Near-Optimal Linear Regression under Distribution Shift

Qi Lei

Princeton University

Joint work with

Wei Hu and Jason Lee.

On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) + \text{noise}, \boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x} \sim p_T}[\|f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}\|^2]$

- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$

On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) + \text{noise}, \boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x}\sim p_T}[\|f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}\|^2]$

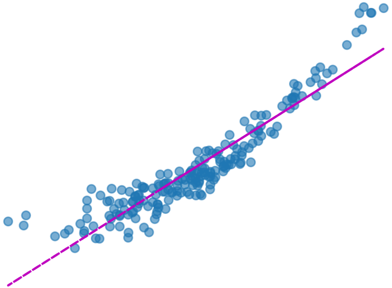
- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$

On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) + \text{noise}, \boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x} \sim p_T}[\| f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x} \|^2]$

- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$

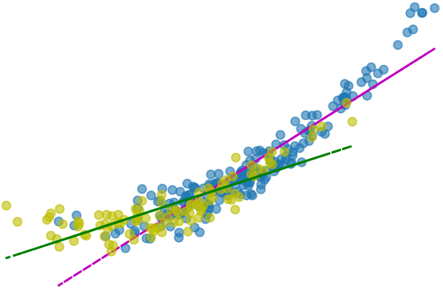


On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) + \text{noise}, \boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x}\sim p_T}[\|f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}\|^2]$

- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$



On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) + \text{noise}, \boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x} \sim p_T}[\|f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}\|^2]$

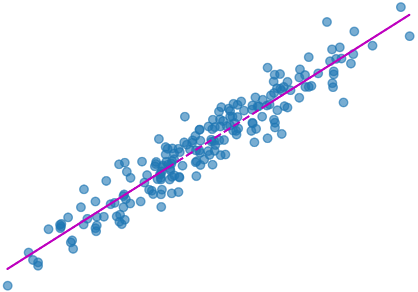
- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$

On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) + \text{noise}, \boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x} \sim p_T}[\| f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x} \|^2]$

- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$

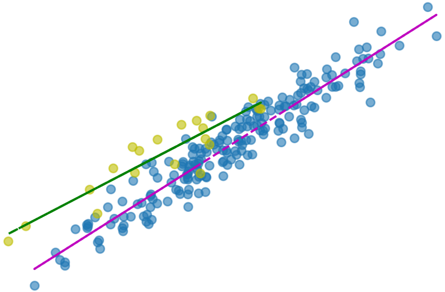


On the Problem of Distributions shift

Setup

- Observations: $y_i = f_S(\boldsymbol{x}_i) +$ noise, $\boldsymbol{x}_i \sim p_S, i = 1, 2, \cdots n_S$
- Estimate a linear model $\hat{\boldsymbol{\beta}}((\boldsymbol{x}_i, y_i)_{i=1}^{n_S})$
- Tested on: $\mathbb{E}_{\boldsymbol{x}\sim p_T}[\|f_T(\boldsymbol{x}) - \hat{\boldsymbol{\beta}}^\top \boldsymbol{x}\|^2]$

- Covariate shift: $f_S = f_T, p_S \neq p_T$
- Model shift: $f_S \neq f_T$ linear, $p_S \neq p_T$



Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{1}$$

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \beta^*, \sigma^2)$,
- measured on source distribution $\boldsymbol{x} \sim p_S$,
- with Gaussian prior $\beta^* \sim \mathcal{N}(0, r^2 I)$.

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \quad (1)$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(x^\top \beta^*, \sigma^2)$,
- measured on source distribution $x \sim p_S$,
- with Gaussian prior $\beta^* \sim \mathcal{N}(0, r^2 I)$. [a]

[a] See e.g. Murphy, K. P. (2012). Machine learning: a probabilistic perspective

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S}\sum_{i=1}^{n_S}(\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2 \quad (1)$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on source distribution $x \sim p_S$,
- with Gaussian prior $\beta^* \sim \mathcal{N}(0, r^2 I)$. [a]

[a] See e.g. Murphy, K. P. (2012). Machine learning: a probabilistic perspective

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \qquad (1)$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on source distribution $\boldsymbol{x} \sim p_S$,
- with Gaussian prior $\beta^* \sim \mathcal{N}(0, r^2 I)$. [a]

[a] See e.g. Murphy, K. P. (2012). Machine learning: a probabilistic perspective

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{1}$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on source distribution $\boldsymbol{x} \sim p_S$,
- with Gaussian prior $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)$. [a]

---

[a] See e.g. Murphy, K. P. (2012). Machine learning: a probabilistic perspective

How to adapt to the target domain?

