Advanced Machine Learning

Domain Adaptation
Non-Ideal World

real world

ideal

time

domain

sampling
Outline

- Domain adaptation.
- Multiple-source domain adaptation.
Domain Adaptation

- Sentiment analysis.
- Language modeling, part-of-speech tagging.
- Statistical parsing.
- Speech recognition.
- Computer vision.

Solution critical for applications.
This Talk

- Domain adaptation
  - Discrepancy
  - Theoretical guarantees
  - Algorithm
  - Enhancements
Domain Adaptation Problem

- **Domains**: source \((Q, f_Q)\), target \((P, f_P)\).

- **Input**:
  - labeled sample \(S\) drawn from source.
  - unlabeled sample \(T\) drawn from target.

- **Problem**: find hypothesis \(h\) in \(H\) with small expected loss with respect to target domain, that is

\[
\mathcal{L}_P(h, f_P) = \mathbb{E}_{x \sim P} \left[ L(h(x), f_P(x)) \right].
\]
Sample Bias Correction Pb

Problem: special case of domain adaptation with

- $f_Q = f_P$.
- $\text{supp}(Q) \subseteq \text{supp}(P)$. 
Related Work in Theory

- Single-source adaptation:
  - relation between adaptation and the $d_A$ distance (Devroye et al. (1996); Kifer et al. (2004); Ben-David et al. (2007)).
  - a few negative examples of adaptation (Ben-David et al. (AISTATS 2010)).
  - analysis and learning guarantees for importance weighting (Cortes, Mansour, and MM (NIPS 2010)).
Related Work in Theory

Multiple-source:

- same input distribution, but different labels (Crammer et al., 2005, 2006).
Distribution Mismatch

Which distance should we use to compare these distributions?
Simple Analysis

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

$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M \mathcal{L}_1(Q, P).$$

**Proof:**

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

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

$$\leq M \sum_x |P(x) - Q(x)|.$$
Example - Zero-One Loss

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

(Mansour, MM, Rostami, COLT 2009)

Definition:

$$\text{disc}(P, Q) = \max_{h, h' \in H} \left| \mathcal{L}_P(h, h') - \mathcal{L}_Q(h, h') \right|.$$ 

- symmetric, triangle inequality, in general not a distance.
- helps compare distributions for arbitrary losses, e.g. hinge loss, or $L_p$ loss.
- generalization of $d_A$ distance (Devroye et al. (1996); Kifer et al. (2004); Ben-David et al. (2007)).
Discrepancy - Properties

- **Theorem**: for $L_q$ loss bounded by $M$, for any $\delta > 0$, with probability at least $1 - \delta$,

\[
\text{disc}(P, Q) \leq \text{disc}(\hat{P}, \hat{Q}) + 4q \left( \hat{R}_S(H) + \hat{R}_T(H) \right) + 3M \left( \sqrt{\log \frac{4}{\delta}} + \sqrt{\log \frac{4}{\delta}} \right).
\]

- **Proof**: Application of McDiarmid’s inequality.
Discrepancy = Distance

**Theorem:** Let $K$ be a universal kernel (e.g., Gaussian kernel) and $H = \{ h \in H_K : \|h\|_K \leq \Lambda \}$. Then, for the $L_2$ loss, discrepancy is a distance over a compact set $X$.

**Proof:** $\Psi : h \mapsto \mathbb{E}_{x \sim P}[h^2(x)] - \mathbb{E}_{x \sim Q}[h^2(x)]$ is continuous for norm $\| \cdot \|_\infty$, thus continuous on $C(X)$.

