

Machine learning in Epidemiological Models

ehsan.karim@ubc.ca

Oct 21, 2020

SPPH 504/007

Ref

- Reference for reading

[What Is Machine Learning: a Primer for the Epidemiologist](#)

[Q Bi](#), KE Goodman, J Kaminsky... - American journal of ..., 2019 - academic.oup.com

Abstract Machine learning is a branch of computer science that has the potential to transform epidemiological sciences. Amid a growing focus on “Big Data,” it offers epidemiologists new tools to tackle problems for which classical methods are not well-suited. In order to critically evaluate the value of integrating machine learning algorithms and existing methods, however, it is essential to address language and technical barriers between the two fields that can make it difficult for epidemiologists to read and assess machine learning studies ...

☆  All 3 versions [Import into BibTeX](#)

Types of Epidemiological models

- $Y \sim A + C$
- Cardiovascular disease (CVD) \sim smoking +
Age + diet + obesity + cholesterol levels +
family history of heart disease

Types of Epidemiological models

- Predictive models
 - Predicting a health outcome,
 - e.g., cardiovascular diseases (CDV)

- Causal or explanatory models
 - Relationship between two health components
 - e.g., association between cardiovascular diseases (CDV) and smoking (smoking)

Analyzing Epidemiological Study data

Historically

- Statistical techniques
 - Logistic regression
 - Parametric regression (more restrictive)

Big data era

- Large datasets available
 - Large number of variables / observation
 - Conventional statistical techniques struggle

Machine learning (ML)

Followings are examples of machine learning?

- Stepwise regression
 - Uses AIC as a criterion
- LASSO
 - Uses cross validation to select parameters
- Answer differs depending on who you ask.

Oversimplified definition of ML

General characteristics:

- Prioritizes predictive accuracy
- Less focus on inference/hypothesis testing
- Handles big-data
 - Large number of observations?
 - Large number of covariates?
- What about causal models?

Terminologies - 1

- Outcome variable: **CDV**
 - Output
 - Dependent variable
 - Predicted variable
- Possible values of outcome: **CVD vs. no CVD**
 - Label
- Rare outcome: **CVD rare**
 - Imbalanced data
- Outcome group with highest frequency
 - Majority class
- Outcome group with lowest frequency
 - Minority class

Terminologies - II

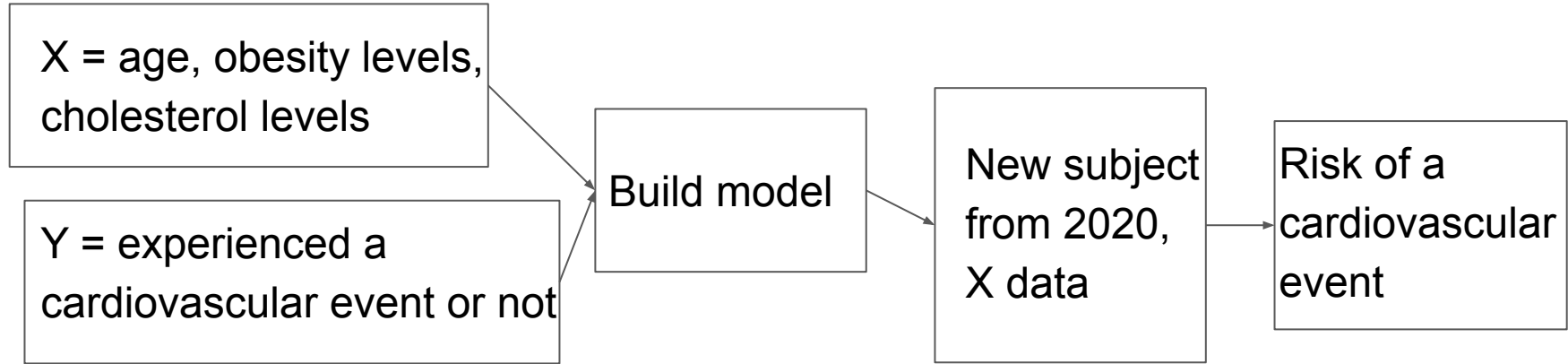
- Independent variables / covariates: age, diet, obesity, cholesterol levels. family history of heart disease
 - Attribute
 - Feature
 - Predictor
 - Field
 - Explanatory variable
 - Predictor variable
- Exposure variable: smoking
 - Input
- # of covariates
 - Dimensionality

