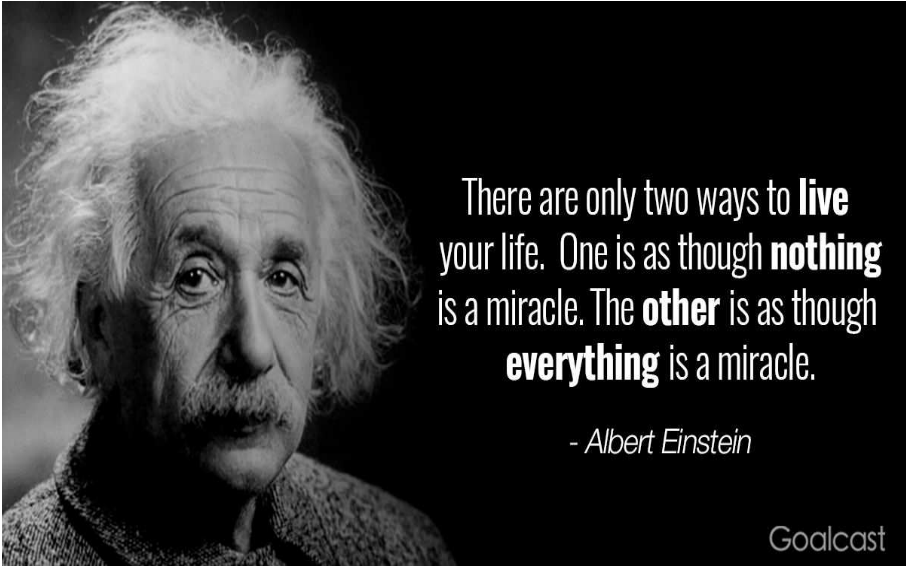
Principles of Statistical Machine Learning
Overview of the 7 Wheels of SML

Fokoué Ernest, PhD
Professor of Statistics

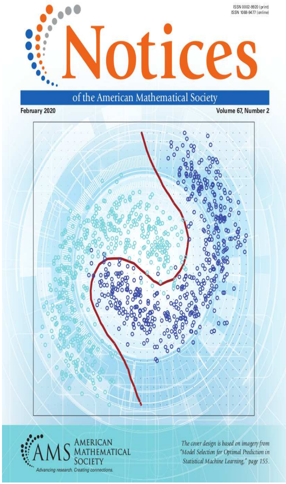
@ErnestFokoue

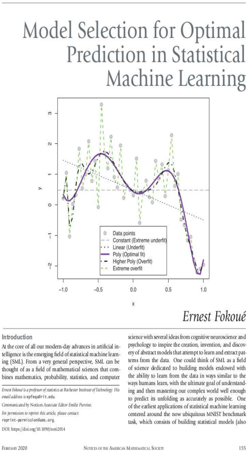
To understand God's thoughts, one must study statistics, the measure of His purpose
Florence Nightingale

Miraculous Universe!



Notices of the American Mathematical Society





Model Selection for Optimal Prediction in Statistical Machine Learning

Ernest Fokoué

Introduction

At the core of all our modern-day advances in artificial intelligence is the emerging field of statistical machine learning (SML). From a very general perspective, SML can be thought of as a field of mathematical sciences that combines mathematics, probability, statistics, and computer science with several ideas from cognitive neuroscience and psychology to inspire the creation, invention, and discovery of abstract models that attempt to learn and extract patterns from the data. One could think of SML as a field of science dedicated to building models endowed with the ability to learn from the data in ways similar to the ways humans learn, with the ultimate goal of understanding and then mastering our complex world well enough to predict its unfolding as accurately as possible. One of the earliest applications of statistical machine learning centered around the now ubiquitous MNIST benchmark task, which consists of building statistical models (also

Ernest Fokoué is a professor of statistics at Rochester Institute of Technology. His email address is epfeqa@rit.edu.

Communicated by Notices Associate Editor Emilie Purvine.

For permission to reprint this article, please contact: reprint-permission@ams.org.

