Domain adaptation under target and conditional shift

Presenter: Kun Zhang

joint work with
Bernhard Schölkopf, Krikamol Muandet, Zhikun Wang

Dept. Empirical Inference
Max Planck Institute for Intelligent Systems
Tübingen, Germany
Why domain adaptation

- Traditional supervised learning:
  \[ P_{XY}^{te} = P_{XY}^{tr} \]

- might not be the case in practice:
Outline

- **Problem**: domain adaptation from \((x^{tr}, y^{tr})\) to \(x^{te}\)
- **Possible causal models** for domain adaptation & solutions
  - Target shift
  - Location-scale **conditional shift**
  - Location-scale **generalized target shift** (target + conditional shift)
Possible situations for domain adaptation: When $X \rightarrow Y$

- **covariate shift**
  
  (Shimodaira00; Sugiyama et al.08; Huang et al.07, Gretton et al.08...)

- **domain**
  
  $\xrightarrow{} X \xrightarrow{} Y$

- **domain**
  
  $\xrightarrow{} X \xrightarrow{} Y$

- **no clue as to find** $P_{Y|X}^{te}$

- **true function** $f(x)$
  
  - training data
  - test data
  - $f_{tr}(x)$
  - $f_{te}(x)$
Possible situations for domain adaptation: When $Y \rightarrow X$

- $Y$ is usually the cause of $X$ (especially for classification)

- Target shift (TarS)

- Conditional shift (ConS)

- Generalized target shift (GeTarS)
**Target shift**

- $P_{Y}^{te} \neq P_{Y}^{tr}$, but $P_{X|Y}^{te} = P_{X|Y}^{tr}$, and furthermore

- **richness**: the support of $P_{Y}^{tr}$ is richer

- **invertibility**: only one $P_{Y} \xrightarrow{P_{X|Y}^{tr}} P_{X}^{te}$

- find the learning machine on test domain by importance reweighting

$$R[P^{te}, \theta, l(x, y, \theta)] = \mathbb{E}_{(X,Y) \sim P^{te}}[l(x, y, \theta)] = \int P_{XY}^{tr} \cdot \frac{P_{XY}^{te}}{P_{XY}^{tr}} \cdot l(x, y, \theta) dx dy$$

$$= \mathbb{E}_{(X,Y) \sim P^{tr}} \cdot \frac{P_{Y}^{te}}{P_{Y}^{tr}} \cdot \frac{P_{X|Y}^{te}}{P_{X|Y}^{tr}} \cdot l(x, y, \theta) dx dy,$$

$$\triangleq \beta^*(y) \triangleq \gamma^*(x, y) \equiv 1!$$

$$\hat{R} = \frac{1}{m} \sum_{i=1}^{m} \beta^*(y_i^{tr}) \gamma^*(x_i^{tr}, y_i^{tr}) l(x_i^{tr}, y_i^{tr}; \theta) = \frac{1}{m} \sum_{i=1}^{m} \beta^*(y_i^{tr}) \cdot l(x_i^{tr}, y_i^{tr}; \theta)$$

- ratio $\beta^*(y)$ can be estimated by min. $\mathcal{D}(P_{X}^{te}, \int P_{Y}^{tr} \beta(y) P_{X|Y}^{tr} dy)$: difficult!
Kernel mean embedding

(Smola et al. 07; Gretton et al. 07; Song et al. 09)

- \( P_X \) has a unique embedding \( \mu[P_X] = \mathbb{E}_{X \sim P_X}[\psi(X)] \) for characteristic kernels.
- Conditional embedding of \( P_{Y|X} \) is an operator from \( \mathcal{F}_X \) to \( \mathcal{G}_Y \): \( U_{Y|X} = C_{YX} C_{XX}^{-1} \); \( C_{YX} \) and \( C_{XX} \) are uncentered cross- and auto-covariance operators.
- \( \mu[P_X] = U_{X|Y} \cdot \mu[P_Y] \).
- \( \hat{U}_{X|Y} = \Psi(L + \lambda I)^{-1} \Phi^T \).

Feature map: \( \psi(x_i) = k(x_i, \cdot) \)
\( \Psi = [\psi(x_1), ..., \psi(x_m)] \),
\( K = \Psi^T \Psi \).