- $\text{disc}(P, Q) = 0$ implies $\Psi(h) = 0$ for all $h \in H$ since:
  $$\forall h, h' \in H, \quad \left| \mathbb{E}_{x \sim P}[(h'(x) - h(x))^2] - \mathbb{E}_{x \sim Q}[(h'(x) - h(x))^2] \right| = 0.$$  
- since $H$ is dense in $C(X)$, $\Psi = 0$ over $C(X)$.
- thus, $\mathbb{E}_P[f] - \mathbb{E}_Q[f] = 0$ for all $f \geq 0$ in $C(X)$.
- this implies $P = Q$. 
Theoretical Guarantees

- Two types of questions:
  - difference between average loss of hypothesis $h$ on $P$ versus $Q$?
  - difference of loss (measured on $P$) between hypothesis $h$ obtained when training on $(\hat{Q}, f_Q)$ versus hypothesis $h'$ obtained when training on $(\hat{P}, f_P)$?
Generalization Bound

Notation:

- $\mathcal{L}_Q(h^*_Q, f_Q) = \min_{h \in H} \mathcal{L}_Q(h, f_Q)$.
- $\mathcal{L}_P(h^*_P, f_P) = \min_{h \in H} \mathcal{L}_P(h, f_P)$.

Theorem: Assume that $L$ obeys the triangle inequality, then the following holds:

$$\mathcal{L}_P(h, f_P) \leq \min_{h_Q, h_P \in H} \left\{ \mathcal{L}_Q(h, h_Q) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \right\} + \min\{\mathcal{L}_Q(h_Q, h_P), \mathcal{L}_P(h_Q, h_P)\}.$$

(Mansour, MM, Rostamizadeh (COLT 2009) + MM addition)
Proof

\[ \mathcal{L}_P(h, f_P) \leq \min_{h_P \in H} \left\{ \mathcal{L}_P(h, h_P) + \mathcal{L}_P(h_P, f_P) \right\} \]  
(triangle ineq.)

\[ \leq \min_{h_P \in H} \left\{ \mathcal{L}_Q(h, h_P) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \right\} \]  
(def. of discrepancy)

\[ \leq \min_{h_Q, h_P \in H} \left\{ \mathcal{L}_Q(h, h_Q) + \mathcal{L}_Q(h_Q, h_P) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \right\}. \]  
(triangle ineq.)

\[ \leq \min_{h_Q, h_P \in H} \left\{ \mathcal{L}_Q(h, h_Q) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) + \min\{\mathcal{L}_Q(h_Q, h_P), \mathcal{L}_P(h_Q, h_P)\} \right\}. \]  
(combining inequalities)
Some Natural 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),

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

- Bound of (Ben-David et al., NIPS 2006) or (Blitzer et al., NIPS 2007): always worse in these cases.
Regularized ERM Algorithms

Objective function:

\[ F_Q(h) = \lambda \|h\|_K^2 + \hat{R}_Q(h), \]

where \(K\) is a PDS kernel;
\(\lambda > 0\) is a trade-off parameter; and
\(\hat{R}_Q(h)\) is the empirical error of \(h\).

- broad family of algorithms including SVM, SVR, kernel ridge regression, etc.
Guarantees for Reg. ERM

(Cortes & MM (TCS 2013))

Theorem: let $K$ be a PDS kernel with $K(x, x) \leq R^2$ and $L$ a loss function such that $L(\cdot, y)$ is $\mu$-Lipschitz. Let $h'$ be the minimizer of $F_{\hat{P}}$ and $h$ that of that $F_{\hat{Q}}$, then, for all $(x, y) \in X \times Y$,

$$|L(h'(x), y) - L(h(x), y)| \leq \mu R \sqrt{\frac{\text{disc}(\hat{P}, \hat{Q}) + \mu \eta_H(f_P, f_Q)}{\lambda}},$$

where

$$\eta_H(f_P, f_Q) = \inf_{h \in H} \left\{ \max_{x \in \text{supp}(\hat{P})} |f_P(x) - h(x)| + \max_{x \in \text{supp}(\hat{Q})} |f_Q(x) - h(x)| \right\}. $$
Proof

By the property of the minimizers, there exist subgradients such that

\[ 2\lambda h' = -\delta R_{\hat{P}}(h') \]

\[ 2\lambda h = -\delta R_{\hat{Q}}(h). \]

