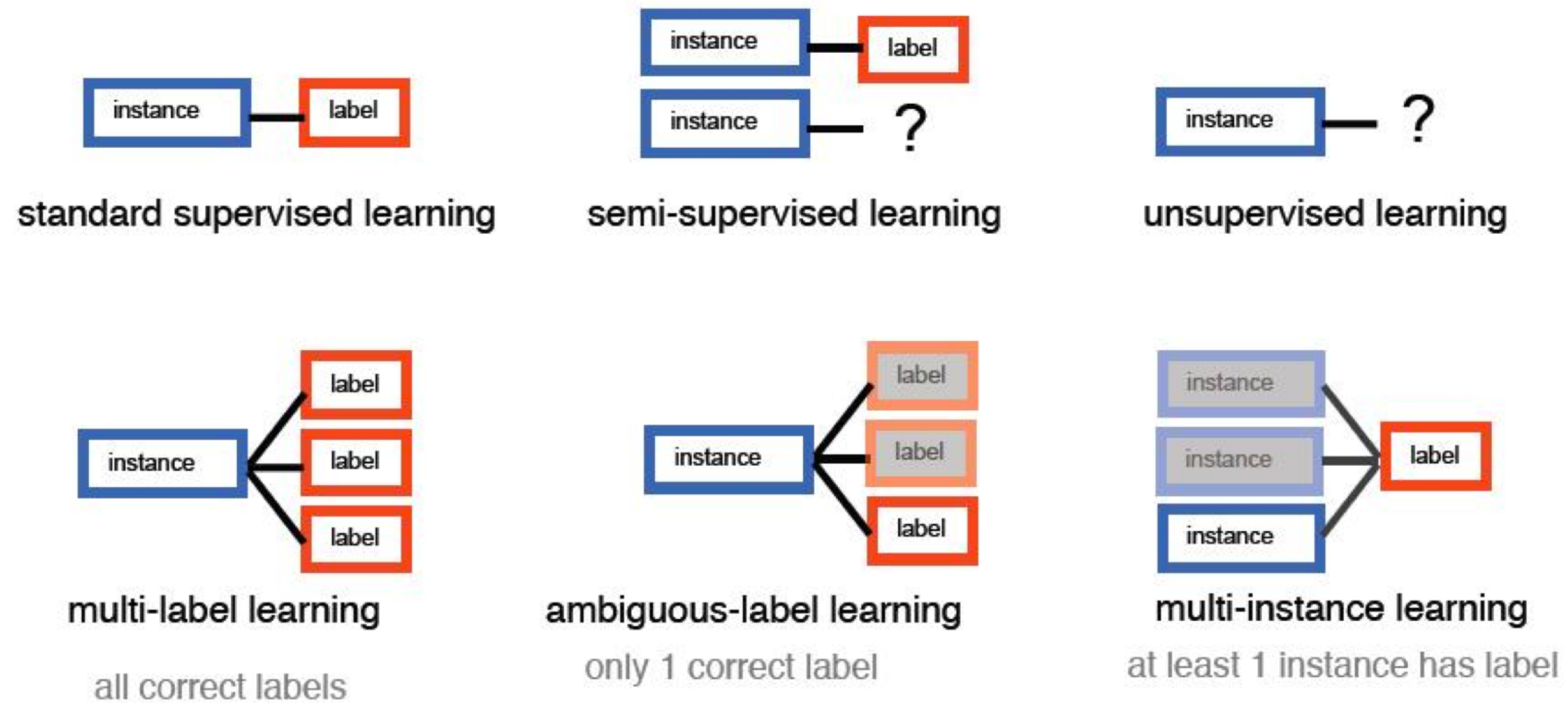
判别性聚类

（Discriminative Clutering）

Recall: 机器学习任务分类



References

- Linli Xu, James Neufeld, Bryce Larson, Dale Schuurmans, **Maximum Margin Clustering**, NIPS2002
  - https://papers.nips.cc/paper/2602-maximum-margin-clustering

- F. Bach and Z. Harchaoui. **DIFFRAC : a discriminative and flexible framework for clustering**, *Advances in Neural Information Processing Systems (NIPS) 20,* 2007.
  - https://www.di.ens.fr/~fbach/diffrac_nips.pdf
  - http://www.di.ens.fr/~fbach/diffrac_nips_poster.pdf
  - http://www.di.ens.fr/~fbach/diffrac/index.htm

