Transfer Learning on Generalized Linear Regression

Mathias Bourel

Universidad de la Republica

Jairo Cugliari

Univ Lumière Lyon 2

August 7th 2023
JSM '23 – Toronto

Landscape of Transfer Learning (TL)

Basic definitions

domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$

task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

. Pan, S.J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345-1359.

Landscape of Transfer Learning (TL)

Basic definitions

domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$

task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

Given $\{\mathcal{D}_S, \mathcal{T}_S, \mathcal{D}_T, \mathcal{T}_T\}$, TL aims to help improve the learning of $f_T(\cdot)$ using information in $\mathcal{D}_S, \mathcal{T}_S$ where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

. Pan, S.J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345-1359.

Landscape of Transfer Learning (TL)

Basic definitions

domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$

task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

Given $\{\mathcal{D}_S, \mathcal{T}_S, \mathcal{D}_T, \mathcal{T}_T\}$, TL aims to help improve the learning of $f_T(\cdot)$ using information in $\mathcal{D}_S, \mathcal{T}_S$ where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

3 main research issues

what, **how** and **when** to transfer ?

. Pan, S.J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345-1359.

Landscape of Transfer Learning (TL)

Basic definitions

domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$

task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$

Given $\{\mathcal{D}_S, \mathcal{T}_S, \mathcal{D}_T, \mathcal{T}_T\}$, TL aims to help improve the learning of $f_T(\cdot)$ using information in $\mathcal{D}_S, \mathcal{T}_S$ where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$.

3 main research issues

what, **how** and **when** to transfer ?

settings

inductive, transductive & unsupervised TL

approaches

instance, feature, parameter, RK

. Pan, S.J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345-1359.

The linear regression (LR) setting

Task : regression ($\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \mathbb{R}$, $f(\cdot) = E[Y|X = \cdot]$)

- $\hat{f}(x)$ is an estimate of $f(x)$
- $\hat{y} = \hat{f}(x)$ is a prediction of y
- $\mathcal{R} = \mathbb{E}[(y - \hat{y})^2]$ (L_2 risk)

. Obst et al. (2022). Improved linear regression prediction by transfer learning. CSDA, vol. 174.

The linear regression (LR) setting

Task : regression ($\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \mathbb{R}$, $f(\cdot) = E[Y|X = \cdot]$)

- $\hat{f}(x)$ is an estimate of $f(x)$
- $\hat{y} = \hat{f}(x)$ is a prediction of y
- $\mathcal{R} = \mathbb{E}[(y - \hat{y})^2]$ (L_2 risk)

Gain of using the alternative predictor $\hat{f}_\square(x)$

$$\Delta\mathcal{R}(x) = \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_\square(x))^2]$$

. Obst et al. (2022). Improved linear regression prediction by transfer learning. CSDA, vol. 174.

The linear regression (LR) setting

Task : regression ($\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \mathbb{R}$, $f(\cdot) = E[Y|X = \cdot]$)

- $\hat{f}(x)$ is an estimate of $f(x)$
- $\hat{y} = \hat{f}(x)$ is a prediction of y
- $\mathcal{R} = \mathbb{E}[(y - \hat{y})^2]$ (L_2 risk)

Gain of using the alternative predictor $\hat{f}_\square(x)$

$$\Delta\mathcal{R}(x) = \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_\square(x))^2]$$

Linear model ($f(x) = x^\top\beta$, fixed design)

Data : $(X_{v,i}, Y_{v,i}), i = 1, \ldots, N_v$ with $v \in \{S, T\}$ and $N_T \ll N_S$

$$Y_v = X_v\beta_v + \epsilon_v, \qquad \mathbb{E}\epsilon_v = 0, \mathbb{E}\|\epsilon_v\|^2 = \sigma_v^2, \qquad \hat{\beta}_v = (X_v^\top X_v)^{-1}X_v^\top Y_v, \qquad v \in S, T$$

. Obst et al. (2022). Improved linear regression prediction by transfer learning. CSDA, vol. 174.

