Machine learning methods for causal inference

Kyungseon Lee, Yeonho Jung

December 9, 2022

Seoul National University

Outline

Introduction

Introduction(1/2)

1. What is the purpose of using **Propensity Score** ?
   - Several methods exist when applying to propensity scores
     - Covariate adjustment, Matching
     - Stratification or subclassification and **Weighting**
   - Use the estimated propensity score in an observational study

Definition (Propensity score)

The probability of receiving a treatment conditional on a set of observed covariates, $e(X_i) = Pr(Z_i = 1|X_i)$

2. **Inverse Probability weighting : IPW**
   - To be designed **similar to randomized experiments**
     - To make the groups more comparable
     - Can be an alternative for **matching** :
       : Using all data not discarding unmatched sets
   - Aim to generate a **pseudo-population** : the treatment is independent of confounders

Introduction(2/2)

Estimating Propensity score and its weight

1. **Estimated Propensity Score : $\hat{e}(X_i)$**
   - $e(X_i)$ is unknown in an observational study
   - Estimation from the data
     - Logistic regression (normally)
     - Several **Machine learning techniques**
       : Classification and Regression Tree(CART), Random Forest, bagged CART, boosted CART
2. **Inverse Probability weighting : IPW**
   - For the ATE estimation, IPW ?
     - $\widehat{\Delta}_{IPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i \cdot Y_i}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)\cdot Y_i}{1-\hat{e}_i}$
   - Because of traits of IPW varing from 0 to 1,
     - Estimation based on SIPW (Stabilized IPW),
     - $\widehat{\Delta}_{SIPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i \cdot Y_i}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)\cdot Y_i}{1-\hat{e}_i}$

Main Framework

Main Framework(1/3)

'Propensity scores are generally estimated using logistic regression'

- **Parametric models require assumptions regarding variable selection** : functional form, distributions and so on...
- The use of **machine learning** methods as an alternative to logistic regression

'Machine learning'

- **Extract the relationship** b/w outcome and predictor through an algorithm without an a *priori* data model
  : Contrary to assuming a data model with parameters
- **Decision tree** $\rightarrow$ CART and Pruned CART
- **Ensemble methods**
  : Bagged CART, , Boosted CART, Random Forest

Main Framework(2/3)

Compare 'Logistic Regression' with 'Machine learning methods' using 'Propensity score and its weighting'

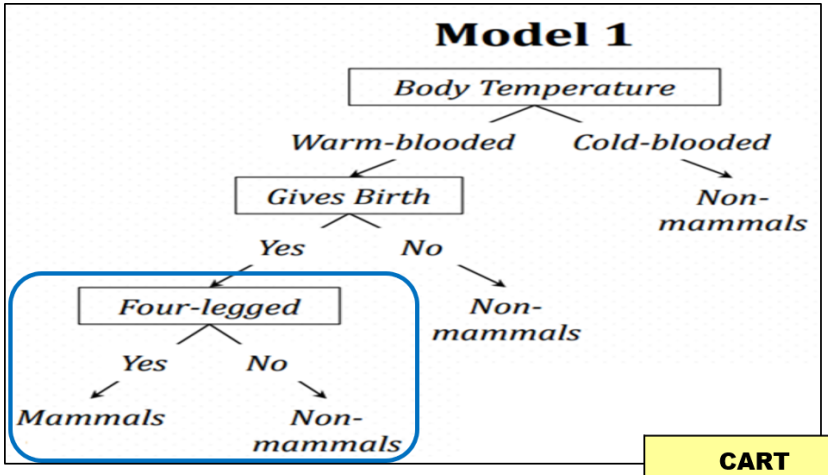
1. **Logistic regression** as parametric models
2. **CART and pruned CART**
   - Within each node of the tree, data will have similar probabilities
   - Sensitive to over-fitting for CART when too many nodes in a tree
   - Pruning, reducing the number of tree splits : less sensitive to noise and generalize to new data
3. **Random Forest** as ensemble method
   - Bagging : randomly sub-sampling from the data set
   - Decision trees : created by using a different subset of data points from the data set
   - The most voted values are chosen

