Original Report:
Methods and Interventions
with Precision Medicine
Approaches

Machine Learning Methods for Precision Medicine Research Designed to Reduce Health Disparities: A Structured Tutorial

Sanjay Basu, MD, PhD[1,2,3]; James H. Faghmous, PhD[4];
Patrick Doupe, PhD[5]

Precision medicine research designed to reduce health disparities often involves studying multi-level datasets to understand how diseases manifest disproportionately in one group over another, and how scarce health care resources can be directed precisely to those most at risk for disease. In this article, we provide a structured tutorial for medical and public health researchers on the application of machine learning methods to conduct precision medicine research designed to reduce health disparities. We review key terms and concepts for understanding machine learning papers, including supervised and unsupervised learning, regularization, cross-validation, bagging, and boosting. Metrics are reviewed for evaluating machine learners and major families of learning approaches, including tree-based learning, deep learning, and ensemble learning. We highlight the advantages and disadvantages of different learning approaches, describe strategies for interpreting "black box" models, and demonstrate the application of common methods in an example dataset with open-source statistical code in *R. Ethn Dis*. 2020;30(Suppl 1):217-228; doi:10.18865/ed.30.S1.217

Keywords: Machine Learning; Precision Medicine; Health Disparities; Gradient Boosting Machines; Random Forest; Deep Learning

[1] Research and Analytics, Collective Health, San Francisco, CA
[2] Center for Primary Care, Harvard Medical School, Boston, MA
[3] School of Public Health, Imperial College London, London, UK
[4] Independent Researcher, Los Angeles, CA
[5] Zalando ES, Berlin, Germany

Address correspondence to Sanjay Basu, MD, PhD; 635 Huntington Avenue, Second Floor; Boston, MA 02115; 617.432.2222; sanjay_basu@hms.harvard.edu

Introduction

Precision medicine research designed to reduce health disparities often involves studying multi-level datasets to understand how diseases manifest disproportionately in one group over another, and how scarce health care resources can be directed precisely to those most at risk for disease.[1] Appropriate application of machine learning methods can help ensure that we maximize the clinical utility of tools we develop by taking into account each individual patient's biology, lifestyle, and environment.

In this article, we provide a structured tutorial for medical and public health researchers on applying machine learning methods in precision medicine research designed to reduce health disparities. Machine learning refers to using algorithmic approaches—commonly referred to as learners, predictive models, or estimators—to categorize data or predict an outcome. The term machine learning encompasses a wide array of methods with a common goal: to link inputs to an output accurately by repeatedly refining rules governing how input data relate to the output result. The rules are refined through analyzing multiple data subsets and sequentially, incrementally improving the rules being learned. Here, we describe the most common machine learning approaches in use today, highlighting their advantages and disadvantages, their potential uses, and situations in which they may be avoided. We demonstrate their application in an example dataset with open-source statistical code described in this article.

Methods

Example Problem

Suppose we have developed an intervention through which patients would receive an environmental home scan service that tests for, and removes, potentially hazardous contaminants that are related to their disease. Also suppose that we are constrained and can only refer those patients who are at high-risk of suffering from environmental contamination. One way to identify high-risk patients would be through a biomarker indicating possible exposure to specific environmental contaminants. We want to know therefore, which biomarker helps allocate services from several biomarkers that have been proposed to screen patients. Each biomarker, however, interacts in a complex biological pathway and has multi-level interactions with clinical,

demographic, social, economic, and community-level factors, including a person's co-morbid conditions, age, housing conditions, and the community's environmental contamination. A biomarker may indicate contamination only among certain people with a key residential housing background, and may only be useful for screening disease among those; furthermore, a single biomarker's level may be determined through multiple mechanisms, such that an environmental scan is only appropriate when other factors affecting the biomarker level are factored in. For example, C-reactive protein level may be high from inflammatory atherosclerotic disease processes or due to specific environmental pollutants,[2] but only in one case would the high level indicate that an environmental contaminant scan would be warranted.

Machine learning refers to using algorithmic approaches—commonly referred to as learners, predictive models, or estimators—to categorize data or predict an outcome.

Now suppose that we have a database of patient features, including a series of biomarkers thought to indicate contaminant exposure, patient demographics and clinical features, and patient outcomes data revealing which patients actually had contamination in their homes and which patients did not. Our goal is to design a machine learning algorithm that will learn from this historical data and accurately direct referrals for future patients, optimizing the referral process to precisely refer those patients most likely to benefit from the intervention.

To guide our approach to studying this problem, we accompany this article with a simulated dataset and statistical code in *R* (available for download at https://github.com/sanjaybasu/MLforPMHD).

Key Terms and Concepts

Supervised vs Unsupervised Learning

There are two major categories of machine learners. Supervised learners learn the relations between input and outcomes data, then predict future outcomes based on future patient's input data, and so require patient features as inputs and a disease outcome. Standard logistic regression is a supervised learning example, as it maps covariates onto the probability of an outcome. Supervised learners will be the exclusive focus of this article. "Unsupervised" learners learn how to cluster or categorize only input data, naturally grouping similar data types, such as through factor analysis.[3] In the above example research problem, a supervised learner would be one that uses patient features to predict who will have contamination in their home, while an unsupervised learner would be one that clusters patients into groups based on similar features (eg, similar demographics, similar biomarkers, etc.).

Overfitting, Regularization and Cross-Validation

Overfitting refers to the process by which a learner not only captures the general relationships between input and output data, but wrongly captures random noise in the dataset (Figures 1-3), which prevents the learner from making accurate and generalizable predictions when applied in the future. In the above example research problem, overfitting might occur if a learner only learns that people who are between 4.1 and 6.9 years old with high values on three biomarkers are at risk for having contamination, and fails to identify that a 4.0 year old or a 7.0 year old with similarly elevated biomarkers are also at risk. (Figures 1 and 2)

To increase generalizability and prevent overfitting, we try to follow Occam's Razor, which reminds us that we don't want to produce a complicated algorithm to do something if we can have a simpler algorithm that does our chosen task well. In addition to being more generalizable, a simpler algorithm may require less input data or complicated data transformations, may be more interpretable, and may be more easily fixed if predictions turn out to be erroneous.

Regularization refers to processes that help enforce Occam's Razor when we're training a machine learner, particularly when the input data variables are correlated, as is often the case with patient clinical data, lab values (biomarkers), and diagnoses and social variables. The two most common regularization processes are called: 1) L1 regularization, or "LASSO" (which stands for least absolute shrinkage and selection operator), and 2) L2 regularization, or "ridge" regularization.