Terminologies - III

- Model: **Logistic regression**
 - Classifier
 - Estimator
- Model fitting algorithm: **stepwise based on AIC**
 - Learner
- Domain
 - Range of possible values
- Error measure: **MSE**
 - Loss function

Classification of ML

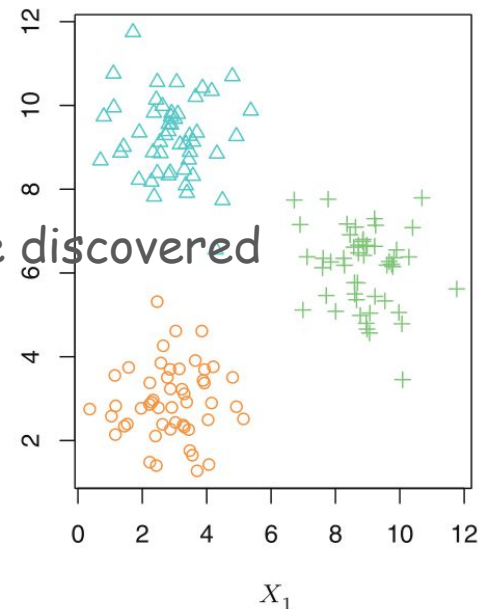
- Supervised learning
 - Values of outcome (label) are known
 - Logistic regression



2019

Classification of ML

- Supervised learning
 - Values of outcome (label) are known
 - Logistic regression
- unsupervised learning
 - Values of outcome (label) are unknown, needs to be discovered
 - Clustering
 - Without being an expert, grouping data
- Semi-supervised learning
 - In between
 - a training dataset
 - with both labeled and unlabeled data
 - Medical imaging is a great application



A clustering data set involving 3 groups

Classification of Supervised learning

- Classification
 - Prediction of categorical outcomes
 - Logistic regression
- Regression
 - Prediction of continuous outcomes
 - Linear regression

Other classifications

- Generative

- Computes the conditional probability indirectly via joint probability
- Models $P(X,Y)$ first
- then decomposes as $P(Y|X) = P(X,Y)/P(X)$
 - Hidden markov model / naive bayes

- Discriminative [epidemiologists mostly use this]

- Models the conditional probability
- $P(Y|X)$
 - Logistic / linear regression

- Reinforced learning

- Self-adaptive, learns gradually/iteratively to improve performance

Popular ML algorithms

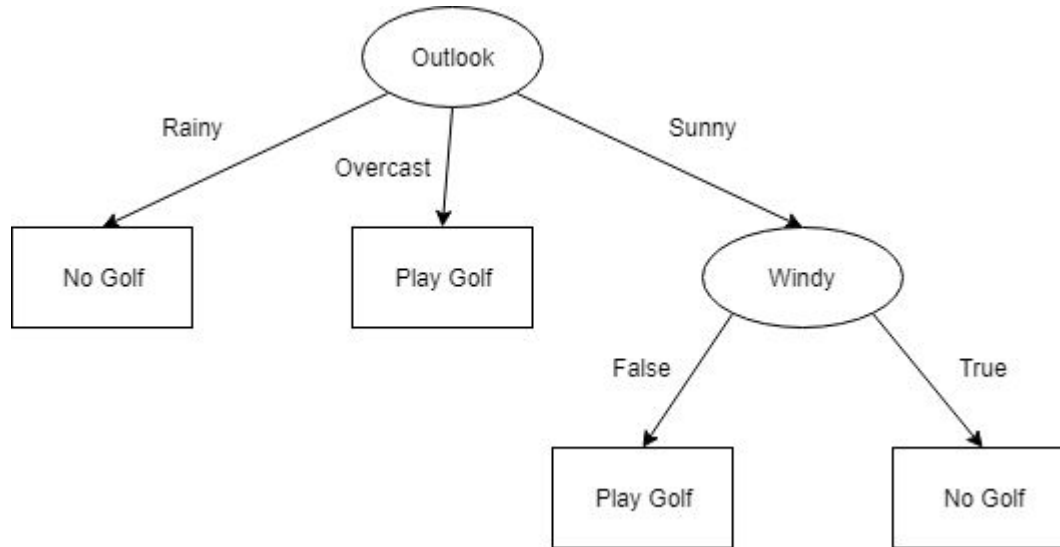
- Decision tree
 - Regression and classification trees
- Shrinkage methods
 - lasso
- Ensemble methods
 - Super learner