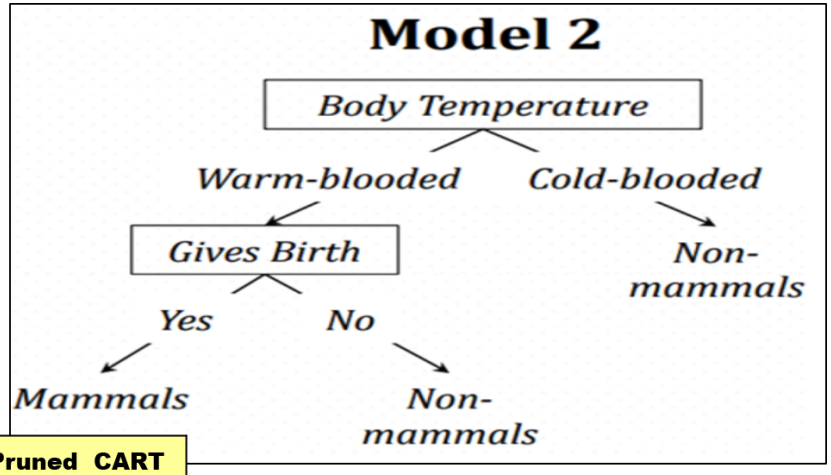
Main Framework(3/3)

Evaluate performance of propensity score (fitting) methods in the paper.

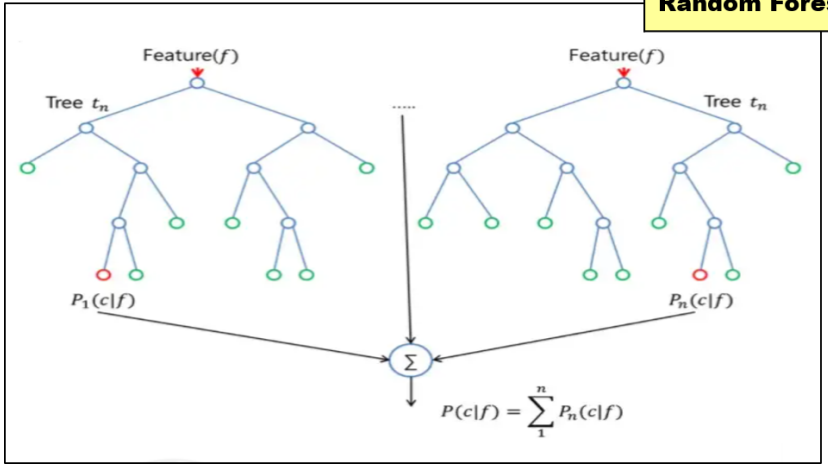
- **ASAM** : Average Standardized absolute mean difference,
  a measure of covariate balance.
  : lower ASAM means the treatment and the controlled are similar.
- **Bias** : the percentage difference from the true treatment effect.
- **SE** : standard error of the effect estimate.
- **Weight** : the distribution of weights for the untreated observations.

