Women in Mathematics Directed Reading Program: Transfer learning from a statistical perspective

Mentee: Kun Zhu, Samantha Wallis and Xinran Zheng

Mentor: Jie Jian

Department of Statistics and Actuarial Science
Faculty of Mathematics
University of Waterloo

Dec 2023

Outline

- Overview of Accomplishments
  - Our reading list
  - Gained knowledge
- Motivation: Why do we need transfer learning?
  - General ideas of transfer learning
  - Motivating examples
  - How can transfer learning assist traditional statistical models?
- Basic concepts and methods in transfer learning
- One specific transfer learning method: Instance Weighting Strategy
  - Measuring the discrepancy between distributions: MMD
  - Instance weighting strategy: KMM
  - Examples
- Summary
  - Difficulties we encounter and how to solve them
  - Future directions
  - Our understanding on EDI (equity, diversity, and inclusion)

Overview of Accomplishments

Our reading list

- Survey paper 1 (1 week):
  Torrey, Lisa, and Jude Shavlik. "Transfer learning." Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI global, 2010. 242-264.
- Survey paper 2 (5 weeks):
  Zhuang, Fuzhen, et al. "A comprehensive survey on transfer learning." Proceedings of the IEEE 109.1 (2020): 43-76.
- Maximum Mean Discrepancy (1.5 week):
  Gretton, Arthur, et al. "A kernel method for the two-sample-problem." Advances in neural information processing systems 19 (2006).
- Kernel Mean Matching (1.5 week):
  Huang, Jiayuan, et al. "Correcting sample selection bias by unlabeled data." Advances in neural information processing systems 19 (2006).

Overview of Accomplishments

Gained knowledge

- Foundation in transfer learning:
  Basic concepts & method
- Advantages of transfer learning compared to traditional statistical methods
- Practical applications of transfer learning
- Typical methods used in transfer learning:
  MMD (Maximum Mean Discrepancy), KMM (Kernel Mean Matching)

Motivation: Why do we need transfer learning?

General ideas of transfer learning

The goal of transfer learning is to leverage knowledge from a *source domain* to improve the learning performance in a *target domain*.

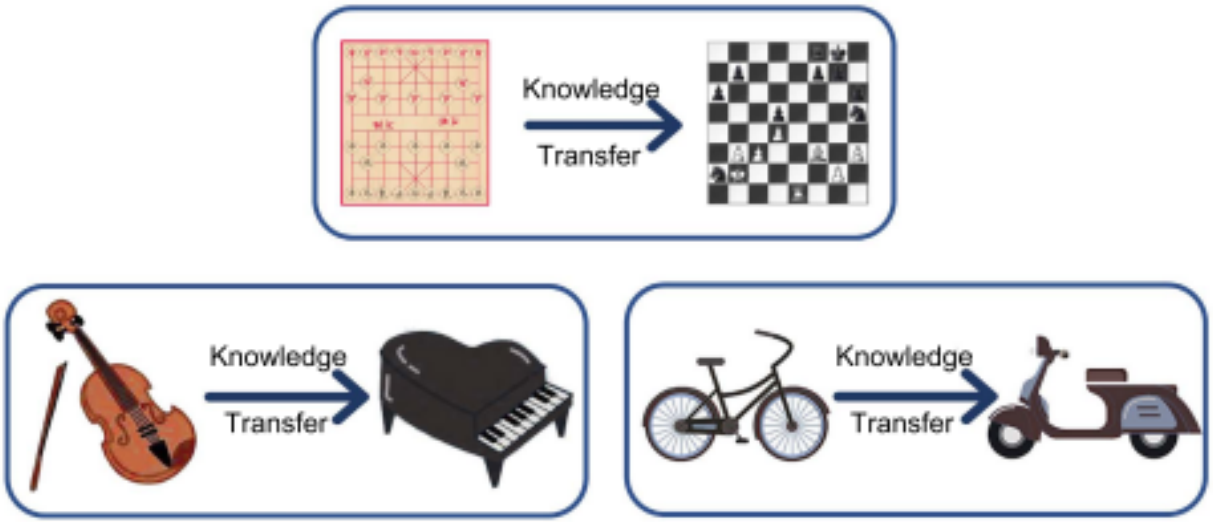


Figure: Intuitive examples of transfer learning, Zhuang et al. (2020)

Motivation: Why do we need transfer learning?