Thus,

\[ 2\lambda \| h' - h \|^2 = -\langle h' - h, \delta R_{\hat{P}}(h') - \delta R_{\hat{Q}}(h) \rangle \]

\[ = -\langle h' - h, \delta R_{\hat{P}}(h') \rangle + \langle h' - h, \delta R_{\hat{Q}}(h) \rangle \]

\[ \leq R_{\hat{P}}(h) - R_{\hat{P}}(h') + R_{\hat{Q}}(h') - R_{\hat{Q}}(h) \]

\[ \leq 2\text{disc}(\hat{P}, \hat{Q}) + 2\mu \eta_H(f_P, f_Q). \]
For any hypothesis $h_0$, we can write:

$$2\lambda \|h' - h\|_K^2 \leq (\mathcal{L}_\hat{P}(h, f_P) - \mathcal{L}_\hat{P}(h, h_0)) - (\mathcal{L}_\hat{P}(h', f_P) - \mathcal{L}_\hat{P}(h', h_0))$$

$$+ (\mathcal{L}_\hat{P}(h, h_0) - \mathcal{L}_\hat{Q}(h, h_0)) - (\mathcal{L}_\hat{P}(h', h_0) - \mathcal{L}_\hat{Q}(h', h_0))$$

$$+ (\mathcal{L}_\hat{Q}(h, h_0) - \mathcal{L}_\hat{Q}(h, f_Q)) - (\mathcal{L}_\hat{Q}(h', h_0) - \mathcal{L}_\hat{Q}(h', f_Q)).$$

Next, by the Lipschitzness, the following holds:

$$(\mathcal{L}_\hat{P}(h, f_P) - \mathcal{L}_\hat{P}(h, h_0)) - (\mathcal{L}_\hat{P}(h', f_P) - \mathcal{L}_\hat{P}(h', h_0)) \leq 2\mu \mathbb{E}_{x \sim \hat{P}}[\|f_P(x) - h_0(x)\|]$$

$$(\mathcal{L}_\hat{Q}(h, h_0) - \mathcal{L}_\hat{Q}(h, f_Q)) - (\mathcal{L}_\hat{Q}(h', h_0) - \mathcal{L}_\hat{Q}(h', f_Q)) \leq 2\mu \mathbb{E}_{x \sim \hat{Q}}[\|f_Q(x) - h_0(x)\|].$$

Since $h_0$ is in $H$, we have

$$(\mathcal{L}_\hat{P}(h, h_0) - \mathcal{L}_\hat{Q}(h, h_0)) - (\mathcal{L}_\hat{P}(h', h_0) - \mathcal{L}_\hat{Q}(h', h_0)) \leq 2 \text{disc}(\hat{P}, \hat{Q}).$$
Guarantees for Reg. ERM

(Cortes & MM (TCS 2013))

**Theorem:** Let $K$ be a PDS kernel with $K(x, x) \leq R^2$ and $L$ the $L_2$ loss bounded by $M$. Then, for all $(x, y)$,

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

where

$$\delta = \min_{h \in H} \left\| \mathbb{E}_{x \sim \hat{Q}} [(h(x) - f_Q(x)) \Phi_K(x)] - \mathbb{E}_{x \sim \hat{P}} [(h(x) - f_P(x)) \Phi_K(x)] \right\|_K.$$

- For $f_P = f_Q = f$,
  - $\delta \leq R \varepsilon$ if $f$ is $\varepsilon$-close to $H$ on samples.
  - $\delta = 0$ for a Gaussian kernel and $f$ continuous.
Proof

For any hypothesis $h_0$, we can write as for previous result:

$$2\lambda \|h' - h\|_K^2 \leq (\mathcal{L}_{\hat{P}}(h, f_P) - \mathcal{L}_{\hat{P}}(h, h_0)) - (\mathcal{L}_{\hat{P}}(h', f_P) - \mathcal{L}_{\hat{P}}(h', h_0))$$

$$+ (\mathcal{L}_{\hat{P}}(h, h_0) - \mathcal{L}_{\hat{Q}}(h, h_0)) - (\mathcal{L}_{\hat{P}}(h', h_0) - \mathcal{L}_{\hat{Q}}(h', h_0))$$

$$+ (\mathcal{L}_{\hat{Q}}(h, h_0) - \mathcal{L}_{\hat{Q}}(h, f_Q)) - (\mathcal{L}_{\hat{Q}}(h', h_0) - \mathcal{L}_{\hat{Q}}(h', f_Q)).$$