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\Delta\mathcal{R}(x) = \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2]$$

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

- Gain for any new x if H is known ! ! !

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

- Gain for any new x if H is known!!!
- Since H is a quadratic form, the eigen structure $(\lambda_j, e_j, j = 1, \dots, D)$ is key

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

- Gain for any new x if H is known ! ! !
- Since H is a quadratic form, the eigen structure $(\lambda_j, e_j, j = 1, \dots, D)$ is key
  - [max|min] gain = $[\lambda_{max}|\lambda_{min}]$ eigen value of H in the direction of $[e_{max}|e_{min}]$

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

- Gain for any new x if H is known ! ! !
- Since H is a quadratic form, the eigen structure $(\lambda_j, e_j, j = 1, \dots, D)$ is key
  - [max|min] gain = $[\lambda_{max}|\lambda_{min}]$ eigen value of H in the direction of $[e_{max}|e_{min}]$
  - No possible positive transfer if $\lambda_{max} < 0$, while if $\lambda_{min} > 0$ transfer is positive $\forall x$

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

- Gain for any new x if H is known!!!
- Since H is a quadratic form, the eigen structure $(\lambda_j, e_j, j = 1, \dots, D)$ is key
  - [max|min] gain = $[\lambda_{max}|\lambda_{min}]$ eigen value of H in the direction of $[e_{max}|e_{min}]$
  - No possible positive transfer if $\lambda_{max} < 0$, while if $\lambda_{min} > 0$ transfer is positive $\forall x$
  - Let $\mathrm{Tr}(A)$: trace of A, $\Sigma_v = (1/N_v) X_v^\top X_v)$, then positive gain if

$$N_S \geq \frac{\sigma_S^2 \mathrm{Tr}(\Sigma_S^{-1})}{(\sigma_T^2/N_T)\mathrm{Tr}(\Sigma_T^{-1}) - \|\beta_S - \beta_T\|^2}$$

Ex. 1) Non-adaptive transfer : $\hat{f}_{\square}(x) = \hat{f}_S(x)$

What happens if we estimate $\hat{f}_S(x) = x^\top \hat{\beta}_S$ and predict $\hat{y}_T = \hat{f}_S(x)$?

$$\begin{aligned}\Delta\mathcal{R}(x) &= \mathbb{E}[(y_T - \hat{f}_T(x))^2] - \mathbb{E}[(y_T - \hat{f}_{\square}(x))^2] \\ &= x^\top \underbrace{(\sigma_T^2 (X_T^\top X_T)^{-1} - \sigma_S^2 (X_S^\top X_S)^{-1} - (\beta_T - \beta_S)(\beta_T - \beta_S)^\top)}_{:=H} x\end{aligned}$$

- Gain for any new x if H is known!!!
- Since H is a quadratic form, the eigen structure $(\lambda_j, e_j, j = 1, \dots, D)$ is key
  - [max|min] gain = $[\lambda_{max}|\lambda_{min}]$ eigen value of H in the direction of $[e_{max}|e_{min}]$
  - No possible positive transfer if $\lambda_{max} < 0$, while if $\lambda_{min} > 0$ transfer is positive $\forall x$
  - Let $\mathrm{Tr}(A)$: trace of A, $\Sigma_v = (1/N_v) X_v^\top X_v)$, then positive gain if

$$N_S \geq \frac{\sigma_S^2 \mathrm{Tr}(\Sigma_S^{-1})}{(\sigma_T^2/N_T)\mathrm{Tr}(\Sigma_T^{-1}) - \|\beta_S - \beta_T\|^2}$$

In general H will be indefinite and its spectral components difficult to estimate

A test of positive gain : $\mathcal{H}_o) \Delta\mathcal{R}(x) \leq 0$ vs $\mathcal{H}_a) \Delta\mathcal{R}(x) > 0$

