Analysis of Kernel Mean Matching under Covariate Shift

Yao-Liang Yu and Csaba Szepesvári

University of Alberta

ICML 2012, Edingburgh

June 29, 2012

Learning from sample

Supervised learning

Given *i.i.d.* sample $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \subseteq \mathcal{X} \times \mathcal{Y}$, learn function $f : \mathcal{X} \mapsto \mathcal{Y}$ that predicts the label Y "well" on the test set $\{X_i^{\mathrm{te}}, Y_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$.

Learning from sample

Supervised learning

Given *i.i.d.* sample $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \subseteq \mathcal{X} \times \mathcal{Y}$, learn function $f : \mathcal{X} \mapsto \mathcal{Y}$ that predicts the label Y "well" on the test set $\{X_i^{\mathrm{te}}, Y_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$.

Well-studied provided that $\mathrm{P}_{\mathrm{te}}(x, y) = \mathrm{P}_{\mathrm{tr}}(x, y)$.

Learning from sample

Supervised learning

Given *i.i.d.* sample $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \subseteq \mathcal{X} \times \mathcal{Y}$, learn function $f : \mathcal{X} \mapsto \mathcal{Y}$ that predicts the label Y "well" on the test set $\{X_i^{\mathrm{te}}, Y_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$.

Well-studied provided that $\mathrm{P}_{\mathrm{te}}(x, y) = \mathrm{P}_{\mathrm{tr}}(x, y)$.

What if $\mathrm{P}_{\mathrm{te}}(x, y) \neq \mathrm{P}_{\mathrm{tr}}(x, y)$, but nontrivially related?

Learning from sample

Supervised learning

Given *i.i.d.* sample $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \subseteq \mathcal{X} \times \mathcal{Y}$, learn function $f : \mathcal{X} \mapsto \mathcal{Y}$ that predicts the label Y "well" on the test set $\{X_i^{\mathrm{te}}, Y_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}$.

Well-studied provided that $\mathrm{P}_{\mathrm{te}}(x, y) = \mathrm{P}_{\mathrm{tr}}(x, y)$.

What if $\mathrm{P}_{\mathrm{te}}(x, y) \neq \mathrm{P}_{\mathrm{tr}}(x, y)$, but nontrivially related?

Covariate Shift (Shimodaira, 2000)

$$\mathrm{P}_{\mathrm{tr}}(y|x) = \mathrm{P}_{\mathrm{te}}(y|x).$$

Previous work

To name a few:

- Huang et al. (2007): kernel mean matching;
- Sugiyama et al. (2008): Kullback-Leibler importance;
- Bickel et al. (2009): logistic regression;
- Kanamori et al. (2012): least-squares;
- Cortes et al. (2008): distributional stability;
- Ben-David et al. (2007) and Blitzer et al. (2008): domain adaptation.

Two books and one monograph:

- *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*, Masashi Sugiyama and Motoaki Kawanabe, MIT, 2012
- *Density Ratio Estimation in Machine Learning*, Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori, Cambridge, 2012
- *Dataset Shift in Machine Learning*, Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer and Neil D. Lawrence, MIT, 2008

The Problem Studied

Predict the mean

Under the covariate shift assumption, construct

$$\hat{f}\left(\{X_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}}; \{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}\right)$$

that approximates $\mathbb{E}(Y^{\mathrm{te}})$ well.

How well?

Can we get a parametric rate, *i.e.* $\mathcal{O}\left(\sqrt{\frac{1}{n_{\mathrm{tr}}} + \frac{1}{n_{\mathrm{te}}}}\right)$?

Why is it interesting?

Relevance

- Given classifiers $\{f_j\}$ trained on $\{(X_i^{\mathrm{tr}}, Z_i^{\mathrm{tr}})\}$, want to rank them based on how well they do on the test set $\{(X_i^{\mathrm{te}}, Z_i^{\mathrm{te}})\}$.
  Fix j and let
$$Y_i^{\mathrm{tr}} = \ell(f_j(X_i^{\mathrm{tr}}), Z_i^{\mathrm{tr}}), \;\; Y_i^{\mathrm{te}} = \ell(f_j(X_i^{\mathrm{tr}}), Z_i^{\mathrm{te}}).$$
- Model-selection/cross-validation under covariate shift.
- Helps understanding the least-squares estimation problem.

Isn't the problem just "trivial"?

Under the covariate shift assumption, the regression function

$$m(x) := \int_{\mathcal{Y}} y \, \mathrm{P}_{\mathrm{tr}}(\mathrm{d}y|x) = \int_{\mathcal{Y}} y \, \mathrm{P}_{\mathrm{te}}(\mathrm{d}y|x)$$

remains unchanged. Estimate $m(\cdot)$ on $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}$ and "plug-in":

$$\hat{y} = \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \hat{m}(X_i^{\mathrm{te}}).$$

Isn't the problem just "trivial"?