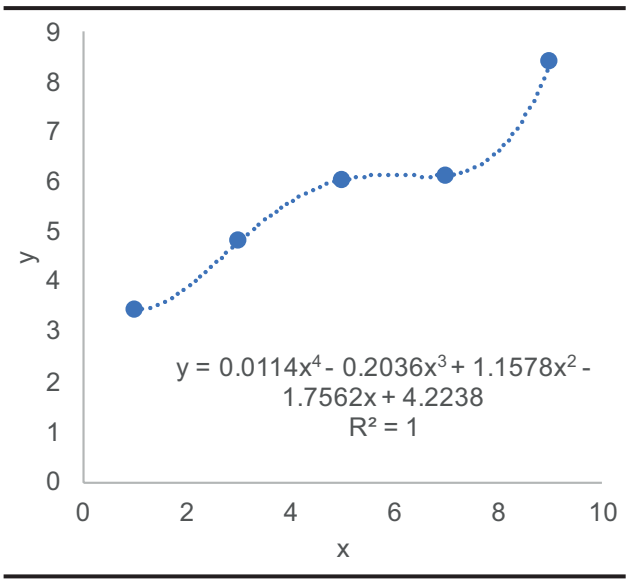

Figure 1

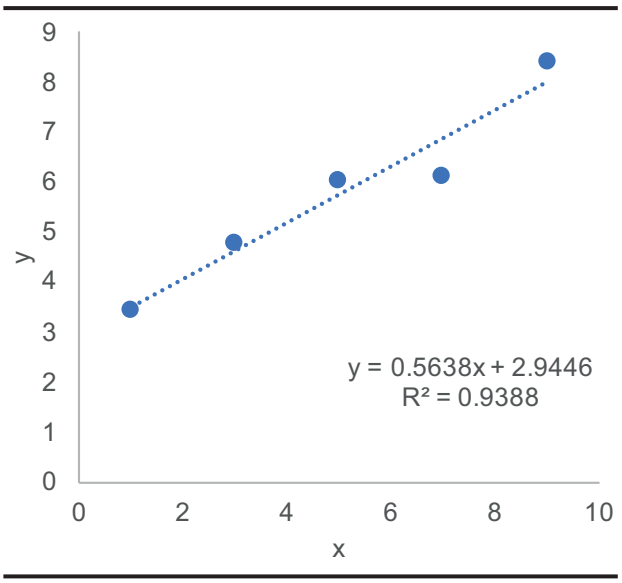

Figure 2

Figures 1-2: Key concepts for machine learning. Overfitting. Suppose we have sampled some variable x and outcome y in a field experiment, such as x variable "age" and y variable "cardiovascular event risk". The estimator (fitted curve) in Figure 1 may appear to be "better" by a performance metric such as the R^2 (the coefficient of determination, reflecting the proportion of variance in y that is explained by the model; this equals 1 because the estimator has perfectly fit the data), but we would not expect the curved estimator to reliably predict outcome y given some values of variable x or even fit the data very well if we were to repeat the experiment, because the curve has fitted random error in the dataset. By contrast, the estimator (line) in Figure 2 may have poorer performance on a metric such as the R^2, but does a better job of capturing the general relationship between outcome y and variable x, as the 'true' model generating the relationship between x and y in this case was a simple linear model with random noise added, not a complex polynomial. Adapted from.[12]

LASSO tends to select one correlated variable among others, like choosing a representative variable for many related variables. In a standard logistic regression, for example, LASSO will include one correlated predictor variable in the final equation and set the regression coefficients for its related variables to zero; this is operationalized by penalizing the absolute sum of the regression coefficients. By contrast, ridge regularization of a logistic regression tends to set the regression coefficients of correlated variables to an equal value, under the premise that correlated variables should be similar and extreme values may be based on outliers; this is operationalized by penalizing the squared sum of the regression coefficients.[4]

In practice, we split the dataset into two subsets. We use the "validation" (test) subset only once, after the learner has been fine-tuned, to assess the estimator's generalizability. We repeatedly sample the other subset, the training subset, in a cross-validation process (Figure 3). In cross-validation, we must choose how much to regularize the learner by selecting a penalty parameter that determines how much to choose one correlated variable over others (via LASSO) or restrict correlated variables' coefficients to be similar (via ridge). With each sample of training data, we train the learner by finding the penalty parameter value that gives the lowest mean-squared error between prediction from the learner and observed. The statistical code accompanying this tutorial illustrates how to implement this regularization approach through cross-validation in *R*.[4]

Bagging and Boosting

Maximizing the learner's predictive performance is a central machine learning principal. In an averaging strategy called bagging, we train many learners on many training data subsamples, assuming that repeating the process many times will settle to a good average prediction. We take the average of a big bag filled with learners, each treated equally.[5] In an alternative boosting strategy, we first train a single learner, then learn from the first learners' errors to improve the learner in the next round of sampling, and so on.[6,7] Bagging can help avoid getting stuck down a wrong path when a bigger improvement could be found by looking more broadly across possible ways to model the data. Conversely, simpler problems such as predicting an out-

