1994)。Relief 的计算已隐然具有距离度量学习的意味（第五讲的度量学习）。周志华《机器学习》第 11 章 11.2 节，式 (11.4)，p.250。

包裹式选择：LVW

包裹式选择：把最终学习器的性能作为特征子集的评价准则，为学习器“量身定做”

Require: 数据集 D ；特征集 A ；学习算法 $\mathcal{L}$ ；控制参数 T

```
1:  $E \leftarrow \infty$ ,  $d \leftarrow |A|$ ,  $A^* \leftarrow A$ ,  $t \leftarrow 0$ 
2: while  $t < T$  do
3:   随机产生特征子集  $A'$ ,  $d' \leftarrow |A'|$ 
4:    $E' \leftarrow$  在  $A'$  上交叉验证估计  $\mathcal{L}$  的误差
5:   if  $E' < E$ , 或  $E' = E$  且  $d' < d$  then
6:      $t \leftarrow 0$ ,  $E \leftarrow E'$ ,  $d \leftarrow d'$ ,  $A^* \leftarrow A'$ 
7:   else
8:      $t \leftarrow t + 1$ 
9:   end if
10: end while
```

Ensure: 特征子集 A^*

- 拉斯维加斯方法框架下用随机策略搜索子集
- 误差更小，或误差相同而特征更少，则保留
- 连续 T 轮未更新则停止

最终性能通常优于过滤式，但需多次训练学习器，计算开销大得多。

记号 LVW: Las Vegas Wrapper (Liu & Setiono, 1996)。特征数很多、 T 很大时可能长时间达不到停止条件。周志华《机器学习》第 11 章 11.3 节，图 11.1，pp.250–252。

03

嵌入式选择与 L_1 正则化

Embedded Selection & L_1 Regularization

岭回归与 LASSO · 为什么 L_1 得到稀疏解 · 近端梯度下降

嵌入式选择：岭回归与 LASSO

把特征选择融入学习器的训练：在同一个优化过程中完成

以平方误差为损失的线性回归，特征多、样本少时易过拟合，加入正则化项：

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

L_2 正则化：岭回归（ridge regression）

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

L_1 正则化：LASSO

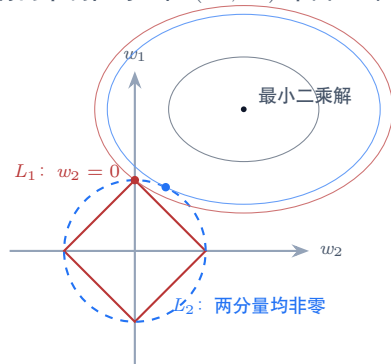
- 两者都有助于降低过拟合风险（第二讲的正则化）
- L_1 范数还带来一个额外的好处：更易于获得稀疏解， $\mathbf{w}$ 的非零分量更少

要点 $\mathbf{w}$ 稀疏 $\Rightarrow$ 只有非零分量对应的特征出现在模型中： L_1 正则化的学习就是嵌入式特征选择。

记号 $\lambda > 0$ ：正则化参数；LASSO：Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996)。施加稀疏约束最自然的是 L_0 范数（非零分量个数，并非真正的范数），但它不连续、难以优化，常用 L_1 近似。周志华《机器学习》第 11 章 11.4 节，式 (11.5)–(11.7)，p.252；Bishop (2006) 第 3.1.4 节。

为什么 L_1 正则化得到稀疏解

x 只有两个属性时，在 (w_1, w_2) 平面上画出平方误差项与正则化项的等值线



- 解在约束域上取：平方误差等值线由内向外扩大，最先接触约束域的点
- L_1 等值线是菱形，交点常落在坐标轴上： w_1 或 w_2 为 0
- L_2 等值线是圆，交点常落在某个象限中：两个分量都非零

问题 $\|w\|_1$ 在 0 处不可导，不能直接用梯度下降。LASSO 如何求解？

记号 示意图，参照周志华《机器学习》第 11 章 11.4 节，图 11.2, p.253; Bishop (2006) 第 3.1.4 节，图 3.4。实线菱形为 L_1 范数等值线，虚线圆为 L_2 范数等值线；红、蓝椭圆分别为与菱形、圆接触的平方误差等值线。