Under the covariate shift assumption, the regression function

$$m(x) := \int_{\mathcal{Y}} y \, \mathrm{P}_{\mathrm{tr}}(\mathrm{d}y|x) = \int_{\mathcal{Y}} y \, \mathrm{P}_{\mathrm{te}}(\mathrm{d}y|x)$$

remains unchanged. Estimate $m(\cdot)$ on $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}$ and "plug-in":

$$\hat{y} = \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \hat{m}(X_i^{\mathrm{te}}).$$

Theorem (Smale & Zhou, 2007; Sun & Wu, 2009)

w.p. $1-\delta$, $\left| \frac{1}{n_{\mathrm{te}}} \sum_{i=1}^{n_{\mathrm{te}}} \hat{m}(Y_i^{\mathrm{te}}) - \mathbb{E} Y^{\mathrm{te}} \right| \leq \sqrt{\frac{1}{2n_{\mathrm{te}}} \log \frac{4}{\delta}} + \sqrt{B} C_1 n_{\mathrm{tr}}^{-\frac{3\theta}{12\theta+16}}$

Dependence on n_{tr} is not nice. Algorithm needs to know θ.

A Naive Estimator?

Observe that

$$\mathbb{E}(Y^{\mathrm{te}}) = \int_{\mathcal{X}} m(x)\, \mathrm{P}_{\mathrm{te}}(\mathrm{d}x) = \int_{\mathcal{X}} \beta(x) m(x)\, \mathrm{P}_{\mathrm{tr}}(\mathrm{d}x),$$

where $\beta(x) := \frac{\mathrm{dP}_{\mathrm{te}}}{\mathrm{dP}_{\mathrm{tr}}}(x)$ is the Radon-Nikodym derivative.

A Naive Estimator?

Observe that

$$\mathbb{E}(Y^{\mathrm{te}}) = \int_{\mathcal{X}} m(x)\,\mathrm{P}_{\mathrm{te}}(\mathrm{d}x) = \int_{\mathcal{X}} \beta(x) m(x)\,\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x),$$

where $\beta(x) := \frac{\mathrm{dP}_{\mathrm{te}}}{\mathrm{dP}_{\mathrm{tr}}}(x)$ is the Radon-Nikodym derivative.

Estimate $m(x)$ from $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}$, and estimate $\mathrm{P}_{\mathrm{tr}}(x)$ from $\{X_i^{\mathrm{tr}}\}$, $\mathrm{P}_{\mathrm{te}}(x)$ from $\{X_i^{\mathrm{te}}\}$ respectively. Density estimation is not easy.

A Naive Estimator?

Observe that

$$\mathbb{E}(Y^{\mathrm{te}}) = \int_{\mathcal{X}} m(x)\, \mathrm{P_{te}}(\mathrm{d}x) = \int_{\mathcal{X}} \beta(x) m(x)\, \mathrm{P_{tr}}(\mathrm{d}x),$$

where $\beta(x) := \frac{\mathrm{dP_{te}}}{\mathrm{dP_{tr}}}(x)$ is the Radon-Nikodym derivative.

Estimate $m(x)$ from $\{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}$, and estimate $\mathrm{P_{tr}}(x)$ from $\{X_i^{\mathrm{tr}}\}$, $\mathrm{P_{te}}(x)$ from $\{X_i^{\mathrm{te}}\}$ respectively. Density estimation is not easy.

Why not estimate $\beta(x)$ directly? Much work is devoted into this.

A Better Estimator?

Kernel Mean Matching (Huang et al., 2007)

$$\hat{\beta}^* \in \arg\min_{\hat{\beta}_i} \left\{ \hat{L}(\hat{\beta}) := \left\| \frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \hat{\beta}_i \Phi(X_i^{\text{tr}}) - \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \Phi(X_i^{\text{te}}) \right\|_{\mathcal{H}} \right\}$$

$$\text{s.t.} \quad 0 \leq \hat{\beta}_i \leq B,$$

where $\Phi : \mathcal{X} \mapsto \mathcal{H}$ denotes the *canonical* feature map, $\mathcal{H}$ is the RKHS induced by the kernel k and $\| \cdot \|_{\mathcal{H}}$ stands for the norm in $\mathcal{H}$.
Standard quadratic programming.

Better?

$$\hat{y}_{KMM} := \frac{1}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} \hat{\beta}_i^* Y_i^{\text{tr}}$$

The population version

$$\hat{\beta}^* \in \arg\min_{\hat{\beta}} \left\| \int_{\mathcal{X}} \Phi(x)\hat{\beta}(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x) - \int_{\mathcal{X}} \Phi(x)\mathrm{P}_{\mathrm{te}}(\mathrm{d}x) \right\|_{\mathcal{H}}$$

$$\text{s.t.} \quad 0 \le \hat{\beta} \le B.$$

At optimum we always have

$$\int_{\mathcal{X}} \Phi(x)\hat{\beta}^*(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x) = \int_{\mathcal{X}} \Phi(x)\mathrm{P}_{\mathrm{te}}(\mathrm{d}x).$$

The question is whether

$$\int_{\mathcal{X}} m(x)\hat{\beta}^*(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x) \stackrel{?}{=} \mathbb{E}\, Y^{\mathrm{te}} = \int_{\mathcal{X}} m(x)\beta(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x).$$

