Multi-Task Learning for HIV Therapy Screening

Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, Tobias Scheffer
HIV Therapy Screening

- Usually combinations (3-6 drugs) out of around 17 antiretroviral drugs administered.
- Effect of combinations on virus similar but not identical.
- Scarce training data available from treatment records.

Challenge: Prediction of therapy outcome from genotypic information.
Multi-Task Learning

- Several related prediction problems (tasks).
  - Not necessarily identical conditional $p(y|x)$ of label given input.
  - Usually, some conditionals are similar.

- Challenge:
  - Use all available training data and account for the difference in distributions across tasks.

- HIV therapy screening:
  - Can be modeled as multi-task learning problem.
  - Drug combinations (tasks) have similar but not identical effect on the virus.
Overview

- Motivation.
  - HIV therapy screening.
  - Multiple tasks with differing distributions.

- Multi-task learning by distribution matching.
  - Problem Setting.
  - Density ratio matches pool to target distribution.
  - Discriminative estimation of matching weights.

- Case study:
  - HIV therapy screening.
Multi-Task Learning – Problem Setting

Target distribution

\[ p(x, y | t) \]

Labeled target data
Multi-Task Learning – Problem Setting

- **Goal**: Minimize loss under target distribution.
  
  \[ E_{(x,y) \sim p(x,y|t)}[\ell(f(x), y)] \]

---

**Target distribution**

\[ p(x, y|t) \]

**Labeled target data**
Multi-Task Learning – Problem Setting

- Goal: Minimize loss under target distribution.

\[ \mathbb{E}_{(x,y) \sim p(x,y|t)}[\ell(f(x), y)] \]

Target distribution

\[ p(x, y|t) \]

Labeled target data
Multi-Task Learning – Problem Setting

- Goal: Minimize loss under target distribution.
  - $E_{(x, y) \sim p(x, y | t)}[\ell(f(x), y)]$

Target distribution

- $p(x, y | t)$
  - Labeled target data

 Auxiliary distributions

- $p(x, y | z = 1)$
- $p(x, y | z = 2)$
- $p(x, y | z = 3)$
  - Labeled auxiliary data
Multi-Task Learning – Problem Setting

- **Goal**: Minimize loss under target distribution.

  \[ \mathbb{E}_{(x,y) \sim p(x,y|t)} [\ell(f(x), y)] \]

---

**Target distribution**

\[ p(x, y|t) \]

**Labeled target data**

**Auxiliary distributions**

\[ p(x, y|z = 1) \]
\[ p(x, y|z = 2) \]
\[ p(x, y|z = 3) \]

**Labeled auxiliary data**
Multi-Task Learning – Problem Setting

- Goal: Minimize loss under target distribution.
  - $\mathbb{E}_{(x,y) \sim p(x,y|t)}[\ell(f(x), y)]$
Multi-Task Learning

- Goal: Minimize loss under target distribution.

\[
\mathbb{E}_{(x,y) \sim \text{Target}}[\ell(f(x), y)] \neq \mathbb{E}_{(x,y) \sim \text{Pool}}[\ell(f(x), y)]
\]

Target distribution

- \( p(x, y | t) \)

Labeled target data

Pool distribution

- \( \sum_z p(z) p(x, y | z) \)

Sum over all tasks \( z \), including target \( t \)
Distribution Matching

- Goal: Minimize loss under target distribution.

\[ E_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = E_{(x,y) \sim \text{Pool}}[r_t(x, y)\ell(f(x), y)] \]
**Distribution Matching**

- **Goal**: Minimize loss under target distribution.

\[
E_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = E_{(x,y) \sim \text{Pool}}[r_t(x, y) \ell(f(x), y)]
\]

- Expected loss under target distribution
- Expectation over training pool
- Rescale loss for each pool example

Labeled target data
Distribution Matching

- Goal: Minimize loss under target distribution.

\[ E_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = E_{(x,y) \sim \text{Pool}} \left[ r_t(x, y) \ell(f(x), y) \right] \]
Distribution Matching

- Goal: Minimize loss under target distribution.

\[ \mathbb{E}_{(x, y) \sim \text{Target}}[\ell(f(x), y)] = \mathbb{E}_{(x, y) \sim \text{Pool}}\left[ 2 \ell(f(x), y) \right] \]
Distribution Matching

- **Goal:** Minimize loss under target distribution.

\[
\mathbb{E}_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = \mathbb{E}_{(x,y) \sim \text{Pool}} \left[ \begin{array}{cc} 0 & \ell(f(x), y) \end{array} \right]
\]

Target distribution: \( p(x, y|t) \)

Pool distribution: \( \sum_z p(z)p(x, y|z) \)
Estimation of Density Ratio

Goal: Minimize loss under target distribution.

\[ E_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = E_{(x,y) \sim \text{Pool}} \left[ \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} \ell(f(x), y) \right] \]
Estimation of Density Ratio

- **Goal:** Minimize loss under target distribution.

\[
\mathbb{E}_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = \mathbb{E}_{(x,y) \sim \text{Pool}} \left[ \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} \ell(f(x), y) \right]
\]

- **Theorem:**

\[
\frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} = \frac{p(t|x,y)}{p(t)}
\]

- Potentially high-dimensional densities
- One binary conditional density
Estimation of Density Ratio

- **Goal:** Minimize loss under target distribution.

\[ \mathbb{E}_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = \mathbb{E}_{(x,y) \sim \text{Pool}} \left[ \sum_z \frac{p(x,y|t)}{p(z)p(x,y|z)} \ell(f(x), y) \right] \]