Motivating examples

- **Medical Imaging:** Training data can be expensive to obtain due to requirement of special medical equipment & doctor expertise. What we can do is to train a model on a general medical image dataset and fine-tuning for a specific medical condition (e.g., tumor detection)
- **Bio-informatics:** Knowledge in one organism can be transferred to another organism in biological sequence analysis and gene expression analysis
- **Transportation:** Traffic scene images from the same location can suffer from variation due to difference in weather & light conditions

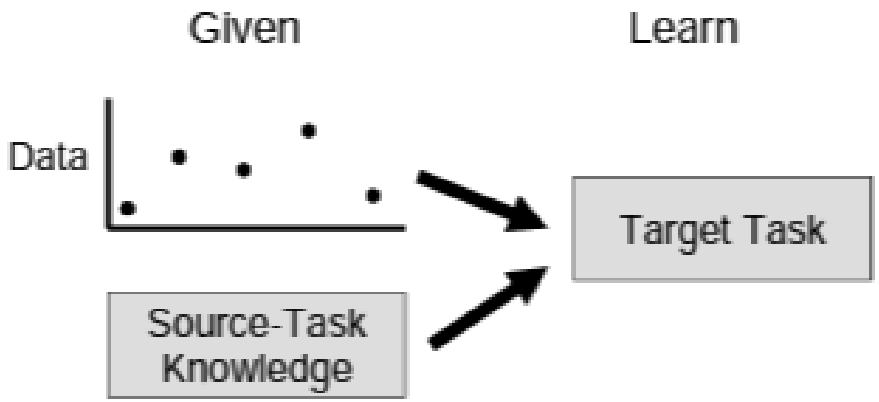


Figure: Knowledge from related tasks used in addition to the standard training data, Torrey and Shavlik (2010)

Motivation: Why do we need transfer learning?

How can transfer learning assist traditional statistical models?

- **Knowledge Transfer:** knowledge from one task can be shared to another
- **Limited Data Handling:** in scenarios with limited labeled data, knowledge can be leveraged from a larger dataset that is related
- **Initialization Benefit**: faster convergence can be achieved when the model weights are initialized with the parameters from the pre-trained model
- **Efficiency in High Dimensions:** provide automatic feature extraction to assist with handling of complex inputs