The population version

$$\hat{\beta}^* \in \arg\min_{\hat{\beta}} \left\| \int_{\mathcal{X}} \Phi(x)\hat{\beta}(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x) - \int_{\mathcal{X}} \Phi(x)\mathrm{P}_{\mathrm{te}}(\mathrm{d}x) \right\|_{\mathcal{H}}$$

$$\text{s.t.} \quad 0 \le \hat{\beta} \le B.$$

At optimum we always have

$$\int_{\mathcal{X}} \Phi(x)\hat{\beta}^*(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x) = \int_{\mathcal{X}} \Phi(x)\mathrm{P}_{\mathrm{te}}(\mathrm{d}x).$$

The question is whether

$$\int_{\mathcal{X}} m(x)\hat{\beta}^*(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x) \stackrel{?}{=} \mathbb{E}\, Y^{\mathrm{te}} = \int_{\mathcal{X}} m(x)\beta(x)\mathrm{P}_{\mathrm{tr}}(\mathrm{d}x).$$

Yes, if

- $m \in \mathcal{H}$, or;
- k is characteristic (Sriperumbudur et al., 2010).

The empirical version

Assumption (Continuity assumption)

The Radon-Nikodym derivative $\beta(x) := \frac{\mathrm{dP_{te}}}{\mathrm{dP_{tr}}}(x)$ *is well-defined and bounded from above by* $B < \infty$.

Assumption (Compactness assumption)

$\mathcal{X}$ *is a compact metrizable space,* $\mathcal{Y} \subseteq [0, 1]$*, and the kernel* k *is continuous, whence* $\|k\|_\infty \leq C^2 < \infty$.

The empirical version

Assumption (Continuity assumption)

The Radon-Nikodym derivative $\beta(x) := \frac{\mathrm{dP_{te}}}{\mathrm{dP_{tr}}}(x)$ *is well-defined and bounded from above by* $B < \infty$.

Assumption (Compactness assumption)

$\mathcal{X}$ *is a compact metrizable space,* $\mathcal{Y} \subseteq [0, 1]$*, and the kernel* k *is continuous, whence* $\|k\|_\infty \leq C^2 < \infty$.

Theorem

Under our assumptions, if $m \in \mathcal{H}$*, then w.p.* $1 - \delta$,

$$\left| \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \hat{\beta}_i Y_i^{\mathrm{tr}} - \mathbb{E} Y^{\mathrm{te}} \right| \leq (1 + 2C\|m\|_{\mathcal{H}}) \cdot \sqrt{2\left(\frac{B^2}{n_{\mathrm{tr}}} + \frac{1}{n_{\mathrm{te}}}\right) \log \frac{6}{\delta}}.$$

More refined result (more realistic?)

Theorem

Under our assumptions, if

$$\mathcal{A}_2(m, R) := \inf_{\|g\|_{\mathcal{H}} \le R} \|m - g\|_{\mathscr{L}^2_{P_{tr}}} \le C_2 R^{-\theta/2},$$

then w.p. $1 - \delta$,

$$\left| \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \hat{\beta}_i Y_i^{tr} - \mathbb{E} Y^{te} \right| \le \mathcal{O}\left(n_{tr}^{-\frac{\theta}{2(\theta+2)}} + n_{te}^{-\frac{\theta}{2(\theta+2)}}\right).$$

Remarks

- As $\theta \to \infty$, we recover the parametric rate;
- The algorithm (KMM) does not need to know θ.

A pessimistic result

Theorem

Under our assumptions, if

$$\mathcal{A}_\infty(m, R) := \inf_{\|g\|_\mathcal{H} \le R} \|m - g\|_\infty \le C_\infty (\log R)^{-s},$$

then (for n_{tr} and n_{te} large),

$$\left| \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} \hat{\beta}_i Y_i^{\mathrm{tr}} - \mathbb{E}\, Y^{\mathrm{te}} \right| \le \mathcal{O}(\log^{-s} \frac{n_{\mathrm{tr}} \cdot n_{\mathrm{te}}}{n_{\mathrm{tr}} + n_{\mathrm{te}}}).$$

The logarithmic decay is satisfied for C^∞ kernels (such as the Gaussian kernel) when $m \notin \mathcal{H}$, under mild conditions.

Conclusion

Summary

For the problem of predicting the mean under covariate shift,

- the KMM estimator enjoys parametric rate of convergence when $m \in \mathcal{H}$;
- more generally, the KMM estimator converges at $\mathcal{O}(n_{\text{tr}}^{-\frac{\theta}{2(\theta+2)}} + n_{\text{te}}^{-\frac{\theta}{2(\theta+2)}})$;
- on the negative side, the KMM estimator converges at $\mathcal{O}(\log^{-s} \frac{n_{\text{tr}} \cdot n_{\text{te}}}{n_{\text{tr}} + n_{\text{te}}})$ if k does not interact well with m.

Future work

- Lower bounds?
- Extension to least-squares estimation.