▶ 基本思想

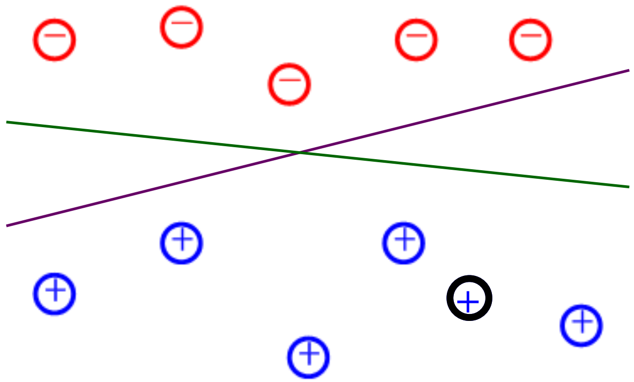
- 非监督的SVM（Unsupervised SVMs）
  - 比SVM更难计算
  - 将约束放松为凸函数——半正定规划（Semidefinite program）

▶ 背景：两类 SVM

- 有监督的*分类任务*：
  - 带标注的数据 → 线性判别

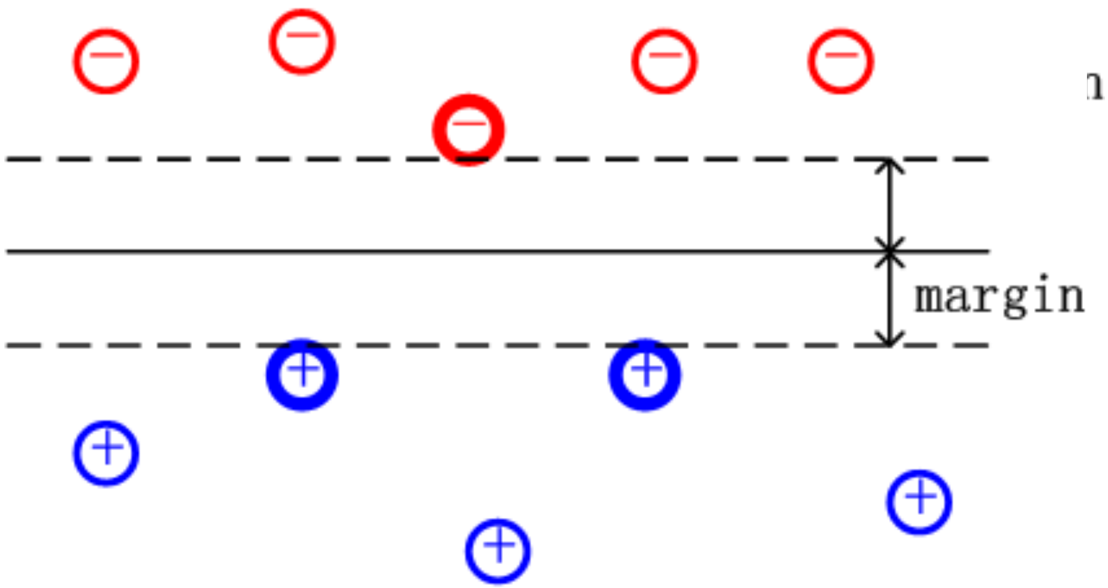
$$\mathbf{w}\cdot\mathbf{x}-b=0 \qquad \hat{y}=\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}-b)$$