DOI: https://doi.org/10.1090/noti2014

February 2020 Notices of the American Mathematical Society 155

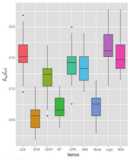


Figure 5. Predictive performances on the ionosphere data.

machines endowed with an inherent capability to predict accurately and precisely. In this paper, we have explored the niceties and subtleties of such a goal and have demonstrated that it requires a hefty dose of care and caution and definitely calls upon a solid theoretical understanding of learnability along with a lot of artlike practical common sense. Anyone who has done practical data science knows beyond a shadow of a doubt that data has a mind of its own, and tends to resist the temptation to seek a holy grail or a unified field, or any paradigm that works perfectly all the time. Practical data science almost always forces the practitioner to solve the problem at hand as thoroughly and as idiosyncratically as possible rather than seek a one-size-fits-all method that works well everywhere. At the heart of what we suggested throughout this paper is the theoretical result known as the no free lunch theorem, which reveals, both implicitly and explicitly, that the theoretical bounds studied extensively by experts do not really help much when it comes to practically selecting the optimal predictive model. Optimal predictive modelling is, and may always be, both a science and an art, requiring both mathematical and statistical rigor along with practical computational common sense.

References

[1] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, 1973:199–213. MR0483125
[2] Barbieri M and Berger JO. Optimal predictive model selection, Ann. Statist., (32):870–897, 2004. MR2065192
[3] Breiman L. Bagging predictors, Machine Learning, (24):123–140, 1996.
[4] Breiman L. Random forests, Machine Learning, (45):5–32, 2001. MR3874153
[5] Clarke B, Fokoue E, Zhang H. Principles and Theory for Data Mining and Machine Learning, first edition, Springer Texts in Statistics, Springer-Verlag, 2009. MR2839778
[6] Csató L, Fokoué E, Opper M, Schottky B, Winther O. Efficient approaches to Gaussian process classification. In: Leen TK, Solla SA, Müller K-R, eds. Advances in Neural Information Processing Systems, number 12 of 12. MIT Press; 2000.
[7] Devroye L, Györfi L, Lugosi G. A Probabilistic Theory of Pattern Recognition, Stochastic Modelling and Applied Probability, Springer New York, 1997.
[8] Domingos P. A unified bias-variance decomposition for zero-one and squared loss, AAAI/IAAI, AAAI Press, 564–569, 2000.
[9] Fokoué E. Estimation of atom prevalence for optimal prediction. In: Prediction and Discovery, Contemporary Mathematics, vol. 443. American Mathematical Society; 2007:103–129. MR2433288
[10] Hoerl A and Kennard R. Ridge regression: biased estimation for non-orthogonal problems, Technometrics, (12):55–67, 1970.
[11] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial, Statist. Sci., 14(4):382–417, 1999. MR1765176
[12] Kohavi R and Wolpert DH. Bias plus variance decomposition for zero-one loss functions. In: Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3–6, 1996. 1996:275–283.
[13] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review, 65–386, 1958. MR0122606
[14] Schapire RE and Freund Y. Boosting: Foundations and Algorithms, The MIT Press, 2012. MR2920188
[15] Schwarz G. Estimating the dimension of a model, The Ann. Statist., (6):461–464, 1978. MR0468014
[16] Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, J. Roy. Statist. Soc. Ser. B (Methodological), 39(1):44–47, 1977. MR501454
[17] Stone M. Cross-validatory choice and assessment of statistical predictions, J. Roy. Statist. Soc. Ser. B (Methodological), 111–147, 1974. MR356377
[18] Tibshirani R. Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. Ser. B (Methodological), 58(1):267–288, 1996. MR1379242
[19] Vapnik VN. The Nature of Statistical Learning Theory, Springer, 2000. MR1719582

Credits

All figures are courtesy of the author.
Author photo is courtesy of Rick Scoggins.

Ernest Fokoué

168 Notices of the American Mathematical Society Volume 67, Number 2

I am infinitely grateful to God for the blessing of publishing a featured paper in the Notices of the American Mathematical Society. To God be the Glory! Amen! Alleluia!

