最小二乘估计似乎不相关回归和迁移学习简介

线性回归模型

$$y = X\beta + e, \qquad E(e) = 0, \qquad \mathrm{Cov}(e) = \sigma^2 I$$

最小二乘估计 $\hat{\beta} = (X'X)^{-1}X'y$ 为 β 的最小二乘估计

优化问题 $Q(\beta) = \|e\|^2 = \|y - X\beta\|^2 = (y - X\beta)'(y - X\beta)$

评价指标：均方误差 $MSE = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 / N = \sum_{i=1}^{N}(Y_i - X_i\hat{\beta})^2 / N$

无偏性 有效性 相合性 一致最小方差无偏估计
Uniformly minimum variance unbiased estimate UMVUE

高斯－马尔可夫定理(Gauss-Markov Theorem)

- 在线性回归模型中，如果误差满足同方差且互不相关，则回归系数的最佳线性无偏估计(BLUE, Best Linear unbiased estimator)就是普通最小二乘估计。
- 高斯**高斯－马尔可夫条件**
  - $\mathrm{E}(\varepsilon_i) = 0$ ，$\forall i$（零均值），
  - $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$ ，$\forall i$（同方差），
  - $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ ，$\forall i \neq j$（不相关）。
- 最佳指的是方差最小
- C-R下界 $Var_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)}$
- fisher 信息量 $J(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln f(x|\theta)\right]^2\right\} = -E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\right\}$
- 是衡量一个无偏估计器是否有效的重要工具，也就是说，给定一个无偏估计器，我们可以利用C-R下界去判断这个估计器是否是最优的。

- 进一步若 $\varepsilon \sim N(0, \sigma^2 I_N)$ 普通最小二乘估计是一致最小方差无偏估计

似不相关回归模型

假设共有n个线性回归模型，每个方程共有T个观测值，T>n，在第i个方程中，共有K_i个解释变量。

第i个方程可以写为

$$\underbrace{y_i}_{T\times 1} = \underbrace{X_i}_{T\times K_i}\ \underbrace{\beta_i}_{K_i\times 1} + \underbrace{\varepsilon_i}_{T\times 1} \quad (i=1,\ 2,\cdots,\ n)$$

将所有方程叠放在一起可得

$$y \equiv \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{nT\times 1} = \underbrace{\begin{pmatrix} X_1 & & 0 \\ & X_2 & \\ & & \ddots \\ 0 & & X_n \end{pmatrix}}_{nT\times \sum_{i=1}^{n} K_i} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}}_{\sum_{i=1}^{n} K_i\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{nT\times 1} \equiv X\beta+\varepsilon$$

误差的协方差矩阵

$$\Omega \equiv \mathrm{Var}\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}=\mathrm{E}\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}\begin{pmatrix}\varepsilon_1' & \varepsilon_2' & \cdots & \varepsilon_n'\end{pmatrix}=\mathrm{E}\begin{pmatrix}\varepsilon_1\varepsilon_1' & \varepsilon_1\varepsilon_2' & \cdots & \varepsilon_1\varepsilon_n'\\ \varepsilon_2\varepsilon_1' & \varepsilon_2\varepsilon_2' & \cdots & \varepsilon_2\varepsilon_n'\\ & & \vdots & \\ \varepsilon_n\varepsilon_1' & \varepsilon_n\varepsilon_2' & \cdots & \varepsilon_n\varepsilon_n'\end{pmatrix}_{nT\times nT}$$

假设同一方程不同期的扰动项不存在自相关，且方差也相同，记第i个方程的方差为σ_{ii}。则协方差矩阵Ω中主对角线上的第$(i，i)$个矩阵为

$$\mathrm{E}\left(\varepsilon_i\varepsilon_i'\right)=\sigma_{ii}\mathrm{I}_T$$

假设不同方程的扰动项之间存在同期相关（如果是横截面数据，则指的是不同方程对应的扰动项之间存在相关性），即

$$\mathrm{E}\left(\varepsilon_{it}\varepsilon_{js}\right)=\begin{cases}\sigma_{ij}, & t=s\\ 0, & t\neq s\end{cases}$$

综合以上结果可知

$$\Omega=\begin{pmatrix}\sigma_{11}I_T & \sigma_{12}I_T & \cdots & \sigma_{1n}I_T\\ \sigma_{21}I_T & \sigma_{22}I_T & \cdots & \sigma_{2n}I_T\\ \vdots & \vdots & & \vdots\\ \sigma_{n1}I_T & \sigma_{n2}I_T & \cdots & \sigma_{nn}I_T\end{pmatrix}=\begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n}\\ \vdots & \vdots & & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}\end{pmatrix}\otimes I_T=\Sigma\otimes I_T$$