近端梯度下降

近端梯度下降 (Proximal Gradient Descent, PGD): 求解 $\min_x f(x) + \lambda \|x\|_1$, ∇f 满足 L -Lipschitz 条件

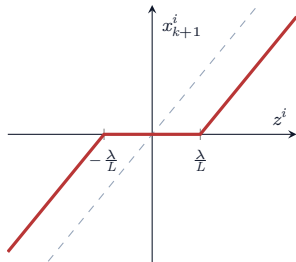
由 L -光滑性, f 在 x_k 附近有二次上界, 梯度下降一步等价于最小化

$$\hat{f}(x) = \frac{L}{2} \left\| x - \left(x_k - \frac{1}{L} \nabla f(x_k) \right) \right\|_2^2 + \text{const}$$

加上 L_1 项: 先算 $z = x_k - \frac{1}{L} \nabla f(x_k)$, 再求

$$x_{k+1} = \arg \min_x \frac{L}{2} \|x - z\|_2^2 + \lambda \|x\|_1$$

各分量互不影响, 有闭式解 (右图)



$$x_{k+1}^i = \begin{cases} z^i - \lambda/L, & z^i > \lambda/L \\ 0, & |z^i| \leq \lambda/L \\ z^i + \lambda/L, & z^i < -\lambda/L \end{cases}$$

记号 L -Lipschitz 条件: $\|\nabla f(x') - \nabla f(x)\|_2 \leq L\|x' - x\|_2$ (书中式 (11.9) 两边带平方, 右端应为 L^2); 由此 $f(x) \leq f(x_k) + \nabla f(x_k)^\top (x - x_k) + \frac{L}{2} \|x - x_k\|_2^2$. $|z^i| \leq \lambda/L$ 的分量被置零, 这就是稀疏的来源。周志华《机器学习》第 11 章 11.4 节, 式 (11.8)–(11.14), pp.253–254。

04

稀疏表示与字典学习

Sparse Representation & Dictionary Learning

稀疏表示 · 字典学习的目标 · 交替优化

稀疏表示

把数据集 D 看成矩阵：每行一个样本，每列一个特征

特征选择	许多列与任务无关，整列去除
稀疏表示	矩阵中有很多零元素，但 不成整行整列

例：文档分类中每个文档是一行、每个字是一列。仅《现代汉语常用字表》就有 3500 列，而一个文档只用到其中一小部分字

稀疏表示的好处

- 常使问题更接近**线性可分**（线性支持向量机在文本数据上表现好的原因之一）
- 稀疏矩阵有高效的存储方法

问题 给定的数据集是稠密的，能否把它转化为“恰当稀疏”的表示？这需要先学出一个“字典”。

记号 稀疏表示 (sparse representation)；书中以汉字为例：基于《现代汉语常用字表》(3500 字) 可能恰当稀疏，基于《康熙字典》(47035 字) 可能过度稀疏。周志华《机器学习》第 11 章 11.5 节，pp.254–255。

字典学习

为稠密样本找到合适的字典，并将样本转化为稀疏表示：字典学习（dictionary learning），亦称稀疏编码

$$\min_{\mathbf{B}, \{\alpha_i\}_{i=1}^m} \sum_{i=1}^m \|\mathbf{x}_i - \mathbf{B}\alpha_i\|_2^2 + \lambda \sum_{i=1}^m \|\alpha_i\|_1, \quad \text{s.t. } \|\mathbf{b}_j\|_2 \leq 1$$

$\mathbf{B} \in \mathbb{R}^{d \times k}$ 字典矩阵； k 为词汇量

$\alpha_i \in \mathbb{R}^k$ 样本 $\mathbf{x}_i$ 的稀疏表示

第一项 由 α_i 很好地重构 $\mathbf{x}_i$

第二项 希望 α_i 尽量稀疏

变量交替优化

1. 固定 $\mathbf{B}$ ，对每个样本求 α_i ：

$$\min_{\alpha_i} \|\mathbf{x}_i - \mathbf{B}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1$$