Objectives and Elements of this Module

- *Prerequisites*
  - *Basic probabilistic and statistical concepts*
  - *Rudimentary ideas of vector-matrix algebra and a bit of calculus*
  - *Basic understanding of algorithmics and complexity*
- *Objectives*
  - *Discover the seven (7) wheels of statistical machine learning*
  - *Get acquainted with the basic vocabulary of statistical machine learning*
  - *Explore some basic concepts and principles of SML and some tools thereof*
- *Resources*
  - *Articles: Notices of the AMS*
  - *Datasets: UC Irvine*
  - *Websites: R Project*

On the Landscape of Statistical Machine Learning

- *Applications: Sharpen your intuition and your commonsense by questioning things, reading about interesting open applied problems, and attempt to solve as many problems as possible*
- *Methodology: Read and learn about the fundamental of statistical estimation and inference, get acquainted with the most commonly used methods and techniques, and consistently ask yourself and others what the natural extensions of the techniques could be.*
- *Computation: Learn and master at least two programming languages. I strongly recommend getting acquainted with R*

  http://www.r-project.org
- *Theory: "Nothing is more practical than a good theory" (Vladimir N. Vapnik). When it comes to data mining and machine learning and predictive analytics, those who truly understand the inner workings of algorithms and methods always solve problems better.*

Note that in this case, a degenerate multinomial is not a good sign.

On the 7 Wheels of Statistical Machine Learning I

I came up with the concept of the seven (7) wheels of statistical machine learning (SML) upon noticing after several years of experience in the field, that these themes tended to almost always adorn all my SML activities.

1. *Wheel #1 - Data Exploration and Discovery:*
   - *What kind of informal insights into the underlying phenomenon can be gleaned from the data?*
   - *Distributional insights?*
   - *The 5 Vs of Data? (Variety, Volume, Velocity, Veracity, Value/Validity)*

$$\mathscr{D}_n = \{(\mathbf{x}_i, y_i) \overset{iid}{\sim} p_{\mathbf{xy}}(\mathbf{x}, y),\ \mathbf{x}_i \in \mathscr{X}, y_i \in \mathscr{Y},\ i = 1, \cdots, n\}, \tag{1}$$

where all pairs $(\mathbf{x}_i, y_i) \in \mathscr{X} \times \mathscr{Y}$, *and* $p_{\mathbf{xy}}(\mathbf{x}, y)$ *is the probability density function associated with the probability measure* $\mathbb{P}$ *on their Cartesian product* $\mathscr{Z} \equiv \mathscr{X} \times \mathscr{Y}$.

2. *Wheel #2 - Function Spaces and Hypothesis Spaces: What kind of abstract mathematical model can be represent and fit the data? What kind of function/hypothesis spaces seem to be suggest by the partial or complete view of the data?*

$$\mathscr{H}(\Phi) = \Big\{ f : \mathscr{X} \to \mathscr{Y} \mid \exists \mathrm{w}_0 \in \mathbb{R}, \mathbf{w} \in \mathscr{F} : \forall \mathbf{x} \in \mathscr{X}, \quad f(\mathbf{x}) = \mathsf{sign}\left(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + \mathrm{w}_0\right) \Big\}, \tag{2}$$

where $\Phi : \mathscr{X} \longrightarrow \mathscr{F}$ is a mapping that projects each input $\mathbf{x}$ up to a high dimensional feature space $\mathscr{F}$, thereby allowing the corresponding machine the capacity to capture nonlinear decision boundaries.

3. *Wheel #3 - Loss Functions and Theoretical Definition of Learning:* *Theoretical Risk Minimization! Zero One Loss, Squared Error Loss, Exponential Loss, Cross Entropy Loss, Hinge Loss, Huber Loss, Epsilon Insensitive Loss*

$$\begin{aligned} R(f) &= \mathbb{E}[\mathcal{L}(Y, f(X))] \\ &= \int_{\mathscr{X} \times \mathscr{Y}} \mathcal{L}(y, f(\mathbf{x})) p_{\mathbf{xy}}(\mathbf{x}, y) d\mathbf{x} dy, \end{aligned} \tag{3}$$

where A loss function $\mathcal{L}(\cdot, \cdot)$ is a nonnegative bivariate function $\mathcal{L} : \mathscr{Y} \times \mathscr{Y} \longrightarrow \mathbb{R}_+$, such that given $a, b \in \mathscr{Y}$, the value of $\mathcal{L}(a, b)$ measures the discrepancy between a and b, or the deviance of a from b, or the loss incurred from using b in place of a. Like

$$\mathcal{L}(y, f(\mathbf{x})) = \mathbb{1}(y \neq f(\mathbf{x})) = \begin{cases} 0 & \textit{if} \quad y = f(\mathbf{x}), \\ 1 & \textit{if} \quad y \neq f(\mathbf{x}). \end{cases} \tag{4}$$

