Learning from Noise:

Applying Sample Complexity for Social Science Research

Perry Carter[*] Dahyun Choi[†]

March 21, 2024

What constitutes “good enough" data for social scientists? Statistical learning provides a bridge between concepts and complex empirical realities while many concepts of interest to scholars are both highly multidimensional and inherently fuzzy, making classification error-prone. In this article, we consider the “*Probably Approximately Correct*” model, which takes advantage of researcher-specified bounds on labeling error to guarantee the sample size required for a minimum level of accuracy. We develop a simulation-based approach to use PAC model and provide the `scR` R package, offering a computationally efficient way for applied researchers to implement the proposed methods. we aim to improve standard practice by providing a general-purpose tool to validate the quality of measures when fuzzy measurement boundaries make generating large amounts of data with ground truth labels infeasible.

Keywords: Noisy Learning, Sample Complexity Bounds, Vapnik-Chervonenkis Dimension, Measurement
Word Count: 2916

[*]Ph.D. Candidate, Department of Politics, Princeton University.Email: pjcarter@princeton.edu
[†]Ph.D. Candidate, Department of Politics, Princeton University. Email: dahyunc@princeton.edu

1. Introduction

What constitutes “good enough" data? The majority of concepts of interest in political science are not directly observable and therefore require the production of labeled data in order to test hypotheses. There has been extensive work on the use of hand-coding or crowdsourced annotations (e.g., Tian and Zhu 2015; Benoit et al. 2016; Carlson and Montgomery 2017; Ying, Montgomery, and Stewart 2022; Miller, Linder, and Mebane 2018), but since the collection of human-labeled data is expensive and time-consuming, researchers have attempted to circumvent the cost by using machine learning to extrapolate from a relatively small training set to a larger set of unlabeled data (e.g., Barberá et al. 2021). However, despite the explosion in interest in machine learning, the issue of how data quality affects measurement and inference has received relatively little attention in the political science literature.

A typical implicit assumption in most applications is that learning algorithms have access to a noise-free oracle[1] for training examples of the target concept. However, many concepts in political science are subject to conceptual ambiguity or “stretching” (Collier and Mahon 1993). Such concepts are typically both high-dimensional and have ambiguous boundaries, making it difficult to specify explicit conditions for inclusion *a priori*.

Given this inherent complexity of measurement tasks in social science, we propose the application of an approach based on the Probably Approximately Correct (PAC) Model, which takes advantage of researcher-specified bounds on conceptual complexity and labeling error to guarantee the sample size necessary to achieve a minimum level of accuracy with a precise level of confidence. We demonstrate the feasibility of this approach through a simulation-based method, implemented in a companion *R* package. This method allows researchers to determine appropriate bounds under a variety of

[1]That is, a mechanism generating accurately labeled examples from a known probability distribution.

sampling regimes for commonly used machine learning models. Our method thus provides a simple analog to power calculations for experimental work, which has allowed applied researchers to assess the feasibility of applying a chosen model at the design stage. We additionally incorporate the estimation of the Vapnik-Chervonenkis (VC) dimension, or "capacity", of the target concept, allowing researchers to evaluate theoretically how much data is needed for their design before making costly investments in data collection.

The aim of this paper is therefore to provide a general-purpose tool for assessing the sample size needed to achieve adequate performance from statistical learning algorithms in applied social science. This is an important topic that provides a foundation for the supervised learning tasks prevalent in political science using a principled approach analogous to the hypothesis testing framework for statistical inference. Moreover, when researchers intend to estimate causal effects of latent concepts as a downstream application, as in the persistent debate over the causal effects of democracy (Acemoglu et al. 2019), measurement error has direct implications for the power of the corresponding hypothesis test. By providing a straightforward way to assess the consequences of sample size for measurement, we therefore offer a means to improve the accuracy of power analysis for causal inference.

