学习理论 ב

——如何从欠标注数据中学习

马锦华

机器学习方法分类

- 无监督学习
  - 聚类（clustering）、主成分分析（PCA）、等距映射（Isomap）等
- 无监督与有监督之间的学习方法？？
  - 半监督学习（Semi-Supervised Learning ）、迁移学习（Transfer Learning）、弱监督学习（Weakly-Supervised Learning）等
- 有监督学习
  - 线性判别分析（LDA）、支持向量机（SVM）、决策树（decision tree）、卷积神经网络（CNN）等

半监督学习

Semi-supervised learning

基于南京邮电大学汪云云老师课件修改

目录

CONTENT

01 半监督学习简介

02 学习方法

03 学习原理

04 研究挑战

1

INTRODUCTION OF SEMI-SUPERVISED LEARNING

半监督学习简介

1 简介

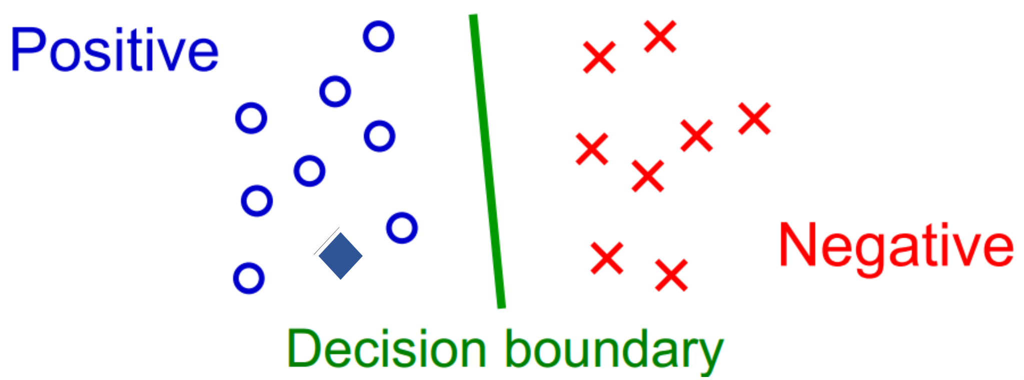
有监督学习

$\mathbf{X} = \{(x_i, y_i)\}_{i=1}^{n}$ $x_i \in R^d$ $y_i \in \{+1, -1\}$

色泽	根	敲声	好瓜
青绿	蜷缩	浊想	是
乌黑	蜷缩	浊想	是
青绿	硬挺	清脆	否
乌黑	稍卷	沉闷	否

类标号

→ 寻找分类函数 $f(x)$



在传统的分类学习中，每个数据样本都给定类标号指导分类学习，因此称为监督型分类学习（supervised classification）。

1 简介

收集数据样本容易，但获得样本类别标记相对困难

- 计算机辅助医学图像分析
- 垃圾邮件检测
- web网页推荐
- 。。。

获得样本类别标记需要人力、物力、财力。

如何同时利用无标记样本?

Label everything then double check every label

1 简介

半监督分类(semi-supervised classification)： 同时利用大量无标记样本和少量有标记样本进行分类学习，以获得比仅利用有标记样本的监督分类学习更好的分类性能。

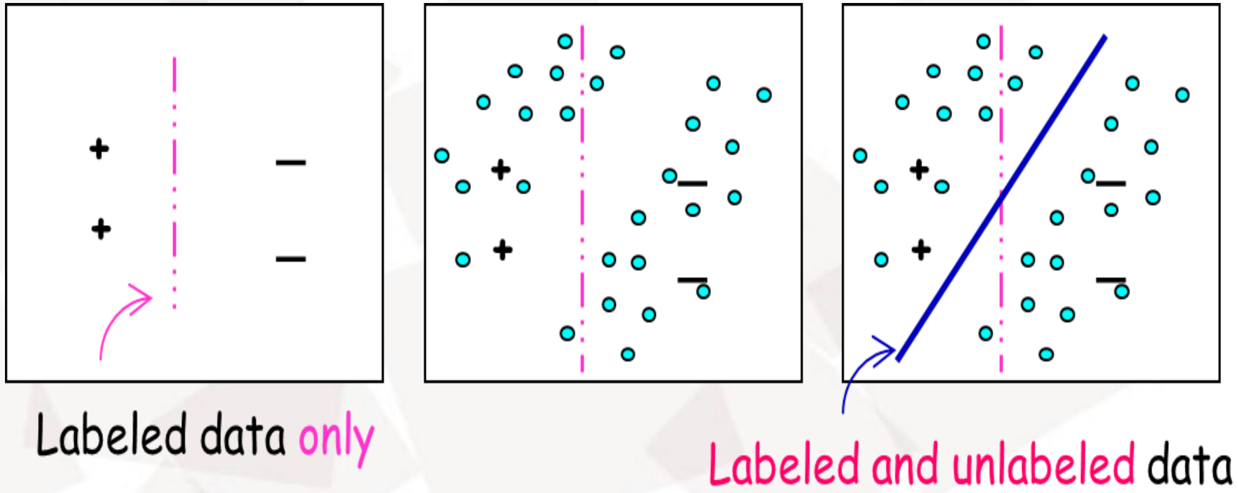
$\mathbf{X} = \{X_l, X_u\}$

$X_l = \{(x_i, y_i)\}_{i=1}^{l}$

$X_u = \{(x_j)\}_{j=1}^{u}$

$x_i \in R^d$

$y_i \in \{+1, -1\}$



综述论文或书籍：

- O. Chapelle et al. Semi-supervised learning. MIT Press Cambridge, 2006.
- X. Zhu and A. Goldberg. Introduction to semi-supervised learning. Morgan & Claypool Publishers, 2009.
- Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. Knowledge and Information Systems, 24(3):415–439, 2010.
- Z.-H. Zhou. Disagreement-based semi-supervised learning. Acta Automatica Sinica. Invited Survey. Nov. 2013.

2

METHODS

学习方法

2 学习方法

◆ 现有半监督分类方法可大致分为四大类：

– 生成式方法

– 大间隔半监督分类方法

– 基于图的半监督分类方法

– 基于差异性的方法

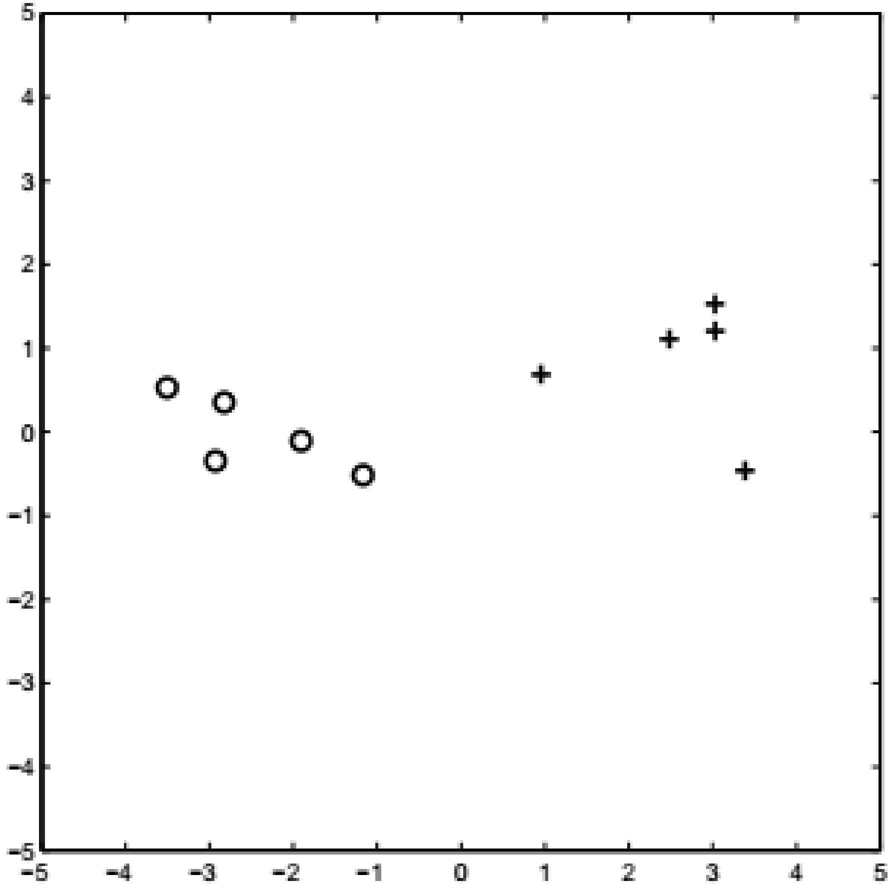
• 生成式（generative）方法

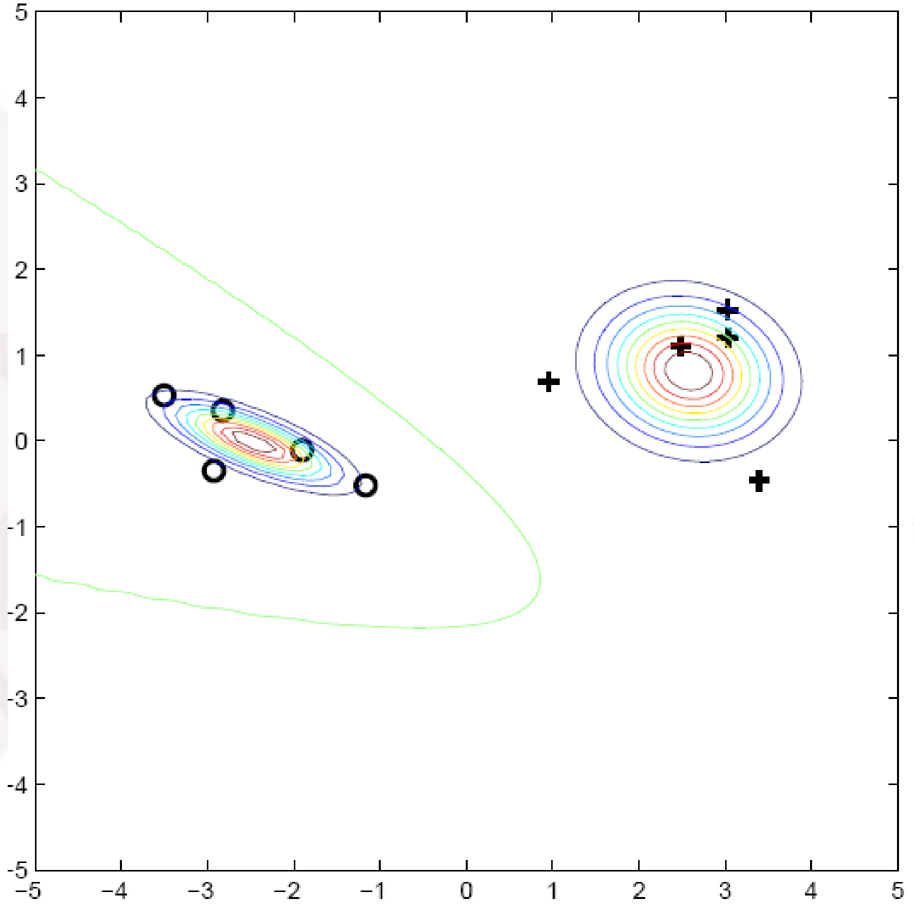
– 假设有标记和未标记数据由同一概率模型生成，以生成式模型作为分类模型，利用期望最大化(Expectation-Maximization，EM)算法估计模型参数以及样本属于每类的概率。

–不同方法的区别在于所采用的生成式模型不同，如混合高斯模型 (mixture Gaussian model)，朴素贝叶斯模型 (naïve Bayes model)等。

- Shahshahani, B. & Landgrebe, D. (1994). The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5): 1087-1095.
- Nigam, K., McCallum, A. K., Thrun, S. & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2):103-134.

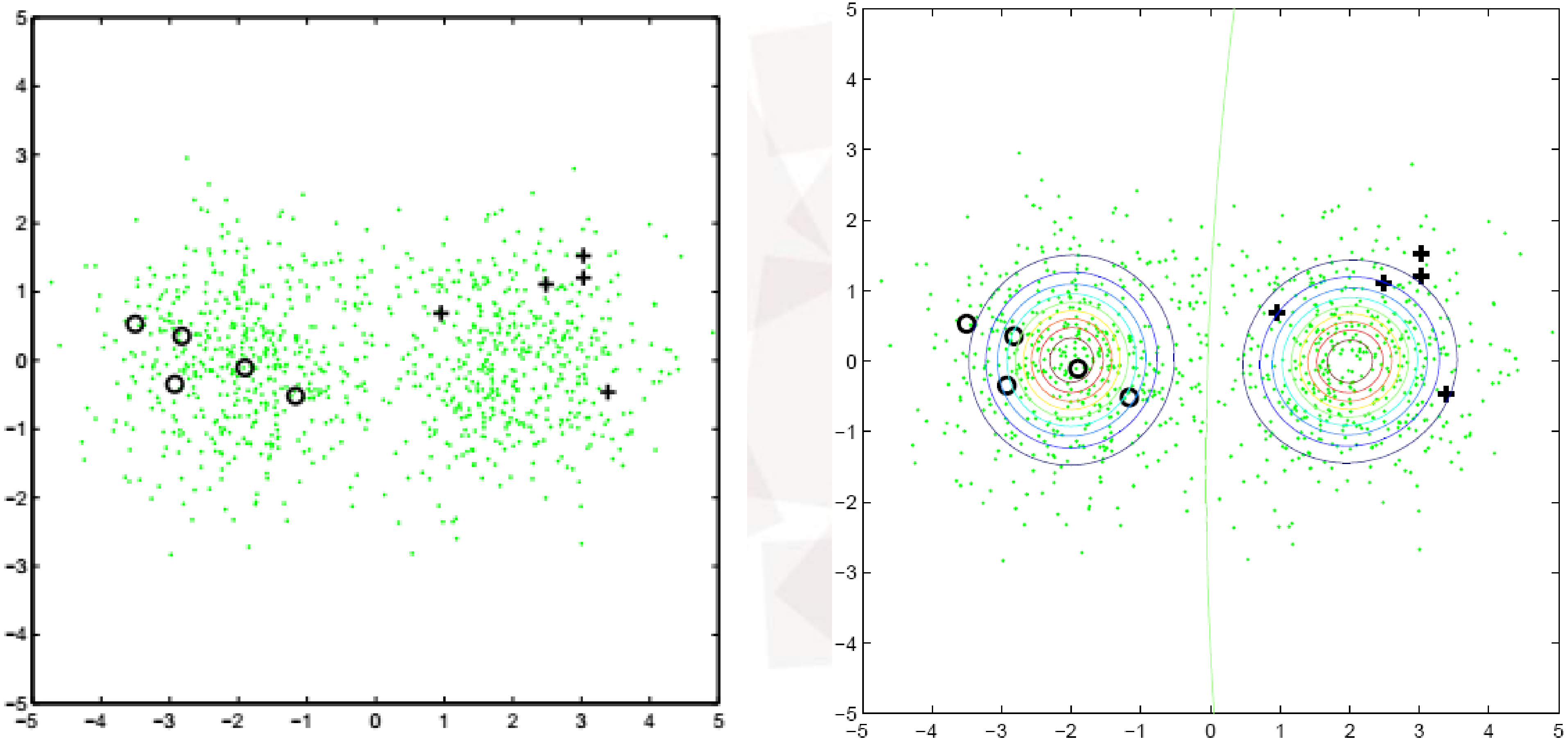
生成式方法





混合高斯模型

生成式方法



- Xiaojin Zhu, Semi-supervised learning Tutorial, Univ. Wisconsin, Madison, ICML, 2007.