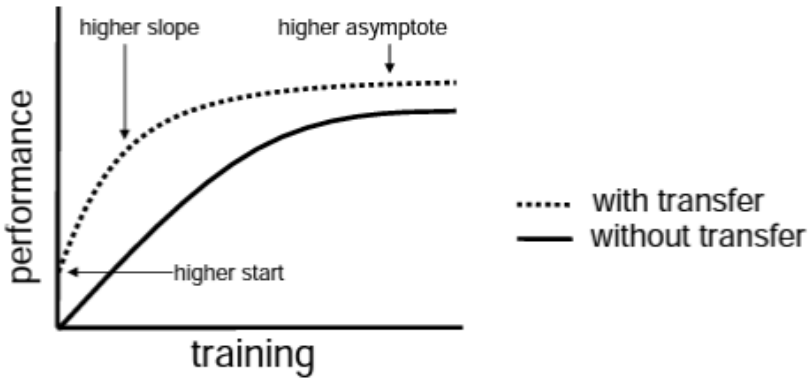


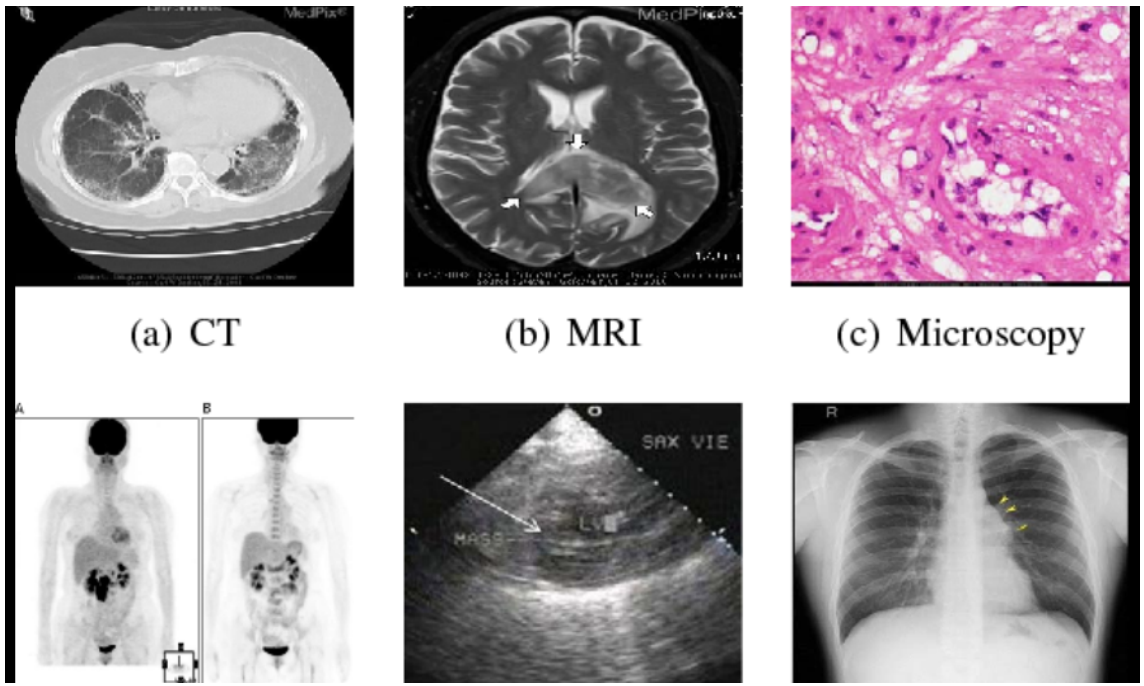
Figure: Learning performance can be improved via several ways, Torrey and Shavlik (2010)

Basic concepts and methods in transfer learning

Basic Concepts

From Zhuang et al. (2020):

- **Definition 1:** A domain D is composed of two parts, i.e., a feature space $\mathcal{X}$ and a marginal distribution $\mathrm{P}(\mathcal{X})$. In other words, $D = \{\mathcal{X}, \mathrm{P}(\mathcal{X})\}$.
- **Definition 2:** A task $\mathcal{T}$ consists of a label space $\mathcal{Y}$ and a decision function f, i.e., $\mathcal{T} = \{\mathcal{Y}, \mathrm{f}\}$. The decision function f is an implicit one, which is expected to be learned from the sample data.



Basic concepts and methods in transfer learning

Basic Concepts

From Zhuang et al. (2020):

- **Definition 3*:**
  - Given:
    - Some observation(s) corresponding to $m^S \in \mathbb{N}^+$ source domain(s) and task(s) (i.e., $\{(D_{S_i}, T_{S_i})|i = 1, , m^S\}$)
    - Some observation(s) about $m^T \in \mathbb{N}^+$ target domain(s) and task(s) (i.e., $\{(D_{T_j}, T_{T_j})|i = 1, , m^T\}$)
  - Transfer learning utilizes the knowledge implied in the source domain(s) to improve the performance of the learned decision functions $f^{T_j}(j = 1, , m^T)$on the target domain(s).

*Based on S.J. Pan and Q. Yang definition from "A survey on transfer learning"

Basic concepts and methods in transfer learning

Basic methods

From Zhuang et al. (2020):

- **Data-based interpretation**: Focus on transferring the knowledge via the adjustment and the transformation of data.
  - A. Instance Weighting Strategy

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y)\sim P^T}\left[\mathcal{L}(\mathbf{x},y;f)\right] &= \mathbb{E}_{(\mathbf{x},y)\sim P^S}\left[\frac{P^T(\mathbf{x},y)}{P^S(\mathbf{x},y)}\mathcal{L}(\mathbf{x},y;f)\right] \\ &= \mathbb{E}_{(\mathbf{x},y)\sim P^S}\left[\frac{P^T(\mathbf{x})}{P^S(\mathbf{x})}\mathcal{L}(\mathbf{x},y;f)\right].\end{aligned}$$