由于Ω不是单位矩阵，故用OLS估计这个多方程系统y＝Xβ＋ε不是最有效率的。假设Ω已知，则GLS是最有效率的估计方法：

$$\min\ Q(\beta)=\min\,(y-X\beta)'\Sigma^{-1}(y-X\beta).$$

假设n=2

- $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$ 即 $y=X\beta+\varepsilon$

$$y = X\beta + e, \qquad E(e) = 0, \qquad \mathrm{Cov}(e) = \sigma^2\Sigma \tag{4.3.1}$$

的参数 β, σ^2 的估计问题，其中 $\Sigma > 0$.

因为假设了 $\Sigma > 0$, 故存在惟一的正定对称阵 $\Sigma^{\frac{1}{2}}$. 用 $\Sigma^{-\frac{1}{2}}$ 左乘 (4.3.1), 并记 $\tilde{y} = \Sigma^{-\frac{1}{2}} y$, $\tilde{X} = \Sigma^{-\frac{1}{2}} X$, $u = \Sigma^{-\frac{1}{2}} e$, 则得到

$$\tilde{y} = \tilde{X}\beta + u, \qquad E(u) = 0, \qquad \mathrm{Cov}(u) = \sigma^2 I, \tag{4.3.2}$$

这就化为以前讨论过的情形了.

对模型 (4.3.2) 用最小二乘法求 β 的 LS 解，即解 $Q(\beta) = \|\tilde{y} - \tilde{X}\beta\|^2$ 的最小值问题. 等价地，解

$$\min\ Q(\beta) = \min\,(y - X\beta)'\Sigma^{-1}(y - X\beta). \tag{4.3.3}$$

正则方程组为

$$X'\Sigma^{-1}X\beta = X'\Sigma^{-1}y, \tag{4.3.4}$$

于是， β 的 LS 解为

$$\beta^* = (X'\Sigma^{-1}X)^- X'\Sigma^{-1}y, \tag{4.3.5}$$

称为广义最小二乘解. 特别，当 $\Sigma = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_n^2)$, σ_i^2, $i = 1, \cdots, n$ 已知时，称 β^* 为加权最小二乘解.