ML algorithm 1: Decision tree

- Decision tree
 - A flow-chart like structure
 - A decision support tool
 - Referred to as CART
 - Classification and regression trees
 - Covers
 - Classification (categorical outcome)
 - Regression (continuous outcome)
 - Easy to understand
 - Flexible to incorporate non-linear effects automatically
 - No need to specify higher order terms / interactions
 - Unstable, prone to overfitting, suffers from high variance

ML algorithm I: Decision tree

Ref: <https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-understanding-decision-trees-adb2165ccab7>



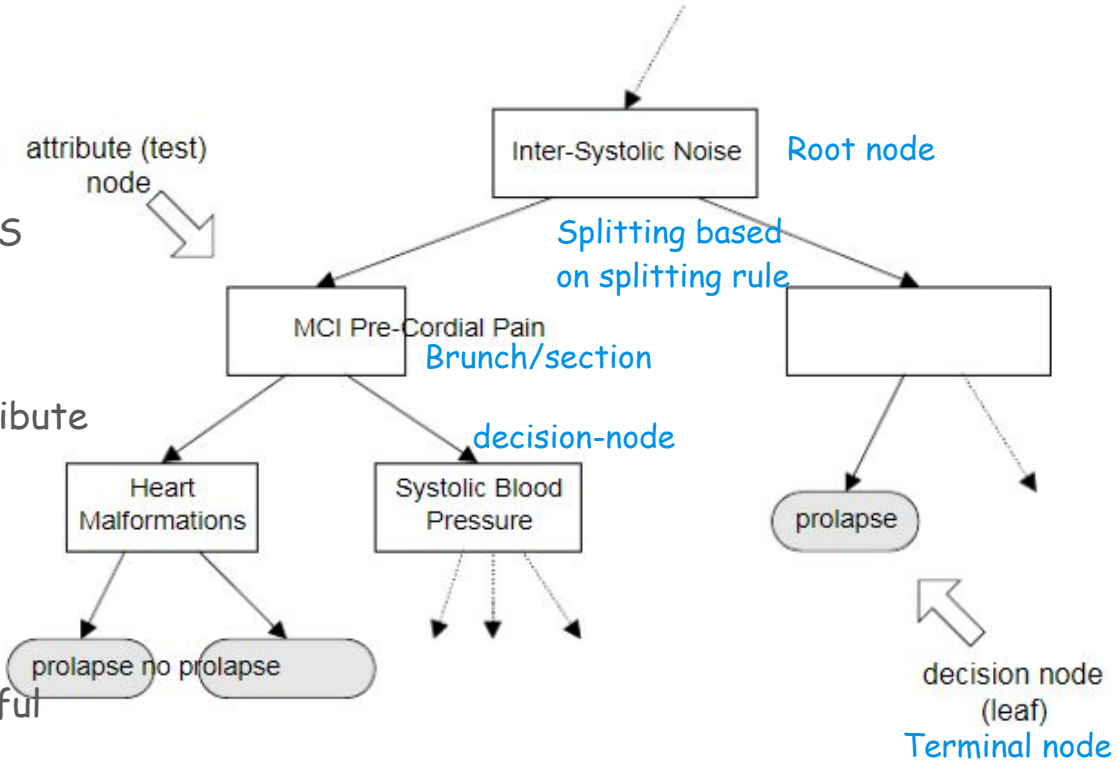
In medical decision making (classification, diagnosing, etc.) there are many situations where decision must be made effectively and reliably. Conceptual simple decision making models with the possibility of automatic learning are the most appropriate for performing such tasks. Decision trees are a reliable and effective decision making technique that provide high classification accuracy with a simple representation of gathered knowledge and they have been used in different areas of medical decision making. In the paper we present the basic ...

