Learning with Imperfect Data

Mehryar Mohri
Courant Institute and Google
mohri@cims.nyu.edu

Joint work with:
Yishay Mansour (Tel-Aviv & Google) and Afshin Rostamizadeh (Courant Institute).
Standard Learning Assumptions

- IID assumption.
- Same distribution for training and test.
- Distributions fixed over time.
Modern Large-Scale Data Sets

Real-world applications:
- Sample points are not drawn IID.
- Training sample is biased.
- Training points with uncertain labels.
- Multiple training sources.
- Distribution may drift with time.

These problems must be addressed for learning to be effective.
Domain Adaptation - Problem

Input:
• Labeled data from source domain.
• Unlabeled data from target domain.

Problem: use labeled and unlabeled data to derive hypothesis $h$ with good performance on target domain.
• Thus, harder generalization problem than standard learning problem!
Domain Adaptation - Examples

- Sentiment analysis:
  - appraisal information for some domains, e.g., movies, books, music, restaurants.
  - but no labeled information for travel.

- Language modeling, part-of-speech tagging, parsing.

- Speech recognition.

- Computer vision.
Related Work

- **Single-source adaptation:**
  - language modeling, probabilistic parsers, maxent models: source domain used to define a prior.
  - relation between adaptation and the $d_A$ distance [Ben-David et al. (2006) and Blitzer et al. (2007)].

- **Multiple-source:**
  - same input distribution, but different labels [Crammer et al. (2005, 2006)].
  - theoretical analysis and method for multiple-source adaptation [Mansour et al. (2008)].
This Talk

- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments
Learning Set-up

- **Distributions:** source $Q$, target $P$.
- **Target function(s):** $f$, or $f_Q$ and $f_P$.
- **Input:** labeled sample drawn from $Q$, unlabeled sample drawn from $P$.
- **Problem:** find hypothesis $h$ with small expected loss with respect to distribution $P$,

$$
\mathcal{L}_P(h, f) = \mathbb{E}_{x \sim P} \left[ L(h(x), f(x)) \right].
$$
Which distance should we use to compare these distributions?
Simple Analysis

**Proposition**: assume that the loss $L$ is bounded by $M$, then

$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M l_1(Q, P).$$

**Proof:**

$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| = \left| \mathbb{E}_Q[L((h(x), f(x))] - \mathbb{E}_P[L((h(x), f(x))] \right|$$

$$= \left| \sum_x (Q(x) - P(x)) L((h(x), f(x))) \right|$$

$$\leq M \sum_x |Q(x) - P(x)|.$$

But, is this bound informative?
Example - 0/1 Loss

\[ |\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| = |Q(a) - P(a)| \]
**d_A** distance

**Definition:**

\[ d_A(Q_1, Q_2) = \sup_{a \in A} |Q_1(a) - Q_2(a)|. \]

where \( A \) is a set of regions or subsets of \( X \) [Devroye et al. (1996), Kifer et al. (2004)], Ben-David et al. (2007), Blitzer et al. (2007)].

For 0/1 loss, the natural choice is the set of all possible disagreement regions:

\[ A = H \Delta H = \{|h' - h| : h, h' \in H\}. \]
Discrepancy Distance

Definition:

\[ \text{disc}(Q_1, Q_2) = \max_{h, h' \in H} \left| \mathcal{L}_{Q_1}(h', h) - \mathcal{L}_{Q_2}(h', h) \right|. \]

- Relationship with discrepancy in combinatorial contexts [Chazelle (2000)].
- \( d_A \) is a special case, 0-1 loss.
- Helps compare distributions for other losses, e.g. hinge loss, \( L_p \) loss.
- Symmetric, verifies triangle inequality, in general not a distance.
Discrepancy - Properties