  - 分类规则：



哪个分类器更好？

最大间隔线性判别

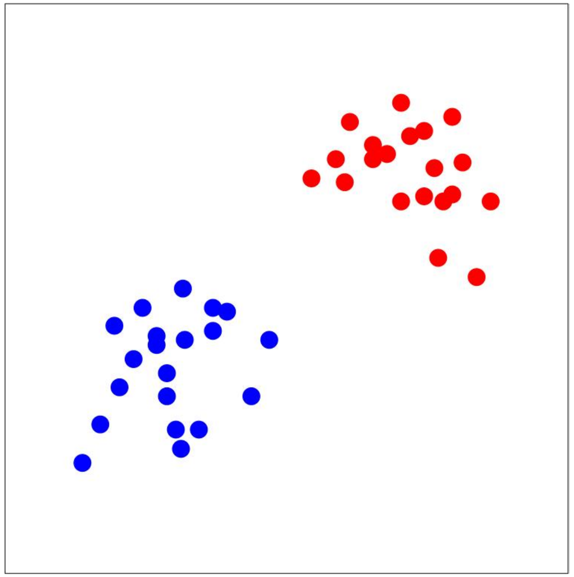


选择最大下面函数的线性判别器 $\mathbf{w}\cdot\mathbf{x}-b=0$

$$\min_{\langle \mathbf{x}_i, y_i\rangle} dist(\langle \mathbf{x}_i, y_i\rangle,\ \text{Plane } \mathbf{w}\cdot\mathbf{x}-b=0)$$

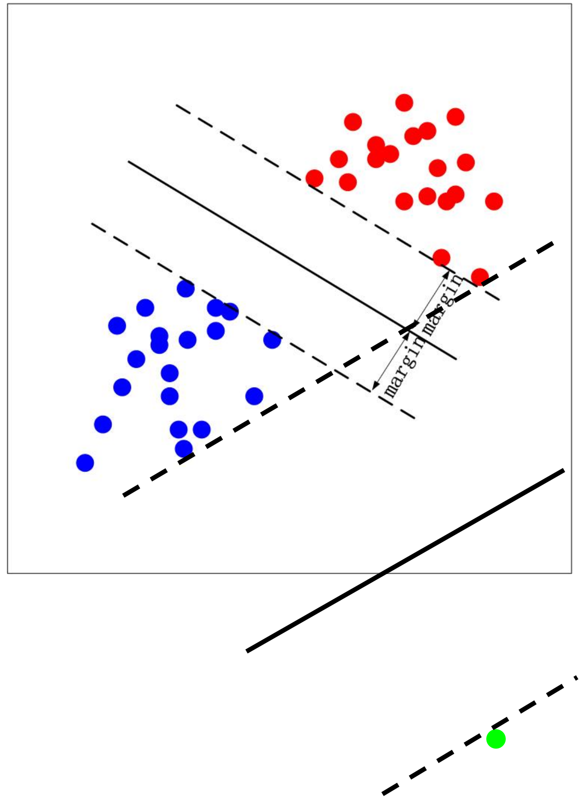
非监督 学习

- 给定非标注的数据，如何判断类别？
  - 将样本分组 — 聚类（clustering）



思想：最大间隔聚类

- 给定非标注的数据，找到最大间隔的超平面
- 对数据进行聚类
- 约束：类别平衡
  - 不同类别大小不同



如何推导非监督的SVM?

在课程课件中为： $Q(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$

两类情况：

1. 从监督算法开始

$$\begin{aligned} \gamma^{*-2} &= \min_{\mathbf{w},b,\boldsymbol{\epsilon}} \|\mathbf{w}\|^2 + C\boldsymbol{\epsilon}^\top \mathbf{e} \quad \text{subject to} \quad y^i(\mathbf{w}^\top \phi(\mathbf{x}^i) + b) \geq 1 - \epsilon_i, \forall_{i=1}^N, \ \boldsymbol{\epsilon} \geq 0 \\ &= \max_{\boldsymbol{\lambda}} 2\boldsymbol{\lambda}^\top \mathbf{e} - \langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^\top, \mathbf{y}\mathbf{y}^\top \rangle \quad \text{subject to} \quad 0 \leq \boldsymbol{\lambda} \leq C, \ \boldsymbol{\lambda}^\top y = 0 \end{aligned} \tag{3}$$

给定标签 **y**, 求解

$$\gamma^{*-2} = \max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{e} - \frac{1}{2\beta}\langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}, \mathbf{y}\mathbf{y}^{\mathrm{T}} \rangle$$

$$\text{subject to} \quad 0 \leq \boldsymbol{\lambda} \leq 1$$

间隔平方的倒数

λ：拉格朗日乘子（对应课件中的α）

K: 核矩阵

β：正则参数

如何推导非监督的SVM?