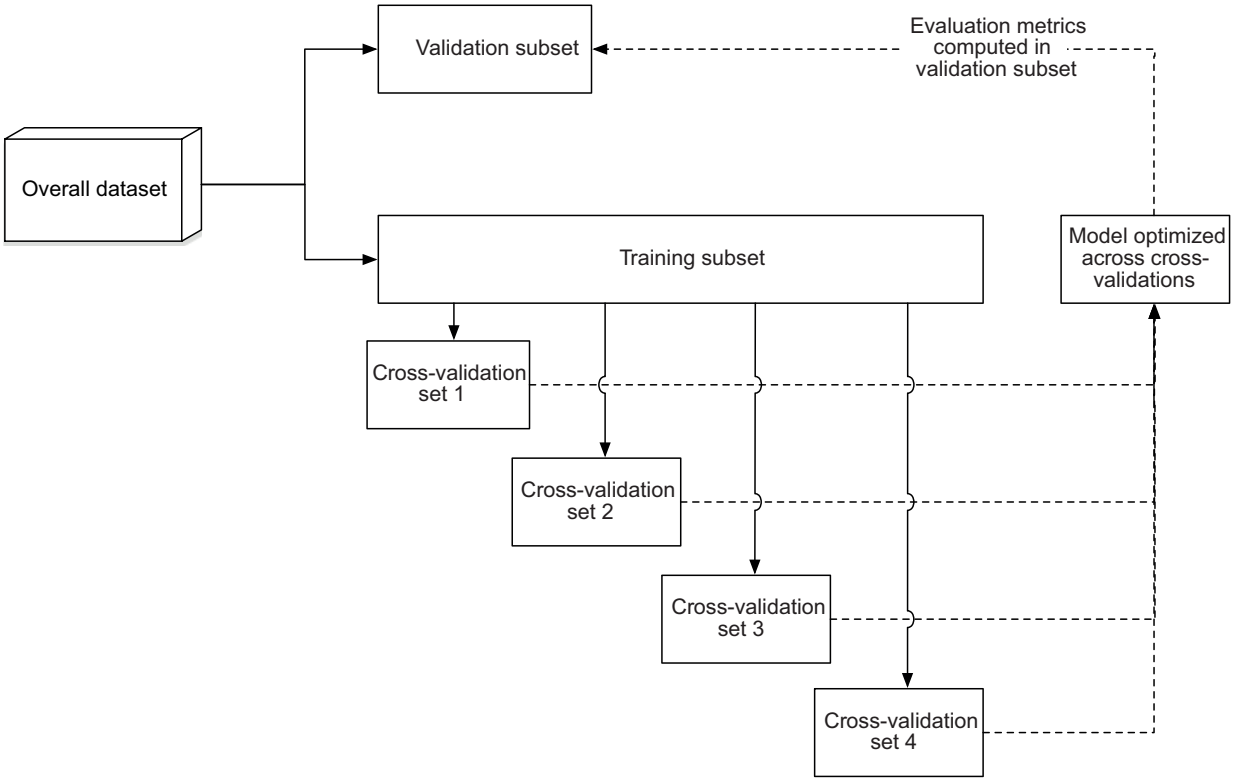

Figure 3. Cross-validation
After each sample of training data is obtained, we train the learner by finding the penalty parameter value that minimizes the mean-squared error between the predictions from the learner and the observed training data subset. We then test the performance on the held-out test subset of the data. Adapted from.[12]

come using variables with few classes, rather than with multi-class or continuous variables, can often be modeled best through boosting.[8,9] Our statistical code illustrates both approaches, which are discussed further below.

Metrics for Rigorous Evaluation

A key study design question is: what marks a good learner in this study and how will we fairly compare our learner to alternatives? The most common machine learning evaluation metric is the C-statistic, or the "area under the receiver operating characteristic (ROC) curve" (Figures 4-5). In our research study example, the C-statistic discriminates, or captures how often a learner will correctly select the higher-risk person among a pair of people, where only one in the pair has the outcome.[10]

In clinical medicine, the ROC plot x-axis corresponds to 1 minus the specificity of a test, and the y-axis to the sensitivity of a test. A machine learner best for reassuring people that they do not have a disease (avoids false negatives) would have a high value on the y-axis of the ROC curve (sensitivity), even if it had a mediocre value on the x-axis (specificity). A machine learner suited best to confirming the presence of a disease (avoids false positives) could have a high value on the x-axis (specificity) even if it has mediocre y-axis values (sensitivity). In these studies, pre-specifying a sensitivity or specificity evaluation metric, rather than just the composite C-statistic, can be important.

To analyze for correct absolute risk estimates, we assess the calibration curve (Figures 4-5), which is a plot of the predicted rate of outcomes among centiles of the validation dataset population (x-axis) against the observed rate of outcomes among those centiles of the validation dataset population (y-axis); a perfect learner will have a 45-degree line between the predicted and observed outcome rates. The Hosmer-Lemeshow test is a common statistical test for evaluating the degree of error between the predicted and observed rates of the outcomes

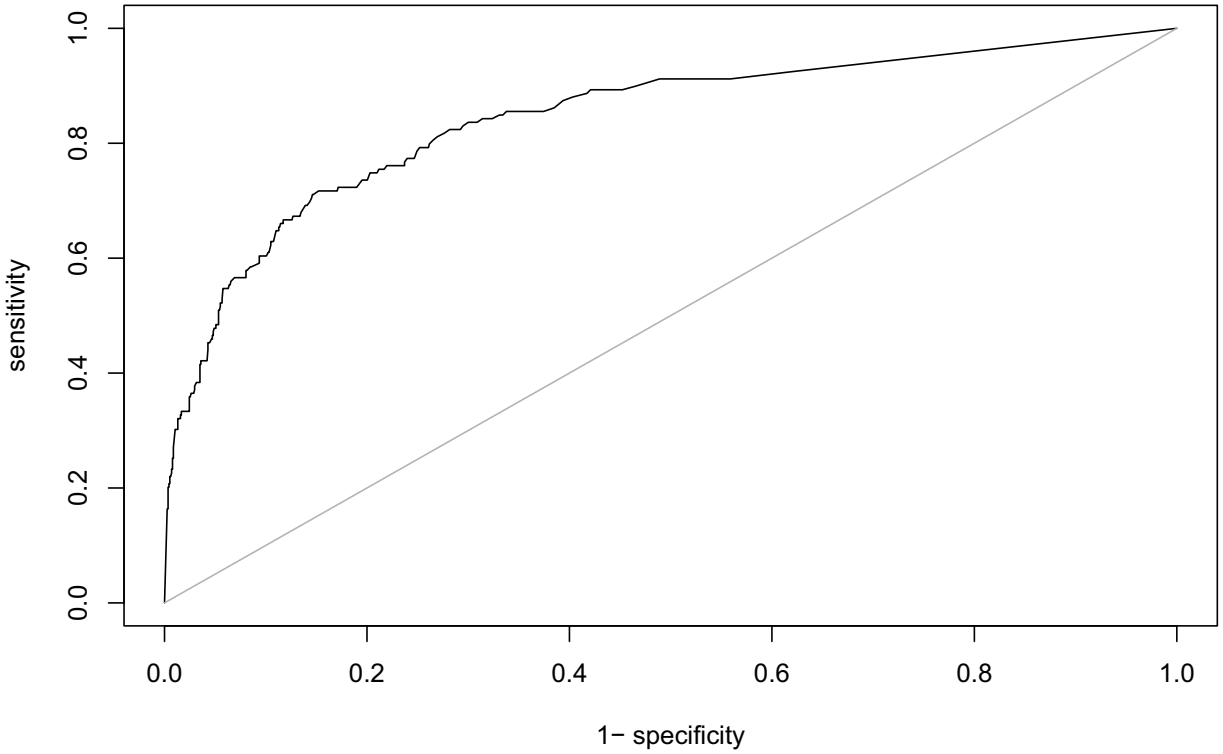

Figure 4. Evaluation metrics
The receiver operating characteristic (ROC) curve, with the true positive rate ('tpr', or sensitivity) on the y-axis and false positive rate ('fpr', or 1-specificity) on the x-axis; the area under the curve is the C-statistic.

plotted in the calibration curve.[11]

Other metrics commonly seen in machine learning literature include a confusion matrix, or contingency table, which is a 2-by-2 table of true and false positive and negative outcomes in the validation dataset (see https://github.com/sanjaybasu/ML-forPMHD for Appendix Table 1). Some machine learning papers also use the metric of accuracy, which is the sum of true positives plus true negatives, divided by the total number of predictions. A third common metric is precision, which is the sum of true positives divided by the sum of true and false positives, which clinical epidemiologists call the positive predictive value. In all cases, the pre-specified evaluation metric should correspond to the ultimate application by selecting a metric that corresponds most to the study's goals, timeframe, effort, and budget.

Major Families of Machine Learning Methods

Below, we present a simplified rubric that covers the major current approaches for training a machine learner.