Next, for the squared loss, we have:

$$\mathcal{L}_{\hat{P}}(h, f_P) - \mathcal{L}_{\hat{P}}(h, h_0) = \mathbb{E}_{x \sim \hat{P}} [(h_0(x) - f_P(x))(2h(x) - f_P(x) - h_0(x))]$$

$$\mathcal{L}_{\hat{P}}(h', f_P) - \mathcal{L}_{\hat{P}}(h', h_0) = \mathbb{E}_{x \sim \hat{P}} [(h_0(x) - f_P(x))(2h'(x) - f_P(x) - h_0(x))].$$

Thus,

$$\mathcal{L}_{\hat{Q}}(h, h_0) - \mathcal{L}_{\hat{Q}}(h, f_Q) - (\mathcal{L}_{\hat{Q}}(h', h_0) - \mathcal{L}_{\hat{Q}}(h', f_Q))$$

$$-2 \mathbb{E}_{x \sim \hat{Q}} [(h_0(x) - f_Q(x))(h(x) - h'(x))].$$
Proof

- As for previous theorem, we have

\[ (\mathcal{L}_{\hat{P}}(h, h_0) - \mathcal{L}_{\hat{Q}}(h, h_0)) - (\mathcal{L}_{\hat{P}}(h', h_0) - \mathcal{L}_{\hat{Q}}(h', h_0)) \leq 2 \operatorname{disc}(\hat{P}, \hat{Q}). \]

- Thus, \( 2\lambda \| h' - h \|^2_K \leq 2 \operatorname{disc}(\hat{P}, \hat{Q}) + 2\Delta \) with:

\[
\Delta = \left\langle h - h', \mathbb{E}_{x \sim \hat{P}}[(h_0(x) - f_P(x))K(x, \cdot)] - \mathbb{E}_{x \sim \hat{Q}}[(h_0(x) - f_Q(x))K(x, \cdot)] \right\rangle
\]

\[
\leq \| h - h' \|^2_K \left\| \mathbb{E}_{x \sim \hat{P}}[(h_0(x) - f_P(x))K(x, \cdot)] - \mathbb{E}_{x \sim \hat{Q}}[(h_0(x) - f_Q(x))K(x, \cdot)] \right\|_K.
\]

- The result follows by solving second-degree inequality.
Empirical Discrepancy

- Discrepancy measure $\text{disc}(\hat{P}, \hat{Q})$ critical term in bounds.
- Smaller empirical discrepancy guarantees closeness of pointwise losses of $h'$ and $h$.
- But, can we further reduce the discrepancy?
Algorithm - Idea

- Search for a new empirical distribution $q^*$ with same support:

$$q^* = \text{argmin}_{\text{supp}(q) \subseteq \text{supp}(\hat{Q})} \text{disc}(\hat{P}, q).$$

- Solve modified optimization problem:

$$\min_{h} F_{q^*}(h) = \sum_{i=1}^{m} q^*(x_i) L(h(x_i), y_i) + \lambda \|h\|_K^2.$$
Case of Halfspaces
**Min-Max Problem**

**Reformulation:**

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

- game theoretical interpretation.
- gives lower bound:

\[
\max_{h, h' \in H} \min_{\hat{Q}' \in Q} |L_{\hat{P}}(h', h) - L_{\hat{Q}'}(h', h)| \leq \min_{\hat{Q}' \in Q} \max_{h, h' \in H} |L_{\hat{P}}(h', h) - L_{\hat{Q}'}(h', h)|.
\]
Problem:

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

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.

\[
\Pi_{H \Delta H} (m_0 + n_0)
\]
Algorithm - 1D
Regression - L2 Loss

Problem:

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

\[
\min_{\hat{Q}' \in Q} \max_{\|w\| \leq 1, \|w'\| \leq 1} \left| \frac{\mathbb{E}[((w' - w)^{\top} x)^2]}{\hat{P}} - \frac{\mathbb{E}[((w' - w)^{\top} x)^2]}{\hat{Q}'} \right|
\]

\[
= \min_{\hat{Q}' \in Q} \max_{\|w\| \leq 1, \|w'\| \leq 1} \left| \sum_{x \in S} (\hat{P}(x) - \hat{Q}'(x))[(w' - w)^{\top} x]^2 \right|
\]

\[
= \min_{\hat{Q}' \in Q} \max_{\|u\| \leq 2} \left| \sum_{x \in S} (\hat{P}(x) - \hat{Q}'(x))[u^{\top} x]^2 \right|
\]

\[
= \min_{\hat{Q}' \in Q} \max_{\|u\| \leq 2} \left| u^{\top} \left( \sum_{x \in S} (\hat{P}(x) - \hat{Q}'(x))xx^{\top} \right) u \right|.
\]
Regression - L2 Loss

Problem equivalent to

$$
\min_{\|z\|_1=1} \max_{\|u\|=1} |u^T M(z) u|,
$$

with:

$$
M(z) = M_0 - \sum_{i=1}^{m_0} z_i M_i,
$$

$$
M_0 = \sum_{x \in S} P(x) xx^T
$$

$$
M_i = s_i s_i^T
$$

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

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

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

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

\[
\mathbf{M}(\mathbf{z}) = \sum_{\mathbf{x} \in S} \hat{\mathbf{P}}(\mathbf{x}) \mathbf{x} \mathbf{x}^\top - \sum_{i=1}^{m_0} z_i \mathbf{s}_i \mathbf{s}_i^\top.
\]
Regression - L2 Loss

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

$$
\min_{z, \lambda} \quad \lambda \\
\text{subject to} \quad \lambda I - M(z) \succeq 0 \\
\lambda I + M(z) \succeq 0 \\
1^\top z = 1 \land z \succeq 0,
$$

with:

$$
M(z) = M_0 - \sum_{i=1}^{m_0} z_i M_i \\
M_0 = K^{1/2} \text{diag}(P(s_1), \ldots, P(s_{p_0})))K^{1/2} \\
M_i = K^{1/2} I_i K^{1/2}.
$$
Discrepancy Min. Algorithm

(Cortes & MM (TCS 2013))

- Convex optimization:
  - cast as semi-definite programming (SDP) prob.
  - efficient solution using smooth optimization.
- Algorithm and solution for arbitrary kernels.
- Outperforms other algorithms in 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.
Experiments

Fig. 11. Results with “easy-to-learn” biasing scheme: Relative MSE performance of (1): Optimal (in black); (2): KMM (in blue); (3): KLIEP (in orange); (4): Uniform (in green); (5): Two-Stage (in brown); and (6): DM (in red). Errors are normalized so that the average MSE of Uniform is 1.
Enhancement

(Cortes, MM, and Muñoz (2014))

- **Shortcomings:**
  - discrepancy depends on maximizing pair of hypotheses.
  - DM algorithm too conservative.

- **Ideas:**
  - finer quantity: *generalized discrepancy*, hypothesis-dependent.
  - reweighting depending on hypothesis.
Choose $Q_h$ such that objectives are unif. close:

$$\lambda \|h\|^2_K + \mathcal{L}_{Q_h}(h, f_Q)$$

$$\lambda \|h\|^2_K + \mathcal{L}_{\tilde{P}}(h, f_P).$$

Ideally:

$$Q_h = \arg\min_q |\mathcal{L}_q(h, f_Q) - \mathcal{L}_{\tilde{P}}(h, f_P)|.$$ 

Using convex surrogate $H''$:

$$Q_h = \arg\min_q \max_{h'' \in H''} |\mathcal{L}_q(h, f_Q) - \mathcal{L}(h, h'')|.$$
Optimization

(Cortes, MM, and Muñoz (2014))

\[
\mathcal{L}_{Q_h}(h, f_Q) = \arg\min_{l \in \{\mathcal{L}_q(h, f_Q) : q \in \mathcal{F}(S_X, \mathbb{R})\}} \max_{h'' \in H''} |l - \mathcal{L}_{\hat{P}}(h, h'')|
\]

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

\[
= \frac{1}{2} \left( \max_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') \right).
\]

Convex optimization problem (loss jointly convex):

\[
\min_h \lambda \|h\|_K^2 + \frac{1}{2} \left( \max_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\hat{P}}(h, h'') \right).
\]
Convex Surrogate Hyp. Set

Choice of $H''$ among balls

$$B(r) = \{ h'' \in H | \mathcal{L}_q(h'', f_Q) \leq r^p \}.$$

Generalization bound proven to be more favorable than DM for some choices of radius $r$.

