Uncoupled Regression from Pairwise Comparison Data

Liyuan Xu[1,2], Junya Honda[1,2], Gang Niu[2], Masashi Sugiyama[2,1]
[1]The University of Tokyo [2]RIKEN

Abstract

- Introduce a novel uncoupled regression problem with pairwise comparison data.
- Propose new empirical risk minimization methods to solve the problem.
- Propose two estimators for a risk for general marginal target distributions.

Problem Setting

Motivating Example

Sensitive Data:

- e.g. Salary, Number of crimes committed before,...
- People won't give an explicit label.
- Containing sensitive data leads to the risk of security breach.

Goal: To build a prediction model from marginal distribution of sensitive data

Uncoupled Regression

Ordinary Regression:

- Target Y generated from feature X as $Y = f(X) +$.
- Learn a model $\hat{f}$ from coupled data $D_{X,Y} = \{(x_i, y_i)\}$ generated from joint distribution $P_{X,Y}$.

Uncoupled Regression:

- Unlabeled data $D_X = \{x_i\}$ generated from distribution $P_X(x) = \int P_{X,Y}(x, y)dy$.
- Target $D_Y = \{y_i\}$ generated from marginal distribution $P_Y(y) = \int P_{X,Y}(x, y)dx$.
- We try to learn a model $\hat{f}$ to predict Y from X.

Since no correspondence in D_X and D_Y,
problem is ill-posed without any further assumption.

Pairwise Comparison Data

Di□cult to get sensitive data but easier to get their order.

- People might be willing to give order information.

Pairwise Comparison Data $D_R = \{x_i^+, x_i^-\}$

- Consists of pairs of features $\{x_i^+, x_i^-\}$ such that

$$y(x_i^+) > y(x_i^-),$$

where $y(x_i^+), y(x_i^-)$ are the target values for x_i^+, x_i^-, respectively.

- Generation Process:
  1. Generate two samples $(X, Y), (X, Y)$ from joint distribution $P_{X,Y}$.
  2. If $Y \geq Y$, $X^+ = X$ and $X^- = X$. If not, the opposite holds.
- Let P_{X^+}, P_{X^-} be the distributions of each comparison data.

Problem

Learn a model $\hat{f}$ from Unlabeled data D_X, Target values D_Y, and Comparative Data D_R

Related Work

[Carpentier and Schlueter, 2016]

Uncoupled regression with one-dimensional features and monotonic target function f.

- Requires features to be one-dimensional and target function f to be monotonic.
- Involves complex optimization in learning.

Our Problem

- Applicable to features with multi-dimensions and non-monotonic f.
- Easy to implement. (Can be reduced to minimization of a convex function.)

Algorithm

Empirical Risk Minimization Principle

- Construct an unbiased estimator of risk (e.g. l_2-risk) from coupled data.
- Minimize the unbiased risk estimator to learn a model $\hat{f}$.

Distribution of Comparison Data

Lemma Let F_Y be the cumulative distribution function of Y. Then,

$$E_{X^+}[f(X^+)] = 2E_{X,Y}[F_Y(Y)f(X)],$$
$$E_{X^-}[f(X^-)] = 2E_{X,Y}[(1 - F_Y(Y))f(X)],$$

Therefore, if $F_Y(y) = y$ (i.e. marginal distribution P_Y is uniform on [0, 1]),

$$E_{X,Y}[(Y - f(X))^2] = E_Y[Y^2] + E_X[(f(X))^2] - 2E_{X,Y}[Y f(X)]$$
$$= E_Y[Y^2] + E_X[(f(X))^2] - E_{X^+}[f(X^+)]$$
$$\xleftarrow{\text{Unbiased}} E_Y[Y^2] + \frac{1}{|D_X|}\sum_{x_i \in D_X}(f(x_i))^2 - \frac{1}{|D_R|}\sum_{(x_i^+, x_i^-) \in D_R} f(x_i^+)$$

- $E_Y[Y^2]$ does not depend on f and can be ignored in optimization.
- $\{x_i^-\}$ can be used for variance reduction.
- However, we cannot construct unbiased estimators for all marginal distributions.

→ Propose two methods to construct estimators with small bias.