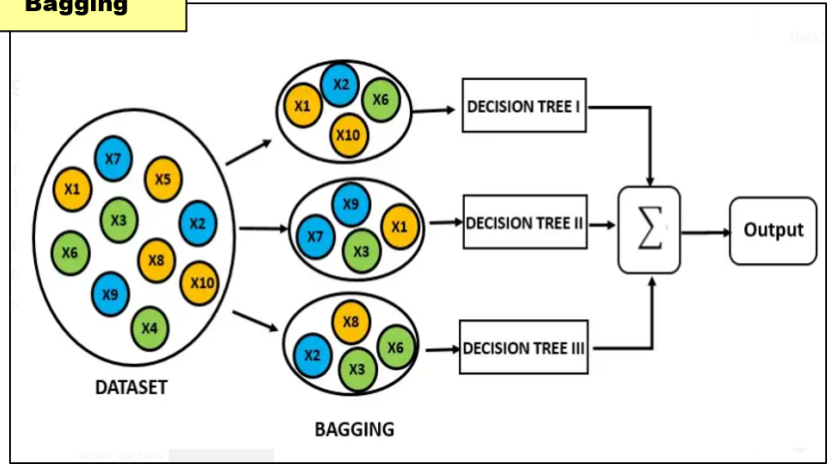
Several models for estimation





CART	Pruned CART
Random Forest	Bagging





Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

Dataset

- Chemical Dataset
  - ▶ We estimate ATE and confirm if poisox is the cause of mortality.
  - ▶ The columns consist of **age, sex, blood difference, poisox and mortality**.
- Korea National Health and Nutrition Examination Survey Dataset
  - ▶ We determine whether drinking is the cause of Hepatitis.
  - ▶ The columns consist of **gender, age, drinking status, AST, ALT, BMI and Hepatitis**.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

The paper explains how good propensity score estimation using machine learning is by comparing the **treatment effect estimation** of the simulated datasets of the 7 scenarios with the **true treatment effect value**.

- A: additivity and linearity (main effects only)
- B: mild non-linearity (one quadratic term)
- C: moderate non-linearity (three quadratic terms)
- D: mild non-additivity (three two-way interaction terms)
- E: mild non-additivity and non-linearity (three two-way interaction terms and one quadratic term)
- F: moderate non-additivity (ten two-way interaction terms)
- **G: moderate non-additivity and non-linearity (ten two-way interaction terms and three quadratic terms).**

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

The paper explains how good propensity score estimation using machine learning is by comparing the **treatment effect estimation** of the datasets with the **true treatment effect value**.

▶ Since we cannot know the true treatment effect of the **real life datasets**,

▶ We evaluate the performance by obtaining only the

- ASAM
- weight distribution
- propensity score distribution

for each models among the performance metrics implemented in the paper.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

ASAM: Average Standardized Absolute Mean difference.

$$ASAM_j = \left| \frac{\sum_{i=1}^{N} W_i Z_i C_{ij}}{N_t} - \frac{\sum_{i=1}^{N} W_i (1 - Z_i) C_{ij}}{N_C} \right| / \sigma_j$$

$$\sigma_j = \frac{\sum_{i=1}^{N} \left(W_i Z_i C_{ij} - \sum_{i=1}^{N} W_i Z_i C_{ij} / N_t \right)^2}{N_t}$$

- C_j : the jth covariate.
- W_i : the ith weight.
- Z_i : the ith treatment.
- σ_j : the standard deviation of the jth covariate in the treatment group.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

ASAM: Average Standardized Absolute Mean difference.

$$ASAM_j = \left| \frac{\sum_{i=1}^{N} W_i Z_i C_{ij}}{N_t} - \frac{\sum_{i=1}^{N} W_i (1 - Z_i) C_{ij}}{N_C} \right| / \sigma_j$$

$$\sigma_j = \frac{\sum_{i=1}^{N} \left(W_i Z_i C_{ij} - \sum_{i=1}^{N} W_i Z_i C_{ij} / N_t \right)^2}{N_t}$$

- A lower ASAM indicates that the treatment and comparison groups are more similar with respect to the given covariates.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

Metric	Method[1]	Scenario[4]						
		A	B	C	D	E	F	G
ASAM[2]	LGR	0.041	0.042	0.058	0.056	0.061	0.068	0.094
	CART	0.159	0.148	0.143	0.171	0.162	0.15	0.143
	PRUNE	0.175	0.164	0.148	0.182	0.173	0.161	0.151
	BAG	0.132	0.127	0.121	0.144	0.141	0.119	0.112
	RFRST	0.08	0.076	0.076	0.089	0.086	0.077	0.075
	BOOST	0.068	0.065	0.067	0.073	0.071	0.065	0.067