  - B. Feature Transformation Strategy

$$\mathrm{MMD}(X^S,X^T) = \left\|\frac{1}{n^S}\sum_{i=1}^{n^S}\Phi(\mathbf{x}_i^S) - \frac{1}{n^T}\sum_{j=1}^{n^T}\Phi(\mathbf{x}_j^T)\right\|_{\mathcal{H}}^2.$$

Basic concepts and methods in transfer learning

Basic methods

From Zhuang et al. (2020):

- **Model-based interpretation**:
  - A. Model Control Strategy

$$\min_{f^T} \mathcal{L}^{T,L}(f^T) + \lambda_1 \Omega^{\mathrm{D}}(f^T) + \lambda_2 \Omega(f^T),$$

  - B. Parameter Control Strategy
    - Sharing
    - Restriction
  - C. Model Ensemble Strategy
    - Constructs classifiers on each source domain.
    - Each classifier is assigned a weight through an iterative process.
    - The selected classifiers are ensembles to produce the final predictions.

One specific transfer learning method: Instance Weighting Strategy

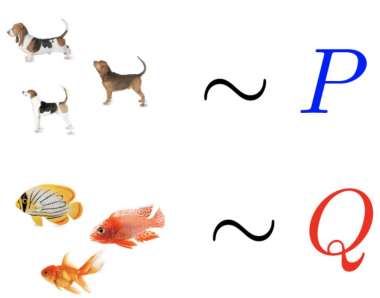
Why Do We Need to Measure the Discrepancy Between Distributions?

- **Domain Adaptation:**
  - **Handle Distribution Shift** The distribution of the source domain may differ from that of the target domain, which degrades the performance of the model
  - **Adapt Model to New Domain** Adapt the model accordingly to improve its performance
- **Domain-Invariant Features Learning**
  - **Generalization** Learn features that are invariant across different domains, aiding the model to generalize better to new data
  - **Feature Alignment** Align the feature spaces of the source and target domains, ensuring that the model learns relevant and common features to both
- **Model Evaluation and Selection:**
  - **Model Optimization** Aid in the fine-tuning and optimization of the model for the target domain
  - **Choose Right Transfer Learning Strategy**
- **Avoid Negative Transfer:**
  - **Detect Irrelevance** happens when the source and target domains are too different, which may lead to negative transfer

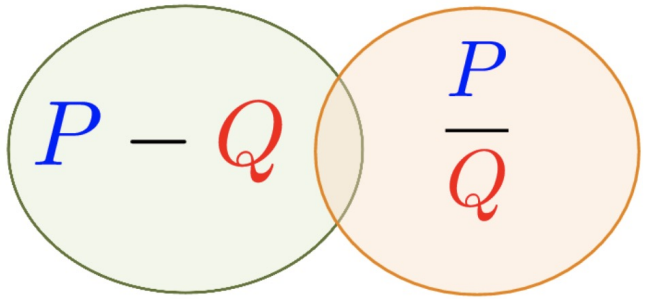
One specific transfer learning method: Instance Weighting Strategy