Tree-based Learners

Two of the most common training approaches are both strategies for building "decision trees" from data. Decision trees are, essentially, flowcharts that guide categorization of a particular patient or population (Figures 6,7,8).[12] Each branch divides the study population into increasingly smaller subgroups that differ in their probability of an outcome of interest or their likelihood of benefitting from a particular intervention.[13] A good decision tree will separate the sampled population into groups that have low within-group variability, but high between-group variability. In the above example research problem, where we are trying to predict whether a patient has home environment contamination, a decision tree might first separate our patient population by neighborhood, then identify which features in the first neighborhood define patients into high-risk vs low-risk groups; those features may differ in the second neighborhood, and the third, and so on. An advantage of tree-based learners is the abil-

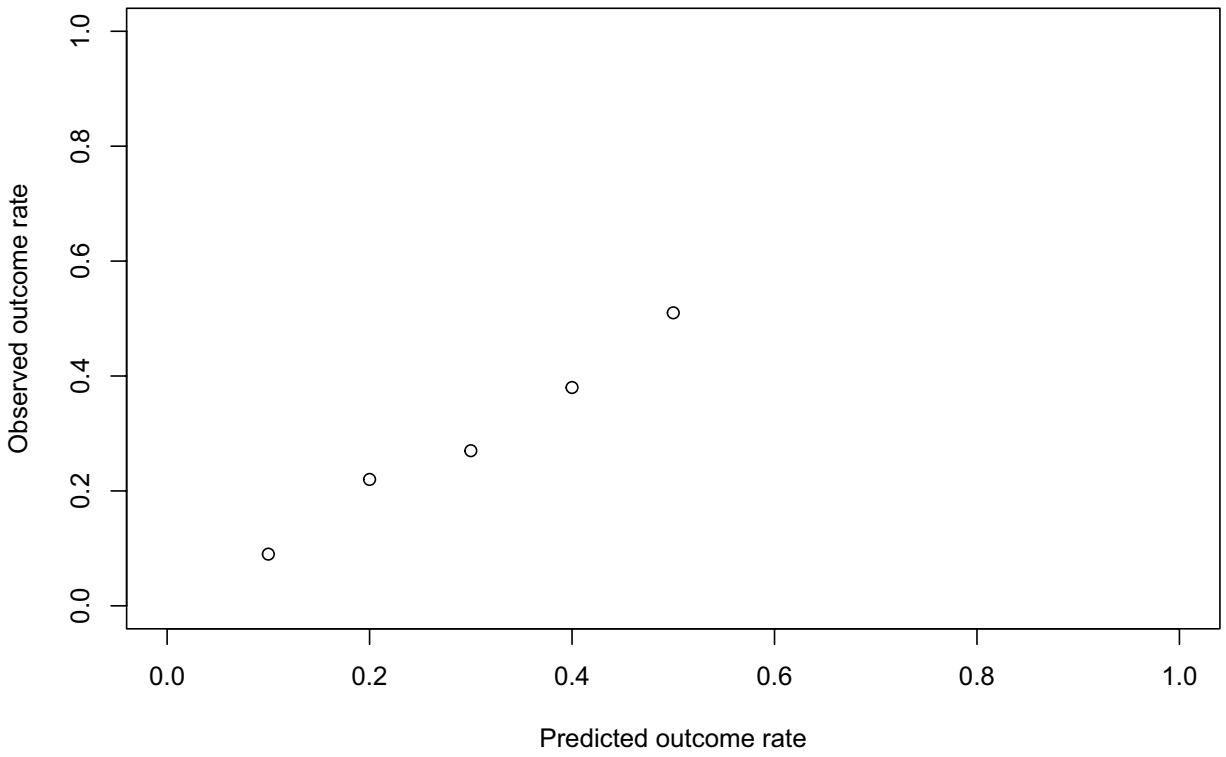

Figure 5. Evaluation metrics
A calibration curve

ity to consider multiple covariates at once, potentially capturing complex interactions between covariates (eg, between neighborhood and other factors) and nonlinearities (since covariates can have different cut-points defining branches, and different outcome rates in one branch than another). However, a limitation is that they are prone to overfitting because the decision tree has over-interpreted noise or random outliers in the data as reflecting a real phenomenon or a real subgroup. Even cross-validation may not detect the over-fitting.[14]

Two common tree-based learners that minimize the risk of overfitting, are gradient boosting machines (GBM) and random forests (RF). In both methods, many trees are grown to subsamples of the training dataset. GBMs average many trees where errors made by the first tree contribute to learning of a more optimal tree in the next iteration (a boosting strategy).[6,7] RFs average a forest composed of many trees, where each tree is independently fitted with a random subset of covariates defining branches (a bagging strategy).[5] As noted above, the bagging strategy will often produce higher discrimination than the boosting strategy in situations where there is a complex outcome being predicted (eg, a categorical or continuous outcome), rather than a simple dichotomous outcome (eg, the absence or presence of a disease or condition).[8,9]

Training learners using tree-based methods can be particularly helpful when trying to identify how different risk factors—in isolation or in combination—contribute to differential overall risk, as is often the case in health disparities research. Additionally, tree-based methods are effective when complex combinations of factors at multiple levels—such as among subcellular, individual, and neighborhood-level factors—interact to determine the overall risk. Trees can also be more useful than standard logistic regression when researchers are trying to predict a rare outcome (eg, a high-cost hospitalization, or a rare but severe disease complication) from many complex interacting factors.[15]

A GBM approach requires researchers to decide several factors that can influence how well a learner

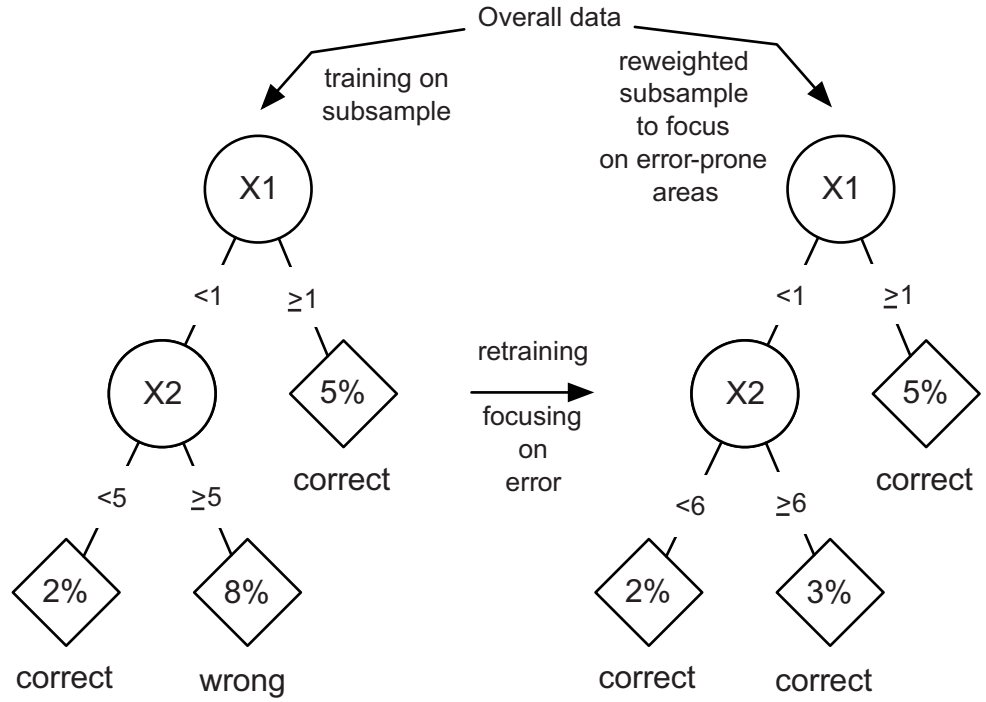

Figure 6. Gradient boosting machines (GBM). Adapted from Doupe et al.[12]