不止于代码

2. 将γ^{*-2}视为y的函数

目标：选择使得间隔平方的倒数**最小**的 **y**

给定 **y**, 求解

$$\gamma^{*-2}(\mathbf{y}) = \max_{\boldsymbol{\lambda}} \ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{e} - \frac{1}{2\beta}\left\langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}, \mathbf{y}\mathbf{y}^{\mathrm{T}} \right\rangle$$

$$\text{subject to} \quad 0 \le \boldsymbol{\lambda} \le 1$$

间隔平方的倒数

y的非凸函数

如何推导非监督的SVM?

3. 将问题转化为比较y 的示性函数

新变量： $M = \mathbf{y}\mathbf{y}^{\mathrm{T}}$

等价关系矩阵

$$M_{ij} = \begin{cases} 1 & \text{if } y_i = y_j \\ -1 & \text{if } y_i \neq y_j \end{cases}$$

给定**y**, 求解

$$\gamma^{*-2}(\mathbf{y}) = \max_{\boldsymbol{\lambda}} \ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{e} - \frac{1}{2\beta}\left\langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}, \mathbf{y}\mathbf{y}^{\mathrm{T}} \right\rangle$$

$$\text{subject to} \quad 0 \leq \boldsymbol{\lambda} \leq 1$$

间隔平方的倒数

如何推导非监督的SVM?

3. 将问题转化为比较y 的示性函数

新变量： $M = \mathbf{y}\mathbf{y}^{\mathrm{T}}$

等价关系矩阵

$$M_{ij} = \begin{cases} 1 & \text{if } y_i = y_j \\ -1 & \text{if } y_i \neq y_j \end{cases}$$

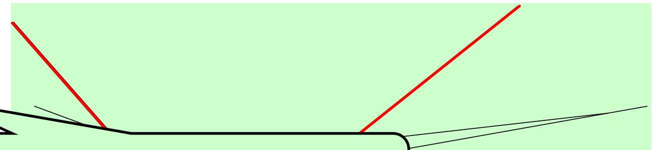
给定M, 求解

$$\gamma^{*-2}(M) = \max_{\boldsymbol{\lambda}} \ \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{e} - \frac{1}{2\beta}\left\langle K \circ \boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}}, M \right\rangle$$

$$\text{subject to} \quad 0 \le \boldsymbol{\lambda} \le 1$$

间隔平方的倒数

注意：M的凸函数



线性函数的最大化函数是凸的

4. 得到带约束的优化问题

求解 M

$$\min_{M} \gamma^{*-2}(M)$$

subject to $0 \le \boldsymbol{\lambda} \le 1$

非凸!

$$M \in \{-1,+1\}^{n\times n}$$

$$M = \mathbf{y}\mathbf{y}^{\mathrm{T}}$$

类别平衡→ $-\varepsilon\mathbf{e} \le M\mathbf{e} \le \varepsilon\mathbf{e}$

$M \in \{-1,+1\}^{n\times n}$ 表示一个等价关系矩阵 iff

$M \succeq 0,\ \mathrm{diag}(M) = \mathbf{e}$

如何推导非监督的SVM?

不 止 于 代 码

4. 得到带约束的优化问题

求解 M

$$
\begin{aligned}
&\min_{M} \gamma^{*-2}(M) \\
&\text{subject to} \quad 0 \le \boldsymbol{\lambda} \le 1 \\
&\qquad M \in \{-1,+1\}^{n\times n} \\
&\qquad M \succeq 0,\ \operatorname{diag}(M) = \mathbf{e} \\
&\qquad -\varepsilon \mathbf{e} \le M\mathbf{e} \le \varepsilon \mathbf{e}
\end{aligned}
$$

$M \in \{-1,+1\}^{n\times n}$ 表示一个等价关系矩阵

iff

$M \succeq 0,\ \operatorname{diag}(M) = \mathbf{e}$

CSDN 不止于代码

5. 放松指示变量，得到一个凸优化问题

求解M

$$
\begin{aligned}
&\min_{M} \gamma^{*-2}(M) \\
&\text{subject to} \quad 0 \le \boldsymbol{\lambda} \le 1 \\
&\qquad\qquad M \in \{-1,+1\}^{n\times n} \\
&\qquad\qquad M \succeq 0,\ \operatorname{diag}(M) = \mathbf{e} \\
&\qquad\qquad -\varepsilon\mathbf{e} \le M\mathbf{e} \le \varepsilon\mathbf{e}
\end{aligned}
$$