On the 7 Wheels of Statistical Machine Learning IV

4. *Wheel #4 - Construction of Learning Machines and Estimators:* *What is your algorithm for constructing the hypothesized (implicit or explicit) learning machine? How does one construct an efficient, stable and hopefully scalable computational scheme/framework for obtaining the empirical realization of the theoretical machine? What are the statistical properties of your learning machine? Bias of your learning machine? Variance of your learning machine? Bias Variance Dilemma? What is the computational complexity of your algorithm?*

$$\begin{aligned}\widehat{f} &= \widehat{f}_{\mathscr{H},n} = \widehat{f}_n = \underset{f\in\mathscr{H}}{\texttt{argmin}}\left\{\widehat{R}_n(f)\right\} \\ &= \underset{f\in\mathscr{H}}{\texttt{argmin}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\mathbf{x}_i))\right\}.\end{aligned} \tag{5}$$

Fundamental result

$$R(\widehat{f}) = \mathbb{E}[(Y-\widehat{f}(\mathbf{x}))^2] = \sigma^2 + \texttt{Bias}^2(\widehat{f}(\mathbf{x})) + \texttt{variance}(\widehat{f}(\mathbf{x})).$$

5. *Wheel #5 - Refinement and Intrinsic Selection: Within the function space and in keeping with Hadamard's wellposedness, how to refine the machine and how to choose the most viable in $\mathscr{H}$? Cross Validation Criterion? Akaike Information Criterion (AIC)? Bayesian Information Criterion (BIC)? MLE? Bayesian?*

$$\widehat{f}_{\mathscr{H},\lambda,n} = \underset{f\in\mathscr{H}}{\texttt{argmin}}\left\{ \widehat{R}_{\mathscr{H},n}(f) + \lambda\Omega_{\mathscr{H}}(f) \right\}, \tag{6}$$

where λ controls the bias-variance trade-off.

$$\boldsymbol{\gamma}^{(\texttt{BIC})} = \underset{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}{\texttt{argmin}}\left\{\texttt{BIC}_n(M_{\boldsymbol{\gamma}})\right\} \tag{7}$$

where the score $\texttt{BIC}_n(M_{\boldsymbol{\gamma}})$ of model $M_{\boldsymbol{\gamma}} \in \mathscr{M}$ is

$$\texttt{BIC}_n(M_{\boldsymbol{\gamma}}) = -2\log L(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}|M_{\boldsymbol{\gamma}};\mathscr{D}_n) + |M_{\boldsymbol{\gamma}}|\log n. \tag{8}$$

6. *Wheel #6 - Empirical Extrinsic Comparison:* No Free Lunch. Given a dataset $\mathscr{D}_n$ and a collection of potential function spaces like $\mathcal{C}$, along with $\mathscr{D}_n^{(\mathtt{s})} = \mathscr{D}_{\mathtt{tr}}^{(\mathtt{s})} \cup \mathscr{D}_{\mathtt{te}}^{(\mathtt{s})}$, one defines

$$
\begin{aligned}
E = (E_{sm}) &= \widehat{R}_{\mathtt{te}}(\hat{f}_m^{(s)}) = \mathtt{te}(\widehat{f}_m^{(\mathscr{D}_{\mathtt{tr}}^{(\mathtt{s})})}) \\
&= \mathtt{Error\ of}\ \ \widehat{f}_m^{(\mathscr{D}_{\mathtt{tr}}^{(\mathtt{s})})}(\cdot) \text{ on } \mathscr{D}_{\mathtt{te}}^{(\mathtt{s})}.
\end{aligned}
$$