生成式方法

模型假定，高斯 or 贝叶斯?

若模型假定错误，利用无标记样本将可能导致分类性能损失。

- F. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In Proceedings of The 15th International Florida Artificial Intelligence Society Conference, Pensacola, Florida, United States, 327-331, 2002
- T. Yang and C.E. Priebe. The Effect of Model Misspecification on Semisupervised Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33 (10): 2093 - 2103, 2011

◆现有半监督分类方法可大致分为四大类：

– 生成式方法
– 大间隔半监督分类方法
– 基于图的半监督分类方法
– 基于差异性的方法

- **大间隔半监督分类方法**
  - 最大化所有样本的类间间隔，从而引导分类边界穿越数据分布的低密度区域。
  - 直传SVM (Transductive SVM, TSVM)、半监督SVM (Semi-Supervised Support Vector Machine, S3VM)等。

- Bennett, K.P., & Demiriz, A. (1999). Semi-supervised support vector machines. In Kearns, M. J., Solla, S. A. & Cohn, D. A. editors, *Advances in Neural Information Processing Systems 11*, MIT Press, Cambridge, MA. 368-374.
- Joachims, T. (1999). Transductive inference for text classification using support vector machines. In: *Proceedings of the 16th International Conference on Machine Learning*, Bled, Slovenia, 200-209.
- Yu-Feng Li, James T. Kwok, and Zhi-Hua Zhou. Semi-supervised learning using label mean. In: Proceedings of the 26th International Conference on Machine Learning (ICML'09), Montreal, Canada, 2009, pp.633-640

SVM

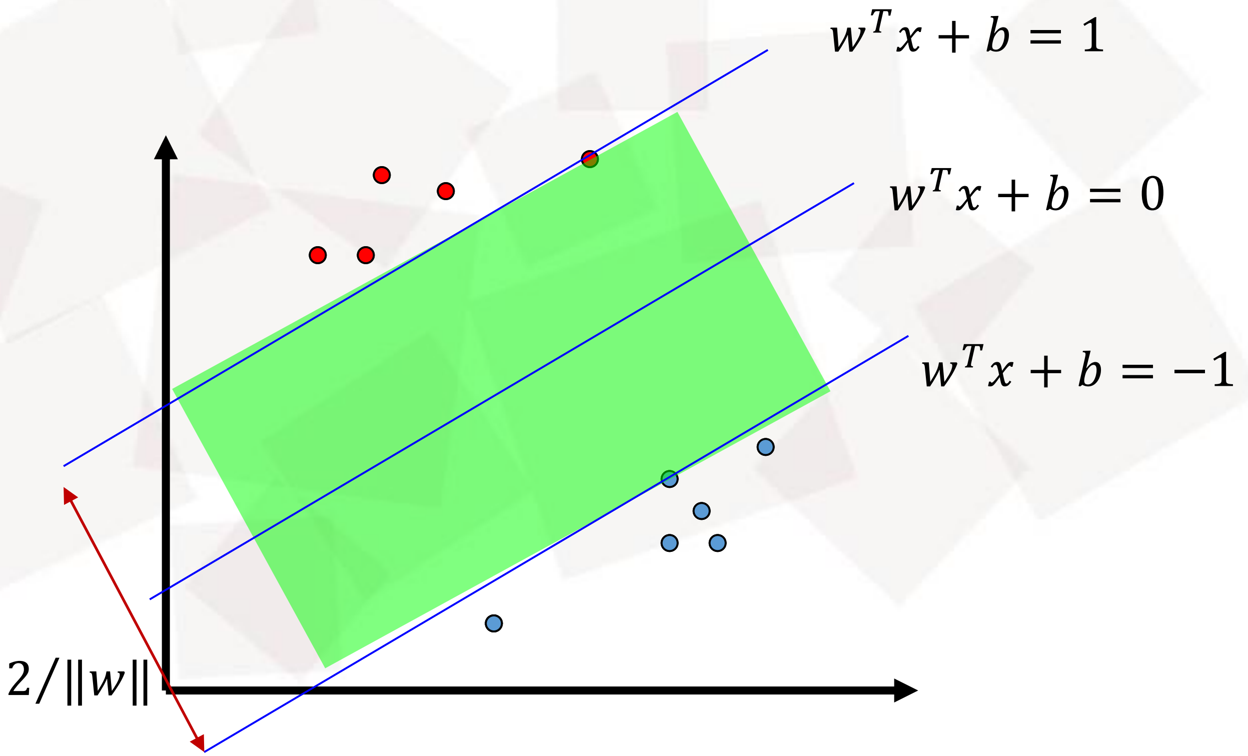
分类函数 $f(x) = w^T x + b$

优化目标

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

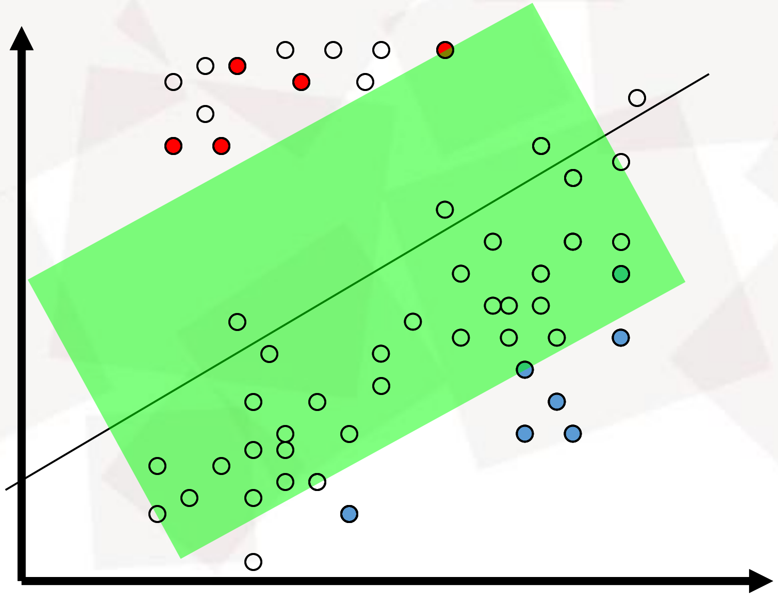
$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$



SVM

最大化有标记样本上的类间间隔



半监督SVM

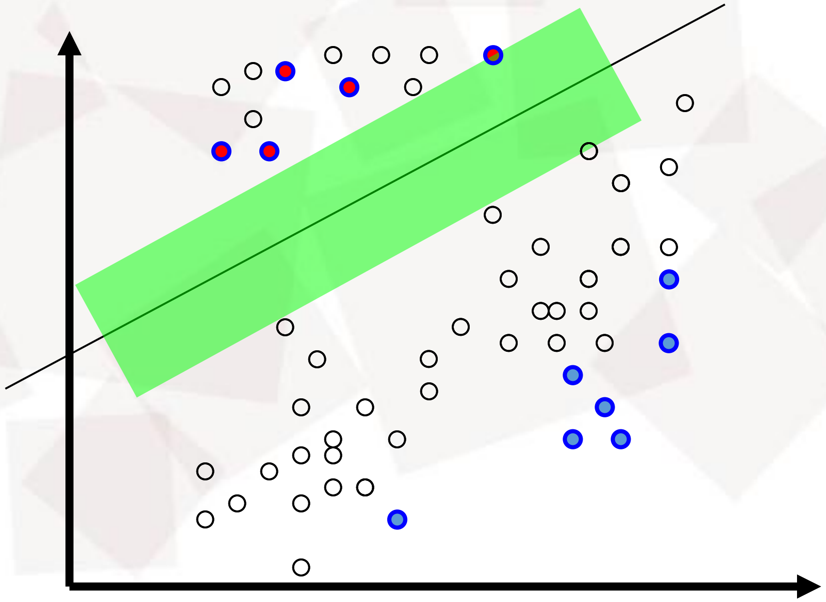
最大化有标记样本上的类间间隔

最大化所有样本上的类间间隔

优化目标

$$\min_{\substack{f\in\mathcal{H}\\ \boxed{\mathbf{y}\in\mathcal{B}}}} \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C_1\sum_{i=1}^{l}\ell(y_i, f(\mathbf{x}_i)) + \boxed{C_2\sum_{j=l+1}^{l+u}\ell(y_j, f(\mathbf{x}_j))}$$

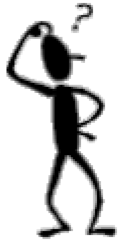
无标记样本的分类损失



半监督S3VM [Joachims, ICML99]工作获得ICML’09十年最佳论文奖

半监督S3VM

S3VM同时寻找分类函数和无标记样本的类别标号，优化目标非凸。



如何求解?

- O. Chapelle, V. Sindhwani, and S. S. Keerthi. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9:203–233, 2008
- O. Chapelle, M. Chi, and A. Zien. A continuation method for semi-supervised SVMs. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 185–192, Pittsburgh, PA, 2006
- O. Chapelle, V. Sindhwani, and S. S. Keerthi. Branch and bound for semi-supervised support vector machines. In B. Scholkopf, J. Platt, and T. Hoffman, editors, ¨ *Advances in Neural Information Processing Systems 19*, pages 217–224. MIT Press, Cambridge, MA, 2007
- Y.-F. Li, J.T. Kwok, and Z.-H. Zhou. Semi-supervised learning using label mean. In *Proceedings of the 26th International Conference on Machine Learning*, pages 633–640, Montreal, Canada, 2009

◆现有半监督分类方法可大致分为四大类：

– 生成式方法

– 大间隔半监督分类方法

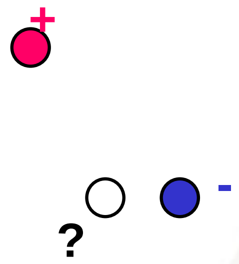
– 基于图的半监督分类方法

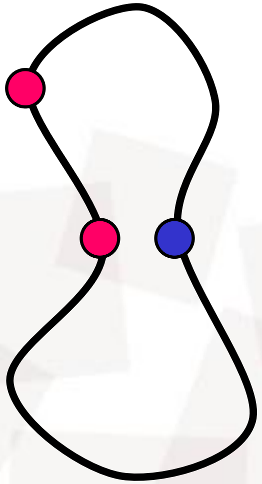
– 基于差异性方法

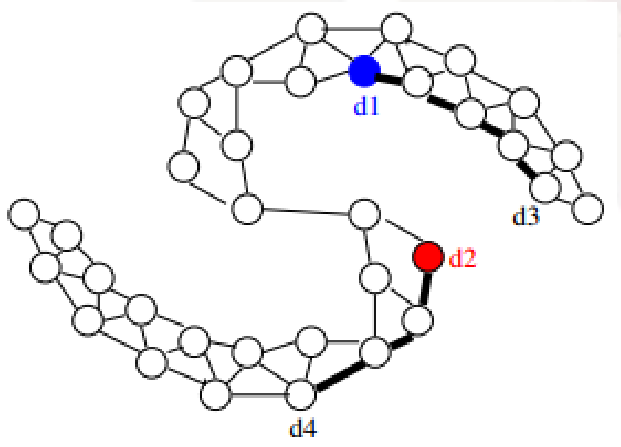
- **基于图的半监督分类方法**
  - 基于有标记和未标记数据间的相似性构建一个流形结构图，然后利用结构图和有标记样本的类别标号获取无标记样本的类别标号。
  - mincut、harmonic和流形正则化（Manifold Regularization）等。

- X Zhu， Z Ghahramani， J Lafferty (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *Twentieth International Conference on International Conference on Machine Learning*, Washington, DC, USA, 912-919
- Wang, F., & Zhang, C. (2006). Label propagation through linear neighborhoods. In: *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 985-992.
- Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7: 2399-2434.