与 LASSO 相同，可用近端梯度下降

2. 固定 α_i ，更新字典： $\min_{\mathbf{B}} \|\mathbf{X} - \mathbf{B}\mathbf{A}\|_F^2$

记号 $\mathbf{b}_j$ ：字典的第 j 列。不约束其范数时，令 $\mathbf{B} \rightarrow c\mathbf{B}$ 、 $\alpha_i \rightarrow \alpha_i/c$ 可使目标无限下降，书中式 (11.15) 未写出此约束。
 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m) \in \mathbb{R}^{d \times m}$ ， $\mathbf{A} = (\alpha_1, \dots, \alpha_m)$ ； $\|\cdot\|_F$ ：Frobenius 范数。周志华《机器学习》第 11 章 11.5 节，式 (11.15)–(11.17)，pp.255–256。

KSVD：逐列更新字典

KSVD (Aharon 等, 2006): 其余各列固定, 同时更新字典的第 i 列 b_i 与 A 的第 i 行 α^i

$$\|X - BA\|_F^2 = \left\| \left(X - \sum_{j \neq i} b_j \alpha^j \right) - b_i \alpha^i \right\|_F^2 = \|E_i - b_i \alpha^i\|_F^2$$

- E_i : 去掉第 i 个原子后的残差, 固定; $b_i \alpha^i$ 是秩一矩阵
- 只看 α^i 非零的那些列 (使用了 b_i 的样本), 得受限残差 E_i^R
- 对 E_i^R 做奇异值分解, 取最大奇异值对应的左、右奇异向量: 分别更新 b_i (单位长度) 与 α^i 的非零元素
- 只更新非零位置, 第一步得到的稀疏模式保持不变

问题 前四节处理的是“属性太多”。另一种常见的困难是标记太少: 大量样本没有标记, 能否利用它们?

记号 秩一近似由最大奇异值给出 (Eckart-Young 定理, 第五讲的 SVD)。初始化字典后反复迭代两步, 即得 B 与 α_i ; 词汇量 k 控制字典规模与稀疏程度。周志华《机器学习》第 11 章 11.5 节, 式 (11.18), pp.256–257; Aharon, Elad & Bruckstein (2006), *IEEE Trans. Signal Processing*, 54(11)。

05

未标记样本

Unlabeled Samples

监督信息的类型 · 主动学习与半监督学习 · 聚类假设与流形假设

监督信息的类型

按训练样本的标记情况，学习可分为以下几类

学习方式	训练数据	例
全监督学习	所有样本都有准确的标记	前几讲的分类与回归
弱监督学习	标记不完全（只有部分样本有标记）、不确切（粗粒度）或不精确（有噪声）	只标图像级类别，不标位置
半监督学习	少量有标记样本 + 大量未标记样本；属于不完全监督	本节
无监督学习	所有样本都没有标记	第六讲的聚类

要点 自监督学习从数据自身构造训练目标（第一讲），不使用人工标记，属于无监督学习。

记号 弱监督的三种类型见 Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53; 周志华《机器学习》第 13 章 13.1 节, pp.293–295。

主动学习与半监督学习

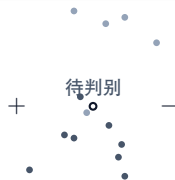
有标记样本 D_l (l 个), 未标记样本 D_u (u 个), $l \ll u$: 只用 D_l , 则 D_u 的信息被浪费

主动学习 (active learning)

用 D_l 训练模型, 挑出最有帮助的瓜询问瓜农, 加入 D_l 重新训练; 用尽量少的“查询”获得尽量好的性能, 引入了额外的专家知识

半监督学习 (semi-supervised learning)