5. 放松指示变量，得到一个凸优化问题

Solve for M

$$\min_{M} \gamma^{*-2}(M)$$

$$\text{subject to} \quad 0 \le \boldsymbol{\lambda} \le 1$$

$$M \in \{-1, +1\}^{n \times n}$$

$$M \succeq 0, \ \operatorname{diag}(M) = \mathbf{e}$$

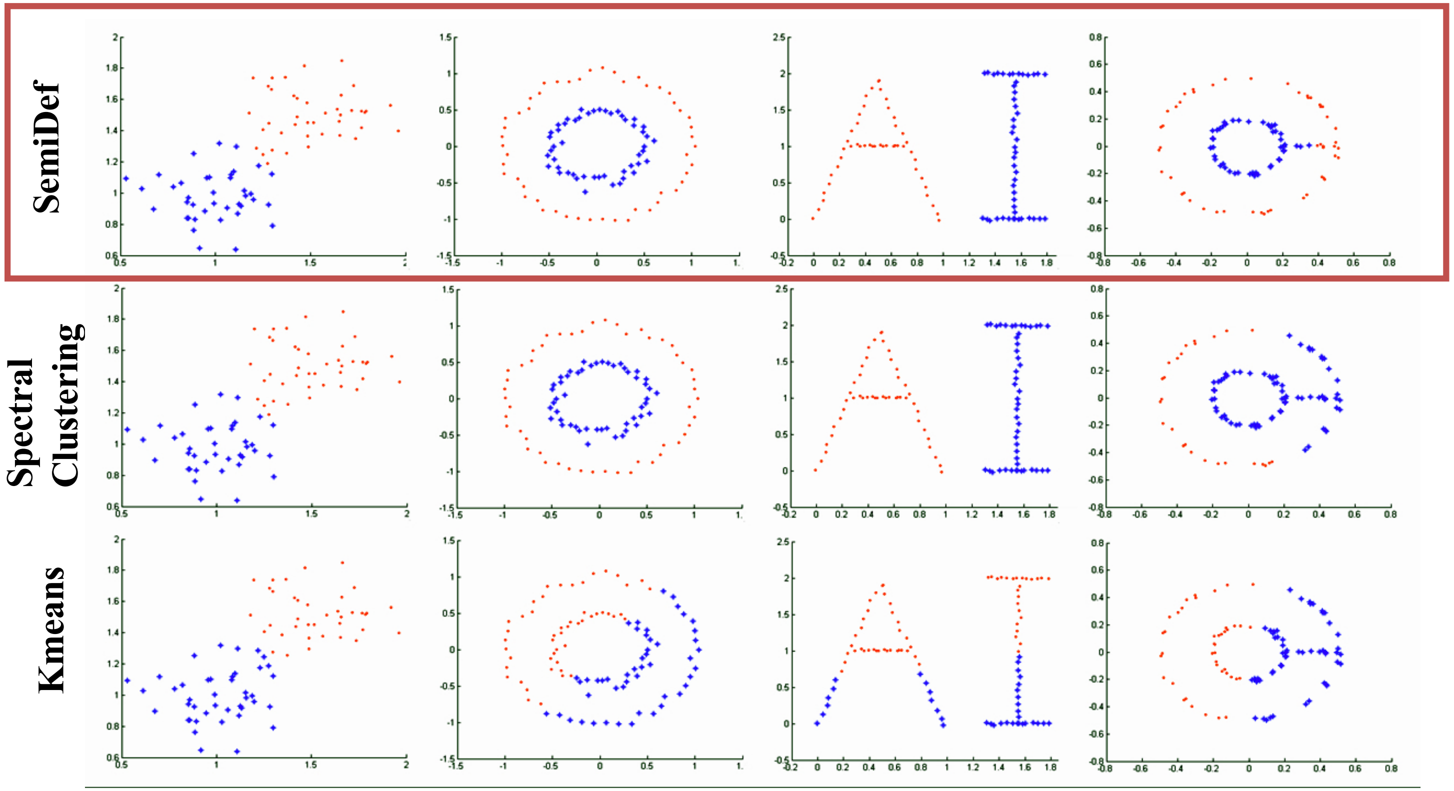
$$-\varepsilon \mathbf{e} \le M\mathbf{e} \le \varepsilon \mathbf{e}$$