基于图的方法







- 在所有样本上构建一个流形结构图，图结点代表数据样本，边权刻画样本之间的相似性或距离。
- 假设在图上相似的样本共享相同的类标号。
- 推断无标记样本的类标号。

harmonic

最小化能量函数

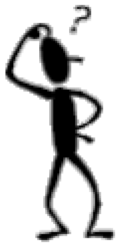
流形图边权

$$\min \quad \sum_{i \sim j} \underline{w_{ij}} (f(x_i) - f(x_j))^2$$

- 设置 $f(x_i)=y_i$, $i=1\ldots l$, $f(x_j)$ 为任意值，$j=1\ldots u$
- 重复更新 $f(x_j)$ 直至收敛

基于图的半监督分类方法[Zhu et al., ICML03]工作获得ICML'13十年最佳论文奖

基于图的方法



流形结构图的构建，K近邻？热核权？

- Yu-Feng Li, Shao-Bo Wang, Zhi-Hua Zhou. Graph quality judgement: A large margin expedition. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16), New York, NY, 2016, pp.1725-1731

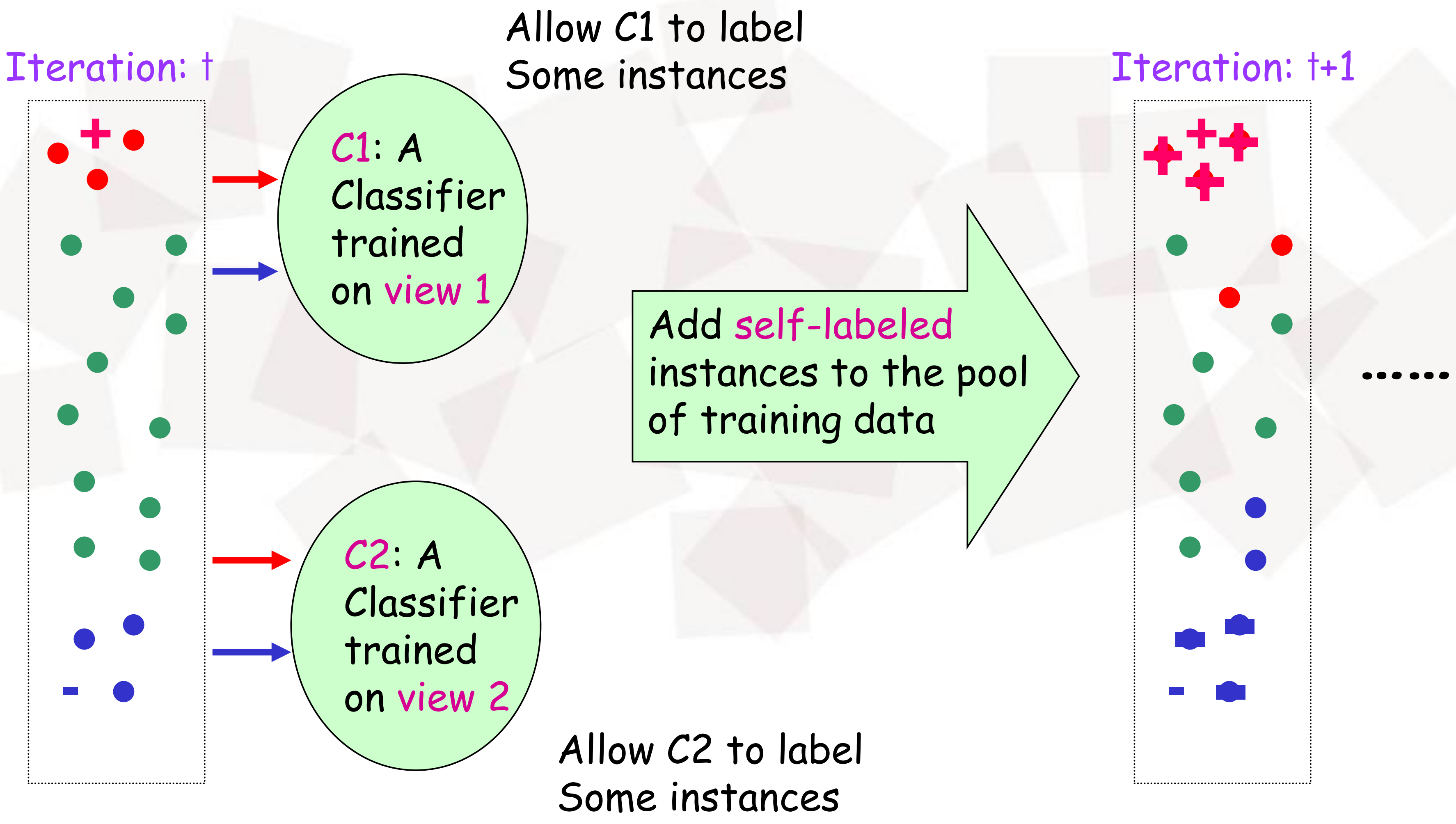
◆ 现有半监督分类方法可大致分为四大类：

– 生成式方法
– 大间隔半监督分类方法
– 基于图的半监督分类方法
– 基于差异性的方法

- **基于差异性的方法**
  - 利用两个或多个分类器，在学习过程中，各分类器挑选若干置信度高的无标记样本进行相互标记，然后更新分类模型。
  - Co-EM、 Tri-training等。

- Nigam, K. & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In: *Proceedings of the 9th ACM International Conference on Information and Knowledge Management*, Mclean, VA, 86-93.
- Zhou, Z.-H., & Li, M. (2005). Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529-1541.

协同训练方法

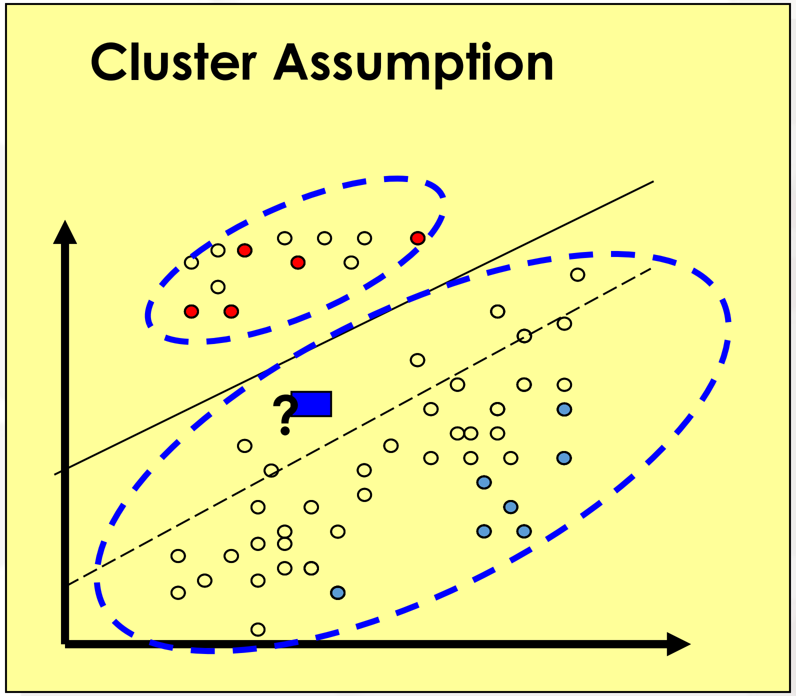

Iteration: t
Allow C1 to label
Some instances
C1: A Classifier trained on view 1
C2: A Classifier trained on view 2
Allow C2 to label
Some instances
Add self-labeled instances to the pool of training data
Iteration: t+1
......

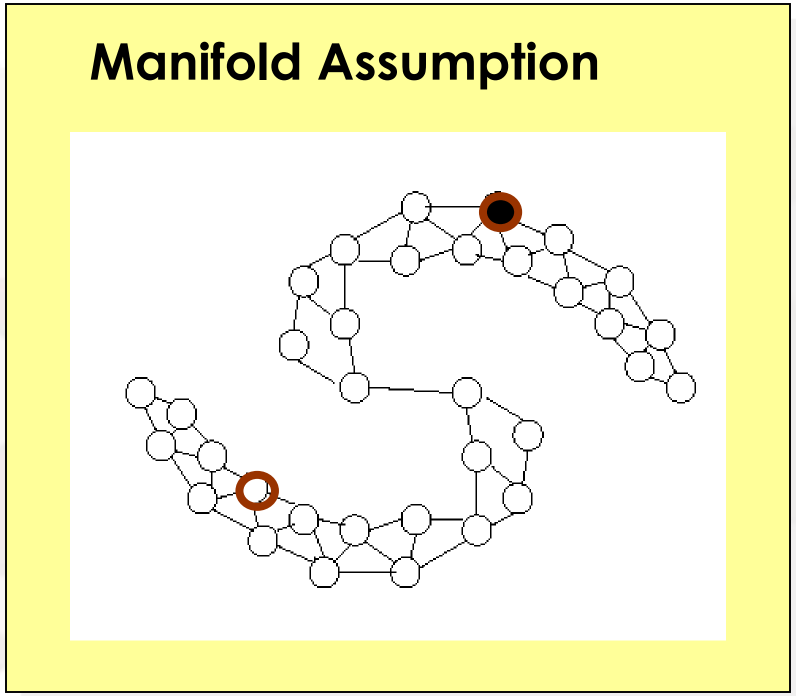
3

PRINCIPLES

学习原理

- 半监督分类利用无标记样本帮助提升学习性能
  - 试图挖掘隐藏在无标记样本中的数据分布信息，并利用该信息指导分类；
  - 为挖掘数据分布信息，必须采用某种数据分布假设；
  - 最常用的数据分布假设为**聚类假设(cluster assumption)**和**流形假设(manifold assumption)**。


Cluster Assumption
?


Manifold Assumption

数据分布假设

- **聚类假设**
  - 假定属于同一聚类的样本有较大可能共享相同的类标号。
  - **等价表述：**分类边界应穿越数据分布的低密度区域，从而使聚类内（高密度区域中）样本被划分在分类边界两侧，也被称为**低密度分割假设** (low-density separation assumption)。
- **流形假设**
  - 假定数据分布在一低维流形上，流形结构可由一无向图表示，图中结点代表样本，边权代表样本间相似性。在流形结构上相似的样本具有相似类标号。

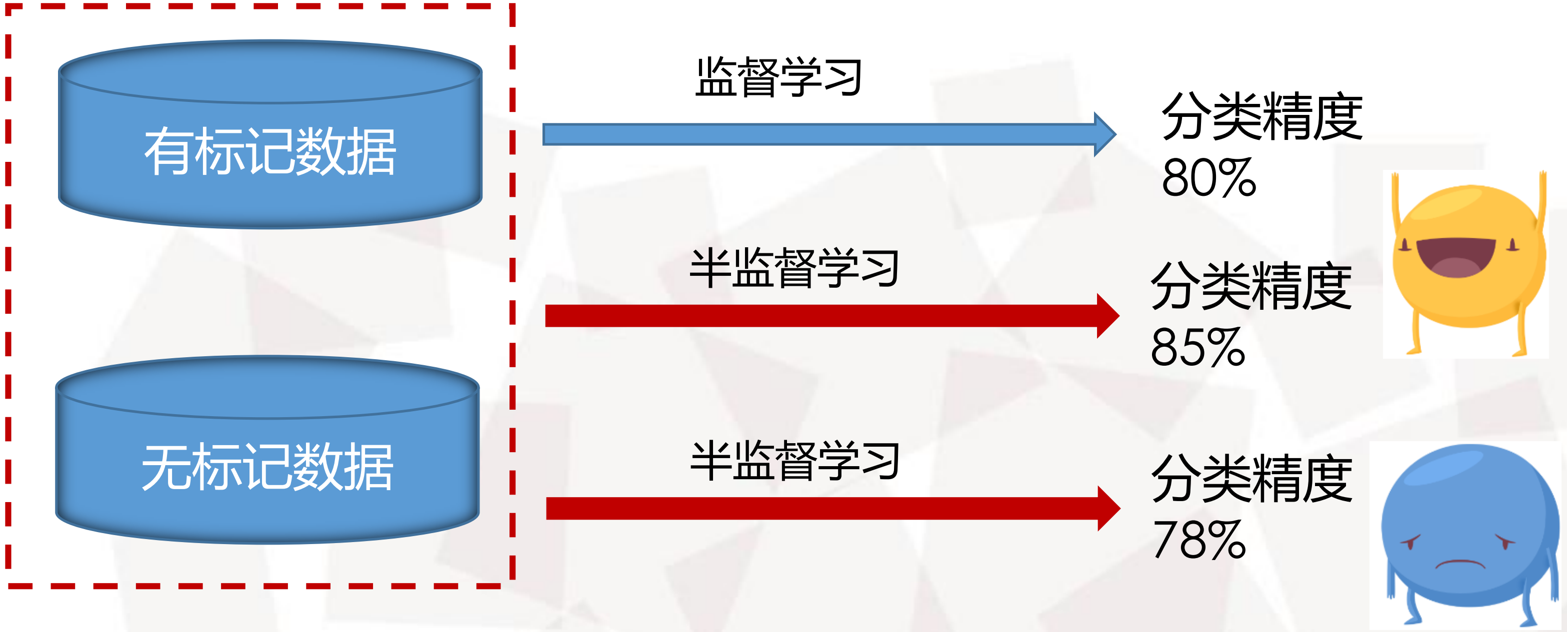
- 大部分现有的半监督分类方法都直接或间接地利用了上述两种假设：
  - **生成式方法：** 直接采用聚类假设
  - **基于支持向量机的方法：** 直接采用低密度分割假设
  - **基于图的方法：** 直接采用流形假设
  - **基于差异性的方法：** 间接利用聚类或流形假设

4

CHALLENGE

研究挑战

安全半监督学习



有时，同时利用无标记样本进行半监督分类学习不会带来性能提升，甚至可能造成性能下降。------------不安全半监督分类（unsafe semi-supervised classification)

安全半监督分类学习

如何实现安全半监督分类学习，使利用无标记样本不会造成性能的显著下降?

- Yu-Feng Li and Zhi-Hua Zhou. Improving semi-supervised support vector machines through unlabeled instances selection. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI'11), San Francisco, CA, 2011, pp.386-391.
- Yunyun Wang and Songcan Chen. Safety-Aware Semi-Supervised Classification. IEEE Trans. Neural Netw. Learning Syst. 24(11): 1763-1772 (2013)
- Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(1):175-188, 2015