Result

Let $\hat{\sigma}_v^2 = \|Y_v - X_v\hat{\beta}_v\|^2/(N_v - D)$ be the estimator of noise variances σ_v^2. Consider $\hat{\rho}$ an estimator of the the quantity $\rho \geq \|\beta_T - \beta_S\|/\sigma_T$. Then, the following test is of approximate level a to test $\mathcal{H}_o)$ against $\mathcal{H}_a)$

$$\mathbb{1}\left(\psi(x) := \frac{\hat{\sigma}_T^2}{\hat{\sigma}_S^2} \frac{x^\top (X_T^\top X_T - \hat{\rho})^{-1} x}{x^\top (X_S^\top X_S)^{-1} x} > q^{1-a}\right),$$

where q^{1-a} is the quantile of order $1-a$ of r.v. $F \sim \mathcal{F}_{N_T-D, N_S-D}$. The associated p-value is

$$p(x) = \mathbb{P}(F > \psi(x))$$

Ex. 2) Transfer by fine-tuning : $\hat{f}_{\square}(x) = \Lambda(\hat{f}_S(x))$

- We use batch gradient descent (GD) of step size α on the $\{\mathcal{D}_T, \mathcal{T}_t\}$ and L_2 error
- At iteration k, the estimator of β can be written as (compare to Chen et al.)

$$\hat{\beta}_k = A^k \hat{\beta}_S + (I - A^k)\hat{\beta}_T, \qquad A = (I - \alpha\Sigma_T), \text{with } \Sigma_T = X_T^\top X_T$$

- Define $\Omega_k = \alpha^{-1}\Sigma_T^{-1}(I_D - A^k)$ and $B = (\beta_T - \beta_S)(\beta_T - \beta_S)^\top$, then the gain is

$$\Delta\mathcal{R}(x) = x^\top \left(\sigma_T^2(\Sigma_T^{-1} - \alpha^2\Omega_k\Sigma_T\Omega_k) - \sigma_S^2 A^k\Sigma_S^{-1}A^k - A^k B A^k\right) x := x^\top H_{\alpha,k} x$$

with a corresponding expression for the test.

. Chen, A., Owen, A.B., & Shi, M. (2013). Data Enriched Linear Regression. Electronic Journal of Statistics, 9, 1078-1112.

Adaptations to Generalized Linear Models (GLM)

The new model is $y = \Psi(x^\top \beta) + \epsilon$. We have to choose,

- a probability distribution for y,
- a loss function (needed for GD, risk),
- an expression of $\Delta\mathscr{R}$.

Adaptations to Generalized Linear Models (GLM)

The new model is $y = \Psi(x^\top \beta) + \epsilon$. We have to choose,

- a probability distribution for y, $y \sim f_\theta \in \mathcal{EF}$, the exponential family,

$$\mathcal{EF} = \{f_\theta : f_\theta(y) = \exp(\theta y - \Psi(\theta)\}$$

- a loss function (needed for GD, risk),
- an expression of $\Delta\mathcal{R}$.

Adaptations to Generalized Linear Models (GLM)

The new model is $y = \Psi(x^\top \beta) + \epsilon$. We have to choose,

- a probability distribution for y, $y \sim f_\theta \in \mathcal{EF}$, the exponential family,

$$\mathcal{EF} = \{f_\theta : f_\theta(y) = \exp(\theta y - \Psi(\theta)\}$$

- a loss function (needed for GD, risk), $B_\Phi(p, q) = \Phi(p) - \Phi(q) - \langle \nabla_\Phi(q), p - q \rangle$
- an expression of $\Delta\mathcal{R}$.