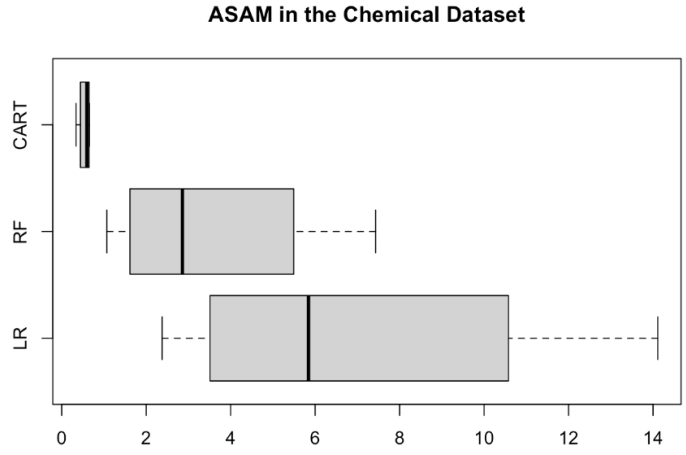
	ASAM in Chemical Dataset
Logistic Regression	7.0443691
Random Forest	3.5544320
CART	0.5460436

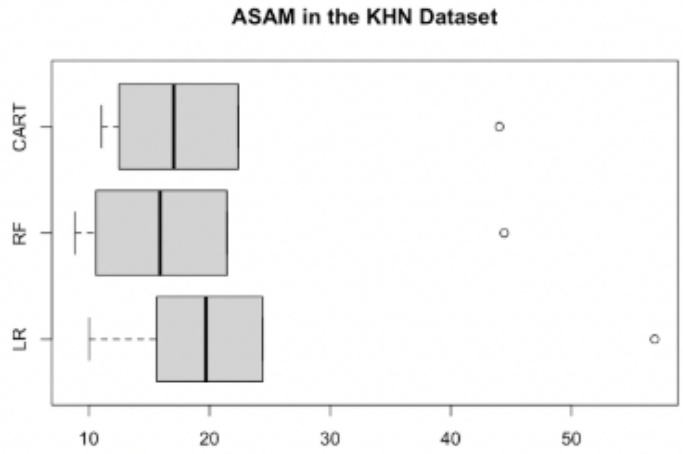
	ASAM in KHN Dataset
logistic regression	24.38727
Random Forest	19.51720
CART	20.66538

- First figure is ASAM table of simulated dataset in the 7 scenarios
- We can see that the ASAM is decreasing when changing from a logistic regression model to a CART model.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

ASAM: Average Standardized Absolute Mean difference.





▶ In the case of logistic regression model, ASAM of some covariates are small. But ASAM of other covariates are larger than others. It means that balance is different by covariate.

▶ It is improved a little bit in the Random Forest model and ASAM of most covariates are decreased in the CART model.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