不依赖外界交互, 自动地利用未标记样本提升学习性能



灰点为未标记样本: 待判别样本与正例“撮合”在一起

要点 未标记样本与有标记样本独立同分布, 它们所包含的数据分布信息对建模大有裨益。

记号 示意图参照周志华《机器学习》第 13 章 13.1 节, 图 13.1, p.294。

半监督学习的基本假设

要利用未标记样本，须假设数据分布与类别标记之间存在联系

聚类假设（cluster assumption）

数据存在簇结构，同一个簇的样本属于同一个类别；常用于分类任务

流形假设（manifold assumption）

数据分布在一个流形结构上，邻近的样本拥有相似的输出值；适用范围更广

	未标记样本	学习目的
纯半监督学习	不是待预测的数据	用于训练中未见的数据（开放世界）
直推学习	恰是待预测的数据	只在这些未标记样本上预测（封闭世界）

要点 两种假设的本质都是“相似的样本拥有相似的输出”。

记号 直推学习（transductive learning）。周志华《机器学习》第 13 章 13.1 节，图 13.2，pp.294–295。

06

生成式方法

Generative Methods

高斯混合模型 · 有标记与未标记的对数似然 · EM 求解

生成式方法

假设所有数据（无论是否有标记）都由同一个潜在的模型生成；未标记数据的标记看作缺失参数

假设样本由高斯混合模型生成，且每个类别对应一个高斯混合成分：

$$p(\mathbf{x}) = \sum_{i=1}^N \alpha_i p(\mathbf{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

由最大后验概率，模型 f 对 x 的预测为

$$f(\mathbf{x}) = \arg \max_{j \in \mathcal{Y}} \sum_{i=1}^N p(y = j \mid z = i, \mathbf{x}) \cdot \underbrace{p(z = i \mid \mathbf{x})}_{\text{不涉及标记}}$$

- $p(y = j \mid z = i, \mathbf{x})$ 需要标记，只能用有标记数据；第 i 类对应第 i 个成分时取 0 或 1
- $p(z = i \mid \mathbf{x})$ 是样本由第 i 个成分生成的后验（第六讲），未标记数据也可用于估计

记号 $z \in \{1, \dots, N\}$ ：样本隶属的高斯混合成分（与第六讲的隐变量 z_j 相同；书中记作 Θ ）； N ：类别数。周志华《机器学习》第 13 章 13.2 节，式 (13.1)–(13.3)，pp.295–296。

用 EM 算法求解

$D_l \cup D_u$ 的对数似然 = 有监督项 + 无监督项

$$\begin{aligned} LL(D_l \cup D_u) = & \sum_{(\mathbf{x}_j, y_j) \in D_l} \ln \left(\sum_i \alpha_i p(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) p(y_j \mid z_j = i, \mathbf{x}_j) \right) \\ & + \sum_{\mathbf{x}_j \in D_u} \ln \left(\sum_i \alpha_i p(\mathbf{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right) \end{aligned}$$

E 步： 计算未标记样本属于各成分的后验 γ_{ji} （第六讲）；**M 步：**

$$\boldsymbol{\mu}_i = \frac{\sum_{\mathbf{x}_j \in D_u} \gamma_{ji} \mathbf{x}_j + \sum_{(\mathbf{x}_j, y_j) \in D_l \wedge y_j = i} \mathbf{x}_j}{\sum_{\mathbf{x}_j \in D_u} \gamma_{ji} + l_i}$$

要点 有标记样本的 γ_{ji} 已知为 0 或 1；模型假设不准时，未标记数据反而有害。

记号 l_i ：第 i 类的有标记样本数； $\boldsymbol{\Sigma}_i$ 的更新式类似， $\alpha_i = \frac{1}{m} (\sum_{\mathbf{x}_j \in D_u} \gamma_{ji} + l_i)$ ；可用有标记数据初始化模型参数。周志华《机器学习》第 13 章 13.2 节，式 (13.4)–(13.8)，p.297。

07

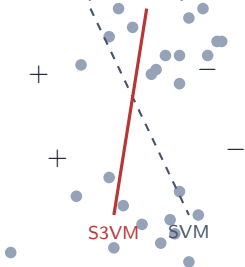
半监督支持向量机

Semi-Supervised SVM

低密度分隔 · TSVM 的目标 · 伪标记与局部搜索

半监督支持向量机：低密度分隔

半监督支持向量机 (Semi-Supervised Support Vector Machine, S3VM)



- 不考虑未标记样本：找最大间隔划分超平面（第三讲）
- 考虑未标记样本：找能分开两类有标记样本、且穿过数据低密度区域的超平面

“低密度分隔” (low-density separation) 是聚类假设在线性超平面划分下的推广。

记号 示意图参照周志华《机器学习》第 13 章 13.3 节，图 13.3，p.298：+、- 为有标记的正、反例，灰色点为未标记样本。

TSVM 的优化目标

TSVM (Transductive Support Vector Machine): 尝试对未标记样本进行各种可能的标记指派

$$\begin{aligned} \min_{\mathbf{w}, b, \hat{\mathbf{y}}, \boldsymbol{\xi}} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C_l \sum_{i=1}^l \xi_i + C_u \sum_{i=l+1}^m \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \dots, l \\ & \hat{y}_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = l+1, \dots, m; \quad \xi_i \geq 0 \end{aligned}$$