迁移学习

主要内容

一 迁移学习简介

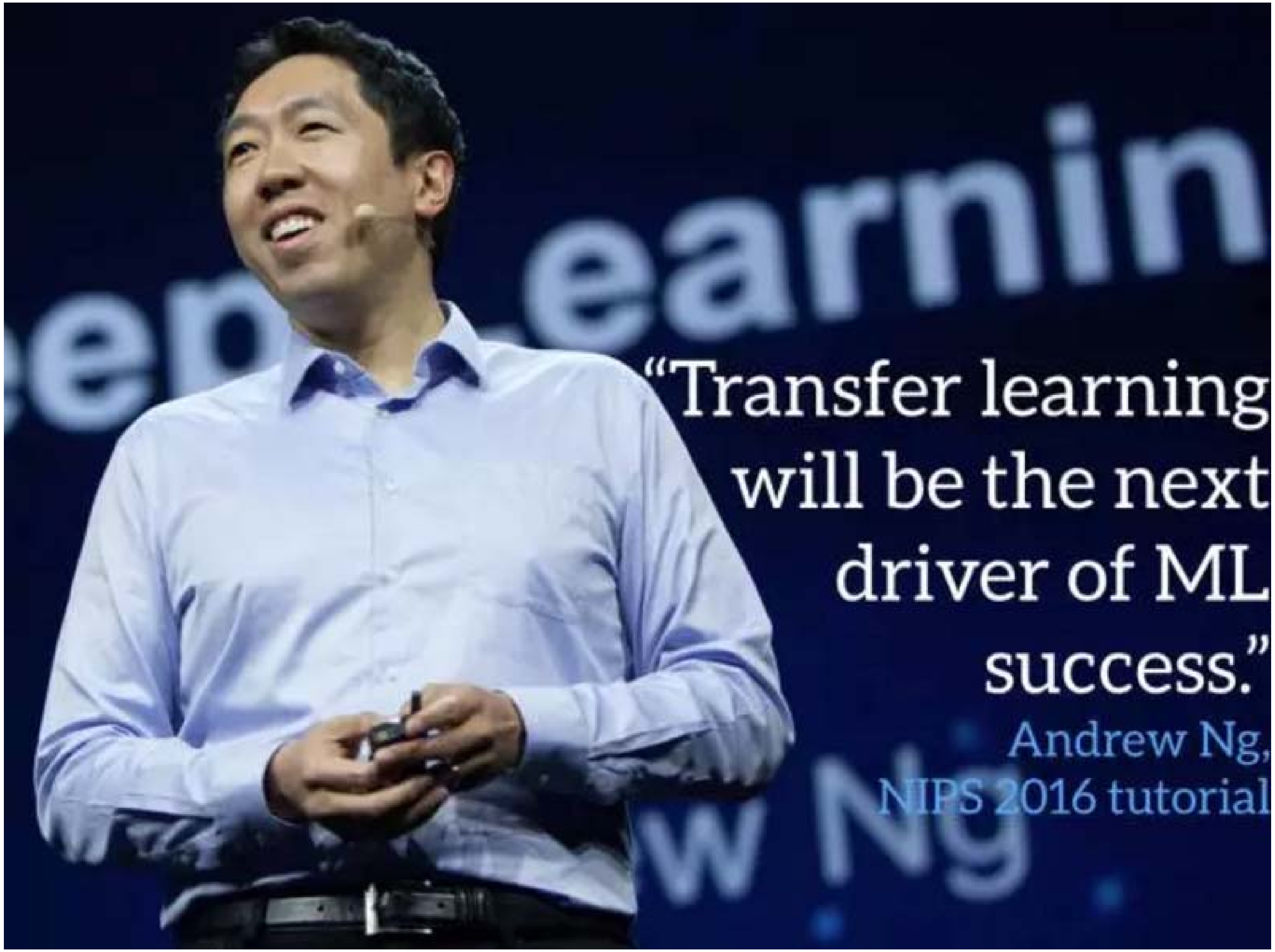
二 代表性研究工作

三 问题与展望

一 迁移学习简介

- 定义与概念
- 迁移学习 vs 传统机器学习

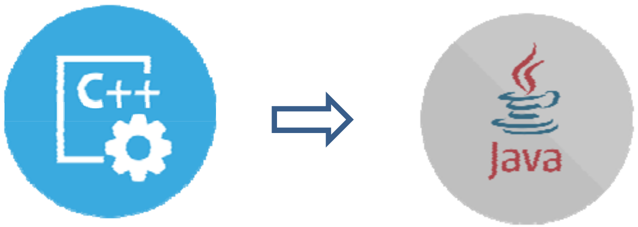
引子

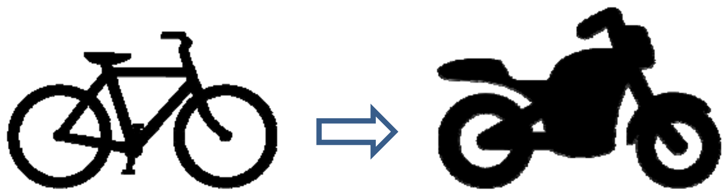


迁移学习将会是引领下一次机器学习热潮的驱动力。——吴恩达,NIPS 2016

迁移学习简介

- 什么是迁移学习？
  - **心理学角度**：人们利用之前的经验和知识进行推理和学习的能力。
  - **机器学习角度**：一个系统将别的相关领域中的知识应用到本应用中的学习模式。[DARPA]
  - 举例：C++→Java；骑自行车→骑摩托车
  - 关键词：举一反三



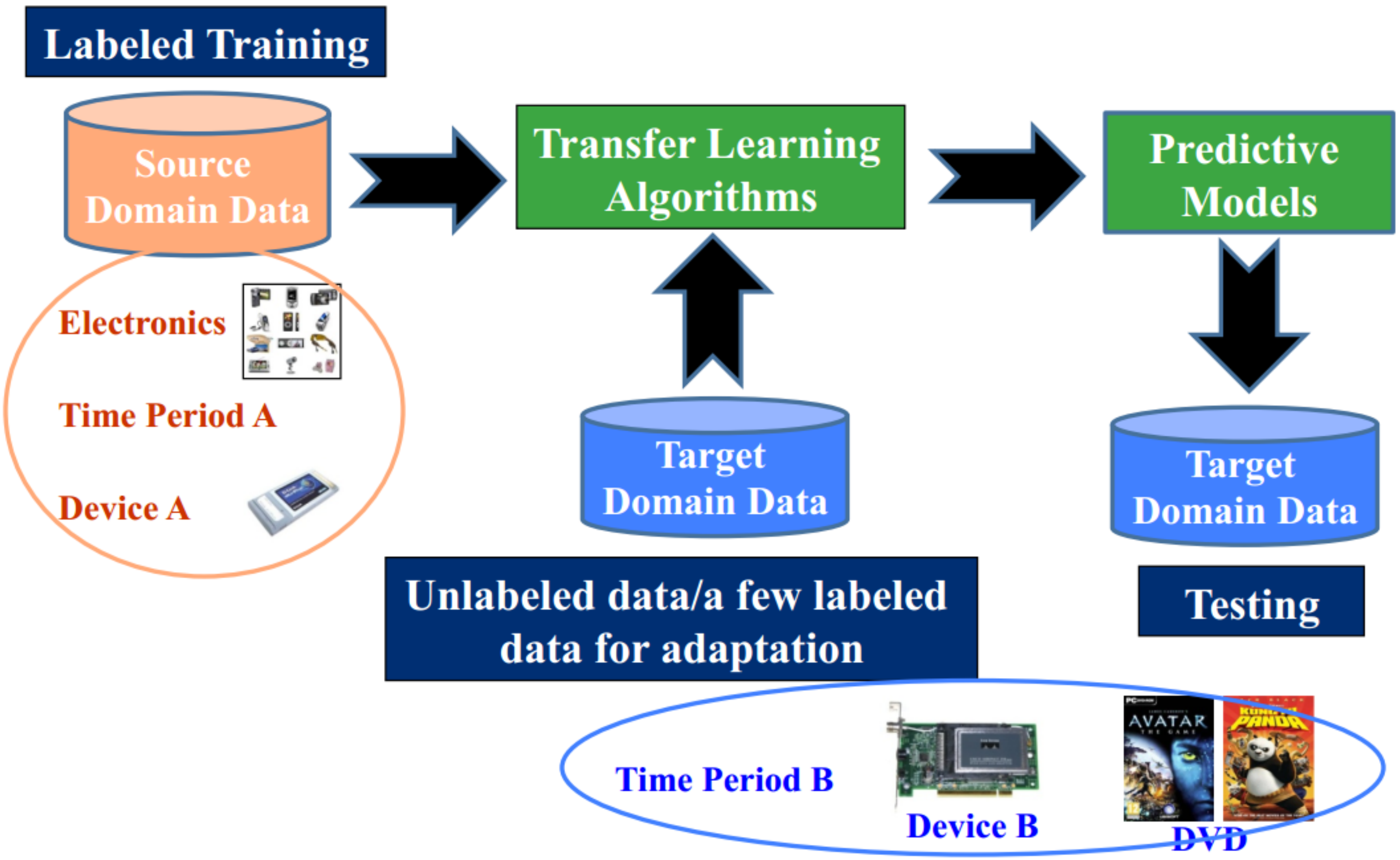


- 迁移学习要解决的问题：
  - 给定一个研究领域和任务，如何利用相似领域进行知识的迁移，从而达成目标？

迁移学习简介

- 为什么要进行迁移学习？
  - 数据的标签很难获取
  - 从头建立模型是复杂和耗时的

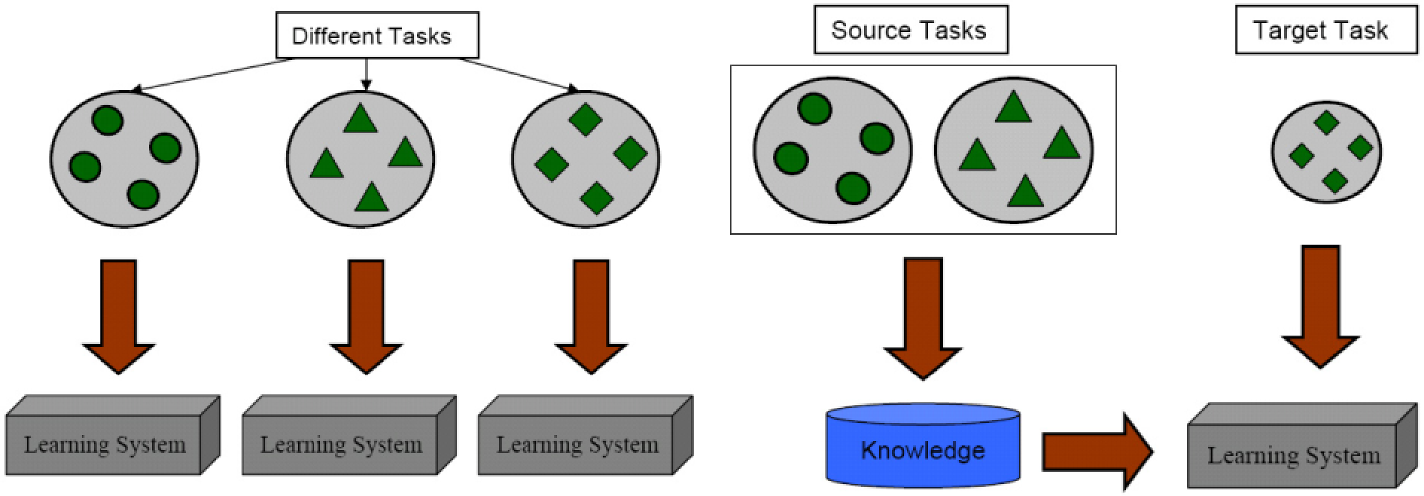
对已有知识的重用是必要的



迁移学习简介

- 迁移学习 vs 传统机器学习

	传统机器学习	迁移学习
数据分布	训练和测试数据同分布	训练和测试数据不需要同分布
数据标签	足够的数据标注	不需要足够的数据标注
建模	每个任务分别建模	可以重用之前的模型



传统机器学习　　迁移学习

迁移学习形式化概念

- 迁移学习常用概念
  - **Domain (域)**：由数据特征和特征分布组成，是学习的主体
    - Source domain (源域)：已有知识的域
    - Target domain (目标域)：要进行学习的域
  - **Task (任务)**：由目标函数和学习结果组成，是学习的结果
- 迁移学习的形式化定义
  - 条件：给定一个源域 $\mathcal{D}_S$ 和源域上的学习任务 $\mathcal{T}_S$, 目标域 $\mathcal{D}_T$ 和目标域上的学习任务 $\mathcal{T}_T$
  - 目标：利用 $\mathcal{D}_S$ 和 $\mathcal{T}_S$ 学习在目标域上的预测函数 $f(\cdot)$ 。
  - 限制条件：$\mathcal{D}_S \neq \mathcal{D}_T$ 或 $\mathcal{T}_S \neq \mathcal{T}_T$

迁移学习

- 应用领域

模式识别
计算机视觉
数据挖掘
语音识别
机器学习
迁移学习
统计学习
自然语言处理
文
A
Google
Translate

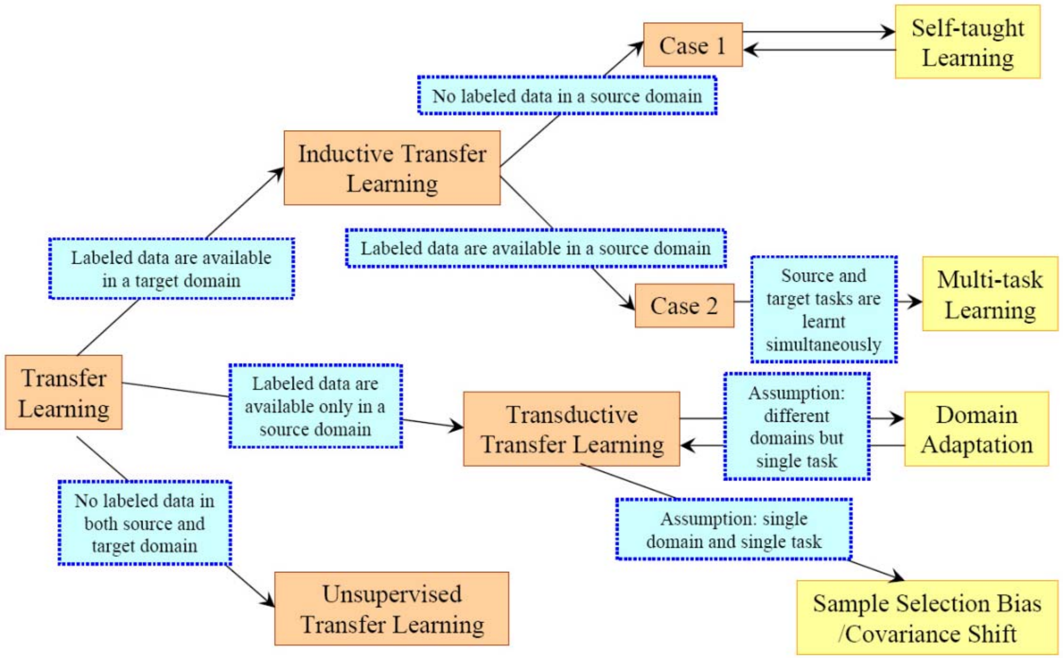
二

代表性研究成果

- 域适配问题
- 深度迁移学习

迁移学习：研究现状

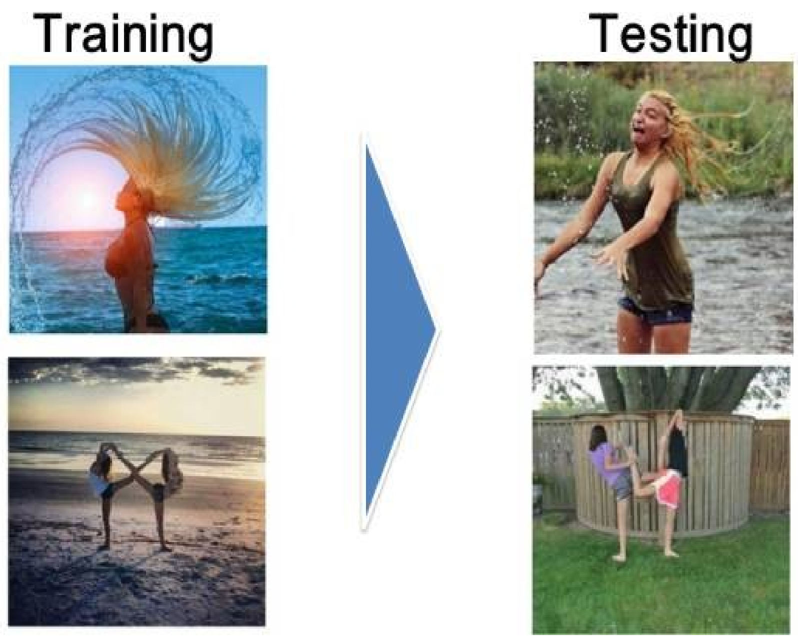
- 迁移学习的热门研究领域
  - 域适配问题 (domain adaptation)
  - 多源迁移学习 (multi-source TL)
  - 深度迁移学习 (deep TL)
  - 异构迁移学习 (heterogeneous TL)