☆ 77 Cited by 369 Related articles All 17 versions Import into BibTeX

ML algorithm I: Decision tree

● Terminologies

- Root node
- Splitting
 - Minimize Residual SS
- Branch
- Sub-node
 - decision node / attribute
- Terminal node/leaf
- Pruning
 - reduces the size of decision trees by removing not so useful sections/nodes



ML algorithm 2: Shrinkage Methods

- Regression (no shrinkage)

- Minimize Residual SS

→
$$\text{RSS} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

- Regression


- Fitted by minimizing Residual Sum of square (RSS)

- $\text{RSS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

- $\text{RSS} = \sum_{i=1}^n (Y_i - \beta_0 + \sum_{j=1}^p \beta_j X_{ij})^2$

- No shrinkage / penalty

ML algorithm 2: Shrinkage Methods

- Ridge 
$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$
 - All coef = non-zero
- Ridge
 - $\text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$
 - $\text{RSS} + \lambda L_2$, where $L_2 = \text{Ridge penalty} = \sum_{j=1}^p \beta_j^2$
 - Here, $\lambda = \text{regularization/tuning parameter}$
 - λ controls the overall strength of the penalty.
- How to estimate lambda?
 - **Cross validation** to minimize error (e.g., **RMSE**).

ML algorithm 2: Shrinkage Methods

- **LASSO** $\longrightarrow \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$
 - Assigns some coef = 0
 - Can be unstable
 - Addresses collinearity
- **LASSO**
 - $\text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$
 - $\text{RSS} + \lambda L_1$, where $L_1 = \sum_{j=1}^p |\beta_j|$

ML algorithm 2: Shrinkage Methods

- Elastic net
 - Compromise (2 parameters): Alpha and lambda
- Elastic net (combination of Ridge and LASSO)
 - $\text{RSS} + (1 - \alpha)\lambda L_2 + \alpha\lambda L_1$
 - $\text{RSS} + \lambda[(1 - \alpha)L_2 + \alpha L_1]$
 - Here, α = The mixing parameter
 - $\alpha = 1$ is the lasso penalty, and
 - $\alpha = 0$ the ridge penalty.
 - α = between 0 and 1 (e.g., 0.2);
 - * the elasticnet/mixing penalty.

ML algorithm 2: Shrinkage Methods

- **Regression** (no shrinkage)

- Minimize Residual SS

$$\text{RSS} = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

- **Ridge**

- All coef = non-zero

- **LASSO**

- Assigns some coef = 0
- Can be unstable
- Addresses collinearity

- **Elasticnet**

- Compromise

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^p \beta_j^2$$

$$\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|$$

How to estimate lambda? Cross validation²³

ML algorithm 2: Shrinkage Methods

- Note: generally speaking:
 - Not okay to select variables from a shrinkage model
 - And then put those variables in non-shrinkage model
 - A common issue in epidemiologic literature
 - Shrinkage parameter determines predictions that is obtained via cross-validation

Invited commentary: variable selection versus shrinkage in the control of multiple confounders

[S Greenland](#) - American Journal of Epidemiology, 2008 - academic.oup.com

After screening out inappropriate or doubtful covariates on the basis of background knowledge, one may still be left with many potential confounders. It is then tempting to use statistical variable-selection methods to reduce the number used for adjustment. Nonetheless, there is no agreement on how selection should be conducted, and it is well known that conventional selection methods lead to confidence intervals that are too narrow and p values that are too small. Furthermore, theory and simulation evidence have found no ...