Risk Approximation Approach

Main Idea

Approximate the expectation $E_{X,Y}[Y f(X)]$
by the linear combination $w_1E_{X^+}[f(X^+)] + w_2E_{X^-}[f(X^-)]$, w_1, w_2 R.

Theorem Let $\hat{f}_{RA}$ be the minimizer of

$$E_Y[Y^2] + \frac{1}{|D_X|}\sum_{x_i \in D_X}(f(x_i))^2 - \frac{1}{|D_R|}\sum_{(x_i^+, x_i^-) \in D_R}\left(w_1 f(x_i^+) + w_2 f(x_i^-)\right). \quad (1)$$

Then, with an adequate condition, with probability $1 - \delta$,

$$R(\hat{f}_{RA}) \leq R(f) + O\left(\sqrt{\frac{\log 1/\delta}{n_U}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{n_R}}\right) + \mathrm{Err}(w_1, w_2)$$

holds, where R is l_2-risk, F_Y is the CDF of Y and Err is defined as

$$\mathrm{Err}(w_1, w_2) = E_Y[|Y - w_1F_Y(Y) - w_2(1 - F_Y(Y))|].$$

- When marginal distribution P_Y is uniform, estimator (1) is unbiased for R.
  - In this case, $\mathrm{Err} = 0$ with $(w_1, w_2) = (1, 0)$.
- Can be generalized to any risk defined based on Bregman divergence.
  - See the paper for the detail.

Referece

- A. Carpentier and T. Schlueter. Learning relationships between data obtained independently. In Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, 2016.

Target Transformation Approach

Main Idea

Transform target Y to $F_Y(Y)$, and minimize the risk on transformed variable:

$$E_{X,Y}[(F_Y(Y) - F_Y(f(X)))^2].$$

Note, marginal distribution of $F_Y(Y)$ is uniform on [0, 1].

Theorem With an appropriate condition, the minimizer $\hat{f}_{TT}$ of

$$E_Y[Y^2] + \sum_{i=1}^{n_U}(F_Y(f(x_i)))^2 - \sum_{i=1}^{n_R}F_Y(f(x_i^+)) \quad (2)$$

satisfies

$$R(\hat{f}_{TT}) \leq R(f) + O\left(\sqrt{\frac{\log 1/\delta}{n_U}}\right) + O\left(\sqrt{\frac{\log 1/\delta}{n_R}}\right) + \Delta_{TT}$$

with probability $1 - \delta$, where R is l_2-risk.

- When marginal is uniform, estimator (2) is unbiased for R, since $F_Y(y) = y$.
- Δ_{TT} depends on the shape of P_Y and noise level.
- The theorem only holds for l_2-risks.

Experiments

Settings

- Used benchmark datasets from the UCI repository.
- Used original features as unlabeled data and sampled 5000 pairs of comparison data.
- Learned linear models and predicted the target value for unlabelled data.

Methods to be compared

- Linear Regression using fully labeled ordinary coupled data.
- Train SVMRank using pairwise comparison data, predict ranking, and predict value by

$$\hat{f}(x) = F_Y^{-1}\left(\frac{\hat{n}(x)}{n_U}\right),$$

where $\hat{n}(x)$ is the predicted rank in the data.

Results

	Supervised Regression	Uncoupled Regression		
Dataset	LR	SVMRank	RA	TT
housing	24.5(5.0)	110.3(29.5)	29.5(6.9)	**22.5(6.2)**
diabetes	3041.9(219.8)	8575.9(883.1)	**3087.3(256.3)**	**3127.3(278.8)**
airfoil	23.3(2.2)	62.1(7.6)	23.7(2.0)	**22.7(2.2)**
concrete	109.5(13.3)	322.9(45.8)	**111.7(13.2)**	139.1(17.9)
powerplant	20.6(0.9)	372.2(34.8)	**21.8(1.1)**	**22.0(1.0)**
mpg	12.1(2.04)	125(15.1)	12.8(2.16)	**10.3(2.08)**
redwine	0.412(0.0361)	1.28(0.112)	**0.442(0.0473)**	0.466(0.0412)
whitewine	0.574(0.0325)	1.58(0.0691)	**0.597(0.0382)**	0.644(0.0414)
abalone	5.05(0.375)	20.9(1.44)	**5.26(0.372)**	5.54(0.424)

Better than SVMRank, and may be better than ordinal supervised learning