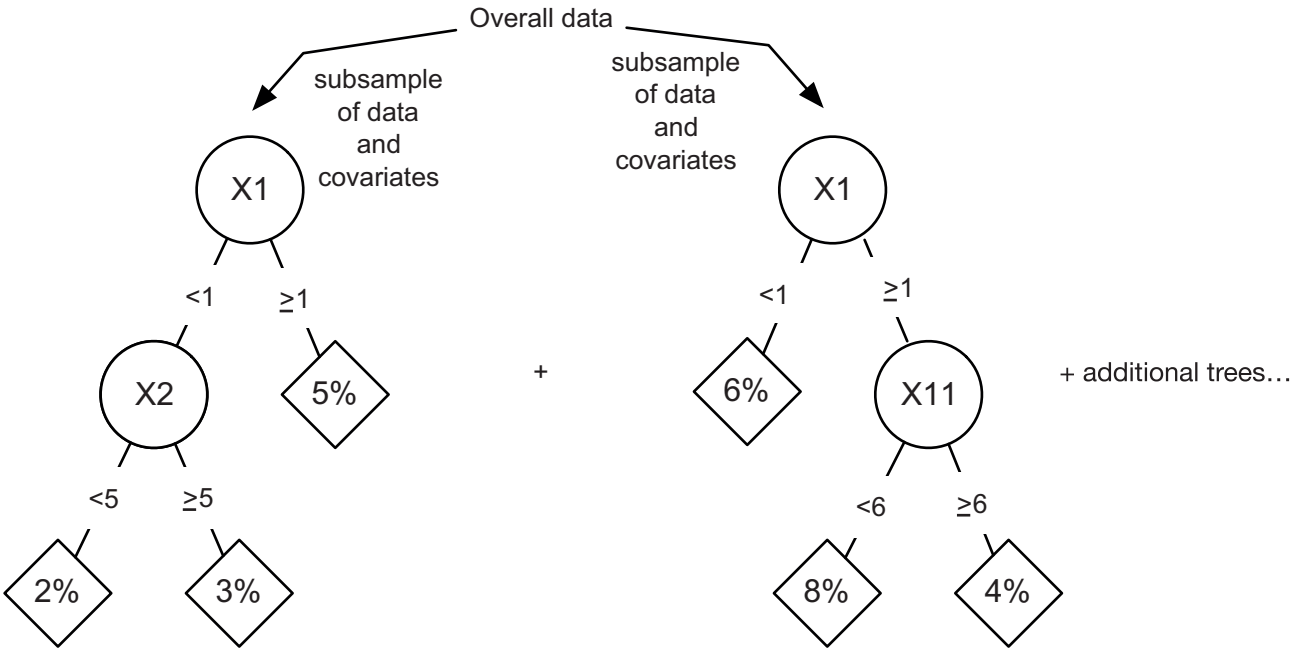

Figure 7. Random forests (RF). The circles display the covariates (X variables) whose values determine each branch point, while the diamonds provide the tree-predicted probability of the outcome under study. Adapted from Doupe et al.[12]

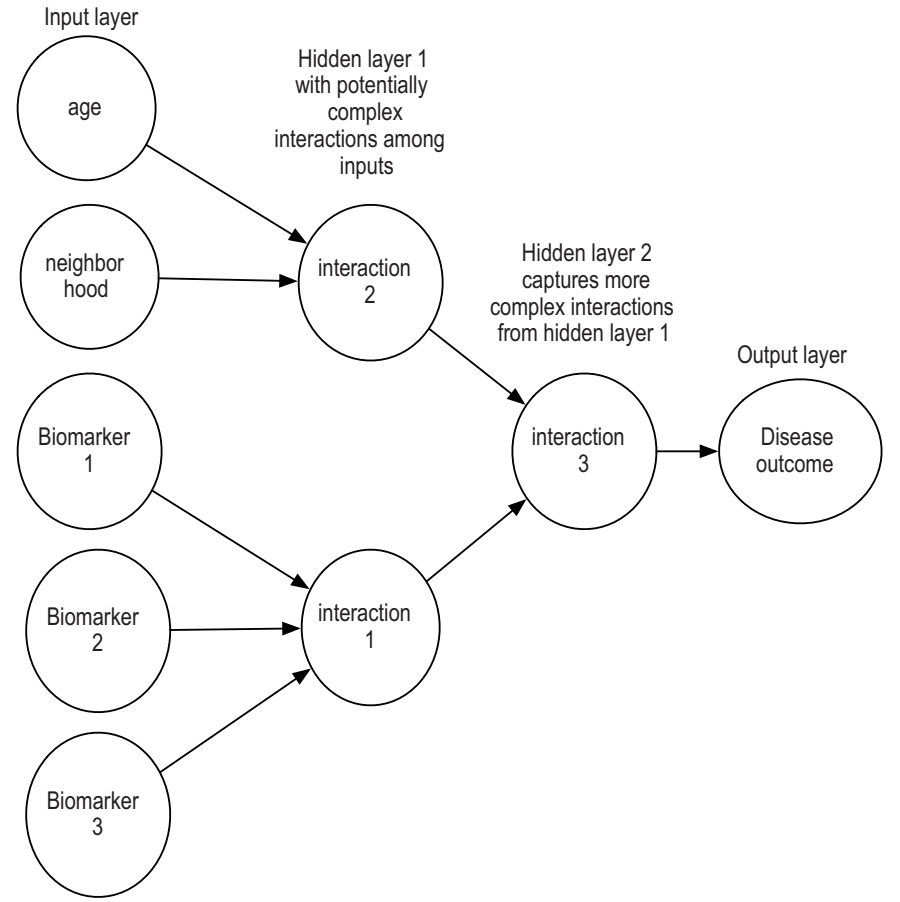

Figure 8. Deep learning neural networks, which are based on a loose caricature of the brain as a series of neurons in which inputs (like features of patients in a dataset) are processed by a layer of neurons, which then inform another layer of neurons, and so forth, until an output is achieved (eg, risk of a disease event). Adapted from Doupe et al.[12]

performs (known as "tuning"): how many trees to average, how deep the trees should be (how many "layers" of branches to have, which determines how many subpopulations to divide the population into), and how quickly the trees should adapt to initial error (the "learning rate") to optimize a performance metric. On the other hand, the RF approach generally produces a reproducible result with maximum discrimination across a wide range of specifications without extensive tuning. Hence, the RF approach is often reasonably favored by researchers new to machine learning. The statistical code (available at https://github.com/sanjaybasu/MLforPMHD) provides examples of both GBMs and RFs applied to the prediction of environmental contamination in a mock dataset and compares their performance to a standard logistic regression.