☆ 🔖 Cited by 236 Related articles All 8 versions

Ensemble methods

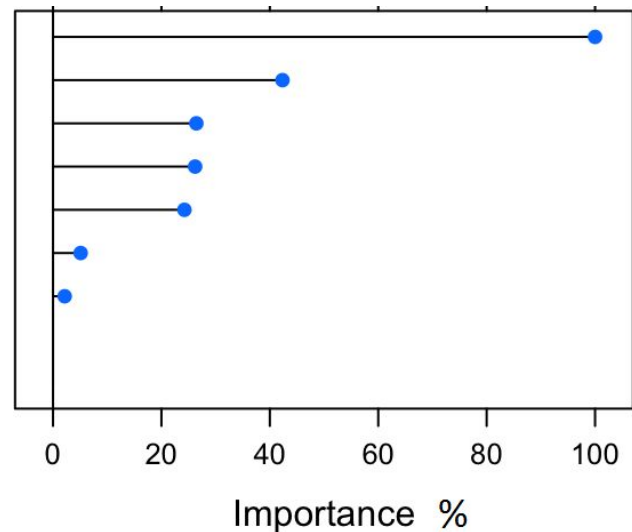
- Uses information from multiple models to improve predictive performance compared to a single model.
- Basic idea
 - Individual models may not be adequate to capture the characteristics of the entire data
 - We can build k models from k samples
 - Once we combine these k models, we should get a better 'full model' in terms of predictive accuracy

Ensemble methods: Type - 1

- Training same model to different samples (of the same data)
 - **Bagging** or bootstrap aggregation
 - independent bootstrap samples ($i = 1, 2, \dots B$ training sets),
 - applies CART on each i (no pruning)
 - Average the resulting predictions
 - Reduces variance as a result of using bootstrap
 - **Boosting**
 - sequentially updated/weighted bootstrap based on previous learning
 - Another variation could be "different models" in different samples
 - **Random forest**
 - Improvement over bagging
 - Tweaks the algorithm so that predictions from all of the sub-nodes have less correlation

Ensemble methods: VIM

- Variable Importance measure (VIM)
 - bagging improves prediction accuracy
 - over prediction using a single tree
 - Loses interpretability
 - as this is an average of many diagrams now
 - But we can get a summary of the importance of each variable
 - VIM_i = sum of the decrease in error when a variable i is used to split
 - Relative $VIM_i = VIM_i / \max(VIM)$ so that VIMs are between 0 and 1
- VIMs are also possible from
 - random forest and
 - boosting