Radius $r$ chosen via cross-validation using small amount of labeled data from target.

Further improvement of empirical results.

(Cortes, MM, and Muñoz (2014))
Conclusion

- Theory of adaptation based on discrepancy:
  - key term in analysis of adaptation and drifting.
  - discrepancy minimization algorithm DM.
  - compares favorably to other adaptation algorithms in experiments.

- Generalized discrepancy:
  - extension to hypothesis-dependent reweighting.
  - convex optimization problem.
  - further empirical improvements.

- Further generalization: (Awasthi, Cortes, MM, 2024).
Outline

- Domain adaptation.

- Multiple-source domain adaptation.
Problem Formulation

- Given distributions and corresponding hypotheses:

\[ h_1 \rightarrow D_1 \]
\[ h_2 \rightarrow D_2 \]
\[ \vdots \]
\[ h_k \rightarrow D_k \]

Each hypothesis performs well in its domain.

\[ \forall i, \mathcal{L}(D_i, h_i, f) \leq \epsilon. \]

Notation: \( \mathcal{L}(D_i, h_i, f) = \mathbb{E}_{x \sim D_i} [L(h_i(x), f(x))]. \)

Loss \( L \) assumed non-negative, bounded, convex and continuous.
Problem Formulation

- The unknown target distribution is a mixture of input distributions.

\[ D_1 \rightarrow \ldots \rightarrow D_k \rightarrow D_T \]

- Task: choose a **hypothesis mixture** that performs well in target distribution.

\[
D_T(x) = \sum_{i=1}^{k} \lambda_i D_i(x)
\]

\[
h_z(x) = \sum_{i=1}^{k} z_i h_i(x)
\]

**convex combination rule**

\[
h_z(x) = \sum_{i=1}^{k} \frac{z_i D_i(x)}{\sum_{j=1}^{k} z_j D_j(x)} h_i(x)
\]

**distribution weighted combination**
Known Target Distribution

For some distributions, any convex combination performs poorly.

- base hypotheses have no error within domain.
- any convex combination has error of $1/2$!

<table>
<thead>
<tr>
<th>distribution weights</th>
<th>hypothesis output</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_T$  $D_0$  $D_1$</td>
<td>$f$  $h_0$  $h_1$</td>
</tr>
<tr>
<td>a  0.5   1   0</td>
<td>a  1   1   0</td>
</tr>
<tr>
<td>b  0.5   0   1</td>
<td>b  0   1   0</td>
</tr>
</tbody>
</table>
Main Results

- Thus, although convex combinations seem natural, they can perform very poorly.

- We will show that distribution weighted combinations seem to define the “right” combination rule.

- There exists a single “robust” distribution weighted hypothesis, that does well for any target mixture.

\[ \forall f, \exists z, \forall \lambda, \mathcal{L}(D_\lambda, h_z, f) \leq \epsilon. \]
Known Target Distribution

- If distribution is known, distribution weighted rule will always do well. Choose: $z = \lambda$.

$$h_\lambda(x) = \sum_{i=1}^{k} \frac{\lambda_i D_i(x)}{\sum_{j=1}^{k} \lambda_j D_j(x)} h_i(x).$$

- Proof:

$$\mathcal{L}(D_T, h_\lambda, f) = \sum_{x \in X} L(h_\lambda(x), f(x)) D_T(x)$$

$$\leq \sum_{x \in X} \sum_{i=1}^{k} \frac{\lambda_i D_i(x)}{D_T(x)} L(h_i(x), f(x)) D_T(x)$$

$$= \sum_{i=1}^{k} \lambda_i \mathcal{L}(D_i, h_i(x), f(x)) \leq \sum_{i=1}^{k} \lambda_i \epsilon = \epsilon.$$
Unknown Target Mixture

- **Zero-sum game:**
  - **NATURE:** select a target distribution $D_i$.
  - **LEARNER:** select a $\mathcal{z}$, i.e. a distribution weighted hypothesis $h_{\mathcal{z}}$.
  - **Payoff:** $\mathcal{L}(D_i, h_{\mathcal{z}}, f)$.
  - Already shown: *game value* is at most $\epsilon$.