Deep Learning

Training neural networks to predict outcomes is deep learning. A neural network is a series of data transformations where the outputs from one series of transformations inform the inputs to the next series (Figures 6,7,8).[12,16] Each transformer (or neuron) in the network takes a weighted combination of inputs reflecting a weighted sum of the observed data (for the first layer of neurons) or from a previous layer of neurons (for the second and subsequent layers of neurons), and produces an output based on a nonlinear transformation function called an activation function, which is like a link function in statistical regression. The weights for each sum plus activation function values determine the output from the network. A standard logistic regression is simply a neural network with a single layer of neurons, in which each transformer multiplies the input covariate by a regression coefficient, and a logistic function is the activation function. Regularization and cross-validation techniques are typically applied to neural networks to prevent overfitting.

There are many ways to adapt a deep learning neural network to complex problems. The simplest deep learning neural network structure, visualized in Figures 6,7,8 is known as a feedforward neural network, in which each layer of neurons is fully connected to the next layer, and information only flows in one direction. By contrast, recurrent neural networks have backward feedback from one layer to a prior layer, which allows for more learning across a time series, and is particularly suited for problems where knowledge of one prediction affects the next prediction (eg, in speech recognition, where the prior word informs the choice of the next word; or in predicting sequential disease events for an individual, where

prior diagnoses inform the probability of a subsequent diagnosis).[17] Additional common types of supervised deep learners include convolutional neural networks, which capture local features of the training dataset then combine locally learned networks to gather a global understanding (eg, for image recognition, where cluster of pixels form one part of a picture, and these parts are combined to recognize the overall image). Some unstructured deep learners include autoencoders, which take noisy data and try to reconstruct the original data, a process known as denoising; and word2vec, a series of two-layer neural networks trained to detect words in context to aid natural language processing.[17–20]

At the time of this writing, deep learning neural networks in the medical and public health context remain limited to image recognition (eg, radiologic detection of abnormalities on X-rays), disease classification problems (eg, reading electronic medical record notes to categorize and classify disease phenotypes in a data-driven manner), and outcome prediction (predicting a clinical outcome based on complex features).[17,20,21] Deep learning may be more effective than tree-based methods for outcome prediction when complex context in text (eg, in clinical notes or other written assessments) must be processed and included as predictors, when extremely complex interactions are thought to exist between covariates that predict an outcome (eg, between multiple-omics markers), or when complex sequential processes must be predicted rather than just a single outcome (eg, hospital admission, discharge diagnosis, then readmission, and mortality). Deep learning can, however, be difficult to implement because researchers have to choose activation functions, network depths (number of layers), and degree of regularization, among other choices in structuring the model to customize it for the task being accomplished. The statistical code referenced earlier in this article provides an example of constructing and tuning a standard feedforward neural network, and includes multiple common options for activation functions, network depths, and regularization processes; the code also enables comparison of network learners to tree-based estimators and standard logistic regression for the example problem of predicting home environmental contamination. Although the statistical code is in *R* to be familiar to epidemiologists and health services researchers, it should be noted that deep learning neural networks are more commonly programmed in Python due to speed and scaling limitations of *R*, and online tutorials for more advanced deep learning for medical applications using Python are recommended.

Ensembles

A key insight from recent machine learning research is that an ensemble of learners can help more closely approximate the truth, even when no underlying base learner is fully correct.[22] Researchers who have little *a priori* theory to favor one type of learner over another can particularly benefit from training an ensemble of learners. To train an ensemble, a researcher will first individually develop learners on the dataset, then use a meta-learner (sometimes called a super learner or stacking method) to combine the predictions of the underlying base learners.[23] Technically, both GBM and RF are ensemble learners because they combine individual decision trees to produce a composite prediction.

Results

Performance Comparison among Methods for the Example Problem

Analysis of the example problem described under the Methods section uses a simulated dataset containing 100 variables that are correlated and inter-dependent in complicated ways. The statistical code referenced in this article allows readers to reproduce the simulated dataset, within which are 40 binary variables, 30 secondary dependent variables correlated with or conditioned upon some of the first 40 variables (reflecting, for example, co-morbidities that often occur together), 20 categorical variables, and 10 continuous variables, some a subset of variables predicting the outcome of home environmental contamination (a dichotomous outcome) with complex interactions and dependencies.

Using the statistical code and approaches described in the Methods, we find that a standard logistic regression—choosing covariates using elastic net regularization to find a balance between LASSO and ridge regression, and thereby address collinearity and reduce overfitting—performed relatively poorly in terms of discrimination, with a C-statistic of .64. A RF learner provided slight improvement with a C-statistic of .69, which was also produced by a GBM learner. A deep learner only produced a slight improvement with a C-statistic of .70. An ensemble of these