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

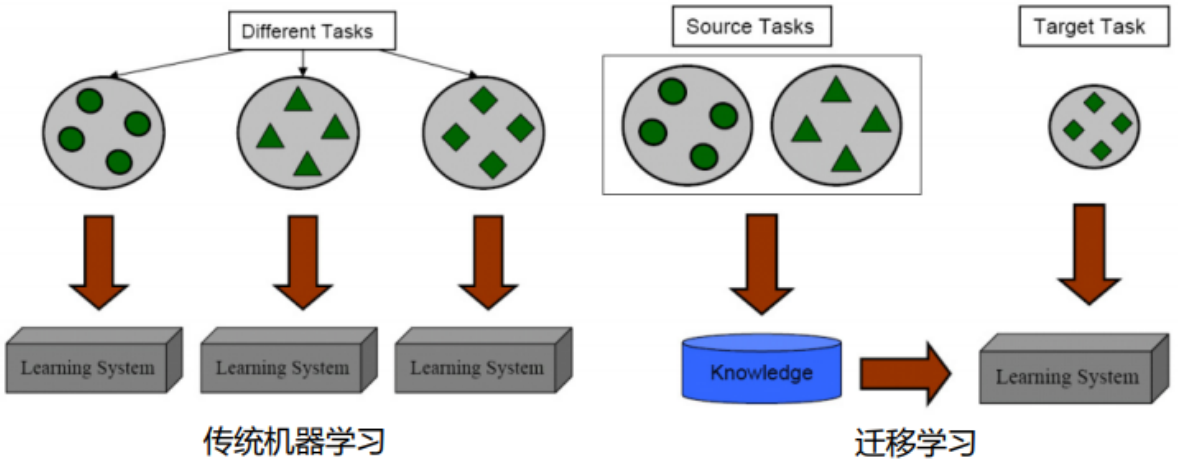
对已有知识的重用是必要的

- 迁移学习与其他已有概念相比，着重强调学习任务之间的**相关性**，并利用这种相关性完成知识之间的迁移。

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

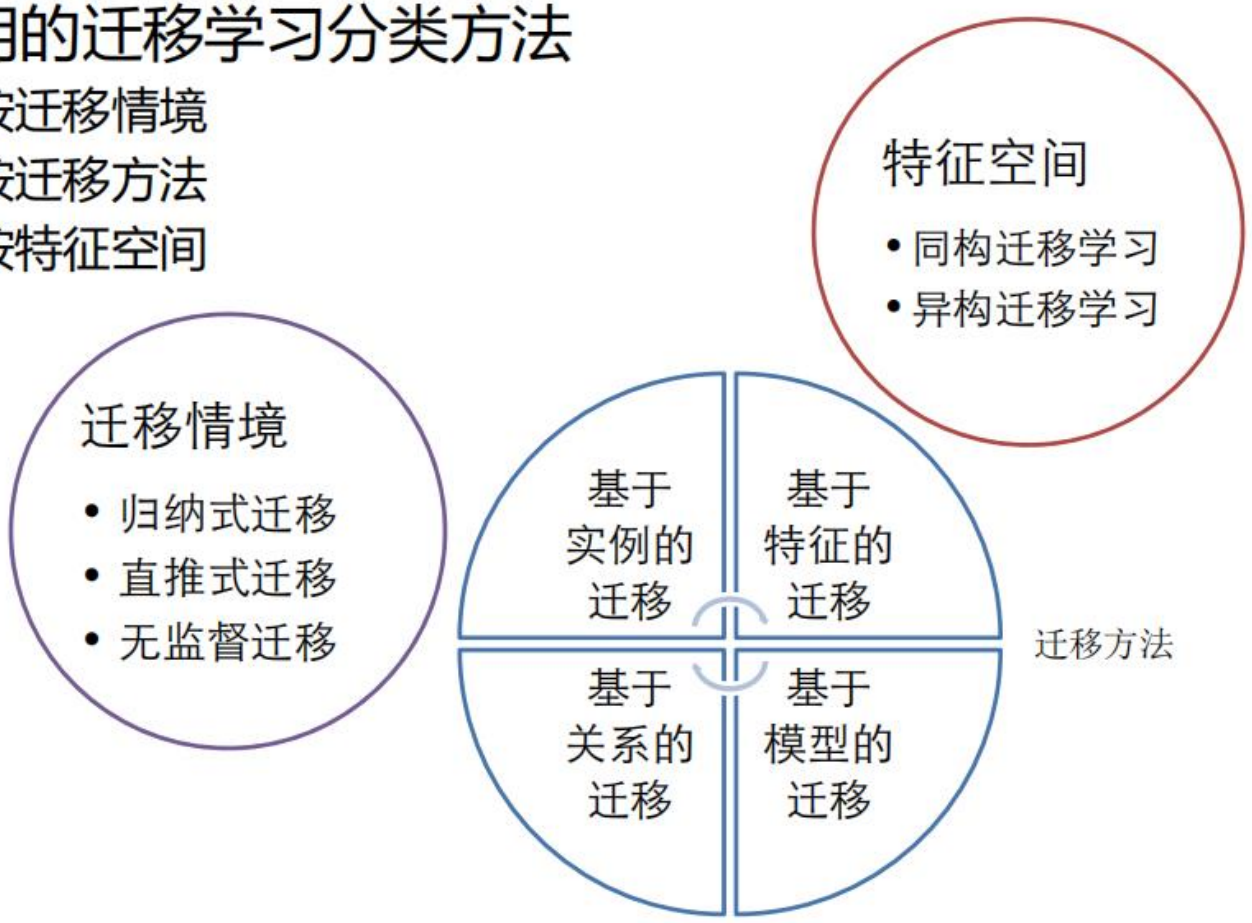


迁移学习的定义

- 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 $\mathcal{T}_S \neq \mathcal{T}_T$

迁移学习的分类方法

- 常用的迁移学习分类方法
  - 按迁移情境
  - 按迁移方法
  - 按特征空间



迁移学习的分类方法

- 按迁移情境分类：

归纳式迁移学习 (inductive transfer learning)

- 源域和目标域的学习任务不同

直推式迁移学习 (transductive transfer learning)