2. Setup

While measurement models are typically optimized for accuracy on observed labeled instances, the ability to make *generalizable* claims beyond the training data is the ultimate goal of statistical learning. Although sample splitting is now widely used to address this concern, it provides limited guarantees for the case when out-of-sample examples are generated from a different distribution than the training data – as, for instance when labelled examples come from a convenience sample, and not a representative random

sample of the population. We formalize the problem as follows (Laird 2012): denote domain $\mathcal{X}$, label sets $\mathcal{Y}$, and concept classes $\mathcal{C}$. $\mathcal{X}$ includes all possible instances that the researcher may want to label and a set $\mathcal{Y}$ includes all possible labels or predictions for a single instance. An instance-label pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is called a labeled instance and a concept is a function $c : \mathcal{X} \to \mathcal{Y}$. We assume that there is an unknown concept[2] c that determines the true labels of instances.

Then, we consider a probability distribution D over χ. We assume that the instances we observe are independent and identically distributed (i.i.d) according to an unknown D. Given that there is an unknown concept c which determines the true label of instances, the set of labeled instances $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is generated by taking $x_i \sim D$ i.i.d and observing the corresponding $y_i = c(x_i)$. Then the true error and empirical error can be defined as follows.

DEFINITION 1. ***True Error:*** *Consider a data-generating distribution D and the true labeling concept c. The true error of a concept h with respect to D is the probability that h makes a mistake.*

$$R(h) = Pr_{x \sim D}[h(x) \neq c(x)] \tag{1}$$

DEFINITION 2. ***Empirical Error:*** *Given a sample set S, the empirical error of a concept h with respect to S is the fraction of instances in S that are incorrectly labeled by h.*

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1(h(x_i) \neq y_i)) \tag{2}$$

Suppose we have a model that produces a hypothesis $h \in \mathcal{H}$, given a sample of N

[2]A concept in this sense is precisely understood as a binary classification rule. This is not as divergent from common social science usage as it may initially appear: for instance, by the concept of "democracy", we mean a set of explicit rules that allow an observer to determine whether a given political system is or is not a *democracy*.

training examples. The algorithm is called *consistent* if for every ϵ and δ, there exists a positive number of training examples N such that for any distribution p^*, we have that

$$P(|R(h) - \hat{R}(h)| > \epsilon) < \delta \tag{3}$$

The *sample complexity* is the minimum value of N for which the equation (3) holds true. We therefore seek to determine the minimum sample size that produces a hypothesis within a specified error tolerance of the true concept with high probability. This is conceptually similar to the widely used approach of power analysis in experimental research and, as we show, has direct implications for power when researchers seek to identify the causal effect of a latent concept.

2.1. Sample Complexity Bounds For Applied Research

In the previous section, it was assumed that all observed instances of the concept were correctly labelled. In practice, this is typically not the case, whether due to data quality or conceptual ambiguity that may cause inter-coder disagreement when data is labelled by hand. We therefore now consider the PAC model in the presence of classification error. In particular, we state a general lower bound on sample complexity given a pre-specified error rate.

A machine is given $\mathcal{H}$, a class of functions including the target h, and accuracy and confidence parameters, ϵ, δ. Then, the machine gains information about the target function by viewing examples labeled by f, subject to i.i.d. misclassification probability $\eta < 0.5$, and attempts to learn the target function according to a generic algorithm A. The goal is to generate outputs with out-of-sample accuracy of at least $1 - \epsilon$ with confidence at least $1 - \delta$; that is, for an algorithm A with error rate $e = |R(h) - \hat{R}(h)|$, we seek to achieve $P_A(e > \epsilon) \leq \delta$. Combining Simon (1993) with Aslam and Decatur (1996),

a general lower bound on sample complexity (SCB) is given by

$$\Omega\left(\frac{VC(F)}{\epsilon(1-2\eta)^2} + \frac{log(1/\delta)}{\epsilon(1-2\eta)^2}\right) \tag{4}$$

where $VC(F)$ indicates the Vapnik–Chervonenkis (VC) dimension, which is a measure of the *capacity* – that is, the underlying complexity – of the target concept[3].