迁移学习：域适配问题

- 域适配问题：
  - domain adaptation; cross-domain learning
  - 问题定义：有标签的源域和无标签的目标域共享相同的特征和类别，但是特征分布不同，如何利用源域标定目标域

    $\mathcal{D}_S \neq \mathcal{D}_T : P_S(X) \neq P_T(X)$

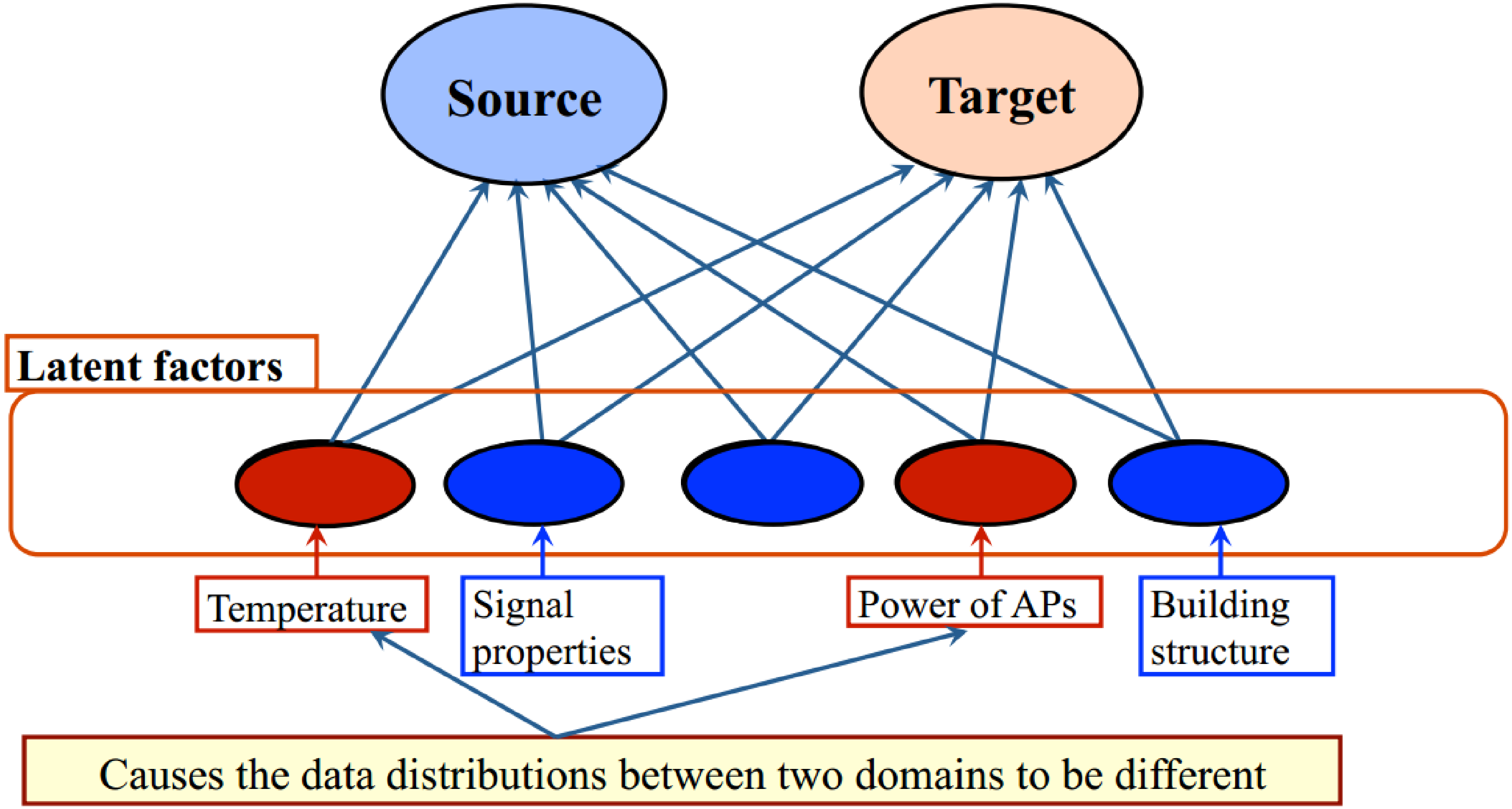


迁移学习：域适配问题

- 域适配问题：
  - 基于特征的迁移方法：
    - Transfer component analysis [Pan, TKDE-11]
    - Geodesic flow kernel [Duan, CVPR-12]
    - Transfer kernel learning [Long, TKDE-15]
    - TransEMDT [Zhao, IJCAI-11]
  - 基于实例的迁移方法：
    - Kernel mean matching [Huang, NIPS-06]
    - Covariate Shift Adaptation [Sugiyama, JMLR-07]
  - 基于模型的迁移方法：
    - Adaptive SVM (ASVM) [Yang et al, ACM Multimedia-07]
    - Multiple Convex Combination (MCC) [Schweikert, NIPS-09]
    - Domain Adaptation Machine (DAM) [Duan, TNNLS-12]

迁移学习：域适配问题

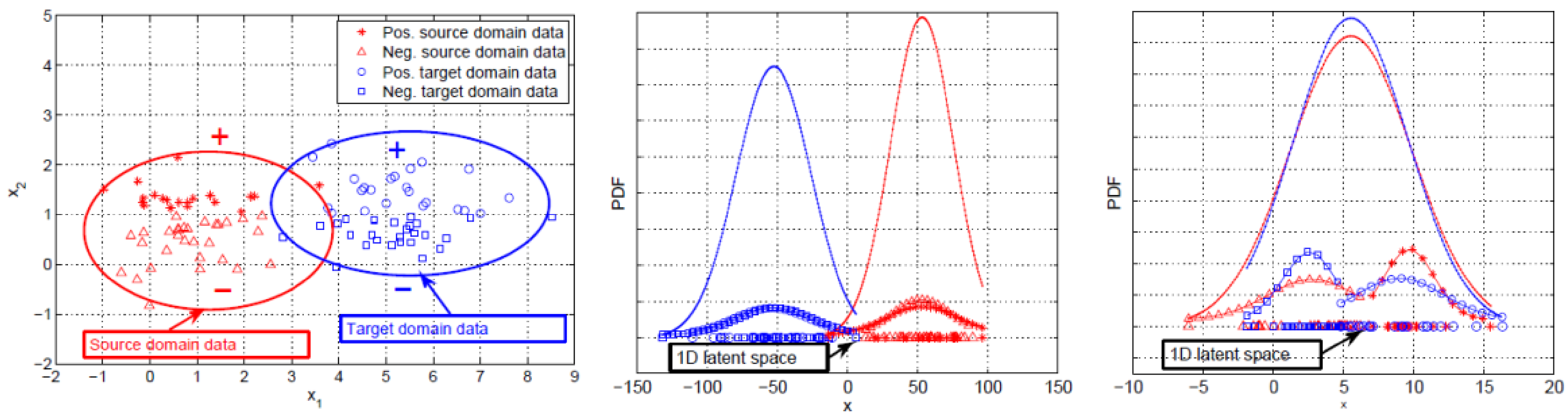
- 迁移成分分析 (TCA, transfer component analysis) [Pan, TKDE-11]
  - 将源域和目标域变换到相同空间，最小化它们的距离



迁移学习：域适配问题

- 迁移成分分析：
  - 优化目标：
$$\min_{\varphi} \quad \text{Dist}(\varphi(\mathbf{X}_S), \varphi(\mathbf{X}_T)) + \lambda\Omega(\varphi)$$
$$\text{s.t.} \quad \text{constraints on } \varphi(\mathbf{X}_S) \text{ and } \varphi(\mathbf{X}_T)$$
  - Maximum mean discrepancy (MMD)

$$\text{Dist}(P(X_S), P(X_T)) = \left\| \frac{1}{n_S}\sum_{i=1}^{n_S} \Phi(x_{S_i}) - \frac{1}{n_T}\sum_{j=1}^{n_T} \Phi(x_{T_j}) \right\|_{\mathcal{H}}$$



Original feature space

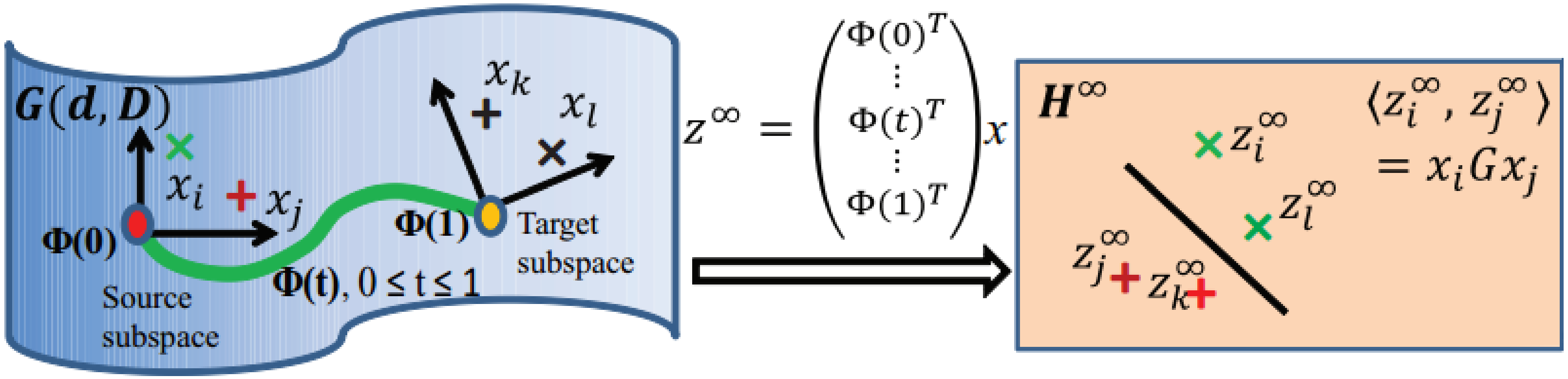
PCA

TCA

迁移学习：域适配问题

- GFK (geodesic flow kernel) [Duan, CVPR-12]
  - 利用流形学习，将数据映射到高维空间中，然后测量其距离，使得源域和目标域差异最大
  - 优化目标： $\boldsymbol{\Phi}(t) = \boldsymbol{P}_{\mathcal{S}}\boldsymbol{U}_1\boldsymbol{\Gamma}(t) - \boldsymbol{R}_{\mathcal{S}}\boldsymbol{U}_2\boldsymbol{\Sigma}(t)$

    $\boldsymbol{P}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{T}} = \boldsymbol{U}_1\boldsymbol{\Gamma}\boldsymbol{V}^{\mathrm{T}}, \quad \boldsymbol{R}_{\mathcal{S}}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{T}} = -\boldsymbol{U}_2\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$
  - 流形正则项：$\mathcal{R}(\mathcal{S},\mathcal{T}) = \frac{1}{\mathrm{d}^*}\sum_{i}^{\mathrm{d}^*}\theta_i\left[KL(\mathcal{S}_i\|\mathcal{T}_i) + KL(\mathcal{T}_i\|\mathcal{S}_i)\right]$

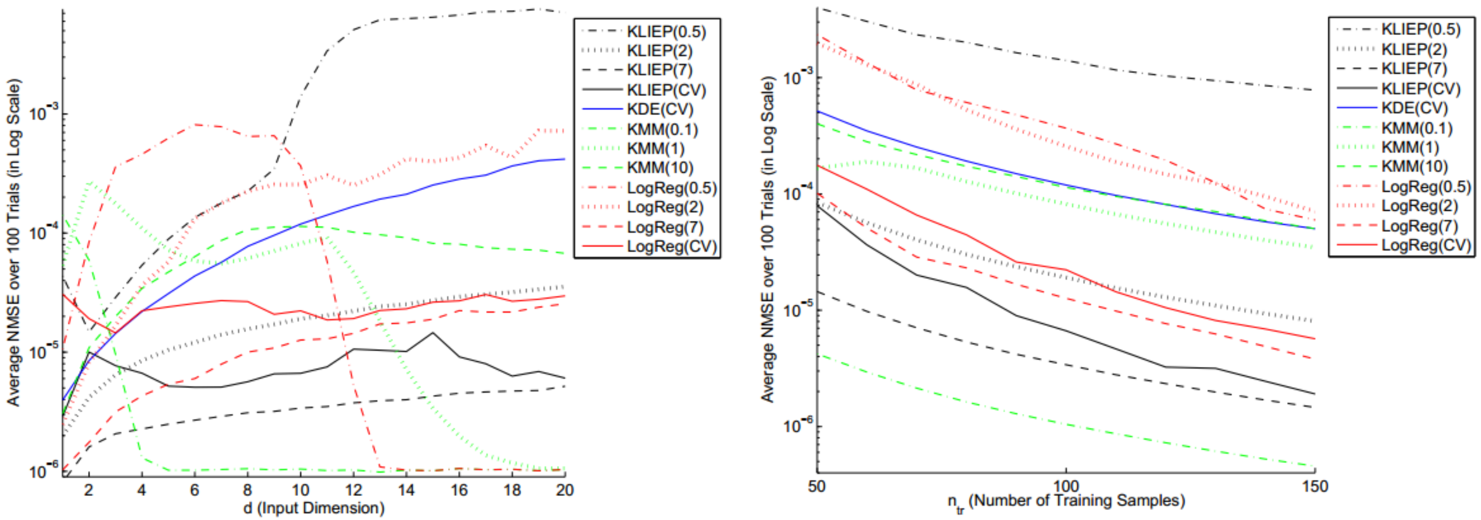


迁移学习：域适配问题

- Covariate Shift Adaptation [Sugiyama, JMLR-07]
  - 采用自然估计法估计源域和目标域的密度比例，然后进行实例权重的分配，最后迁移
  - 优化目标：

$$\underset{\{\alpha_\ell\}_{\ell=1}^{b}}{\text{maximize}} \left[\sum_{j=1}^{n_{\mathrm{te}}} \log \left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_j^{\mathrm{te}}) \right) \right]$$

$$\text{subject to } \sum_{i=1}^{n_{\mathrm{tr}}} \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i^{\mathrm{tr}}) = n_{\mathrm{tr}} \text{ and } \alpha_1, \alpha_2, \ldots, \alpha_b \geq 0.$$



