# LEARNING BOUNDS FOR IMPORTANCE WEIGHTING

Tamas Madarasz & Michael Rabadi

April 15, 2015

# Introduction

- Often, training distribution does not match testing distribution
- Want to utilize information about test distribution
- Correct bias or discrepancy between training and testing distributions

# Importance Weighting

- Labeled training data from source distribution $Q$
- Unlabeled test data from target distribution $P$
- Weight the cost of errors on training instances.
- Common definition of weight for point $x$: $w(x) = P(x)/Q(x)$

# Importance Weighting

- Reasonable method, but sometimes doesn't work
- Can we give generalization bounds for this method?
- When does DA work? When does it not work?
- How should we weight the costs?

# Overview

- Preliminaries
- Learning guarantee when loss is bounded
- Learning guarantee when loss is unbounded, but second moment is bounded
- Algorithm

# Preliminaries: Rényi divergence

For $\alpha \geq 0$, $D_\alpha(P||Q)$ between distributions P and Q

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log_2 \sum_x P(x) \left( \frac{P(x)}{Q(x)} \right)^{\alpha - 1}$$

$$d_\alpha(P||Q) = 2^{D_\alpha(P||Q)} = \left[ \sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \right]^{\frac{1}{\alpha-1}}$$

- Metric of info lost when Q is used to approximate P
- $D_\alpha(P||Q) = 0$ iff $P = Q$

# Preliminaries: Importance weights

Lemma 1:

$$\mathrm{E}[w] = 1 \qquad \mathrm{E}[w^2] = d_2(P||Q) \qquad \sigma^2 = d_2(P||Q) - 1$$

Proof:

$$\mathrm{E}_Q[w^2] = \sum_{x \in X} w^2(x)Q(x) = \sum_{x \in X} \left(\frac{P(x)}{Q(x)}\right)^2 Q(x) = d_2(P||Q)$$

Lemma 2: For all $\alpha > 0$ and $x \in X$,

$$\mathrm{E}_Q[w^2(x)L_h^2(x)] \leq d_{\alpha+1}(P||Q)R(h)^{1-\frac{1}{\alpha}}$$

# Preliminaries: Importance weights

Hölder's Inequality (Jin, Wilson, and Nobel, 2014): Let $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\sum_x |a_x b_x| \leq \left( \sum_x |a_x|^p \right)^{\frac{1}{p}} \left( \sum_x |b_x|^q \right)^{\frac{1}{q}}$$

# Preliminaries: Importance weights

Proof for Lemma 2: Let the loss be bounded by $B = 1$, then

$$E_{x\sim Q}[w^2(x)L_h^2(x)] = \sum_x Q(x)\left[\frac{P(x)}{Q(x)}\right]^2 L_h^2(x) = \sum_x P(x)^{\frac{1}{\alpha}}\left[\frac{P(x)}{Q(x)}\right]P(x)^{\frac{\alpha-1}{\alpha}}L_h^2(x)$$

$$\leq \left[\sum_x P(x)\left[\frac{P(x)}{Q(x)}\right]^{\alpha}\right]^{\frac{1}{\alpha}}\left[\sum_x P(x)L_h^{\frac{2\alpha}{\alpha-1}}(x)\right]^{\frac{\alpha-1}{\alpha}}$$

$$= d_{\alpha+1}(P||Q)\left[\sum_x P(x)L_h(x)L_h^{\frac{\alpha+1}{\alpha-1}}(x)\right]^{\frac{\alpha-1}{\alpha}}$$

$$\leq d_{\alpha+1}(P||Q)R(h)^{1-\frac{1}{\alpha}}B^{1+\frac{1}{\alpha}} = d_{\alpha+1}(P||Q)R(h)^{1-\frac{1}{\alpha}}$$

# Learning Guarantees: Bounded case

$\sup_x w(x) = \sup_x \frac{P(x)}{Q(x)} = d_\infty(P||Q) = M$. Let $d_\infty(P||Q) < +\infty$. Fix $h \in H$. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$|R(h) - \hat{R}_w(h)| \leq M\sqrt{\frac{log\frac{2}{\delta}}{2m}}$$

- M can be very large, so we naturally want a more favorable bound...