- $\hat{\mathbf{y}} = (\hat{y}_{l+1}, \dots, \hat{y}_{l+u})$: 未标记样本的预测标记, $\hat{y}_i \in \{-1, +1\}$, **也是优化变量**
- C_l, C_u : 平衡模型复杂度、有标记与未标记样本重要程度的折中参数

问题 尝试全部标记指派是一个穷举过程, 只有未标记样本很少时才可行。一般情形如何求解?

记号 Joachims (1999); ξ : 松弛向量 (第三讲的软间隔)。周志华《机器学习》第 13 章 13.3 节, 式 (13.9), p.299。

TSVM 的求解：局部搜索

1. 用 D_l 训练一个支持向量机，对 D_u 预测，结果作为伪标记（pseudo-label）
 2. 初始化 $C_u \ll C_l$ ：伪标记可能不准，让有标记样本起更大作用
 3. 求解目标函数；若一对未标记样本 $\hat{y}_i \hat{y}_j < 0$ 、 $\xi_i > 0$ 、 $\xi_j > 0$ 且 $\xi_i + \xi_j > 2$ ，则二者的标记很可能有误：交换后重新求解
 4. 标记调整完成后，令 $C_u = \min\{2C_u, C_l\}$ ，直到 $C_u = C_l$
- 每次交换后目标函数值下降
 - 输出未标记样本的标记，也能对训练中未见的样本预测
 - 伪标记可能类别不平衡：把 C_u 拆为 C_u^+ 、 C_u^- ， $C_u^+ = \frac{u_-}{u_+} C_u^-$

要点 搜索可能出错的每一对未标记样本计算量很大：如何高效优化是半监督 SVM 研究的重点。

记号 u_+ 、 u_- ：基于伪标记当作正、反例使用的未标记样本数（类别不平衡的处理见第五讲）。周志华《机器学习》第 13 章 13.3 节，图 13.4，式 (13.10)，pp.299–300。

08

半监督聚类

Semi-Supervised Clustering

必连与勿连约束 · 约束 k 均值 · 约束种子 k 均值

半监督聚类：两类监督信息

聚类是无监督任务，但现实中往往能获得一些额外的监督信息

必连与勿连约束

必连 (must-link): $(x_i, x_j) \in \mathcal{M}$, 必属于同一个簇

勿连 (cannot-link): $(x_i, x_j) \in \mathcal{N}$, 必不属于同一个簇

⇒ 约束 k 均值

少量有标记样本

$S = \bigcup_{j=1}^k S_j \subset D$, S_j 为隶属于第 j 个簇的样本, $|S| \ll |D|$

⇒ 约束种子 k 均值

要点 两种算法都是第六讲 k 均值的扩展：分配样本时满足约束，或把有标记样本作为固定的“种子”。

记号 半监督聚类 (semi-supervised clustering); 约束关系也用于第五讲的成对约束度量学习。此处的标记指簇标记，不是类别标记。勿连约束集合书中记作 $\mathcal{C}$ ，本课程 $\mathcal{C}$ 已用于簇划分，故记 $\mathcal{N}$ 。周志华《机器学习》第 13 章 13.6 节，pp.307–309。

约束 k 均值

Constrained k -means (Wagstaff 等, 2001): 在 k 均值的分配步骤中检查约束

对每个样本 x_i :

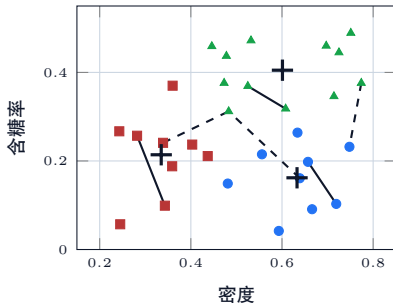
1. 按距离从近到远, 依次尝试各个簇 C_r
2. 若把 x_i 划入 C_r 违背 $\mathcal{M}$ 或 $\mathcal{N}$ 中的约束, 换下一个簇
3. 所有簇都违背约束: 返回错误提示