半正定规划
（Semidefinite program）

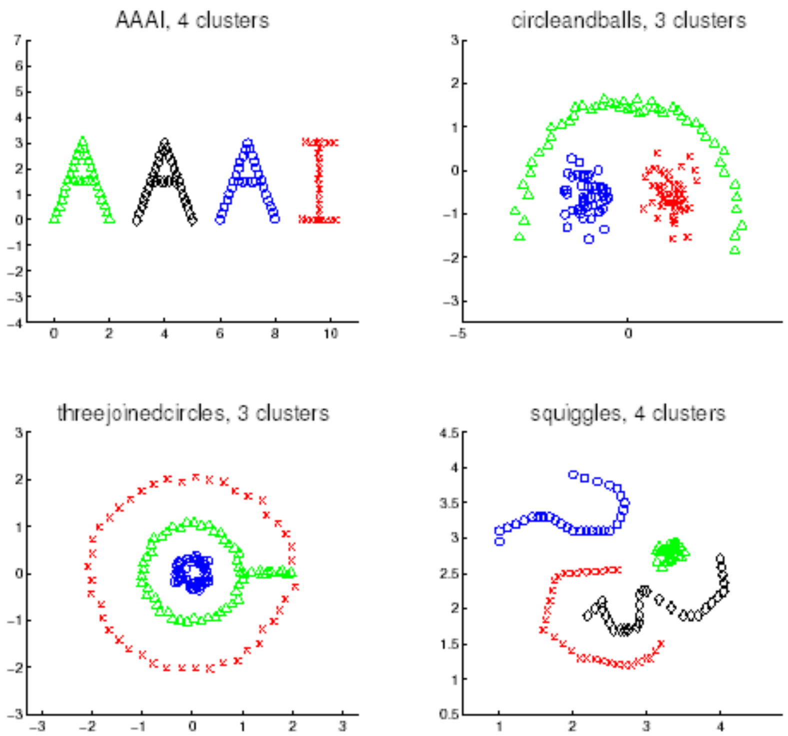
多类聚类推导类似

- 请见文献

Experimental Results



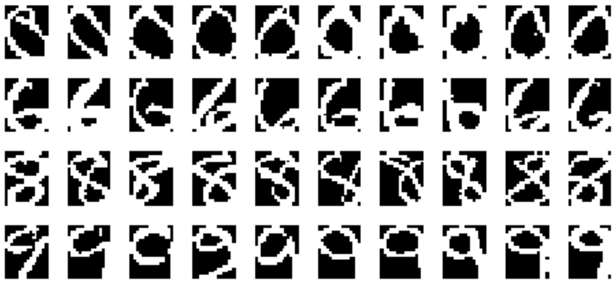
► Experimental Results



Experimental Results

	Semidefi nite	Spectral	Kmeans
4 Letters (AAAI)	0.0	0.0	0.0
Circle & Balls	0.0	0.0	18.8
3 Joined Circles	4.0	4.0	47.3
Squiggles	0.0	0.0	28.3
Digits: 689	3.4	12.0	12.0
Digits: 0689	7.5	9.2	38.3

Percentage of misclassification errors



Digit dataset

DIFFRAC :
a discriminative and flexible framework for clustering

DIFFRAC

- 将非监督SVM中的合页损失（hinge loss）换成L2损失
  - 有解析解

▶ 问题描述

- 给定n个d维数据点 $x_1, \ldots, x_n$ 用矩阵表示为 $X \in \mathbb{R}^{n\times d}$
- 聚类：将$\{1, \ldots, n\}$ 分成k个类别，k>1，用示性矩阵表示为

$$y \in \{0,1\}^{n\times k} \text{ such that } y1_k = 1_n \qquad y_{ij} = \begin{cases} 0 & y_i \neq k \\ 1 & y_i = k \end{cases},$$ 1_k表示全1的k维向量

- 给定$\mathbf{X}$，$\mathbf{y}$，可用岭回归求解参数$\mathbf{w}, b$：

$$J(y, X, \kappa) = \min_{w\in\mathbb{R}^{d\times k},\ b\in\mathbb{R}^{1\times k}} \frac{1}{n}\|y - Xw - 1_n b\|_F^2 + \kappa \ \operatorname{tr}\ w^\top w$$