- **Theorem:**

\[ \sum_z \frac{p(x,y|t)}{p(z)p(x,y|z)} = \frac{p(t|x,y)}{p(t)} \]

- **Intuition of** \( p(t|x,y) \): how much more likely is \((x,y)\) to be drawn from target than from auxiliary density.
Estimation of Density Ratio

- Goal: Minimize loss under target distribution.

\[ E_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = E_{(x,y) \sim \text{Pool}} \left[ \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} \ell(f(x), y) \right] \]

- Theorem:

\[ \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} = \frac{p(t|x,y)}{p(t)} \]

- Intuition of \( p(t|x,y) \): how much more likely is \((x,y)\) to be drawn from target than from auxiliary density.
Estimation of Density Ratio

- **Goal:** Minimize loss under target distribution.

\[ \mathbb{E}_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = \mathbb{E}_{(x,y) \sim \text{Pool}} \left[ \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} \ell(f(x), y) \right] \]

- **Theorem:**

\[
\frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} = \frac{p(t|x,y)}{p(t)}
\]

- **Intuition of** \( p(t|x,y) \): how much more likely is \((x, y)\) to be drawn from target than from auxiliary density.

![Diagram](Image)
Estimation of Density Ratio

- **Goal:** Minimize loss under target distribution.
  \[
  \mathbb{E}_{(x,y) \sim \text{Target}}[\ell(f(x), y)] = \mathbb{E}_{(x,y) \sim \text{Pool}} \left[ \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} \ell(f(x), y) \right]
  \]

- **Theorem:**
  \[
  \frac{p(x,y|t)}{\sum_z p(z)p(x,y|z)} = \frac{p(t|x,y)}{p(t)}
  \]

- **Intuition of** \(p(t|x,y)\): how much more likely is \((x,y)\) to be drawn from target than from auxiliary density.
Prior Knowledge on Task Similarity

- Prior knowledge in task similarity kernel \( k(z, z') \).
- Encoding of prior knowledge in Gaussian prior
  \[
  \mathbf{v} \sim N(0, \Sigma)
  \]
on parameters \( \mathbf{v} \) of a multi-class logistic regression model for the resampling weights.
- Main diagonal entries of \( \Sigma \) set to \( \sigma_v^2 \) (standard regularizer),
- Diagonals of sub-matrices set to \( k(z, z') \rho \sigma_v^2 \).
Distribution Matching – Algorithm

1. **Weight Model:**
   Train Logreg of target vs. auxiliary data with task similarity in $\Sigma$.
   \[
   \text{Over } \mathbf{v}, \text{ maximize } \sum_{(x_i, y_i, z_i) \in \text{Pool}} \log(p(z_i|x_i, y_i, \mathbf{v})) - \mathbf{v}^T \Sigma^{-1} \mathbf{v}
   \]

2. **Target Model:**
   Minimize regularized empirical loss on pool weighted by $\frac{p(t|x_i, y_i, \mathbf{v})}{p(t)}$.
   \[
   \text{For task } t, \text{ over } \mathbf{w}_t, \text{ minimize } \sum_{(x_i, y_i) \in \text{Pool}} \frac{p(t|x_i, y_i, \mathbf{v})}{p(t)} \ell(f(x_i, \mathbf{w}_t), y_i) + \frac{\mathbf{w}_t^T \mathbf{w}_t}{2\sigma_w^2}
   \]

Result of step 1: weight model
Overview

- Motivation.
  - HIV therapy screening.
  - Multiple tasks with differing distributions.

- Multi-task learning by distribution matching.
  - Problem Setting.
  - Density ratio matches pool to target distribution.
  - Discriminative estimation of matching weights.

- Case study:
  - HIV therapy screening.
HIV Therapy Screening – Prediction Problem

- Information about each patient $x$, binary vector
  - of resistance-relevant virus mutations and
  - of previously given drugs.

- Drug combination selected out of 17 drugs.
  - Drug combinations correspond to tasks $z$.

- Target label $y$ (success or failure of therapy).
  - 2 different labelings (virus load and multi-conditional).
HIV Therapy Screening – Data

- Patients from hospitals in Italy, Germany, and Sweden.
  - 3260 labeled treatments.
  - 545 different drug combinations (tasks).
  - 50% of combinations with only one labeled treatment.

- Similarity of drug combinations: task kernel.
  - Drug feature kernel: product of drug indicator vectors.
  - Mutation table kernel: similarity of mutations that render drug ineffective.

- 80/20 training/test split, consistent with time stamps.
Reference Methods

- Independent models (separately trained).
- One-size-fits-all, product of task and feature kernel,
- Hierarchical Bayesian Kernel,
- Hierarchical Bayesian Gaussian Process

- Logistic regression is target model (except for Gaussian process model).
- RBF kernels.
Results – Distribution Matching vs. Other

- Distribution matching always best (17 of 20 cases stat. significant) or as good as best reference method.
- Improvement over separately trained models 10-14%.
Results – Benefit of Prior Knowledge

- The common prior knowledge on similarity of drug combinations does not improve accuracy of distribution matching.
Conclusions

- Multi-task Learning:
  - Multiple problems with different distributions.

- Distribution matching:
  - Weighted pool distribution matches target distribution.
  - Discriminative estimation of weights with Logreg.
  - Training of target model with weighted loss terms.

- Case study: HIV therapy screening.
  - Distribution matching beats iid learning and hier. Bayes.
  - Benefit over separately trained models 10-14%.