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{1}$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on source distribution $\boldsymbol{x} \sim p_S$,
- with Gaussian prior $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)$.

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{1}$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on target distribution $\boldsymbol{x} \sim p_T$,
- with Gaussian prior $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)$.

Interestingly, this does not give us different estimator.

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathsf{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{1}$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on target distribution $\boldsymbol{x} \sim p_T$,
- with Gaussian prior $\boldsymbol{\beta}^* \sim \mathcal{N}(0, r^2 I)$.

Revisit Ridge Regression

Ridge Regression as MAP Inference

$$\hat{\boldsymbol{\beta}}_{\mathrm{RR}} \leftarrow \arg\min_{\boldsymbol{\beta}} \frac{1}{n_S} \sum_{i=1}^{n_S} (\boldsymbol{\beta}^\top \boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \tag{1}$$

$\Leftarrow$ This is the optimal linear estimator

- assuming $y \sim \mathcal{N}(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma^2)$,
- measured on target distribution $\boldsymbol{x} \sim p_T$,
- for the worse-case $\boldsymbol{\beta}^*$, $\|\boldsymbol{\beta}^*\| \leq r$.

This gives the "minimax" linear estimator. We developed a meta algorithm for this mechanism.

A Meta Algorithm

Step 1: Find a sufficient statistic $\hat{\boldsymbol{\beta}}_{\mathsf{SS}}$ for the optimal linear estimator given the obervations

Step 2: Solve the best estimator that is linear in $\hat{\boldsymbol{\beta}}_{\mathsf{SS}}$ in the worse-case setting:

$$\hat{\boldsymbol{\beta}}_{\mathsf{MM}} \leftarrow \underset{\boldsymbol{\beta} \text{ linear in } \hat{\boldsymbol{\beta}}_{\mathsf{SS}}}{\arg\min} \max_{\|\boldsymbol{\beta}^*\| \le r} \mathbb{E}_{\mathsf{noise}, \boldsymbol{x} \sim p_T} (\boldsymbol{x}^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*))^2.$$

A Meta Algorithm

Step 1: Find a sufficient statistic $\hat{\boldsymbol{\beta}}_{\mathsf{SS}}$ for the optimal linear estimator given the obervations

Step 2: Solve the best estimator that is linear in $\hat{\boldsymbol{\beta}}_{\mathsf{SS}}$ in the worse-case setting:

$$\hat{\boldsymbol{\beta}}_{\mathsf{MM}} \leftarrow \underset{\boldsymbol{\beta} \text{ linear in } \hat{\boldsymbol{\beta}}_{\mathsf{SS}}}{\arg\min} \max_{\|\boldsymbol{\beta}^*\| \le r} \mathbb{E}_{\text{noise}, \boldsymbol{x} \sim p_T} (\boldsymbol{x}^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^*))^2.$$

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Results on Different Settings

- Covariate shift under Gaussian sequence model
  1. developed algorithms with only unlabeled target data and/or
  2. with few labeled target data,
  3. show that our algorithm is optimal among all linear estimators, and
  4. is within constant of the best nonlinear estimators under some conditions,
  5. prove a separation result between ours and ridge regression
- Covariate shift with approximation error (nonlinear model)
  1. developed algorithm that is
  2. asymptotically optimal among all linear estimators,
  3. provide a practical way to estimate the selection bias on source distribution,
  4. tested on real dataset
- Small model shift under Gaussian sequence model
  1. developed algorithms that balance model shift and variance,
  2. prove optimality among all linear estimators, and
  3. is within constant of the best nonlinear estimators under some conditions

Thank you!