**Theorem**: the discrepancy distance can be estimated from finite samples for $H$ with finite VC dimension. For $L_q$ loss, $L_q(y, y') = |y - y'|^q$, for any $\delta > 0$, with probability at least $1 - \delta$,

$$
\text{disc}(P, Q) \leq \text{disc}(\hat{P}, \hat{Q}) + 4q \left( \hat{\mathcal{R}}_S(H) + \hat{\mathcal{R}}_T(H) \right) + 3M \left( \sqrt{\frac{\log \frac{4}{\delta}}{2m}} + \sqrt{\frac{\log \frac{4}{\delta}}{2n}} \right).
$$
This Talk

- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments
Theoretical Guarantees

Two types of questions:

- difference between average loss of hypothesis $h$ on $Q$ versus $P$?
- difference of loss between hypothesis $h$ trained on $Q$ and $h'$ trained on $P$. 
Generalization Bound

Notation:

- \( \mathcal{L}_Q(h_Q^*, f) = \min_{h \in H} \mathcal{L}_Q(h, f) \)
- \( \mathcal{L}_P(h_P^*, f) = \min_{h \in H} \mathcal{L}_P(h, f) \)

Theorem: assume that \( L \) obeys the triangle inequality, then the following holds:

\[
\mathcal{L}_P(h, f_P) \leq \mathcal{L}_Q(h, h_Q^*) + \mathcal{L}_P(h_P^*, f_P) + \text{disc}(P, Q) + \mathcal{L}_Q(h_Q^*, h_P^*).
\]
Some Special Cases

- When $h^* = h^*_Q = h^*_P$,

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_Q(h, h^*) + \mathcal{L}_P(h^*, f_P) + \text{disc}(P, Q).$$

- When $f_P \in H$ (consistent case),

$$|\mathcal{L}_P(h, f_P) - \mathcal{L}_Q(h, f_P)| \leq \text{disc}_L(Q, P).$$
Kernel-Based Reg. Algorithms

- Algorithms minimizing objective function:

\[ F_{\hat{Q}}(h) = \lambda \|h\|^2_K + \hat{R}_{\hat{Q}}(h), \]

where \( K \) is a positive definite symmetric kernel, \( \lambda > 0 \) is a trade-off parameter, and \( \hat{R}_{\hat{Q}}(h) \) the empirical error of \( h \).

- family of algorithms including SVMs, SVR, kernel ridge regression, etc.
Guarantees for KBR Algorithms

Theorem: let $K$ be a positive definite symmetric kernel with $\forall x, K(x, x) \leq \kappa$ and the loss s.t. $L(\cdot, y)$ is $\sigma$-Lipschitz. Assume that $f_P \in H$ and that $f_P$ and $f_Q$ coincide on the training sample. Then, for all $x \in X, y \in Y$,

$$|L(h'(x), y) - L(h(x), y)| \leq \kappa \sigma \sqrt{\frac{\text{disc}(\hat{P}, \hat{Q})}{\lambda}}.$$
Guarantees for KBR Algorithms

Theorem: same assumptions but \( f_P \) and \( f_Q \) potentially different on the training sample, \( H \) bounded by \( M \), and \( L \) the square loss; then, for all \( x \in X, y \in Y \),

\[
\left| L(h'(x), y) - L(h(x), y) \right| \leq \frac{2\kappa M}{\lambda} \left( \kappa \delta + \sqrt{\kappa^2 \delta^2 + 4\lambda \text{disc}_L(\hat{P}, \hat{Q})} \right),
\]

with \( \delta^2 = L_{\hat{Q}}(f_Q(x), f_P(x)) \ll 1. \)
Empirical Discrepancy

- Discrepancy distance $\text{disc}(\hat{P}, \hat{Q})$ critical term in bounds.

- Smaller empirical discrepancy guarantees closeness of pointwise losses of $h'$ and $h$.

- But, can we further reduce the discrepancy?
This Talk

- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments
Algorithm - Idea

- The training sample is given, but we can search for a new empirical distribution $\hat{Q}'$ such that

$$\hat{Q}' = \arg\min_{\hat{Q}' \in \mathcal{Q}} \text{disc}(\hat{P}, \hat{Q}'),$$

where $\mathcal{Q}$ is the set of distributions with support $\text{supp}(\hat{Q})$.