Two Typical Classes of Measurements

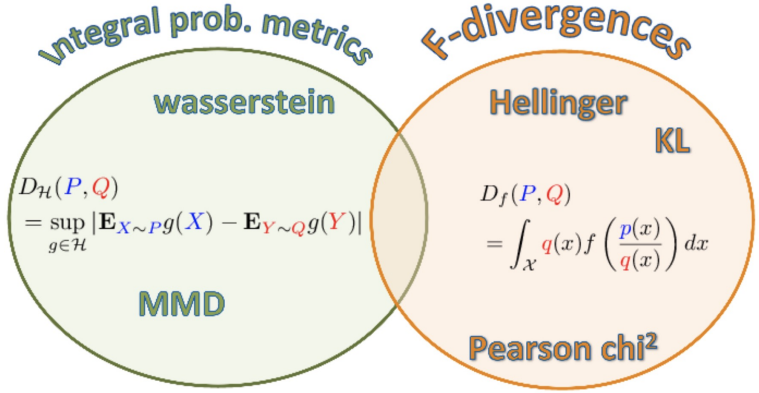
Figure source: Slides by ArthurGretton



(a) Two Samples



(b) Basic Ideas of Measurements



(c) Two Methods of Measurements

One specific transfer learning method: Instance Weighting Strategy

Measuring the discrepancy between distributions: MMD

Definition

Let $(\mathcal{X}, d)$ be a metric space. Let x and y be random variables defined on $\mathcal{X}$, with respective Borel probability measures p and q. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. Given observations $X := \{x_1, ..., x_m\}$ and $Y := \{y_1, ..., y_n\}$ i.i.d. from p and q respectively, we define the maximum mean discrepancy(MMD) as

$$\text{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} (\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)])$$

How to Compute MMD

An Empirical Estimate of MMD is:

$$\text{MMD}\left[\mathcal{F}, X, Y\right] = \left[\frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i, y_j)\right]^{\frac{1}{2}}.$$

One specific transfer learning method: Instance Weighting Strategy

Instance weighting strategy: KMM

From Huang et al. (2006)

- **Two Examples:**
  - **Breast Cancer** most women joined in this test are middle-aged and have lower risk of breast cancer, but the target subjects are in their 20s.
  - **Performing Data Analysis using a Brain Computer Interface** need to adapt the estimator to a new distribution of patterns
- **Key Assumption:** $Pr(x, y)$ and $Pr'(x, y)$ only differ via $Pr(x, y) = Pr(y \mid x)Pr(x)$ and $Pr(y \mid x)Pr'(x)$.
- **Ideas of KMM**

$$\mu(\Pr) := \mathbf{E}_{x \sim \Pr(x)} \left[\Phi(x) \right].$$

Empirical KMM Optimization to find suitable β

$$\underset{\beta}{\text{minimize}} \ \frac{1}{2}\beta^\top K\beta - \kappa^\top \beta \text{ subject to } \beta_i \in [0, B] \text{ and } \left| \sum_{i=1}^{m} \beta_i - m \right| \leq m\epsilon.$$

Summary

Difficulties we encounter and how to solve them

A key challenge we faced was about how to grasp the core idea of theoretical results while avoiding delving too deeply into the details that may hinder our comprehension. Some strategies that helped us are:

- **Focus on the key theorems** - Understanding the foundational concepts and theories prevent us from getting lost in what's beyond our knowledge
- **Use supplementary material** - For a more well-rounded intro into some basic concepts, we took advantage of resources like lecture slides to propel our understanding. The visual aids provided by the paper also lend us more mathematical intuition.
- **Iterative reading** - Plan to read the paper multiple times: skim through first, focus on challenging parts, and prepare questions for group discussion
- **Focus on results applications** - The practical implications of the theoretical findings, as well as the importance of transfer learning, are often revealed through examining these more concrete examples.

Summary

Future directions

- Domain adaptation techniques
- Multi-task learning
- The ethical implications of transfer learning

Summary

Our understanding on EDI (equity, diversity, and inclusion)

Combining diverse backgrounds in STEM is like transfer learning for solving problems – it enriches the dataset of ideas and adapts to new challenges in innovative ways!

> *Your journey in STEM is more than a career choice; it is a path to becoming a change-maker and an inspiration for future generations. All the topics you have learned through this program serve as a solid foundation, empowering you to continue exploring on this journey. I am confident that you will find a path that excites you the most, leading to impactful and meaningful contributions in the field.*
>
> — *Dr. Jiayuan Huang*

Since women are traditionally underrepresented in STEM, we appreciate initiatives like this which prepares us to make a contribution in machine learning and related STEM fields :) Thank you!

Reference

Gretton, A., K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola (2006). A kernel method for the two-sample-problem. *Advances in neural information processing systems 19*.

Huang, J., A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola (2006). Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems 19*.

Torrey, L. and J. Shavlik (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pp. 242–264. IGI global.

Zhuang, F., Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE 109*(1), 43–76.