- **Minimax theorem** (modulo discretization of $\mathcal{z}$): there exists a mixture $\sum_j \alpha_j h_{\mathcal{z}_j}$ of distribution weighted hypothesis that does well for any distribution mixture.
Balancing Losses

- **Brouwer’s Fixed Point theorem**: for any compact, convex, non-empty set $A$ and any continuous function $f: A \to A$, there exists $x$ such that: $f(x) = x$.

- Define mapping $\phi$ by: $[\phi(z)]_i = \frac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z}$.

- By fixed point theorem (modulo continuity):

  \[
  \exists z : \forall i, z_i = \frac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z} \implies \forall i, \mathcal{L}_i^z = \sum_j z_j \mathcal{L}_j^z =: \gamma.
  \]
Bounding Loss

For fixed point $z$,

$$\mathcal{L}(D_z, h_z, f) = \sum_{x \in X} L(h_z(x), f(x)) \left( \sum_{i=1}^k z_i D_i(x) \right)$$

$$= \sum_{i=1}^k z_i \sum_{x \in X} D_i(x) L(h_z(x), f(x))$$

$$= \sum_{i=1}^k z_i \mathcal{L}^z_i = \sum_{i=1}^k z_i \gamma = \gamma.$$ 

Also, by convexity,

$$\gamma = \mathcal{L}(D_z, h_z, f) \leq \sum_{x \in X} \sum_{i=1}^k \frac{z_i D_i(x)}{D_z(x)} L(h_i(x), f(x)) D_z(x) = \sum_{i=1}^k z_i \mathcal{L}(D_i, h_i, f) \leq \epsilon.$$
Thus, \( \gamma \leq \epsilon \) and for any mixture \( \lambda \),

\[
\mathcal{L}(D_\lambda, h_z, f) = \sum_{i=1}^{k} \lambda_i \mathcal{L}(D_i, h_z, f) \leq \sum_{i=1}^{k} \lambda_i \gamma = \gamma \leq \epsilon.
\]
Details

To deal with non-continuity refine hypotheses:

\[ h^n_z(x) = \sum_{i=1}^{k} \frac{z_i D_i(x) + \eta/k}{\sum_{j=1}^{k} z_j D_j(x) + \eta} h_i(x). \]

Theorem: for any target function \( f \) and any \( \delta > 0 \),

\[ \exists \eta > 0, z: \forall \lambda, \mathcal{L}(D_\lambda, h^n_z, f) \leq \epsilon + \delta. \]

If loss obeys triangle inequality:

\[ \forall \delta > 0, \exists z, \eta > 0, \forall \lambda, f \in \mathcal{F}, \mathcal{L}(D_\lambda, h^n_z, f) \leq 3\epsilon + \delta. \]

holds for all admissible target functions.
A Simple Algorithm

A simple constructive algorithm, choose $z$ with uniform weights:

$$
\mathcal{L}(D_\lambda, h_u, f) = \sum_x D_\lambda(x) L \left( \sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x) \right)
$$

$$
= \sum_x \left( \sum_{m=1}^k \lambda_m D_m(x) \right) L \left( \sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x) \right)
$$

$$
\leq \sum_x \frac{\sum_{m=1}^k \lambda_m D_m(x)}{\sum_{j=1}^k D_j(x)} \sum_{i=1}^k D_i(x) L (h_i(x), f(x))
\leq 1
$$

$$
\leq \sum_{i=1}^k \sum_x D_i(x) L (h_i(x), f(x)) = \sum_{i=1}^k \mathcal{L}(D_i, h_i, f) = \sum_{i=1}^k \epsilon_i \leq k\epsilon.
$$
Preliminary Empirical Results

- **Sentiment Analysis** - given a product review (text string), predict a rating (between 1.0 and 5.0).