  – 多类分类有解析解：令 $\Pi_n = I_n - \frac{1}{n}1_n 1_n^\top$

$$w^* = (X^\top \Pi_n X + n\kappa I_n)^{-1} X^\top \Pi_n y \quad \text{and} \quad b^* = \frac{1}{n} 1_n^\top (y - Xw^*)$$

判别性聚类

- 岭回归

$$J(y, X, \kappa) = \min_{w \in \mathbb{R}^{d \times k},\ b \in \mathbb{R}^{1 \times k}} \frac{1}{n} \| y - Xw - 1_n b \|_F^2 + \kappa \ \operatorname{tr}\ w^\top w$$

- 的解为

$$w^* = (X^\top \Pi_n X + n\kappa I_n)^{-1} X^\top \Pi_n y \quad \text{and} \quad b^* = \frac{1}{n} 1_n^\top (y - Xw^*)$$

- 因此判别性聚类的损失为：$J(y, X, \kappa) = \operatorname{tr}\ yy^\top A(X, \kappa)$
- 其中

$$\boxed{A(X, \kappa) = \frac{1}{n} \Pi_n (I_n - X(X^\top \Pi_n X + n\kappa I)^{-1} X^\top) \Pi_n}$$

DIFFRAC

不止于代码

$$A(X,\kappa)=\frac{1}{n}\Pi_n(I_n-X(X^\top\Pi_nX+n\kappa I)^{-1}X^\top)\Pi_n$$

- 优化问题：求解y，使得 $\operatorname{tr} yy^\top A(X,\kappa)$ 最小
- 损失函数只与 $M = yy^\top \in \mathbb{R}^{n\times n}$ （k类等价矩阵）

$$M_{ij}=\begin{cases}0 & y_i \neq y_j\\ 1 & y_i = y_j\end{cases}$$

- 对M进行凸外包近似（Convex outer approximation ）
  - M是半正定的：$M \succcurlyeq 0$
  - M的对角线为1_n: $\operatorname{diag}(M) = 1_n$
  - 如果M最多对应k类，有：$M \succcurlyeq \frac{1}{k}1_n1_n^\top$

- 则有凸集

$$\mathcal{C}_k=\{M\in\mathbb{R}^{n\times n},\ M=M^\top,\ \operatorname{diag}(M)=1_n,\ M\geqslant 0,\ M\succcurlyeq\frac{1}{k}1_n1_n^\top\}$$

▶ 类的最小大小

- 为避免无意义的解（将所有样本点归为一个类），给每个类别定义一个正则：类的最小大小 λ_0
- 行累加和： $M1_n \geqslant \lambda_0 1_n$ 且 $M1_n \leqslant (n-(k-1)\lambda_0)1_n$
- 特征值：类别的大小就是矩阵**M**前k个最大的特征值，因此约束等价于 $\sum_{i=1}^{n} 1_{\lambda_i(M)\geqslant\lambda_0} \geqslant k$ ，其中 $\lambda_1(M),\ldots,\lambda_n(M)$ 为**M**的n个特征值
  - 非凸约束，放松约束为 $\sum_{i=1}^{n}\phi_{\lambda_0}(\lambda_i(M)) \geqslant k$, where $\phi_{\lambda_0}(\kappa) = \min\{\kappa/\lambda_0, 1\}$
- 最终放松后的凸约束为：最小化 $\operatorname{tr} A(X,\kappa)M$
  - $\mathbf{M} = \mathbf{M}^T$
  - $\operatorname{diag}(M) = 1_n,\ M \geqslant 0,\ M \succcurlyeq \frac{1}{k}1_n1_n^{\top}, \sum_{i=1}^{n}\phi_{\lambda_0}(\lambda_i(M)) \geqslant k$