# Overview

- Preliminaries
- Learning guarantee when loss is bounded
- Learning guarantee when loss is unbounded, but second moment is bounded
- Algorithm

# Learning Guarantees: Bounded case

Theorem 1: Fix $h \in H$. For any $\alpha \geq 1$, for any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds:

$$R(h) \leq \hat{R}_w(h) + \frac{2M \log \frac{1}{\delta}}{3m} + \sqrt{\frac{2[d_{\alpha+1}(P||Q)R(h)^{1-\frac{1}{\alpha}} - R(h)^2] \log \frac{1}{\delta}}{m}}$$

# Learning Guarantees: Bounded case

Bernstein's inequality (Bernstein 1946):

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right) \leq \exp\left(\frac{-n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right)$$

when $|x_i| \leq M$.

# Learning Guarantees: Bounded case

Proof of Theorem 1: Let $Z$ be the random variable $w(x)L_h(x) - R(x)$. Then $|Z| \leq M$. Thus, by lemma 2, the variance of Z can be bounded in terms of $d_{\alpha+1}(P||Q)$:

$$\sigma^2(Z) = \mathrm{E}_Q[w^2(x)L_h(x)^2)] - R(h)^2 \leq d_{\alpha+1}(P||Q)R(h)^{1-\frac{1}{\alpha}} - R(h)^2$$

$$\Pr[R(h) - \hat{R}_w(h) > \epsilon] \leq \exp\left(\frac{-m\epsilon^2/2}{\sigma^2(Z) + \epsilon M/3}\right).$$

# Learning Guarantees: Bounded case

Thus, setting $\delta$ to match upper bound, then with probability at least $1 - \delta$

$$R(h) \leq \hat{R}_w(h) + \frac{2M \log \frac{1}{\delta}}{3m} + \sqrt{\frac{M^2 \log^2 \frac{1}{\delta}}{9m^2} + \frac{2\sigma^2(Z) \log \frac{1}{\delta}}{m}}$$

$$= \hat{R}_w(h) + \frac{2M \log \frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma^2(Z) \log \frac{1}{\delta}}{m}}$$

# Learning Guarantees: Bounded case

Theorem 2: Let $H$ be a finite hypothesis set. Then for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for the importance weighting method:

$$R(h) \leq \hat{R}_w(h) + \frac{2M(log|H| + \log\frac{1}{\delta})}{3m} + \sqrt{\frac{2d_2(P||Q)(\log|H| = \log\frac{1}{\delta}}{m}}$$

# Learning Guarantees: Bounded case

Theorem 2 holds when $\alpha = 1$. Note that theorem 1 can be simplified in the case of $\alpha = 1$:

$$R(h) \leq \hat{R}_w(h) + \frac{2M \log \frac{1}{\delta}}{3m} + \sqrt{\frac{2d_2(P||Q) \log \frac{1}{\delta}}{m}}$$

Thus, theorem 2 follows by including the cardinality of H

# Learning Guarantees: Bounded case

Proposition 2: Lower bound. Assume $M < \infty$ and $\sigma^2(w)/M^2 \geq 1/m$. Assume there exists $h_0 \in H$ such that $L_{h_0}(x) = 1$ for all $x$. There exists an absolute constant $c$, $c = 2/41^2$, such that

$$\Pr\left[\sup_{h \in H} |R(h) - \hat{R}_w(h)| \geq \sqrt{\frac{d_2(P||Q) - 1}{4m}}\right] \geq c > 0$$

Proof from theorem 9 of Cortes, Mansour, and Mohri, 2010.