- **4 Domains**: Books, DVDs, Electronics and Kitchen Appliances.

- **Base hypotheses** are trained within each domain (Support Vector Regression).

- We are **not given** the distributions. We model each distribution using a bag of words model.

- We then test the distribution combination rule on known target mixture domains.
Empirical Results

Uniform Mixture Over 4 Domains

MSE

In–Domain
Out–Domain

1 2 3 4 5 6
Empirical Results

2 class

Mixture = \( \alpha \) book + \((1 - \alpha)\) kitchen

\[
\text{MSE} = \text{weighted, linear, book, kitchen}
\]
Conclusion

- Formulation of the multiple source adaptation problem.
- Theoretical analysis for mixture distributions.
- Efficient algorithm for finding distribution weighted combination hypothesis?
- Beyond mixture distributions?
Rényi Divergences

**Definition:** for $\alpha \geq 0$,

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log \sum_x P(x) \left[ \frac{P(x)}{Q(x)} \right]^{\alpha - 1}.$$  

- $\alpha = 1$: coincides with relative entropy.
- $\alpha = 2$: logarithm of expected probability ratio;

$$D_\alpha(P||Q) = \log E_{x \sim P} \left[ \frac{P(x)}{Q(x)} \right].$$

- $\alpha = +\infty$: logarithm of maximum probability ratio;

$$D_\alpha(P||Q) = \log \sup_{x \sim P} \left[ \frac{P(x)}{Q(x)} \right].$$
Extensions - Arbitrary Target

(Mansour, MM, and Rostami, 2009)

Theorem: for any $\delta > 0$ and $\alpha > 1$,

$$\exists \eta, z : \forall P, \mathcal{L}(P, h^\eta_z, f) \leq \left[ d_\alpha(P \parallel Q)(\epsilon + \delta) \right]^{\frac{\alpha-1}{\alpha}} M^{\frac{1}{\alpha}}.$$

$$Q = \left\{ \sum_{i=1}^k \lambda_i D_i : \lambda \in \Delta_k \right\}$$

measured in terms of Rényi divergence,

$$d_\alpha(P, Q) = \left[ \sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \right]^{\frac{1}{\alpha-1}}.$$
Proof

By Hölder’s inequality, for any hypothesis $h$,

$$
\mathcal{L}(P, h, f) = \sum_x \frac{P(x)}{Q^{\alpha-1}(x)} Q^{\frac{1}{\alpha}}(x) L(h(x), f(x))
\leq \left[ \sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \right]^{\frac{1}{\alpha}} \left[ \sum_x Q(x) L^{\frac{\alpha}{\alpha-1}}(h(x), f(x)) \right]^{\frac{\alpha-1}{\alpha}}
= (d_\alpha(P \parallel Q))^{\frac{\alpha-1}{\alpha}} \left[ \mathbb{E}_{x \sim Q} [L^{\frac{\alpha}{\alpha-1}}(h(x), f(x))] \right]^{\frac{\alpha-1}{\alpha}}
= (d_\alpha(P \parallel Q))^{\frac{\alpha-1}{\alpha}} \left[ \mathbb{E}_{x \sim Q} [L(h(x), f(x)) L^{\frac{1}{\alpha-1}}(h(x), f(x))] \right]^{\frac{\alpha-1}{\alpha}}
\leq (d_\alpha(P \parallel Q))^{\frac{\alpha-1}{\alpha}} \left[ \mathcal{L}(Q, h, f) M^{\frac{1}{\alpha-1}} \right]^{\frac{\alpha-1}{\alpha}}.
$$
Other Extensions

(Approximate distributions (estimated):

- similar results shown depending on divergence between true and estimated distributions.

Different source target functions $f_i$:

- similar results when target functions close to $f$ on target distribution.

(Mansour, MM, and Rostami, 2009)
References


References


References