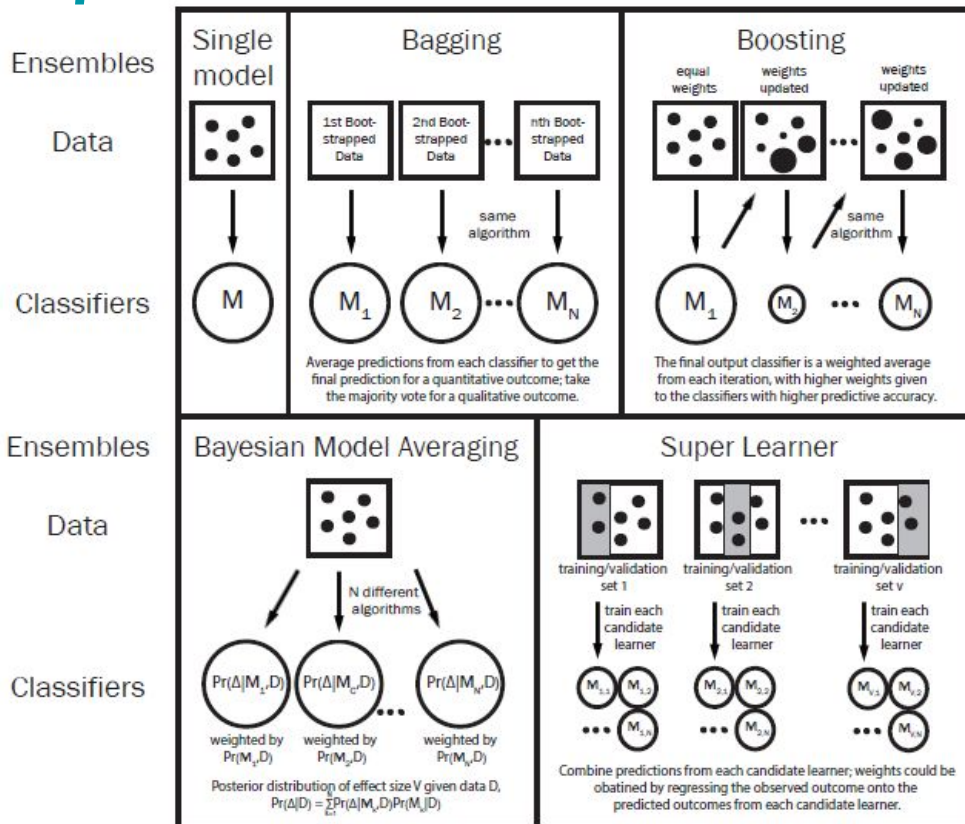


Ensemble methods: Type - II

- Training different models on the same data
 - Bayesian model averaging (BMA) can use wide range of learners
 - Super learner (SL)
 - Large number of candidate learners (CL) with different strengths
 - Parametric (logistic)
 - Non-parametric (CART)
 - Cross-validation: CL applied on training data, prediction made on test data
 - Final prediction uses a weighted version of all predictions
 - Weights = coef of Observed outcome \sim prediction from each CL

Ensemble methods: Types

Differences



Epidemiologic applications

- Causal Modelling: CVD versus smoking

- Propensity score modelling

- Exposure modelling:
smoking ~ features
=> propensity score

- In Logistic regression we need to add
 - interactions and non-linearities

- No need to specify them in CART methods

- Outcome modelling:

$$\text{CVD} \sim \text{smoking} + f(\text{propensity score})$$

Using super prediction modeling to improve high-dimensional propensity score estimation

[R Wyss, S Schneeweiss, M van der Laan... - ..., 2018 - ingentaconnect.com](#)

The high-dimensional propensity score is a semiautomated variable selection algorithm that can supplement expert knowledge to improve confounding control in nonexperimental medical studies utilizing electronic healthcare databases. Although the algorithm can be used to generate hundreds of patient-level variables and rank them by their potential confounding impact, it remains unclear how to select the optimal number of variables for adjustment. We used plasmode simulations based on empirical data to discuss and ...

☆ 57 Cited by 28 Related articles All 8 versions

Can we train machine learning methods to outperform the high-dimensional propensity score algorithm?

[ME Karim, M Pang, RW Platt - Epidemiology, 2018 - ingentaconnect.com](#)

The use of retrospective health care claims datasets is frequently criticized for the lack of complete information on potential confounders. Utilizing patient's health status-related information from claims datasets as surrogates or proxies for mismeasured and unobserved confounders, the high-dimensional propensity score algorithm enables us to reduce bias. Using a previously published cohort study of postmyocardial infarction statin use (1998–2012), we compare the performance of the algorithm with a number of popular machine ...

☆ 57 Cited by 19 Related articles All 7 versions

Epidemiologic applications

- Causal inference: CVD versus smoking
 - High-dimensional propensity score
 - ML versions (LASSO) to add proxy variables
 - Super learner (SL)
 - Causal trees / causal forests
 - Causal discovery to build DAGs automatically

[\[HTML\] Effect Estimation in Point-Exposure Studies with Binary Outcomes and High-Dimensional Covariate Data—A Comparison of Targeted Maximum Likelihood ...](#)

M Pang, T Schuster, KB Filion, ME Schnitzer... - ... international journal of ..., 2016 - degruyter.com

Inverse probability of treatment weighting (IPW) and targeted maximum likelihood estimation (TMLE) are relatively new methods proposed for estimating marginal causal effects. TMLE is doubly robust, yielding consistent estimators even under misspecification of either the treatment or the outcome model. While IPW methods are known to be sensitive to near violations of the practical positivity assumption (eg, in the case of data sparsity), the consequences of this violation in the TMLE framework for binary outcomes have been less ...

☆ 27 Cited by 7 Related articles All 9 versions

Epidemiologic applications

- Big data
 - LASSO for dimension reduction
 - ML on Gene expression data
 - Many covariates compared to observations
 - Forecasting infectious disease
 - Propensity score modelling in the high dimensions
 - Prognostic model building / risk score
 - SL
 - Geo-spatial data
 - Text mining

Future Reading

[BOOK] [An introduction to statistical learning](#)

[G James](#), [D Witten](#), [T Hastie](#), [R Tibshirani](#) - 2013 - Springer

Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine **learning**. The field encompasses many ...

☆  Cited by 5356 [Related articles](#) [All 17 versions](#) [Import into BibTeX](#)

Thanks!

ehsan.karim@ubc.ca

www.ehsank.com