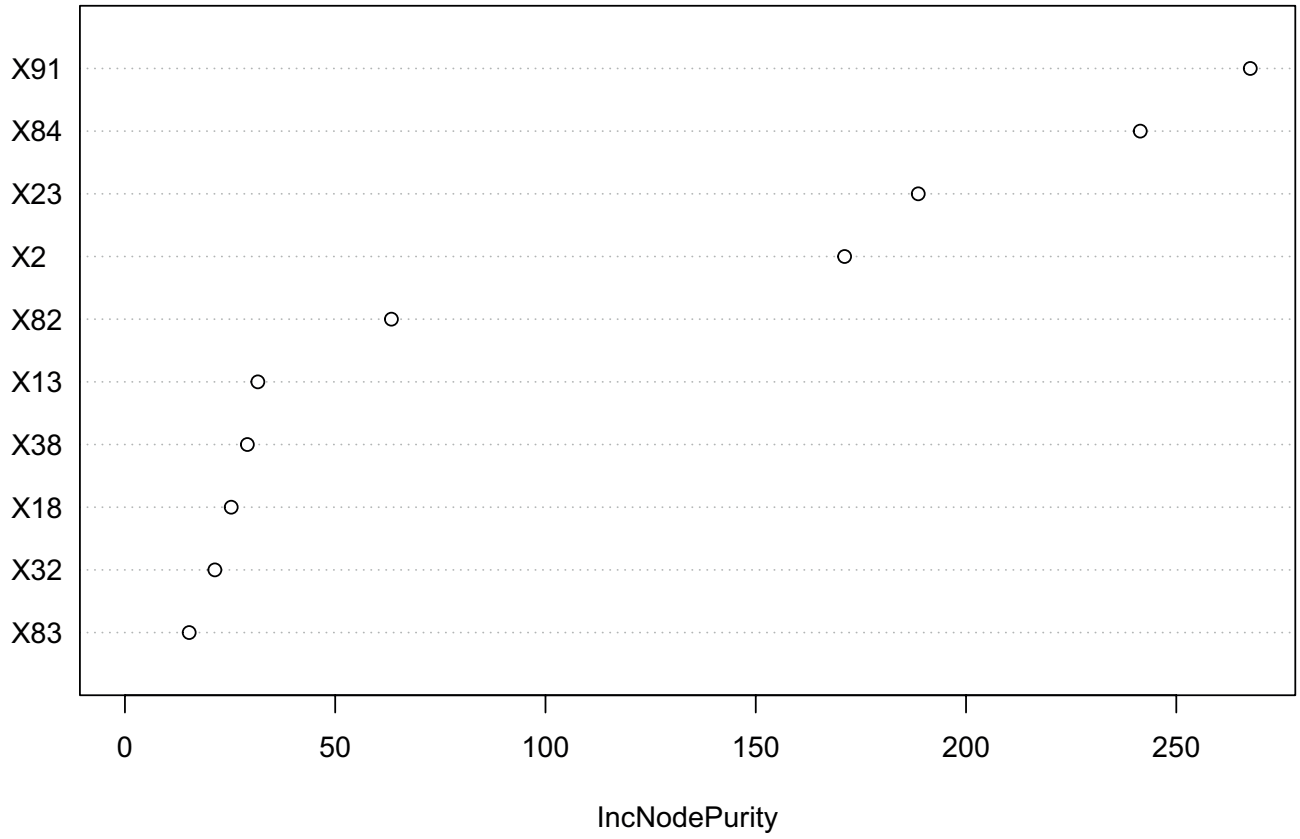

Figure 9. A variable importance plot showing the standardized coefficients from a learner (showing X variable #91 to be particularly important). IncNodePurity refers to the average increase in mean squared error of the model when the corresponding variable is excluded from the model.

learners, however, dramatically improved discrimination with a C-statistic of .86. To further characterize the sensitivity and specificity of the learners, we have included a confusion matrix (contingency table) in the Appendix (available at: https://github.com/sanjaybasu/MLforPMHD). Note that all evaluation metrics are calculated on the validation dataset, not on the training data.

Interpreting Machine Learners: Visualizing inside the Black Box

Two key strategies are available to researchers trying to understand how the learners have interpreted the data to produce useful predictions. A variable importance table or plot is a standard way to display how different variables are included in the learners, clarifying which variables are most influential in prediction (Figures 9,10). The plot displays standardized coefficient magnitudes (ie, Z-scores) across all variables.[23] Standardization is common in public health applications because it allows for correction of highly-skewed variables. Partial dependence plots allow us to visualize how learners relate covariates to outputs.[24] In particular, when covariates go through complex transformations, a partial dependence plot effectively reveals to us how a learner has related values of an individual covariate with the probability of the outcome, revealing important nonlinearities for example (Figures 9,10).

Conclusion

Here, we reviewed key terms and concepts in machine learning, critical methods for evaluation and interpretation of machine learners, and major common types of learners, with accompanying statistical code to demonstrate their application.

As with the presentation of many other types of research, machine learning articles are recommended to follow key principles of good research practice.

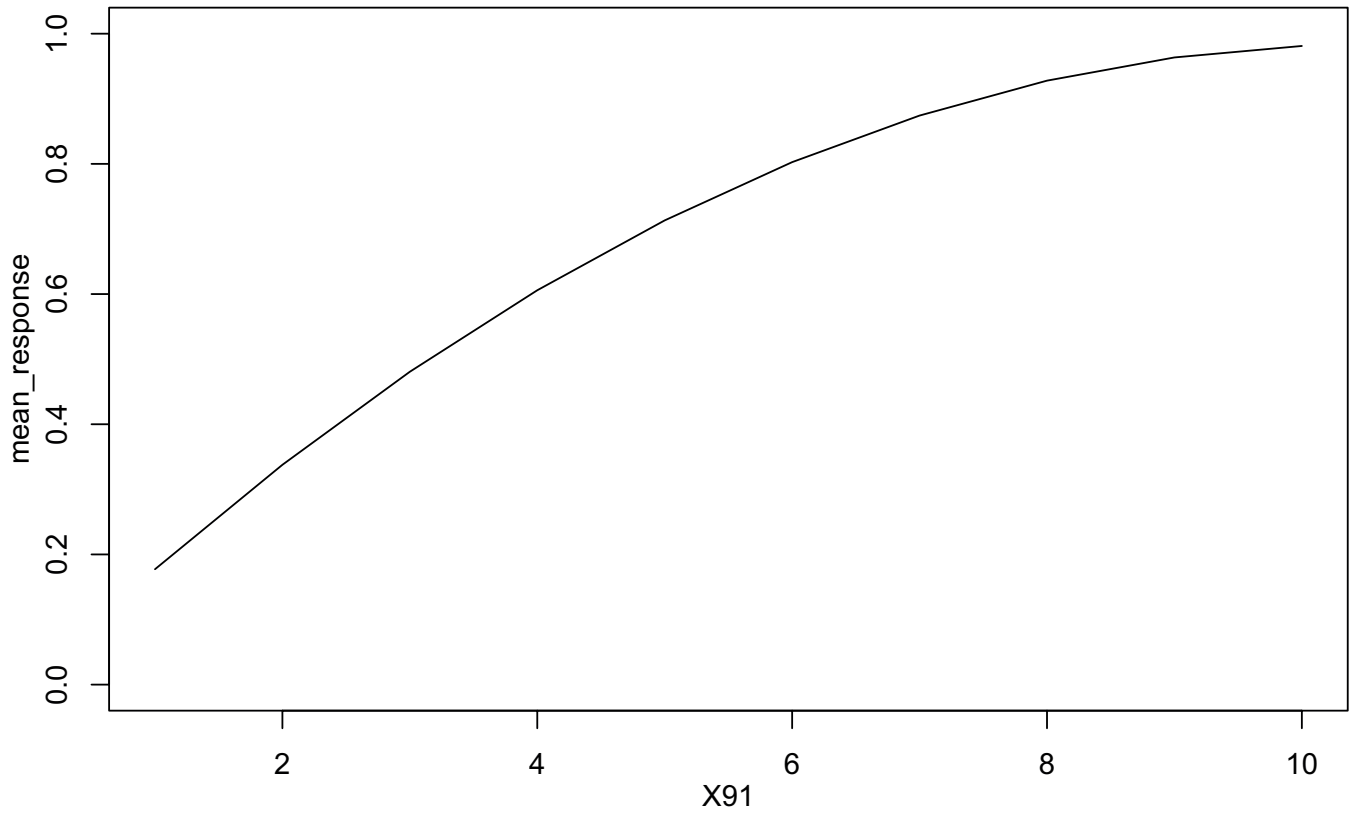

Figure 10. A partial dependence plot showing the relationship between an individual variable (X variable #91 in the example code) and the predicted probability of an outcome (mean_response). Note that all X variables are theoretical and explained further in the code accompanying this manuscript. The mean_response is the y outcome variable reflecting the probability of the outcome from 0 to 1

In closing, we review some of the principles considered most critical for high-quality machine learning papers.[25–27]

First, an important standard for reproducibility and extension in machine learning literature is the sharing of statistical code and underlying data. De-identifying data and sharing the raw statistical code is particularly important given the problem that many researchers' papers have been found to not be reproducible.[28]

Second, it is important for machine learning problems to be pre-specified, so that researchers are not tempted to use the approach purely to produce (potentially false-positive) associations.