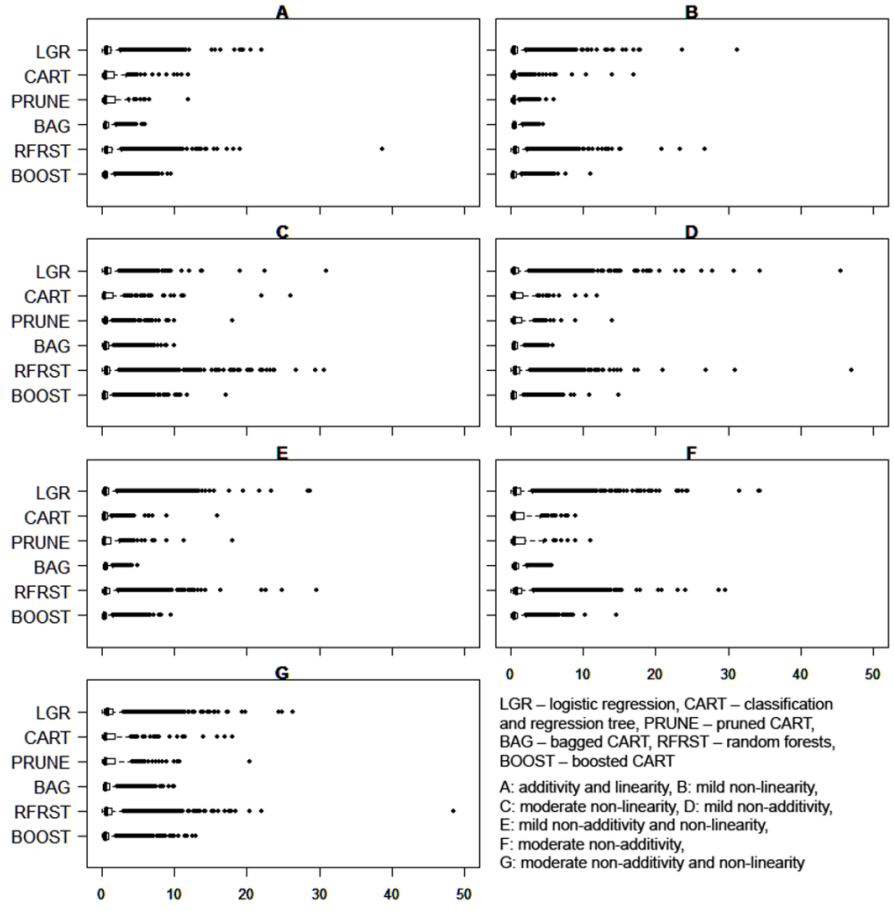
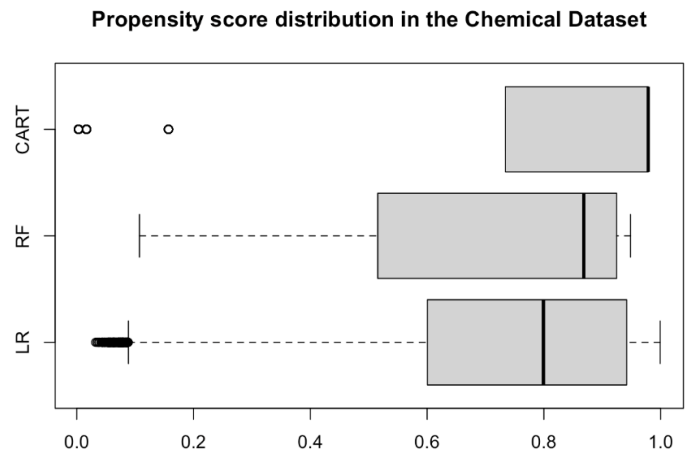
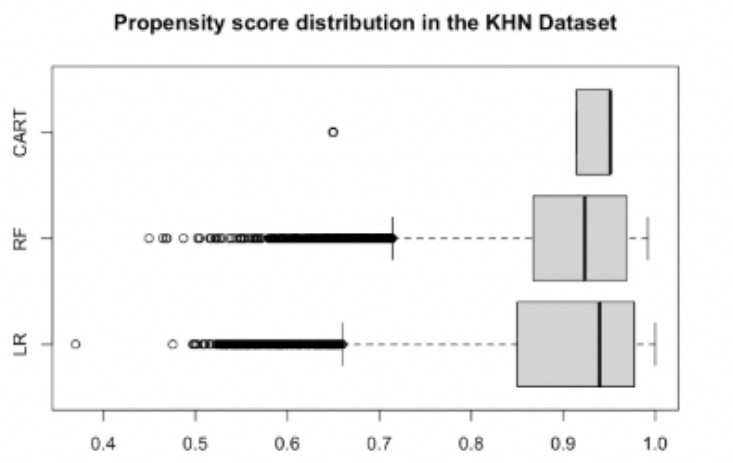


Figure 1: Distribution of Propensity Score Weights for the Comparison Group for Ten Random Datasets of (N=1000)

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

Propensity score distribution.