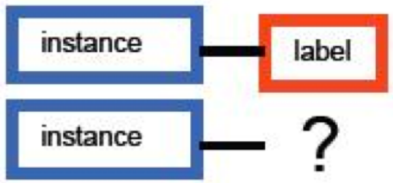
▶ 核

- 矩阵可以只用Gram矩阵 $K = XX^\top$表示为：

$$\boxed{A(K,\kappa) = \kappa\Pi_n(\widetilde{K} + n\kappa I_n)^{-1}\Pi_n}$$

- 其中 $\widetilde{K} = \Pi_n K \Pi_n$数据点**X**中心化后的Gram矩阵。

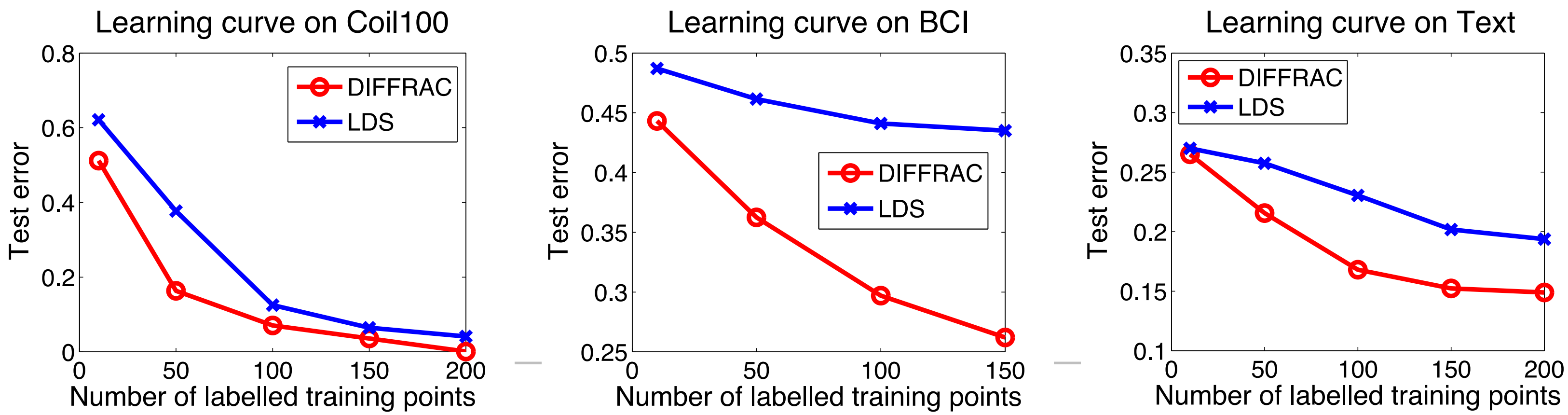
▶和半监督学习结合



semi-supervised learning

- 半监督学习：一部分样本有标签，更多样本没有标签

- 矩阵**M**可以很好的编码监督信息
  - “must-link”约束（正约束）：$M_{ij} = 1$
  - “must-not-link” 约束（负约束）：$M_{ij} = 0$



实验结果

Dataset	K-means	DIFFRAC	RCA
Mnist-linear 0%	**5.6 ± 0.1**	6.0 ± 0.4	
Mnist-linear 20%	4.5 ± 0.3	3.6 ± 0.3	**3.0 ± 0.2**
Mnist-linear 40%	2.9 ± 0.3	2.2 ± 0.2	**1.8 ± 0.4**
Mnist-RBF 0%	5.6 ± 0.2	**4.9 ± 0.2**	
Mnist-RBF 20%	4.6 ± 0.0	**1.8 ± 0.4**	4.1 ± 0.2
Mnist-RBF 40%	4.9 ± 0.0	**0.9 ± 0.1**	2.9 ± 0.1
Isolet-linear 0%	**12.1 ± 0.6**	12.3 ± 0.3	
Isolet-linear 20%	10.5 ± 0.2	**7.8 ± 0.8**	9.5 ± 0.4
Isolet-linear 40%	9.2 ± 0.5	**3.7 ± 0.2**	7.0 ± 0.4
Isolet-RBF 0%	11.4 ± 0.4	**11.0 ± 0.3**	
Isolet-RBF 20%	10.6 ± 0.0	**7.5 ± 0.5**	7.8 ± 0.5
Isolet-RBF 40%	10.0 ± 0.0	**3.7 ± 1.0**	6.9 ± 0.6

▶工具包

- Matlab：http://www.di.ens.fr/~fbach/diffrac/index.htm
- 在线学习版本（Matlab）：
  - http://www.di.ens.fr/willow/research/learningvideotext/