Third, the end-user of a machine learner must be kept in mind. Different audiences need to either be able to interpret a learner, or just use the learner by inputting data and having the learner automatically provide results (eg, at the back-end of an electronic medical record).

Finally, it is critical that researchers using machine learning methods have data empathy, or the perspective that the quality and type of data must correspond well to the type of question being asked and the future utilization of the method. If a particular question cannot be answered well with a small dataset, it is unlikely that the question will be better answered with a larger dataset of the same type and data quality.[4] For example, abundant use of insurance claims or electronic medical record data, which are large and widely available in the medical literature, is problematic for clinical studies, as prediction models will not actually predict the presence of disease, but rather of diagnostic billing codes that may poorly correlate to actual disease (and suffer from selection biases and misclassification errors).

As machine learning methods evolve, abiding by key principles of good machine learning practice will serve to help improve

the utility, trustworthiness, and impact of machine learning.

Acknowledgements

Research reported in this publication was supported by the National Institute on Minority Health And Health Disparities of the National Institutes of Health under Award Numbers U54MD010724 and DP2MD010478. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. SB is employed by Collective Health, which uses machine learning methods to support care navigation programs. SB has previously received grants, stipends or consulting fees, unrelated to the current work, from the US National Institutes of Health, US Centers for Disease and Prevention, US Department of Agriculture's Economic Research Service, Center for Poverty Research, KPMG, Research Triangle International, the Robert Wood Johnson Foundation, Harvard University, Stanford University, and PLOS Medicine.

Conflict of Interest

No conflicts of interest to report.

Author Contributions

Research concept and design: Basu, Doupe; Acquisition of data: Faghmous, Doupe; Data analysis and interpretation: Basu, Doupe; Manuscript draft: Basu, Doupe; Statistical expertise: Basu, Doupe

References

1. Basu S, Sussman JB, Hayward RA. Black-White Cardiovascular Disease Disparities After Target-Based Versus Personalized Benefit-Based Lipid and Blood Pressure Treatment. *MDM Policy Pract*. 2017;2(2);ePub available from https://doi.org/10.1177/2381468317725741 PMID:30288429
2. Li Y, Rittenhouse-Olson K, Scheider WL, Mu L. Effect of particulate matter air pollution on C-reactive protein: a review of epidemiologic studies. *Rev Environ Health*. 2012;27(2-3):133-149. https://doi.org/10.1515/reveh-2012-0012 PMID:23023922
3. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edition. New York: Springer; 2011.
4. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22. https://doi.org/10.18637/jss.v033.i01 PMID:20808728
5. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324
6. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting. *Ann Stat*. 2000;28(2):337-407. https://doi.org/10.1214/aos/1016218223
7. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. https://doi.org/10.1214/aos/1013203451
8. Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. In: *ICML '08: Proceedings of the 25th International Conference on Machine Learning*. 2008:96-103. https://doi.org/10.1145/1390156.1390169
9. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: *ICML '08: Proceedings of the 23rd International Conference on Machine Learning*. 2006;C(1):161-168. https://doi.org/10.1145/1143844.1143865
10. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. https://doi.org/10.1097/EDE.0b013e3181c30fb2 PMID:20010215
11. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. 2004. Wiley Online Library available at https://doi.org/10.1002/0471722146
12. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. *Value Health*. 2019;22(7):808-815. https://doi.org/10.1016/j.jval.2019.02.012 PMID:31277828
13. Quinlan JR. Induction of decision trees. *Mach Learn*. 1986;1(1):81-106. https://doi.org/10.1007/BF00116251
14. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci USA*. 2016;113(27):7353-7360. https://doi.org/10.1073/pnas.1510489113 PMID:27382149
15. Poole S, Grannis S, Shah NH. Predicting emergency department visits. *AMIA Jt Summits Transl Sci Proc*. 2016;2016:438-445. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001776/.
16. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw*. 2015;61:85-117. https://doi.org/10.1016/j.neunet.2014.09.003 PMID:25462637
17. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. *Doctor AI: predicting clinical events via recurrent neural networks*. Presented at 2016 Machine Learning and Healthcare Conference, 2016. Last accessed Oct 1, 2019 from http://arxiv.org/abs/1511.05942.
18. Krizhevsky A, Sutskever I, Hinton GE. *ImageNet Classification with Deep Convolutional Neural Networks*. 2012:1097-1105. Last accessed Oct 1, 2019 from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
19. Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. February 2014. Last accessed Oct 1, 2019 from http://arxiv.org/abs/1402.3722.
20. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2018;25(10):1419-1428; epub ahead of print. https://doi.org/10.1093/jamia/ocy068 PMID:29893864
21. Kachuee M, Fazeli S, Sarrafzadeh M. ECG heartbeat classification: a deep transferable representation. April 2018. Last accessed Oct 1, 2019 from http://arxiv.org/abs/1805.00794.. https://doi.org/10.1109/ICHI.2018.00092
22. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1):e25. https://www.degruyter.com/downloadpdf/j/sagmb.2007.6.issue-1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml https://doi.org/10.2202/1544-6115.1309 PMID:17910531
23. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9(1):307. https://doi.org/10.1186/1471-2105-9-307 PMID:18620558
24. Larsen WA, McCleary SJ. The use of partial residual plots in regression analysis. *Technometrics*. 1972;14(3):781-790. https://doi.org/10.1080/00401706.1972.10488966
25. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J Med Internet Res*. 2016;18(12):e323. https://doi.org/10.2196/jmir.5870 PMID:27986644
26. Boulesteix AL. Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLOS Comput Biol*. 2015;11(4):e1004191. https://doi.org/10.1371/journal.pcbi.1004191 PMID:25905639
27. Tanwani AK, Afridi J, Shafiq MZ, Farooq M. Guidelines to select machine learning scheme for classification of biomedical datasets. In: *Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. LNCS; 2009:128-139. *Lecture Notes in Computer Science*; vol 5483., https://doi.org/10.1007/978-3-642-01184-9_12
28. Ioannidis JPA. Acknowledging and Overcoming Nonreproducibility in Basic and Preclinical Research. *JAMA*. 2017;317(10):1019-1020. https://doi.org/10.1001/jama.2017.0549 PMID:28192565