where

$$
\widehat{R}_{\mathtt{te}}(f) = \frac{1}{|\mathscr{D}_{\mathtt{te}}|} \sum_{j=1}^{n} \mathcal{L}(Y_j, f(X_j)) \mathbb{1}(Z_j \in \mathscr{D}_{\mathtt{te}}). \tag{9}
$$

$$
\mathtt{AVTE}(\widehat{f}) = \frac{1}{S} \sum_{\mathtt{s}=1}^{S} \mathtt{te}(\widehat{f}^{(\mathscr{D}_{\mathtt{tr}}^{(\mathtt{s})})}). \tag{10}
$$

$$
\widehat{f}_{\mathtt{best}}^{(\mathcal{C})} = \underset{\widehat{f} \in \mathcal{C}}{\mathtt{argmin}} \Big\{ \mathtt{AVTE}(\widehat{f}) \Big\}.
$$

On the 7 Wheels of Statistical Machine Learning VII

7. *Wheel #7 - Theoretical assessment and justification: Generalization, Out of Sample Performance, Probabilistic View of Predictive Performance, Probabilistic Inequalities, Confidence Intervals, Hypothesis Testing, Confidence Bounds, VC Theory, VC Bounds, VC Dimension, Rademacher Complexity.*

$$\mathbb{E}[R(\widehat{f}_n) - R^\star] = \underbrace{\mathbb{E}[R(\widehat{f}_n) - R(f^\diamond)]}_{\text{Estimation error}} + \underbrace{\mathbb{E}[R(f^\diamond) - R^\star]}_{\text{Approximation error}} \tag{11}$$

From Vapnik and Chervonenkis, we have the fundamental theorem: For every $f \in \mathscr{H}$, and $n > h$, with probability at least $1 - \eta$, we have

$$R(f) \leq \widehat{R}_{\mathscr{H},n}(f) + \sqrt{\frac{h\left(\log\frac{2n}{h} + 1\right) + \log\left(\frac{4}{\eta}\right)}{n}}.$$

Bias-Variance Trade-off

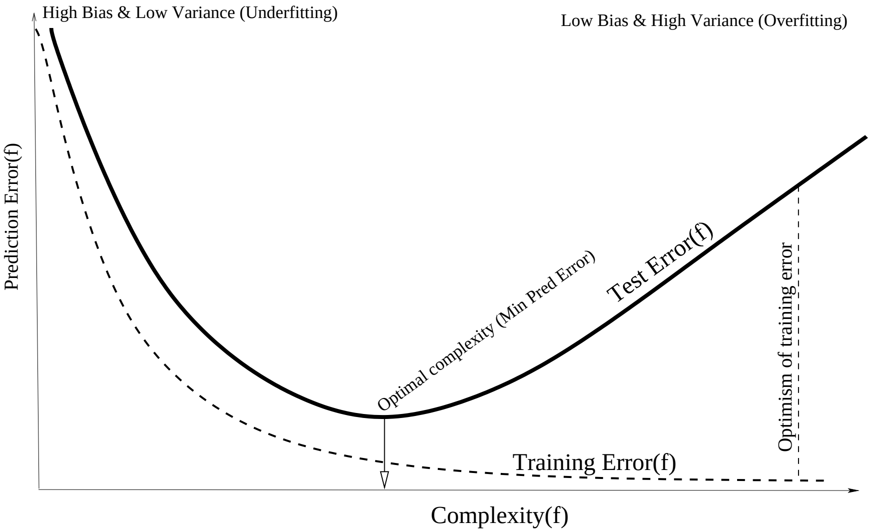


Figure: *Illustration of the qualitative behavior of the dependence of bias versus variance on a tradeoff parameter such as* λ *or* h. *For small values the variability is too high; for large values the bias gets large.*

Cross Validation for Intraspace Model Selection

Algorithm 1 $V-$fold Cross Validation

for $\mathtt{v} = 1$ *to* V **do**

Extract the validation set $\mathscr{D}_{\mathtt{v}} = \{\mathbf{z}_i \in \mathscr{D}_n : \ i \in [1+(\mathtt{v}-1)\times m, \mathtt{v}\times m]\}$