Adaptations to Generalized Linear Models (GLM)

The new model is $y = \Psi(x^\top \beta) + \epsilon$. We have to choose,

- a probability distribution for y, $y \sim f_\theta \in \mathcal{EF}$, the exponential family,

$$\mathcal{EF} = \{f_\theta : f_\theta(y) = \exp(\theta y - \Psi(\theta)\}$$

- a loss function (needed for GD, risk), $B_\Phi(p, q) = \Phi(p) - \Phi(q) - \langle \nabla_\Phi(q), p - q \rangle$
- an expression of $\Delta\mathcal{R}$. Difference of deviances

Example : the logistic regression

Task : regression ($\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \mathbb{R}$, $f(\cdot) = E[\Psi(Y)|X = \cdot]$), Ψ the logit link function

- $\hat{f}(x)$ is an estimate of $f(x)$
- $\hat{y} = \hat{f}(x)$ is a prediction of y
- $\mathcal{R} = \mathbb{E}[l(y, \hat{y})]$, with l the negative binary entropy)

Example : the logistic regression

Task : regression ($\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \mathbb{R}$, $f(\cdot) = E[\Psi(Y)|X = \cdot]$), Ψ the logit link function

- $\hat{f}(x)$ is an estimate of $f(x)$
- $\hat{y} = \hat{f}(x)$ is a prediction of y
- $\mathcal{R} = \mathbb{E}[l(y, \hat{y})]$, with l the negative binary entropy)

Gain of using the alternative predictor $\hat{f}_{\square}(x)$

$$\Delta\mathcal{R}(x) = \mathbb{E}[l(y_T, \hat{y}_T)] - \mathbb{E}[l(y_T, \hat{y}_{\square}(x)]$$

Example : the logistic regression

Task : regression ($\mathcal{X} = \mathbb{R}^D$, $\mathcal{Y} = \mathbb{R}$, $f(\cdot) = E[\Psi(Y)|X = \cdot]$), Ψ the logit link function

- $\hat{f}(x)$ is an estimate of $f(x)$
- $\hat{y} = \hat{f}(x)$ is a prediction of y
- $\mathcal{R} = \mathbb{E}[l(y, \hat{y})]$, with l the negative binary entropy)

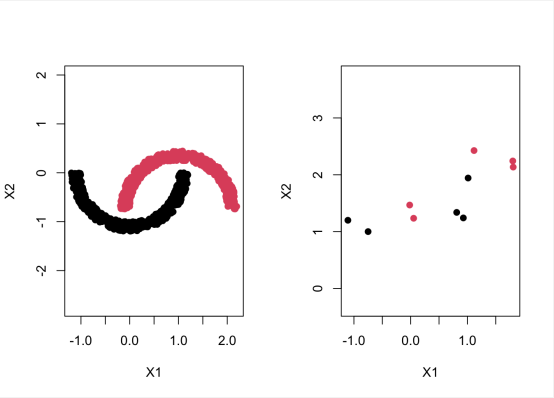
Gain of using the alternative predictor $\hat{f}_{\square}(x)$

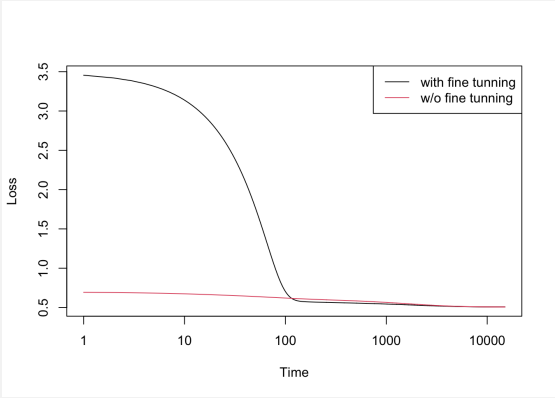
$$\Delta\mathcal{R}(x) = \mathbb{E}[l(y_T, \hat{y}_T)] - \mathbb{E}[l(y_T, \hat{y}_{\square}(x)]$$

Generalized Linear model ($f(x) = \Psi(x^\top\beta)$, fixed design)

Estimation : use IWLS on source and GD (neg binary entropy) on target

Simulated dataset





Conclusion

- The framework for testing transferability is adaptable to generalized linear regression models.
- Strong use of the connection between Exponential Family and Bernstein divergences.
- Further work : gain insight on the gain to obtain a test.