# Overview

- Preliminaries
- Learning guarantee when loss is bounded
- Learning guarantee when loss is unbounded, but second moment is bounded
- Algorithm

# Learning Guarantees: Unbounded case

$d_\infty(P||Q) < \infty$ does not always hold... Assume P and Q follow a Guassian distribution with $\sigma_P$ and $\sigma_Q$ with means $\mu$ and $\mu'$

$$\frac{P(x)}{Q(x)} = \frac{\sigma_P}{\sigma_Q} \exp\left[ - \frac{\sigma_Q^2(x-\mu)^2 - \sigma_P^2(x-\mu')^2}{2\sigma_P^2\sigma_Q^2} \right]$$

Thus, even if $\sigma_P = \sigma_Q$ and $\mu \neq \mu'$, $d_\infty(P||Q) = \sup_x \frac{P(x)}{Q(x)} = \infty$, thus Theorem 1 is not informative.

# Learning Guarantees: Unbounded case

However, the variance of the importance weights is bounded.

$$d_w(P||Q) = \frac{\sigma_Q}{\sigma_P^2\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left[-\frac{2\sigma_Q^2(x-\mu)^2 - \sigma_P^2(x-\mu')^2}{2\sigma_P^2}\sigma_Q^2\right] dx$$

# Learning Guarantees: Unbounded case

Intuition: if $\mu = \mu'$ and $\sigma_P >> \sigma_Q$

- $Q$ provides some useful information about $P$
- But sample from $Q$ only has few points far from $\mu$
- A few extreme sample points would have large weights

Likewise, if $\sigma_P = \sigma_Q$ but $\mu >> \mu'$, weights would be negligible.

# Learning Guarantees: Unbounded case

Theorem 3: Let $H$ be a hypothesis set such that $\mathrm{Pdim}(\{L_h(x) : H \in H\}) = p < \infty$. Assume that $d_2(P||Q) < +\infty$ and $w(x) \neq 0$ for all $x$. Then for $\delta > 0$, with probability at least $1 - \delta$, the following holds:

$$R(h) \leq \hat{R}_w(h) + 2^{5/4} \sqrt{d_2(P||Q)} \sqrt[\frac{3}{8}]{\frac{p \log \frac{2me}{p} + \log \frac{4}{\delta}}{m}}$$

# Learning Guarantees: Unbounded case

Proof outline (full proof in of Cortes, Mansour, Mohri, 2010):

- $\Pr\left[\sup_{h\in H} \frac{\mathrm{E}[L_h]-\hat{\mathrm{E}}[L_h]}{\sqrt{\hat{\mathrm{E}}[L_h^2]}} > \epsilon\sqrt{2+log\frac{1}{\epsilon}}\right] \leq$

  $\Pr\left[\sup_{h\in H, t\in\Re} \frac{\hat{\Pr}[L-h>t]-\Pr[L_h>t]}{\sqrt{\hat{\Pr}[L_h>t]}} > \epsilon\right]$
- $\Pr\left[\sup_{h\in H} \frac{R(h)-\hat{R}(h)}{\sqrt{R(h)}} > \epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right] \leq 4\Pi_H(2m)\exp\left(-\frac{m\epsilon^2}{4}\right)$
- $\Pr\left[\sup h \in H \frac{\mathrm{E}[L_h(x)]-\hat{\mathrm{E}}[L_h(x)]}{\sqrt{\mathrm{E}[L_h^2(x)]}} > \epsilon\sqrt{2+\log\frac{1}{\epsilon}}\right] \leq$

  $4\exp\left(p\log\frac{2em}{p} - \frac{m\epsilon^2}{4}\right)$
- $\Pr\left[\sup_{h\in H} \frac{\mathrm{E}[L_h(x)]-\hat{\mathrm{E}}[L_h(x)]}{\sqrt{\mathrm{E}[L_h^2(x)]}} > \epsilon\right] \leq 4\exp\left(p\log\frac{2em}{p} - \frac{m\epsilon^{8/3}}{4^{5/3}}\right)$
- $|\mathrm{E}[L_h(x)] - \hat{\mathrm{E}}[L_h(x)]| \leq$