Extract the training set $\mathscr{D}_{\mathtt{v}}^c := \mathscr{D}_n \backslash \mathscr{D}_{\mathtt{v}}$

Build the estimator $\widehat{f}^{(-\mathscr{D}_{\mathtt{v}})}(\cdot)$ using $\mathscr{D}_{\mathtt{v}}^c$

Compute predictions $\widehat{f}^{(-\mathscr{D}_{\mathtt{v}})}(\mathbf{x}_i)$ `for` $\mathbf{z}_i \in \mathscr{D}_{\mathtt{v}}$

Compute the validation error for the $\mathtt{v}^{th}$ chunk

$$\widehat{\varepsilon}_{\mathtt{v}} = \frac{1}{|\mathscr{D}_{\mathtt{v}}|} \sum_{i=1}^{n} \mathbb{1}(\mathbf{z}_i \in \mathscr{D}_{\mathtt{v}}) \mathcal{L}(\mathrm{y}_i, \widehat{f}^{(-\mathscr{D}_{\mathtt{v}})}(\mathbf{x}_i))$$

Compute the `CV` score $\mathtt{CV}(\widehat{\mathrm{g}}) = \frac{1}{V} \sum_{v=1}^{V} \hat{\varepsilon}_{\mathtt{v}}$

Example of Cross Validated Size of Neighborhood

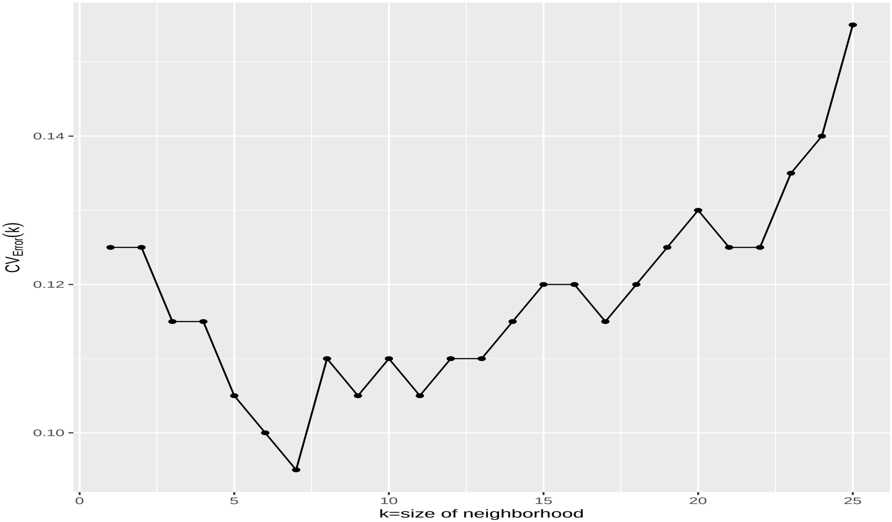


Figure: *Cross Validated Size of Neighborhood on the Lung Dataset*

Algorithm for Extrinsic Predictive Comparisons

Algorithm 2 Stochastic Hold Out for Generalization

for $s = 1$ *to* S **do**

Generate the $\mathtt{s}^{th}$ random split $\mathscr{D}_n$ into $\mathscr{D}_{\mathtt{tr}}^{(s)}$ and $\mathscr{D}_{\mathtt{te}}^{(s)}$

Such that $\mathscr{D}_n = \mathscr{D}_{\mathtt{tr}}^{(s)} \cup \mathscr{D}_{\mathtt{te}}^{(s)}$ and $n = |\mathscr{D}| = \tau|\mathscr{D}_{\mathtt{tr}}^{(s)}| + (1-\tau)|\mathscr{D}_{\mathtt{te}}^{(s)}|$

for $m = 1$ *to* M **do**

Build and refine the m^{th} learning machine $\widehat{f}_m^{(\mathscr{D}_{\mathtt{tr}}^{(s)})}(\cdot)$ using $\mathscr{D}_{\mathtt{tr}}^{(s)}$