更新均值向量, 重复直到均值向量不再更新

例: 必连 (x_4, x_{25}) 、 (x_{12}, x_{20}) 、 (x_{14}, x_{17}) ; 勿连 (x_2, x_{21}) 、 (x_{13}, x_{23}) 、 (x_{19}, x_{23}) ; $k = 3$, 初值

x_6, x_{12}, x_{27}

最终结果 (5 轮迭代)



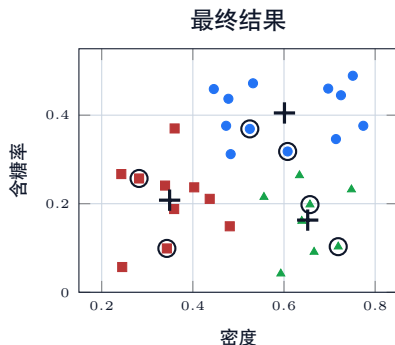
记号 实线: 必连约束; 虚线: 勿连约束; +: 均值向量。例题: 周志华《机器学习》第 13 章 13.6 节, 图 13.7、图 13.8, pp.307–309, 重新计算的最终结果与书一致。

约束种子 k 均值

Constrained Seed k -means (Basu 等, 2002): 用有标记样本作“种子”

1. 用每个 S_j 的平均向量初始化第 j 个簇的均值向量
2. 种子样本直接划入所属的簇，迭代中不改变其簇隶属关系
3. 其余样本划入最近的簇；更新均值向量，直到不再更新

例: $S_1 = \{x_4, x_{25}\}$, $S_2 = \{x_{12}, x_{20}\}$, $S_3 = \{x_{14}, x_{17}\}$;
4 轮后均值向量不再变化



有标记样本的信息利用起来很容易：只需固定它们的簇隶属关系。

记号 圈出的点为种子样本；例题：周志华《机器学习》第 13 章 13.6 节，图 13.9、图 13.10，pp.309–310，重新计算的最终结果与书一致。

09

小结

Summary

本讲要点 · 中英文术语对照 · 课后作业

本讲要点

- 特征选择** 子集搜索（前向、后向、双向，贪心）+ 子集评价（信息增益等）
- 三类方法** 过滤式 Relief（猜中近邻、猜错近邻）；包裹式 LVW（学习器性能）；嵌入式 L_1 正则化
- 稀疏学习** L_1 等值线为菱形，解常落在坐标轴上；近端梯度下降的软阈值；字典学习交替优化
- 监督信息** 全监督、弱监督（含半监督）、无监督；自监督属于无监督
- 半监督** 聚类假设与流形假设；生成式方法（EM）、TSVM（低密度分隔）、约束 k 均值与约束种子 k 均值

中英文术语对照

英文全称	缩写	中文
Relevant Features	Relief	相关特征（过滤式方法）
Las Vegas Wrapper	LVW	拉斯维加斯包裹式方法
Least Absolute Shrinkage and Selection Operator	LASSO	最小绝对收缩选择算子
Proximal Gradient Descent	PGD	近端梯度下降
Semi-Supervised Support Vector Machine	S3VM	半监督支持向量机
Transductive Support Vector Machine	TSVM	直推式支持向量机

课后作业

1. 软阈值

对 $\min_x \frac{L}{2}(x - z)^2 + \lambda|x|$ ($x, z \in \mathbb{R}$) 分 $x > 0$ 、 $x < 0$ 、 $x = 0$ 三种情形讨论，推出近端梯度下降中的闭式解（周志华《机器学习》第 11 章 11.4 节，式 (11.14)，p.254），并说明它为什么产生稀疏解。

2. 生成式半监督学习

写出 $D_l \cup D_u$ 的对数似然，推导 M 步中 μ_i 的更新式，并说明有标记样本在其中的作用相当于 γ_{ji} 取何值。

3. 编程

在西瓜数据集 4.0 上实现约束种子 k 均值，比较不同种子样本对聚类结果的影响。