- can be interpreted as reweighting training points.
Case of Halfspaces
Min-Max Problem

Reformulation:

\[
\hat{Q}' = \arg\min_{\hat{Q}' \in Q} \max_{h, h' \in H} |\mathcal{L}_{\hat{P}}(h', h) - \mathcal{L}_{\hat{Q}'}(h', h)|. 
\]

- game theoretical interpretation.
- gives lower bound:

\[
\max_{h, h' \in H} \min_{\hat{Q}' \in Q} |\mathcal{L}_{\hat{P}}(h', h) - \mathcal{L}_{\hat{Q}'}(h', h)| \leq \min_{\hat{Q}' \in Q} \max_{h, h' \in H} |\mathcal{L}_{\hat{P}}(h', h) - \mathcal{L}_{\hat{Q}'}(h', h)|.
\]
Classification - 0/1 Loss

Problem:

$$\min_{Q'} \max_{a \in H \Delta H} |\hat{Q}'(a) - \hat{P}(a)|$$

subject to \( \forall x \in S_Q, \hat{Q}'(x) \geq 0 \land \sum_{x \in S_Q} \hat{Q}'(x) = 1. \)
Classification - 0/1 Loss

- **Linear program (LP):**

\[
\begin{align*}
\min_{Q'} & \quad \delta \\
\text{subject to} & \quad \forall a \in H \Delta H, \hat{Q}'(a) - \hat{P}(a) \leq \delta \\
& \quad \forall a \in H \Delta H, \hat{P}(a) - \hat{Q}'(a) \leq \delta \\
& \quad \forall x \in S_Q, \hat{Q}'(x) \geq 0 \land \sum_{x \in S_Q} \hat{Q}'(x) = 1.
\end{align*}
\]

- No. of constraints bounded by shattering coefficient \( \Pi_{H \Delta H}(m_0 + n_0) \).
Problem:

\[
\min_{\hat{Q}' \in Q} \max_{h, h' \in H} \left| \frac{E[(h'(x) - h(x))^2]}{\hat{P}} - \frac{E[(h'(x) - h(x))^2]}{\hat{Q}'} \right|.
\]
Regression - L2 Loss

- **Semi-definite program (SDP):** linear hypotheses.

\[
\begin{align*}
\min_{z, \lambda} & \quad \lambda \\
\text{subject to} & \quad \lambda I - M(z) \succeq 0 \\
& \quad \lambda I + M(z) \succeq 0 \\
& \quad 1^\top z = 1 \land z \geq 0,
\end{align*}
\]

where the matrix \( M(z) \) is defined by:

\[
M(z) = \sum_{x \in S} \hat{P}(x)xx^\top - \sum_{i=1}^{m_0} z_i s_i s_i^\top.
\]

**elements of supp(\( \hat{Q} \))**
Regression - L2 Loss

- **SDP**: generalization to $H$ RKHS for some kernel $K$. 

\[
\begin{aligned}
\min_{\mathbf{z}, \lambda} & \quad \lambda \\
\text{subject to} & \quad \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0 \\
& \quad \lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0 \\
& \quad 1^\top \mathbf{z} = 1 \land \mathbf{z} \succeq 0,
\end{aligned}
\]

with:

- $\mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{m_0} z_i \mathbf{M}_i$
- $\mathbf{M}_0 = K^{1/2} \text{diag}(P(s_1), \ldots, P(s_{p_0})) K^{1/2}$
- $\mathbf{M}_i = K^{1/2} \mathbf{I}_i K^{1/2}$. 
This Talk

- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments
Experiments

Classification:

- $Q$ and $P$ Gaussians.
- $H$: halfspaces.
- $f$: interval $[-1, +1]$. 
Experiments

Regression:

SDP solved in about 15s using SeDuMi on 3GHz CPU with 2GB memory.
Conclusion

- **Discrepancy distance**: appears as the ‘right’ measure of difference of distributions for adaptation.

- **Theoretical analysis**: generalization bounds and strong guarantees for a large class of algorithms.

- **Algorithm**: discrepancy minimization algorithms for other loss functions, more efficient large-scale algorithms.