Feature map: \( \phi(y_j) = l(y_j, \cdot) \)
\( \Phi = [\phi(y_1), ..., \phi(y_n)] \),
\( L = \Phi^T \Phi \).

\( \hat{\mu}[P(X)] = \frac{1}{m} \sum_{i=1}^{m} \psi(x_i) \)
\( \mu[P(X)] = \mathbb{E}_{P(X)}[\psi(X)] \)
\( \mu_Y|X = \mathbb{E}_{Y|X=x}[\phi(Y)] \)
Correcting TarS by reweighting target to match covariate with KMM

how to find \( \beta^*(y) = \frac{P_{X|Y}^{te}}{P_Y^{te}} \)?

\[
P_Y^{new} = \beta(y) P_Y^{tr}
\]

\[
P_X^{new} \approx P_X^{te}
\]

\[
\beta(y) \text{ can be estimated by matching } P_X^{new} \text{ with } P_X^{te} : -)
\]

- i.e., minimizing

\[
\left\| \hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{tr}] \right\|^2 = \left\| \hat{U}_{X|Y} \cdot \frac{1}{m} \sum_{i=1}^{m} \beta_i \phi(y_i^{tr}) - \frac{1}{n} \sum_{i=1}^{n} \psi(x_i^{te}) \right\|^2
\]

\[
= \frac{1}{m^2} \beta^T L(L + \lambda_m I)^{-1} K(L + \lambda_m I)^{-1} L \beta - \frac{2}{mn} \frac{1^T}{A} K^c(L + \lambda_m I)^{-1} L \beta + \text{const}
\]

- QP problem: unique solution to \( \beta \)!

- reparameterization such that \( \beta \) is a function of \( y \) and smooth in \( y \): still a QP problem
Correction for TarS: An illustration

\[ x = f(y) + e \]
Conditional shift

- If $P_{X|Y}^{te} \neq P_{X|Y}^{tr}$, possible to determine $P_{Y|X}^{te}$?
- In general, not possible: marginal $P_{X}^{te}$ do not contain enough information to determine $P_{X|Y}^{te}$ (or $P_{Y|X}^{te}$)
- Change in $P_{X|Y}$ must be constrained
Location-scale conditional shift

- For each $y$, domain could change the scale and mean of each feature
- does not change the dependence structure between features
- Assumption $A^{\text{ConS}}$:

  $\exists \mathbf{w}$ and $\mathbf{b}$, such that $P_{X|Y}^{\text{new}} = P_{X|Y}^{\text{te}}$, where $X^{\text{new}} = \mathbf{w}(Y^{tr}) \circ X^{tr} + \mathbf{b}(Y^{tr})$

- Can we find $\mathbf{w}$ and $\mathbf{b}$?
Identifiability & solution

\[ X^{new} = w(Y^{tr}) \circ X^{tr} + b(Y^{tr}) \]

- \( P_{X|Y}^{new} \) is theoretically identifiable under mild conditions (c.f. Theo 2)
- **Necessary** condition: \( P_{X|Y}^{tr}(x|y_i), i = 1, \ldots, C, \) are linearly independent after any LS transformation)

- Estimation: SCG for optimization
  by minimizing \( \|\hat{\mu}[P_{X|Y}^{new}] - \hat{\mu}[P_{X|Y}^{te}]\|^2 = \|\hat{U}[P_{X|Y}^{new}]\hat{\mu}[P_{Y|Y}^{tr}] - \hat{\mu}[P_{Y|Y}^{te}]\|^2, \)
such that \( P_{X|Y}^{new} = P_{X|Y}^{te} \) with \( X^{new} = w(Y^{tr}) \cdot X^{tr} + b(Y^{tr}) \)

- regularization to prefer little change in the conditional

\[ J^{reg} = \frac{\lambda_{LS}}{m} \cdot \|W - 1_m 1_d^T\|^2_F + \frac{\lambda_{LS}}{m} \cdot \|B\|^2_F \]
LS generalized target shift

- **LS-GeTarS:** target-shift + location-scale conditional shift

- $P_{Y}^{te}$ and $P_{X|Y}^{te}$ can be uniquely recovered under mild conditions (c.f. Theo 3)