Compute predictions $\widehat{f}_m^{(\mathscr{D}_{\mathtt{tr}}^{(s)})}(\mathbf{x}_i)$ $\mathtt{for}$ $\mathbf{z}_i \in \mathscr{D}_{\mathtt{te}}^{(s)}$

Compute the test error for the m^{th} learning machine

$$\begin{aligned}\widehat{\varepsilon}_{sm} &= \widehat{R}_{\mathtt{te}}(\hat{f}_m^{(s)}) \\ &= \frac{1}{|\mathscr{D}_{\mathtt{te}}^{(s)}|}\sum_{i=1}^{n} \mathbb{1}(\mathbf{z}_i \in \mathscr{D}_{\mathtt{te}})\mathcal{L}(\mathrm{y}_i, \widehat{f}_m^{(\mathscr{D}_{\mathtt{tr}}^{(s)})}(\mathbf{x}_i))\end{aligned}$$

Example of Extrinsic Predictive Comparisons

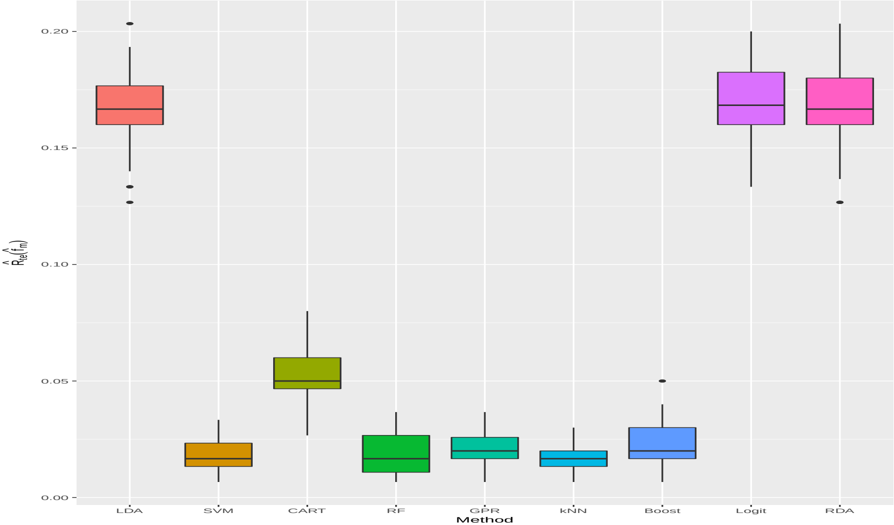
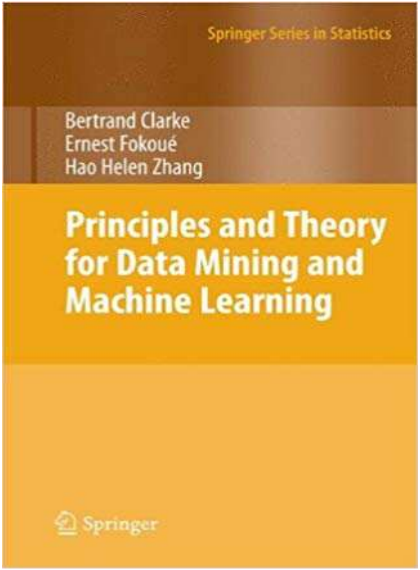


Figure: *Extrinsic Predictive Comparisons on the Banana Dataset*



Clarke, B. and Fokoué, E. and Zhang, H. (2009). Principles and Theory for Data Mining and Machine Learning. Springer Verlag, New York, (ISBN: 978-0-387-98134-5), (2009)

References

[1] *Clarke, B., Fokoué, E. and Zhang, H. H. (2009).* Principles and Theory for Data Mining and Machine Learning. *Springer Verlag, New York, (ISBN: 978-0-387-98134-5), (2009)*

[2] *Vapnik, N. V.(1998).* Statistical Learning Theory. *Wiley, ISBN: 978-0-471-03003-4, (1998)*

[3] *Vapnik, N. V.(2000).* The Nature of Statistical Learning Theory. *Springer, ISBN 978-1-4757-3264-1, (2000)*

[4] *Hastie, T. and Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer, ISBN 978-0-387-84858-7*