(a) When input dimension is changed

(b) When training sample size is changed

迁移学习：域适配问题

- Adaptive SVM (ASVM) [Yang et al, ACM Multimedia-07]
  - 使用SVM模型，在适配和原始模型之间学习“数据函数”，达到模型迁移效果
  - 优化目标：

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.}\quad \xi_i \geq 0,\quad y_i\sum^{M} t_k f_k^a(\mathbf{x}_i) + y_i\mathbf{w}^T\phi(\mathbf{x}_i) \geq 1-\xi_i$$

- Multiple Convex Combination (MCC) [Schweikert, NIPS-09]
  - 对一些域适配的方法做集成学习
  - 优化目标：$F(\mathbf{x}) = \alpha f_T(\mathbf{x}) + (1-\alpha)\frac{1}{|\mathcal{S}|}\sum_{S\in\mathcal{S}} f_S(\mathbf{x})$

迁移学习：域适配问题

- 总结
  - 通常假设源域和目标域的数据有着相同的条件分布，或者在高维空间里，有着相同的条件分布
  - 这个假设是有一定局限性的，无法衡量源域和目标域之间相似性，可能发生负迁移

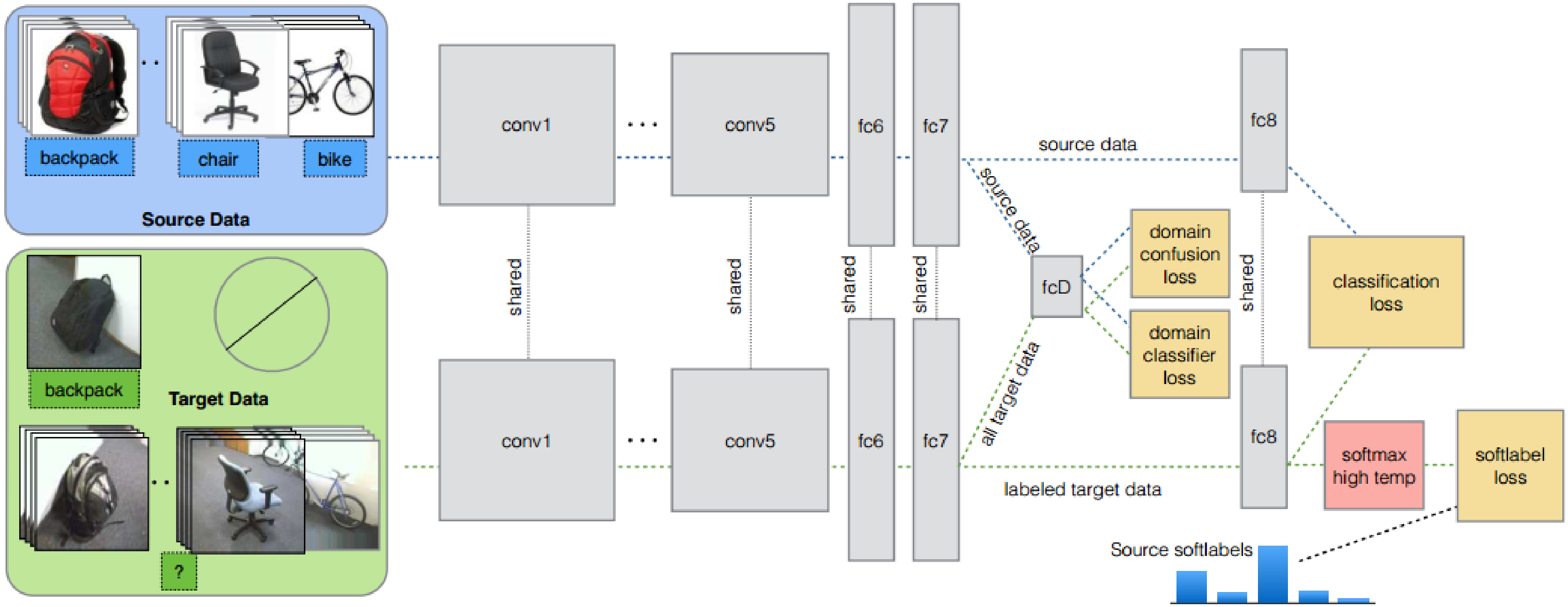
迁移学习：深度迁移学习

- 深度迁移学习
  - 利用深度神经网络的结构进行迁移学习
  - 神经网络学习非线性的特征表示
  - 层次性
  - 使得数据具有不可解释性
  - 表明在数据中具有某些不可变的成分，可以用来迁移
- 代表方法
  - Joint CNN [Tzeng, ICCV-15]
  - SHL-MDNN [Huang, ICASSP-13]
  - Deep Adaptation Network (DAN) [Long, ICML-15]
  - Joint Adaptation Networks [Long, CVPR-13]

迁移学习：深度迁移学习

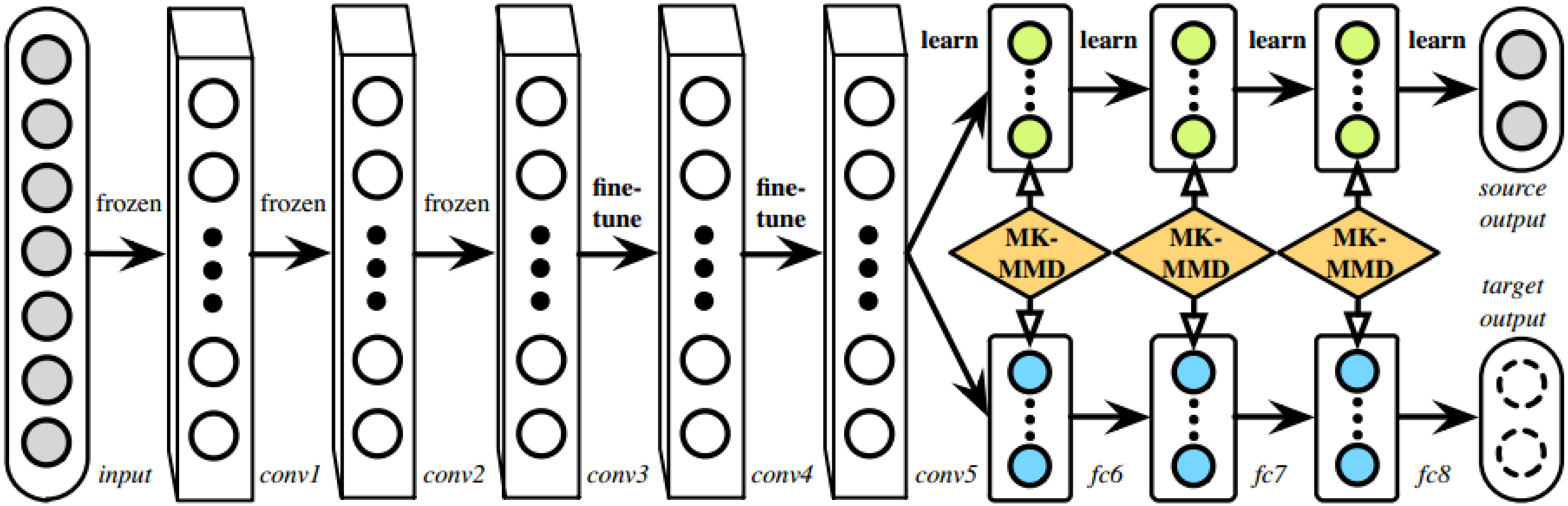
- Joint CNN [Tzeng, ICCV-15]
  - 针对有稀疏标记的目标域数据，用CNN同时优化域之间的距离和迁移学习任务的损失

$$\mathcal{L}(x_S, y_S, x_T, y_T, \theta_D; \theta_{\text{repr}}, \theta_C) = \mathcal{L}_C(x_S, y_S, x_T, y_T; \theta_{\text{repr}}, \theta_C) + \lambda \mathcal{L}_{\text{conf}}(x_S, x_T, \theta_D; \theta_{\text{repr}}) + \nu \mathcal{L}_{\text{soft}}(x_T, y_T; \theta_{\text{repr}}, \theta_C).$$



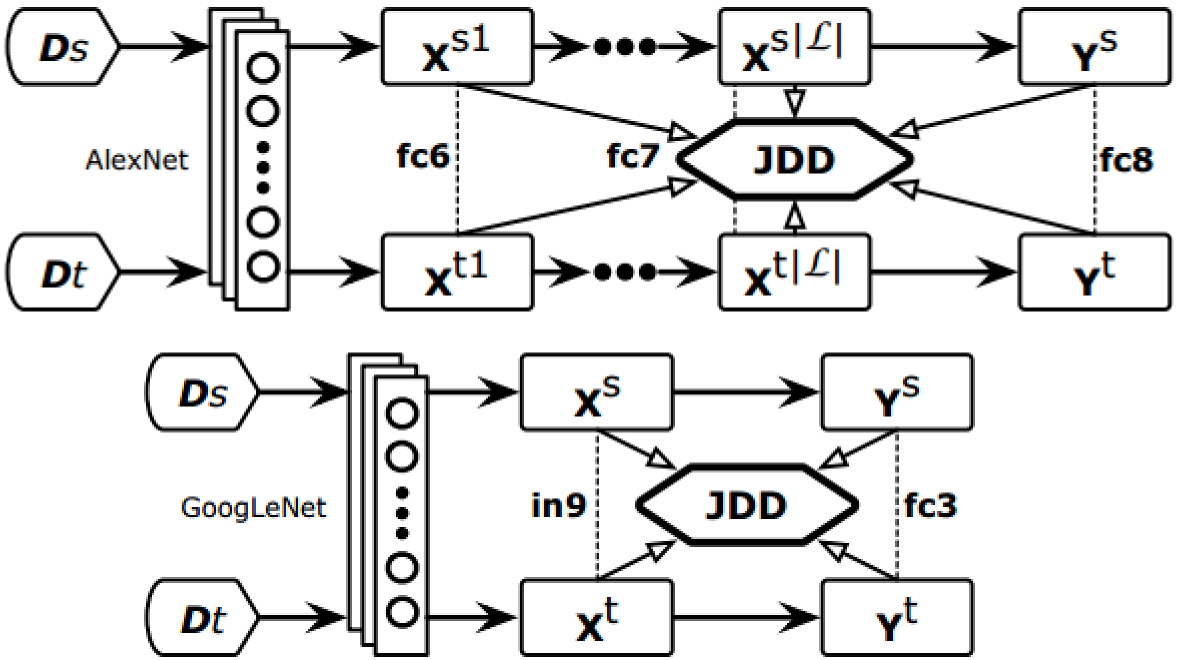
迁移学习：深度迁移学习

- Deep Adaptation Network (DAN) [Long, ICML-15]
  - 将CNN中与学习任务相关的隐藏层映射到再生核希尔伯特空间中，通过多核优化的方法最小化不同域之间的距离



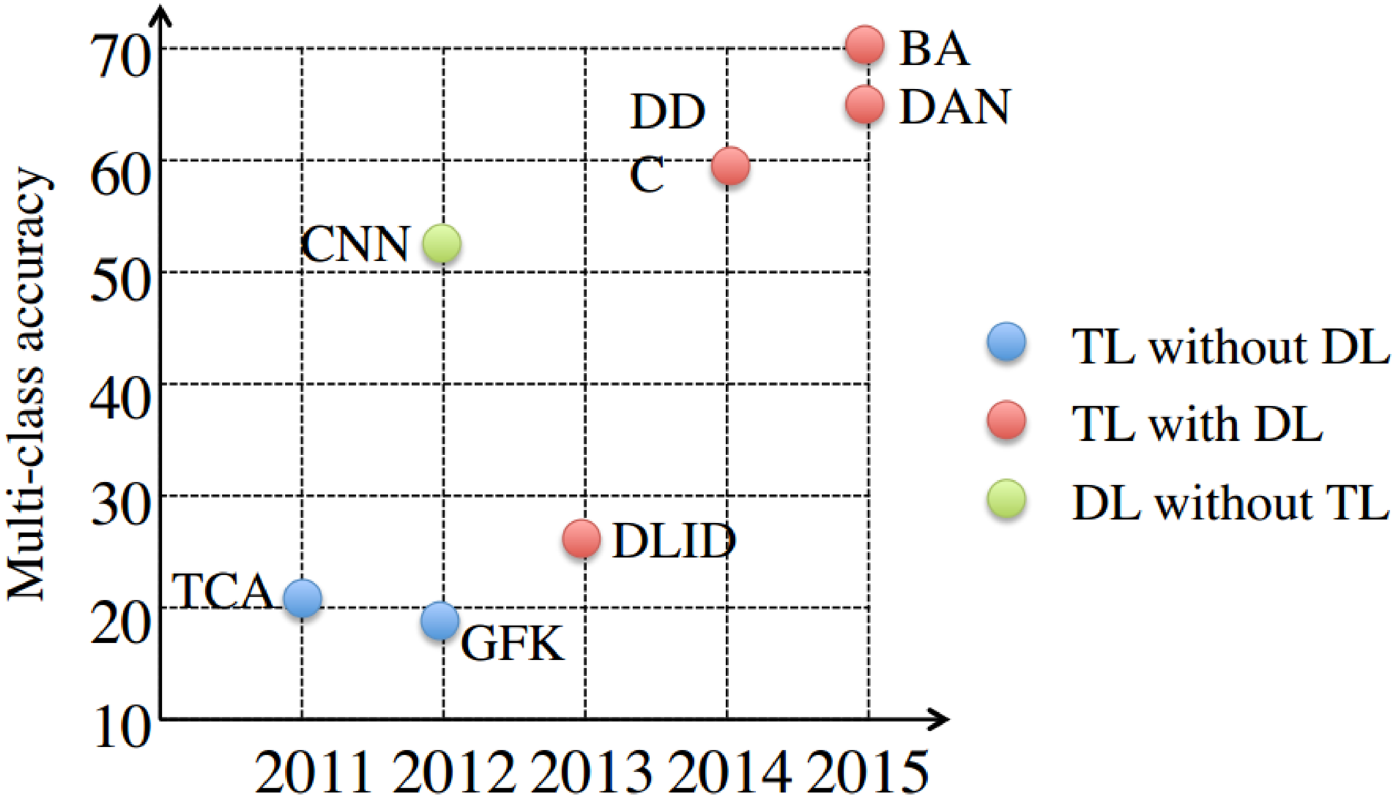
迁移学习：深度迁移学习

- Joint Adaptation Networks [Long, CVPR-15]
  - 提出一种新的联合分布距离度量关系，利用这种关系泛化深度模型的迁移学习能力，从而适配不同领域的数据分布。基于AlexNet和GoogLeNet重新优化了网络结构