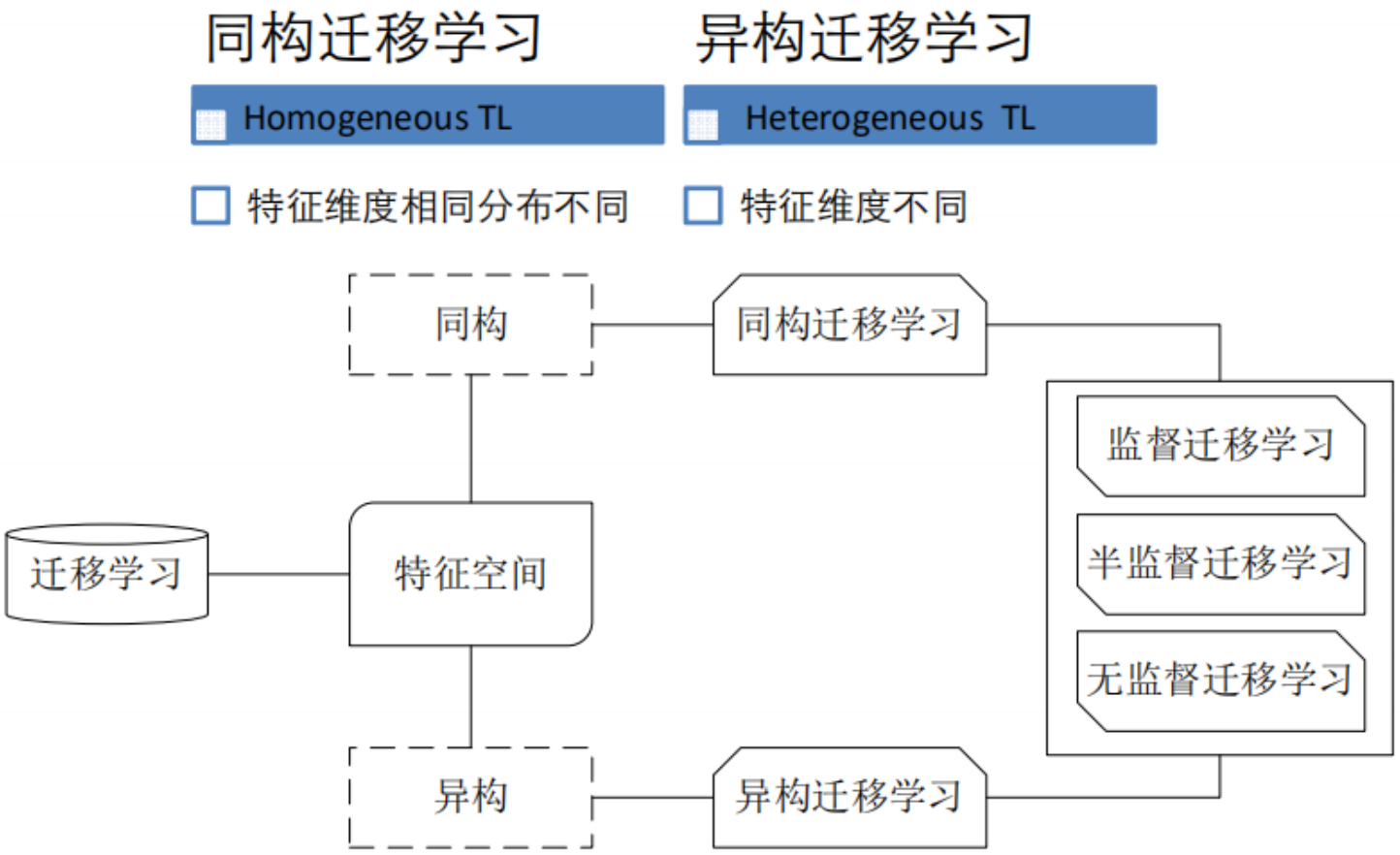
- 源域和目标域不同，学习任务相同

无监督迁移学习 (unsupervised transfer learning)