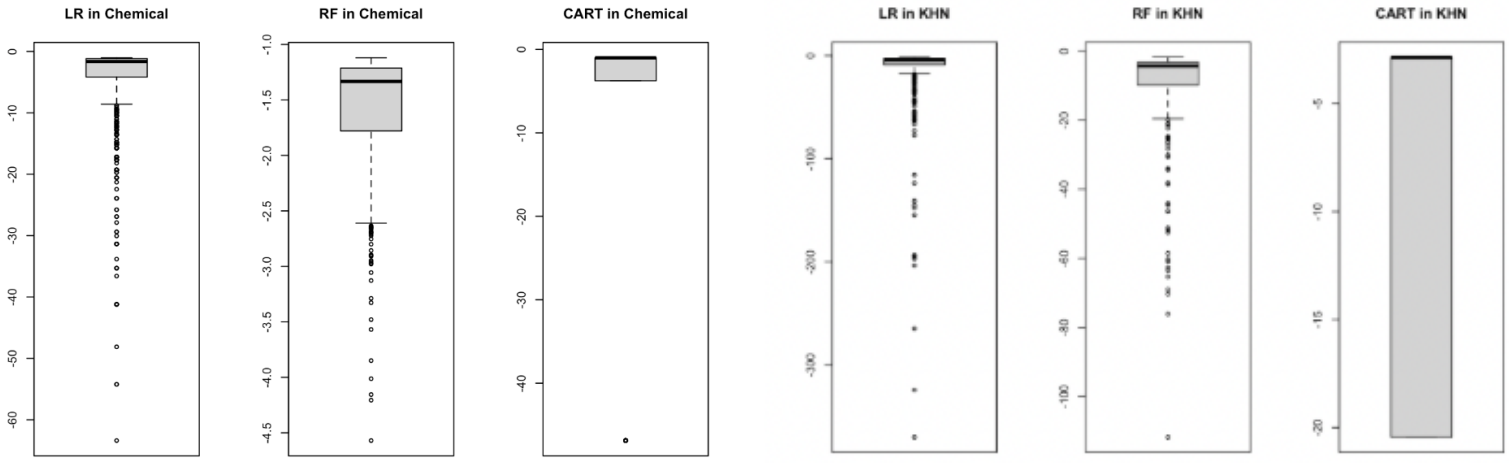


- The closer the propensity score is to 1 or 0, the worse the ATE is.

It can be seen that the propensity score of the logistic regression model has many values close to 1.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

Weight distribution.



▶ The performance of weighting methods can be adversely affected if weights are extreme, as a result of estimated propensity scores that are close to 0 or to 1.

▶ We can see that the weights of the **logistic regression** are much more extreme.

Experiment - Chemical Dataset, Korea National Health and Nutrition Examination Survey Dataset

	ATE table in Chemical Dataset
ATE_ipw_log	-0.1294140
ATE_sipw_log	-0.0056416
ATE_ipw_rf	0.2138551
ATE_sipw_rf	0.0677030
ATE_ipw_cart	-0.0036436
ATE_sipw_cart	0.0371014

	ATE table in KHN Dataset
ATE_ipw_log	0.0063232
ATE_sipw_log	0.0079276
ATE_ipw_rf	0.0056801
ATE_sipw_rf	0.0051109
ATE_ipw_cart	0.0019576
ATE_sipw_cart	0.0019567

- We know that ASAM values of logistic regression model and random forest model are similar, so we can see that ATE of both models are similar too.
- On the other side, ASAM values of logistic regression model and CART model differ greatly. So it is also in ATE values.

Conclusion

Conclusion

- The paper check the performance of new models with **simulated dataset** close to real life. So we check that new models equally apply in the **unbalanced real datasets**.
- In the Chemical dataset and KHN dataset, which were really surveyed, We can check the **performance improvement** and **balanced form** by box plot of various performance metrics.

References

References

- Brian Lee, Justin Lessler, and Elizabeth Stuart. Improving propensity score weighting using machine learning. Statistics in medicine, 29:337–46, 11 20.
- Leo Breiman. Random forests. Machine learning, 45(1):5– 32, 2001.
- Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. Journal of economic surveys, 22(1):31–72, 2008.
- Ariel Linden DrPH. Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. Journal of Evaluation in Clinical Practice, 34(1):3661–3679, 2015.