- Solution: $\min \left\| \hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}] \right\|^2 = \frac{1}{m^2} \beta^T \Omega \tilde{K} \Omega^T \beta - \frac{2}{mn} 1_n^T \tilde{K}^c \beta$

- alternate between optimizing $\beta$ (for TarS) and optimizing $w$ and $b$ (for ConS)
To find the learning machine under LS-GeTarS

- With sample transformation + importance reweighting

\[ X^{\text{new}} = \mathcal{T}(X^{\text{tr}}, Y^{\text{tr}}) \text{ satisfies } P^{\text{new}}_{X|Y} = P^{\text{te}}_{X|Y} \]

\[
R[P^{\text{te}}, \theta, l(x, y, \theta)] = \mathbb{E}_{(X,Y) \sim P^{\text{te}}}[l(x, y, \theta)] = \int P^{\text{tr}}_Y \cdot \beta^*(y) \cdot P^{\text{te}}_{X|Y} \cdot l(x, y, \theta) \, dx \, dy
\]

\[
= \int P^{\text{tr}}_Y \cdot \beta^*(y) \cdot P^{\text{new}}_{X|Y} \cdot l(x, y, \theta) \, dx \, dy, = \mathbb{E}_{(X,Y) \sim P^{\text{new}}_{X,Y}} \beta^*(y) \cdot l(x, y, \theta),
\]

\[
R_{\text{emp}}[P^{\text{te}}, \theta, l(x, y, \theta)] = \frac{1}{m} \sum_{i=1}^{m} \beta^*(y^{\text{tr}}_i) l(x^{\text{new}}_i, y^{\text{tr}}_i, \theta).
\]
Simulation 1: Regression under TarS

\[ x = f(y) + e \]
Simulation 2: Classification under TarS
Simulation 3: Classification under LS-GeTarS
Regression under TarS: Real data

- Cause-effect pair 68
- \( X: \) \# bytes sent by a computer and \( Y: \) \# open http connections at the same time
- \( Y \rightarrow X \)
- TarS greatly improves the prediction performance for \( Y \)

**Time series of \( X \) and \( Y \) and their joint distribution**

Always for training due to large values

Test set 1

2

3

4

**Estimated \( \beta^* \) on the four test sets**

**Prediction performance (MSE) on test data**

<table>
<thead>
<tr>
<th>Test set</th>
<th>Unweight.</th>
<th>CovS</th>
<th>CovS (( q = 0.5 ))</th>
<th>TarS</th>
<th>TarS (( q=0.5 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3789</td>
<td>0.3844</td>
<td>0.3802</td>
<td>0.3310</td>
<td><strong>0.3229</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.0969</td>
<td>0.1126</td>
<td>0.1071</td>
<td>0.0937</td>
<td><strong>0.0887</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.0578</td>
<td>0.0673</td>
<td>0.0659</td>
<td><strong>0.0466</strong></td>
<td>0.0489</td>
</tr>
<tr>
<td>4</td>
<td>0.2054</td>
<td>0.2126</td>
<td>0.2136</td>
<td>0.2008</td>
<td><strong>0.1630</strong></td>
</tr>
</tbody>
</table>
Remote sensing image classification

- Two domains (area 1 & area 2)
- 14 classes

### Misclassification rates by different methods

<table>
<thead>
<tr>
<th>Problem</th>
<th>Unweight</th>
<th>CovS</th>
<th>TarS</th>
<th>LS-GeTarS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR₁ → TS₂</td>
<td>20.73%</td>
<td>20.73%</td>
<td>20.41%</td>
<td>11.96%</td>
</tr>
<tr>
<td>TR₂ → TS₁</td>
<td>26.36%</td>
<td>25.32%</td>
<td>26.28%</td>
<td>13.56%</td>
</tr>
</tbody>
</table>

### # correctly classified points for each class

(a) Domain adaptation from TR₁ to TS₂

(a) Domain adaptation from TR₂ to TS₁
Summary

- Different causal models underlying covariate shift, target shift, conditional shift, and generalized target shift.
- Efficiently solved with kernel mean embedding.
- Nothing comes from nothing: What to transfer?
- Background causal info. helps.