- 源域和目标域均没有标签

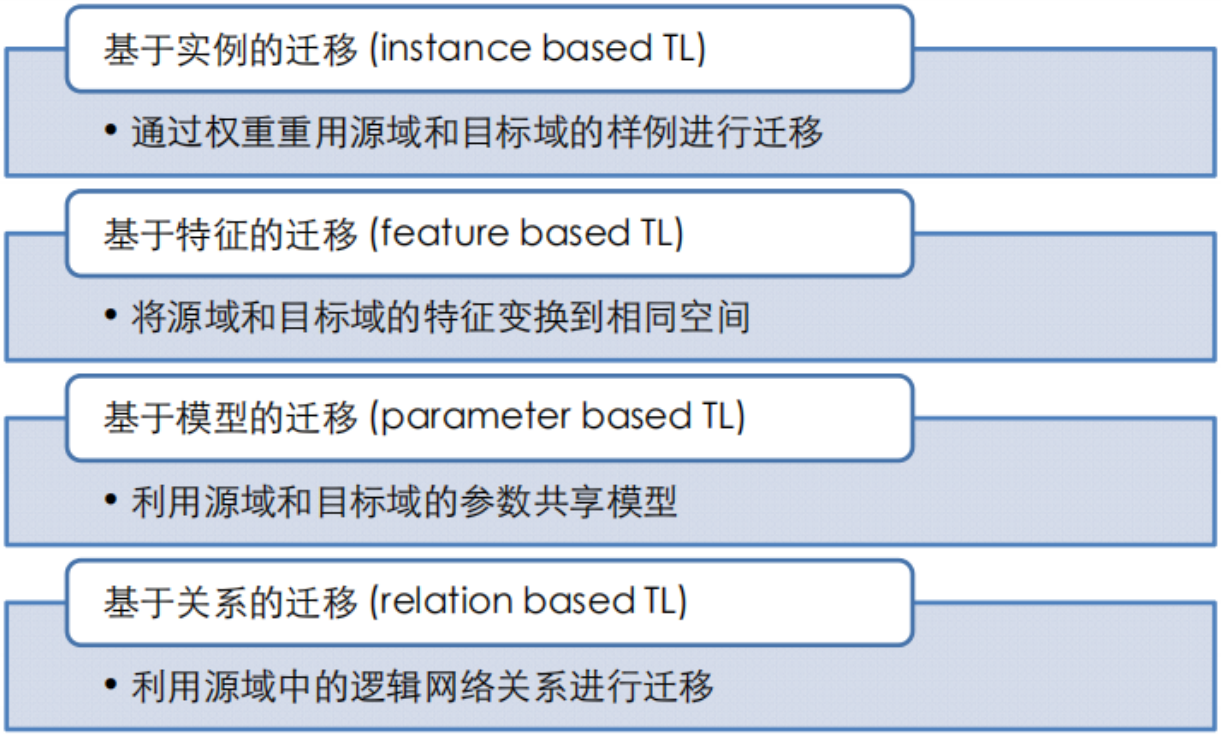
学习情境		源域和目标域	源域和目标域任务
传统机器学习		相同	相同
迁移学习	归纳式迁移/ 无监督迁移	相同	不同但相关
		不同但相关	不同但相关
	直推式迁移	不同但相关	相同

迁移学习的分类方法


同构迁移学习
异构迁移学习
Homogeneous TL
Heterogeneous TL
特征维度相同分布不同
特征维度不同
同构
同构迁移学习
监督迁移学习
半监督迁移学习
无监督迁移学习
迁移学习
特征空间
异构
异构迁移学习

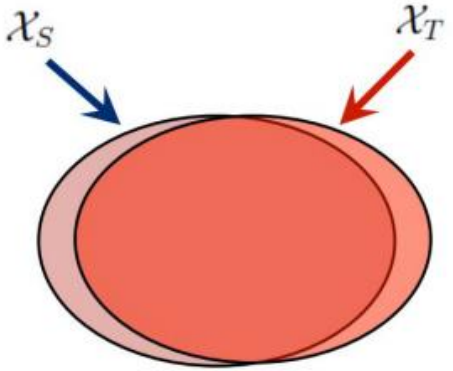
迁移学习的分类方法

- 按迁移方法分类：



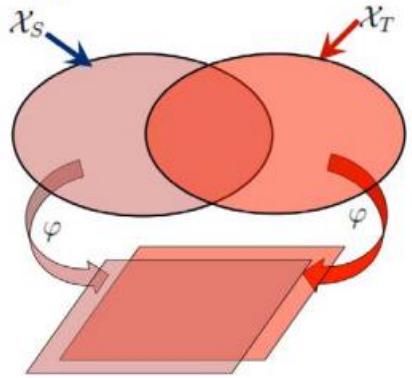
迁移方法

- 基于实例的迁移学习方法
  - 假设：源域中的一些数据和目标域会共享很多共同的特征
  - 方法：对源域进行instance reweighting，筛选出与目标域数据相似度高的数据，然后进行训练学习
  - 代表工作：
    - TrAdaBoost [Dai, ICML-07]
    - Kernel Mean Matching (KMM) [Smola, ICML-08]
    - Density ratio estimation [Sugiyama, NIPS-07]
  - 优点：
    - 方法较简单，实现容易
  - 缺点：
    - 权重选择与相似度度量依赖经验
    - 源域和目标域的数据分布往往不同