迁移学习：深度迁移学习

- 总结：
  - 迁移学习大增强了模型的泛化能力
  - 深度学习可以深度表征域中的知识结构
  - 深度学习+迁移学习还可大有作为



问题与展望

- 迁移学习存在的问题：
  - 负迁移：无法判断域之间的相关性，导致负迁移
  - 缺乏理论支撑：尚未有统一的迁移学习理论
  - 相似度衡量：域之间的相似度通常依赖经验进行衡量，缺乏统一有效的相似度衡量方法
- 已有的基础
  - 负迁移：利用自编码器实现相关度较低的两个域之间的迁移（人脸→飞机）[Tan, AAAI-2017]
  - 理论支撑：利用物理学定律为迁移找到理论保证[Stewart, AAAI-17]
  - 相似度衡量：提出迁移度量学习，寻找行为之间相关性最高的域进行迁移[Al-Halah, ICPR-14]

迁移学习资源

- 综述
  - A survey on transfer learning [Pan and Yang, TKDE-10]
  - A survey of transfer learning [Weiss, Big data-15]
- 开源项目
  - http://www.cse.ust.hk/TL/
- 研究学者
  - Qiang Yang @ HKUST: http://www.cs.ust.hk/~qyang/
  - Sinno Jialin Pan @ NTU: http://www.ntu.edu.sg/home/sinnopan/
  - Fuzhen Zhuang @ ICT CAS: http://www.intsci.ac.cn/users/zhuangfuzhen/
  - Mingsheng Long @ THU: http://ise.thss.tsinghua.edu.cn/~mlong/
  - Lixin Duan @ Amazon: http://www.lxduan.info/
- 会议、期刊
  - 人工智能、机器学习：AAAI，ICML，ICJAI，NIPS，TNNLS，TIST，CVPR
  - 数据挖掘：TKDE，SIGKDD，ACL，WWW, SIGIR

- [DARPA] http://www.darpa.mil/ipto/programs/tl/tl.asp
- [Pan, TKDE-10] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.
- [Dai, ICML-07] Dai W, Yang Q, Xue G R, et al. Boosting for transfer learning[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 193-200.
- [Smola, ICML-08] Gretton A, Smola A, Huang J, et al. Covariate shift by kernel mean matching[J]. Dataset shift in machine learning, 2009, 3(4): 5.
- [Sugiyama, NIPS-07] Sugiyama M, Suzuki T, Kanamori T. Density ratio estimation in machine learning[M]. Cambridge University Press, 2012.
- [Pan, TKDE-11] Pan S J, Tsang I W, Kwok J T, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks, 2011, 22(2): 199-210.
- [Pan, WWW-10] Pan S J, Ni X, Sun J T, et al. Cross-domain sentiment classification via spectral feature alignment[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 751-760.
- [Duan, CVPR-12] Gong B, Shi Y, Sha F, et al. Geodesic flow kernel for unsupervised domain adaptation[C]//Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2066-2073.
- [Long, TKDE-15] Long M, Wang J, Sun J, et al. Domain invariant transfer kernel learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(6): 1519-1532.
- [Zhao, IJCAI-11] Zhao Z, Chen Y, Liu J, et al. Cross-people mobile-phone based activity recognition[C]//IJCAI. 2011, 2011(2011): 2545-2550.
- [Yao, CVPR-10] Yao Y, Doretto G. Boosting for transfer learning with multiple sources[C]//Computer vision and pattern recognition (CVPR), 2010 IEEE conference on. IEEE, 2010: 1855-1862.
- [Davis, ICML-09] Davis J, Domingos P. Deep transfer via second-order markov logic[C]//Proceedings of the 26th annual international conference on machine learning. ACM, 2009: 217-224.
- [Mihakova, AAAI-07] Mihalkova L, Huynh T, Mooney R J. Mapping and revising Markov logic networks for transfer learning[C]//AAAI. 2007, 7: 608-614.
- [Yang, ACM MM-07] Yang J, Yan R, Hauptmann A G. Cross-domain video concept detection using adaptive svms[C]//Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007: 188-197.

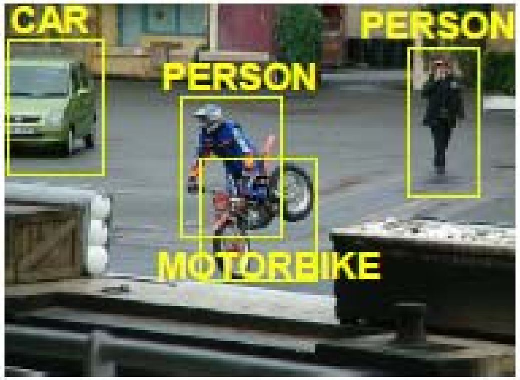
- [Schweikert, NIPS-09] Schweikert G, Rätsch G, Widmer C, et al. An empirical analysis of domain adaptation algorithms for genomic sequence analysis[C]//Advances in Neural Information Processing Systems. 2009: 1433-1440.
- [Duan, TNNLS-12] Duan L, Tsang I W, Xu D, et al. Domain adaptation from multiple sources via auxiliary classifiers[C]//Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 289-296.
- [Xu, ICONIP-12] Xu Z, Sun S. Multi-source transfer learning with multi-view adaboost[C]//International Conference on Neural Information Processing. Springer Berlin Heidelberg, 2012: 332-339.
- [Luo, CIKM-08] Luo P, Zhuang F, Xiong H, et al. Transfer learning from multiple source domains via consensus regularization[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 103-112.
- [Tan, KDD-15] Tan B, Song Y, Zhong E, et al. Transitive transfer learning[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015: 1155-1164.
- [Tan, AAAI-17] Tan B, Qiang Y, et al. Distant domain transfer learning. AAAI 2017.
- [Tzeng, ICCV-15] Tzeng E, Hoffman J, Darrell T, et al. Simultaneous deep transfer across domains and tasks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4068-4076.
- [Huang, ICASSP-13] Huang J T, Li J, Yu D, et al. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 7304-7308.
- [Long, ICML-15] Long M, Cao Y, Wang J, et al. Learning Transferable Features with Deep Adaptation Networks[C]//ICML. 2015: 97-105.
- [Long, CVPR-13] Long M, Wang J, Ding G, et al. Transfer feature learning with joint distribution adaptation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 2200-2207.
- [Stewart, AAAI-17] Stewart R, Ermon S. Label-free supervision of neural networks with physics and domain knowledge[C]//AAAI 2017.
- [Al-Halah, ICPR-14] Al-Halah Z, Rybok L, Stiefelhagen R. What to transfer? High-level semantics in transfer metric learning for action similarity[C]//Pattern Recognition (ICPR), 2014 22nd International Conference on. IEEE, 2014: 2775-2780.[Oquab, CVPR-14] Oquab, Bottou, Laptev, Sivic: Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. CVPR 2014
- [Weiss, Bigdata-15] Weiss K, Khoshgoftaar T M, Wang D D. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 1-40.

Weakly-supervised learning

Credits to Cordelia Schmid & Marco Loog

From Coarse to Fine...

Weakly supervised
large-scale learning



Object detection

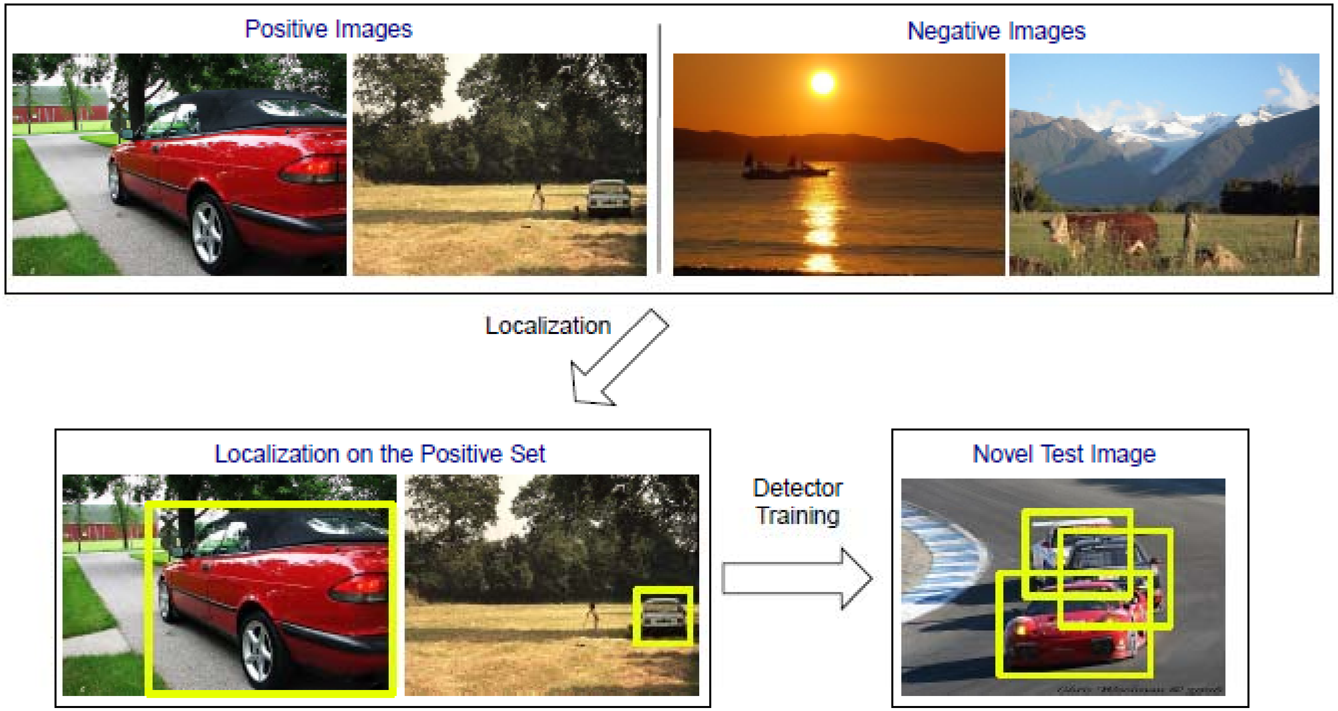
Action recognition

From Coarse to Fine...

- Coarse labeling much easier than detailed
- Given coarse label, can we trace back its "cause" at finer resolution?
  - E.g. given MR images with and without diagnosis of particular brain disease, can we localize voxels in which the disease manifests itself?
  - Given X-rays with and without TB, can we segment regions with TB?

Weakly-supervised learning for images

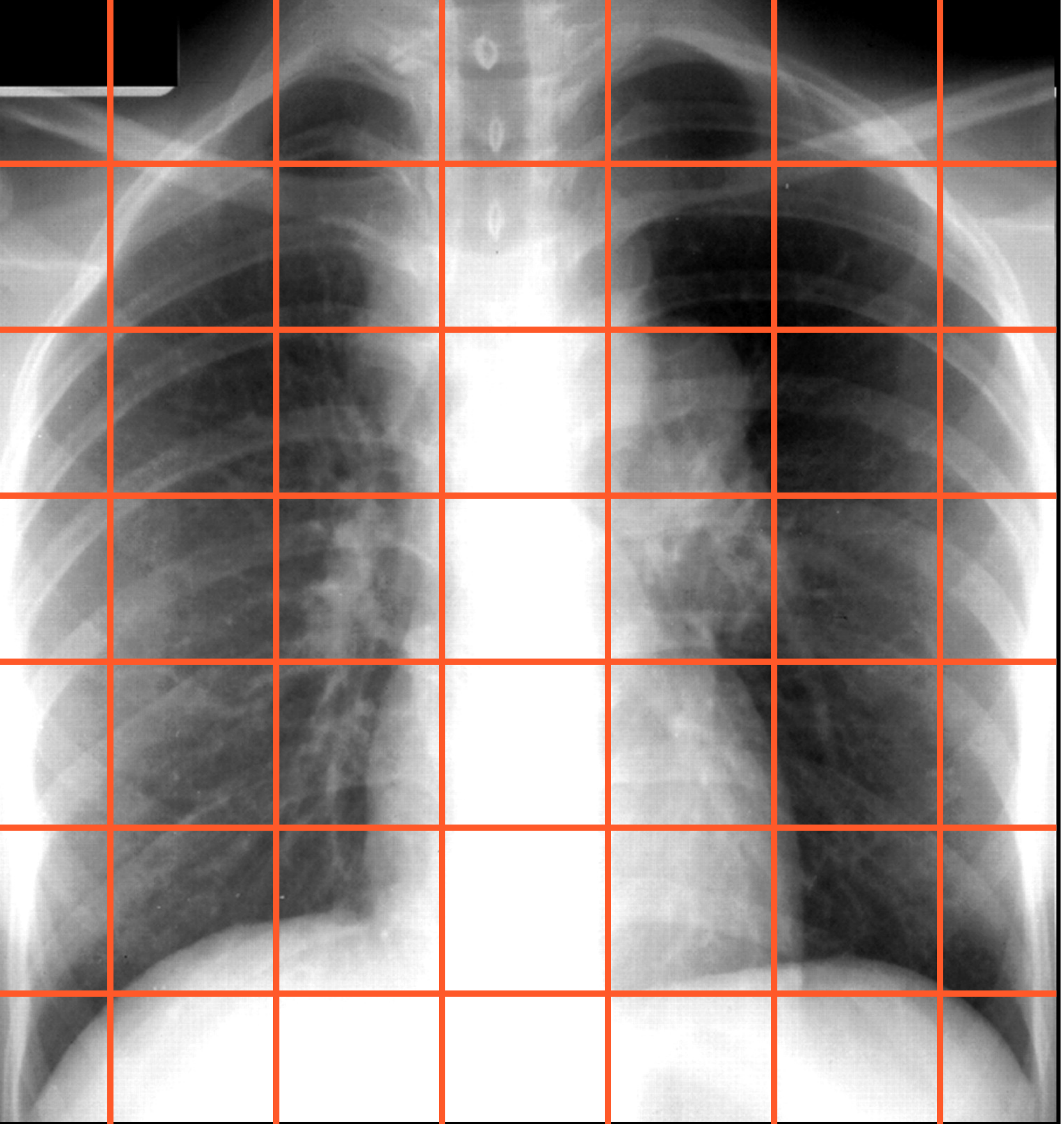
- Given a set of images with positive and negative labels, determine the object region, learn detector
- Avoids costly annotation of object regions