  $2^{5/4}\max\{\sqrt{E[L_h^2(x)]}, \sqrt{\hat{\mathrm{E}}[L_h^2(x)]}\}\sqrt[\frac{3}{8}]{\frac{p\log\frac{2me}{p}+\log\frac{8}{\delta}}{m}}$

Thus, we can show the following:

$$\Pr\Big[\sup_{h\in H}\frac{R(h)-\hat{R}_w(h)}{\sqrt{d_2(P||Q)}}>\epsilon\Big]\le 4\exp\left(p\log\frac{2em}{p}-\frac{m\epsilon^{8/3}}{4^{5/3}}\right).$$

Where $p = \mathrm{Pdim}(\{L_h(x) : h \in H\}$ is the pseudo-dimension of $H'' = \{w(x)L_h(x) : h \in H\}$. Note, any set shattered by $H'$ is shattered by $H''$, since there exists a subset $B$ of a set $A$ that is shattered by $H''$, such that $H'$ shatters $A$ with witnesses $s_i = r_i/w(x_i)$.

# Overview

- Preliminaries
- Learning guarantee when loss is bounded
- Learning guarantee when loss is unbounded, but second moment is bounded
- Algorithm

# Alternative algorithms

We can generalize this analysis to an arbitrary function $u : X \mapsto R, u > 0$. Let $\hat{R}_u(h) = \frac{1}{m}\sum_{i=1}^{m} u(x_i)L_h(x_i)$ and let $\hat{Q}$ be the empirical distribution: Theorem 4: Let $H$ be a hypothesis set such that $\mathrm{Pdim}(\{L_h(x) : h \in H\}) = p < \infty$. Assume that $0 < \mathrm{E}_Q[u^2(x)] < +\infty$ and $u(x) \neq 0$ for all $x$. Then for any $\delta > 0$ with probability at least $1 - \delta$,

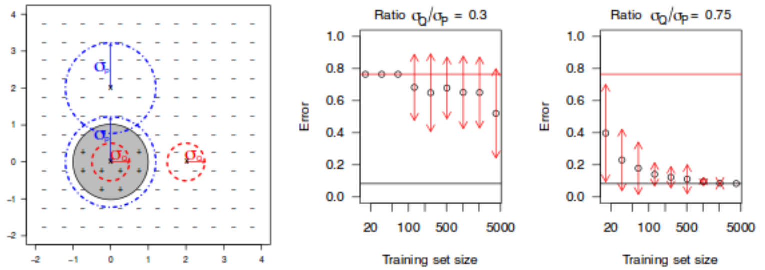
$$|R(h) - \hat{R}_u(h) \leq |E_Q\big[[w(x) - u(x)]L_h(x)\big]|$$

$$+2^{5/4} \max\left(\sqrt{\mathrm{E}_Q[u^2(x)L_h^2(x)]}, \sqrt{\mathrm{E}_{\hat{Q}}[u^2(x)L_h^2(x)]}\right) \sqrt[\frac{3}{8}]{\frac{p \log \frac{2me}{p} + \log \frac{4}{\delta}}{m}}$$

# Alternative algorithms

- Other functions $u$ than $w$ can be used to reweight cost of error
- Minimize upper bound
- $\max\left(\sqrt{\mathrm{E}_Q[u^2]}, \sqrt{\mathrm{E}_{\hat{Q}}[u^2]}\right) \leq \sqrt{\mathrm{E}_Q[u^2]}(1 + O(1/\sqrt{m}))$,
- $\min_{u \in U} \mathrm{E}\Big[|w(x_{-}u(x)|\Big] + \gamma\sqrt{\mathrm{E}_Q[u^2]}$
- Trade-off between bias and variance minimization.

# Alternative algorithms



# Alternative algorithms