迁移方法

- 基于特征的迁移学习方法
  - 假设：源域和目标域仅仅有一些交叉特征
  - 方法：通过特征变换，将两个域的数据变换到同一特征空间，然后进行传统的机器学习
  - 代表工作：
    - Transfer component analysis (TCA) [Pan, TKDE-11]
    - Spectral Feature Alignment (SFA) [Pan, WWW-10]
    - Geodesic flow kernel (GFK) [Duan, CVPR-12]
    - Transfer kernel learning (TKL) [Long, TKDE-15]
  - 优点：
    - 大多数方法采用
    - 特征选择与变换可以取得好效果
  - 缺点：
    - 往往是一个优化问题，难求解
    - 容易发生过适配

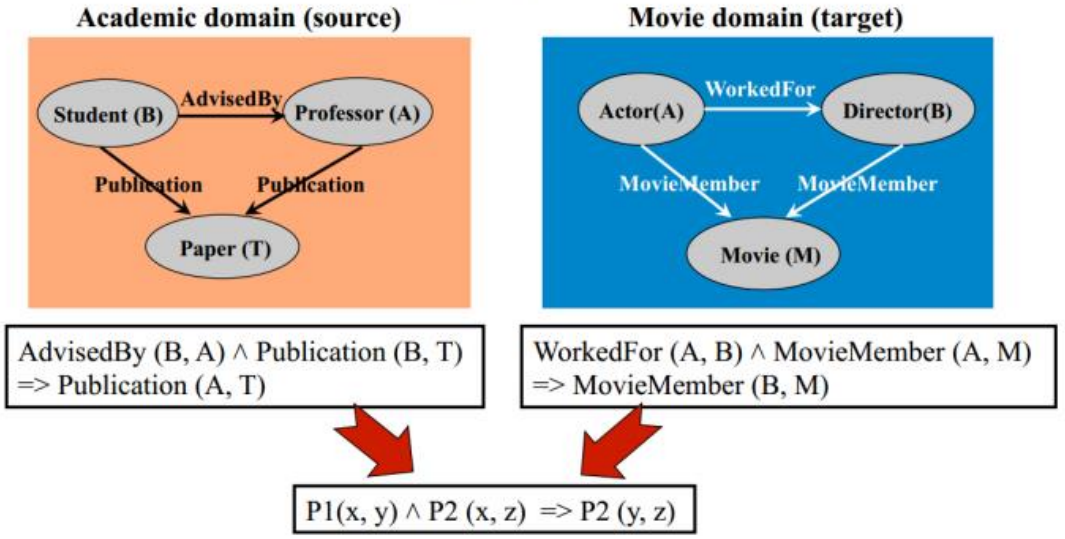


迁移方法

- 基于模型的迁移学习方法
  - 假设：源域和目标域可以共享一些模型参数
  - 方法：由源域学习到的模型运用到目标域上，再根据目标域学习新的模型
  - 代表工作：
    - TransEMDT [Zhao, IJCAI-11]
    - TRCNN [Oquab, CVPR-14]
    - TaskTrAdaBoost [Yao, CVPR-10]
  - 优点：
    - 模型间存在相似性，可以被利用
  - 缺点：
    - 模型参数不易收敛

迁移方法

- 基于关系的迁移学习方法
  - 假设：如果两个域是相似的，那么它们会共享某种相似关系
  - 方法：利用源域学习逻辑关系网络，再应用于目标域上
  - 代表工作：
    - Predicate mapping and revising [Mihalkova, AAAI-07],
    - Second-order Markov Logic [Davis, ICML-09]



研究热点

- 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

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

  - 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

- 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]

研究热点

- 多源迁移学习
  - 问题定义：多个源域和目标域，如何进行有效的域筛选，从而进行迁移？

- 总结：
  - 多源迁移学习可以有效利用存在的多个可用域，综合起来进行迁移，达到较好的效果
  - 如何衡量多个域之间的相关性还是一个问题
  - 对多个域的利用方法也存在一定挑战性

存在问题

- 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]

我的想法

- 将 $Y_1 = X_1\beta_1 + \varepsilon_1$ 当成主要问题，利用 ε_1 和 ε_2 的相关性，利用 $Y_2 = X_2\beta_2 + \varepsilon_2$ 的信息改进 β_1 的估计

方法：ε 服从正态分布时用条件期望改进

ε 服从非正态分布时用典型相关变量改进