Multiple Instance Learning

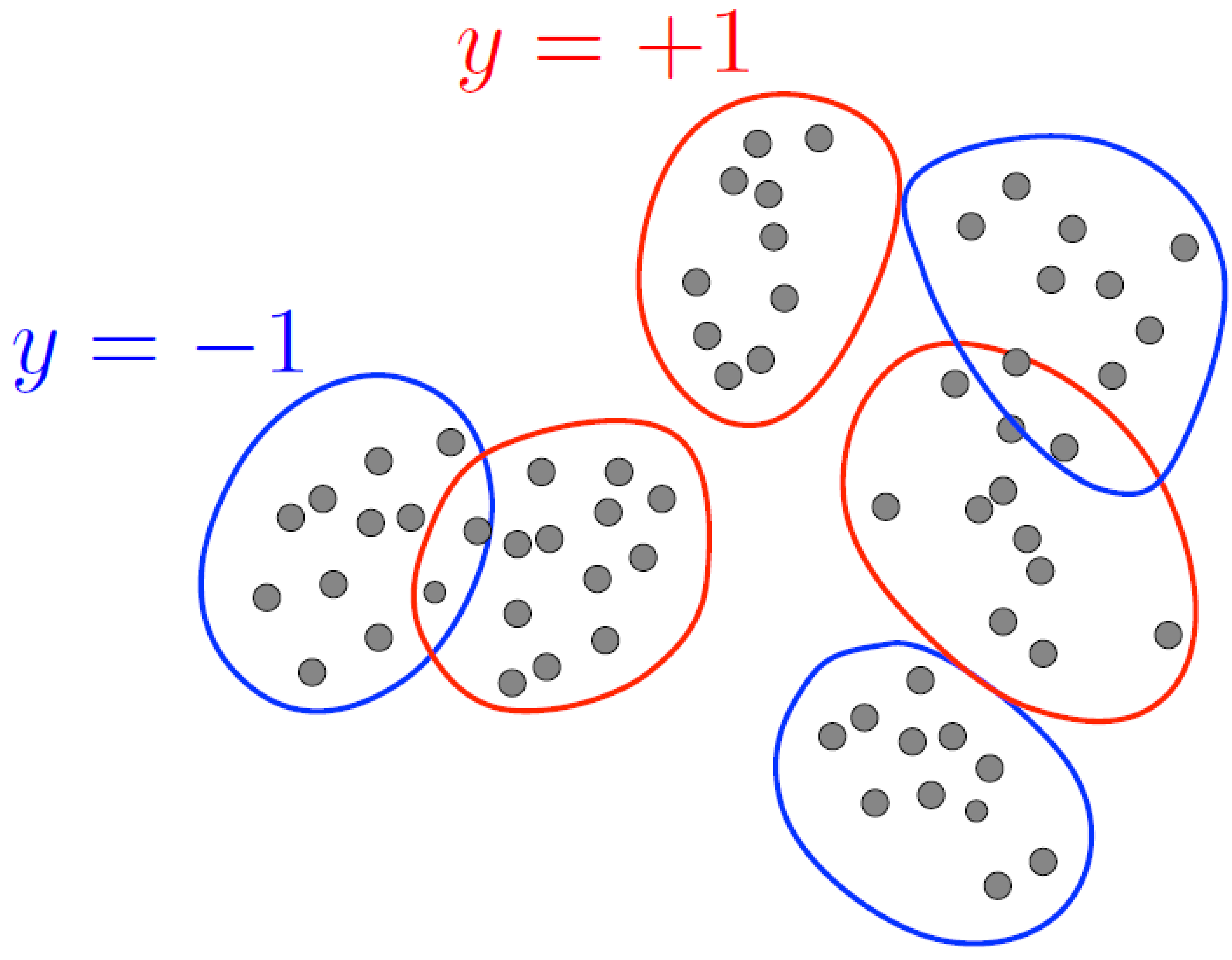
- Represent an object by a collection or [multi-]set of feature vectors, i.e., by means of **multiple instances**
- Sets or so-called bags [of instances] to represent every object, but still a **single label** per object
  - Vectors assumed to be in same feature space
  - Set sizes do not have to be equal

[16,17,18]



E.g. a Bag with Instances

Graphically Speaking



Original Goal is Twofold

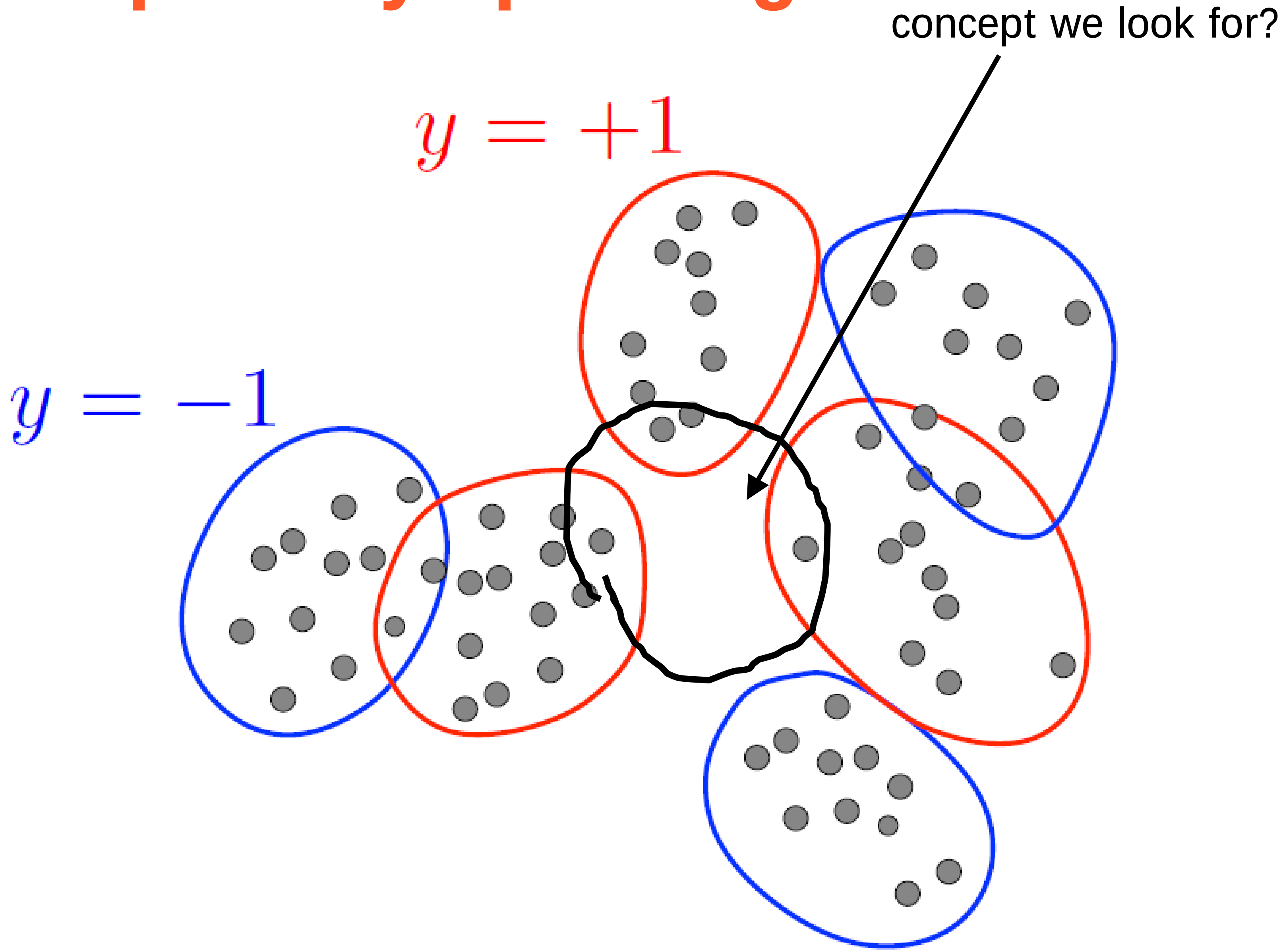
- MIL aims to classify new and previously unseen bags as accurate as possible

- But also : MIL tries to discover a **concept** that determines the positive class
  - Concepts are instances uniquely identifying a class

  - First more easy than second goal
  - Latter is often not considered in MIL literature
  - But latter is needed to go from coarse to fine...

[19,20]

Graphically Speaking



Most Promising MIL Approaches

- MILES and variations
  - Describe a bag by means of similarities to instances
- Dissimilarity-based approaches
  - Describe a bag by means of distances to other bags
- Good for classification, not for coarse to fine
  - Both do not discover something like a concept

[17,21,22]

Some Core Challenges

- Little known about relationship bag labels vs. instance labels : what assumptions are of use?
- To what extent is it at all possible to go to more details? What assumptions are necessary?
- How stable is coarse-to-fine process and how can we stabilize it?

- Overall, little is known at theoretical level...

[20,23,24]

Some More Core Challenges

- Stronger methods needed that really can discover concepts or similarly
- Promising techniques based on convolutional networks [e.g. Somol, Rosmalen, Bazzani, Bency, ...]
  - They use clever design of CNN
  - Relations to CRFs, MRFs?

[25,26,27]

Slightly Biased Reference List

- [1] Chapelle, Schölkopf, Zien, “Semi-supervised learning”, MIT press, 2006
- [2] Zhu, “Semi-supervised learning literature survey”, University of Wisconsin, TR 1530, 2008
- [3] Loog, “Constrained parameter estimation for semi-supervised learning: the case of the nearest mean classifier”, ECML, 2010
- [4] Loog, Jensen, "Constrained Log-Likelihood-Based Semi-supervised Linear Discriminant Analysis", S+SSPR, 2012
- [5] Loog, "Semi-supervised Linear Discriminant Analysis Using Moment Constraints", Partially Supervised Learning, 2012
- [6] Abney, “Understanding the Yarowsky algorithm,” Computational Linguistics, 2004
- [7] Ben-David, Lu, Pal, “Does unlabeled data provably help?”, COLT, 2008
- [8] Cozman, Cohen, “Risks of semi-supervised learning,” in Semi-Supervised Learning. MIT Press, 2006
- [9] Lafferty, Wasserman, “Statistical analysis of semi-supervised regression”, NIPS, 2007
- [10] Singh, Nowak, Zhu, “Unlabeled data: Now it helps, now it doesn't”, NIPS, 2008
- [11] Li, Zhou, "Towards making unlabeled data never hurt", ICML 2011
- [12] Krijthe, Loog, “Implicitly constrained semi-supervised least squares classification,” IDA, 2015
- [13] Loog, “Contrastive Pessimistic Likelihood Estimation for Semi-Supervised Classification”, IEEE TPAMI, 2016
- [14] Krijthe, Loog, “Robust semi-supervised least squares classification by implicit constraints,” PR, 2017
- [15] Krijthe, Loog, “Projected Estimators for Robust Semi-supervised Classification,” ML, 2017
- [16] Amores, “Multiple instance classification: Review, taxonomy and comparative study”, AI, 2013.
- [17] Cheplygina, Tax, Loog, “Multiple instancelearning with bag dissimilarities”, PR, 2015
- [18] Foulds, Frank, “A review of multi-instance learning assumptions”, KER, 2010
- [19] Dietterich, Lathrop, Lozano-Perez, “Solving the multiple instance problem with axis-parallel rectangles”, AI, 1997
- [20] Cheplygina, Sørensen, Tax, de Bruijne, Loog, “Label Stability in Multiple Instance Learning,” MICCAI, 2015
- [21] Chen, Bi, Wang, “MILES: Multiple-instance learning via embedded instance selection”, IEEE TPAMI, 2006
- [22] Tax, Loog, Duin, Cheplygina, Lee, “Bag dissimilarities for multiple instance learning”, SIMBAD, 2011
- [23] Ray, Craven, “Supervised versus multiple instance learning,” ICML, 2005
- [24] Alpydin, Cheplygina, Loog, Tax, “Single-vs. multiple-instance classification,” PR, 2015

Slightly Biased Reference List

- [25] Pevny, Somol, “Using Neural Network Formalism to Solve Multiple-Instance Problems”, arXiv, 2016
- [26] Rosmalen, “Localization Maps: A framework for weakly supervised positive class localization,” MSc thesis, TU Delft, 2016
- [27] Bergamo, et al. “Self-taught object localization with deep networks,” arXiv preprint, 2014
- [28] Pan, Yang, “A survey on transfer learning,”, IEEE TKDE, 2010
- [29] Quionero-Candela, Sugiyama, Schwaighofer, Lawrence, “Dataset shift in machine learning,” The MIT Press, 2009
- [30] Ben-David, Blitzer, Crammer, Kulesza, Pereira, Vaughan, “A theory of learning from different domains,” ML, 2010
- [31] Ben-David, Lu, Pál, “Impossibility theorems for domain adaptation,” AISTATS, 2010
- [32] Ben-David, Urner, “On the hardness of domain adaptation and the utility of unlabeled target samples,” ALT, 2012
- [33] Mansour, Mohri, Rostamizadeh, “Domain adaptation: Learning bounds and algorithms,” COLT, 2009
- [34] Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” J.St.Plan.Inf, 2000
- [35] Cortes, Mohri, “Domain adaptation and sample bias correction theory and algorithm for regression,” Theoretical CS, 2014
- [36] Sugiyama, Krauledat, Müller, “Covariate shift adaptation by importance weighted cross validation,” JMLR, 2007
- [37] Cortes, Mohri, “Domain adaptation and sample bias correction theory and algorithm for regression,” Th.Comp.Sc., 2014
- [38] Swaminathan, Joachims, “The self-normalized estimator for counterfactual learning,” NIPS, 2015
- [39] Cohn, Atlas, Ladner, “Improving generalization with active learning”, ML, 1994
- [40] Cohn, Ghahramani, Jordan, “Active learning with statistical models”, JAIR, 1996
- [41] Settles, “Active Learning Literature Survey”, TR 1648, University of Wisconsin–Madison, 2010
- [42] Tuia, Volpi, et al., “A survey of active learning algorithms for supervised remote sensing image classification”, J-STSP, 2011
- [43] Yang, Loog, “A benchmark and comparison of active learning methods for logistic regression,” arXiv, 2016
- [44] Loog, Yang, “An Empirical Investigation into the Inconsistency of Sequential Active Learning,” ICPR 2016
- [45] Hanneke, “Theory of disagreement-based active learning,” Foundations and Trends in Machine Learning, 2014
- [46] Yang, Hanneke, Carbonell, “A theory of transfer learning with applications to active learning,” ML, 2013
- [47] Kouw, Van der Maaten, Krijthe, Loog, “Feature-Level Domain Adaptation,” JMLR, 2016
- [48] De Bruijne, “Machine learning approaches in medical image analysis: From detection to diagnosis,” MedIA, 2016

The Obligatory Quote?

- *”Practice two things
  in your dealings with disease :
  either help
  or do not harm the patient”*
  – Hippocrates