This bound is pessimistic in the sense that it is designed to hold in the worst-case scenario where the sampling distribution is skewed towards the least informative examples, although it may be attained even under i.i.d. sampling regimes Long (1995). In practice, however, classifiers may perform much better than the bound would suggest with fewer data, such that the bound given here should be seen as a conservative estimate. Further discussion on this point can be found in Appendix A.

2.2. Estimating VC dimension

The bound given by 0(4) depends on the VC dimension of the target concept, which can generally be calculated analytically only for the most straightforward classifiers. To address this challenge, we calculate the sample complexity bound using an estimate that is consistent in simulation parameters rather than the true VC dimension. To do so, we estimate it empirically based on the known relationship between the worst-case generalization error of a classifier and its VC dimension, following McDonald, Shalizi, and Schervish (2011). Formally, the VC dimension of a hypothesis space $\mathcal{H}$ indicates the cardinality of the largest set $\mathcal{S}$ that can be *shattered* by $\mathcal{H}$. Further details can be found in Appendix B.

In Figure 1, we present the results of this estimation procedure for a k-dimensional

[3]Although mismatch between measurement and concept is a potentially serious concern in applied research, we will abstract away from this issue by assuming that the algorithm being used is capable of perfectly learning the target concept given infinite correctly-labeled data, so that we can substitute the unknown $VC(F)$ with the known $VC(A)$ without loss.

linear discriminant classifier, which is known to have a VC dimension of $k + 1$. The y-axis gives an estimated bound on the relationship between empirical risk and sample size for the given classifier, while the x-axis gives the sample size. Since the functional form of this relationship is known up to a constant given the true VC dimension, we can then estimate the VC dimension of any classifier through non-linear regression (Vapnik, Levin, and Le Cun 1994). McDonald, Shalizi, and Schervish (2011) demonstrates that this estimate is consistent in the number of simulations, so that the estimate converges to the true VC dimension given sufficient computational resources.

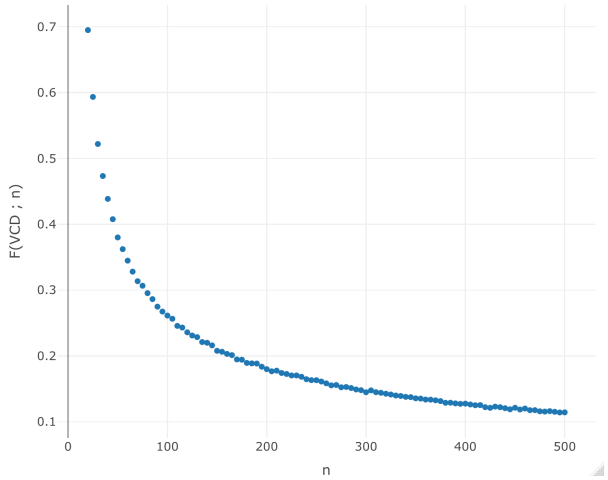

FIGURE 1. Simulation Study for Estimating the VC dimension of a linear discriminant model, $k = 7$

3. Simulation Studies

In order to verify that the sample complexity bounds perform reasonably well for realistic configurations of the parameters, we provide simulation-based analysis similar to Fong and Tyler (2021) and compare the empirical performance with theoretical bounds given the following inputs: (1) a confidence parameter (δ), (2) the accuracy parameter (ϵ) and (3) the misclassification rate η. The simulations are generated by the following process:

Algorithm Simulation-based Analysis

1. Decide on desired accuracy parameters and concept
2. Calculate the VCD of the chosen model using the above estimation procedure
3. Generate a fine grid of points over the k-dimensional feature space
4. Classify these points according to the pre-defined concept
5. Generate observed labels by adding fixed independent random noise with probability η
6. Calculate sample complexity bounds empirically for a range of acceptable error rates
7. Repeat the process according to a range of values of "optimism" parameter (analytic bound corresponds to worst-case sampling)

To facilitate application, we provide an R package, `scR`, that provides a computationally efficient way to implement the proposed methods. In this exercise, we consider a stylized version of the model of "polyarchy" proposed by in Dahl (2008), where the goal is to learn where the cutoff between polyarchies and non-polyarchies lies in a region defined by two latent dimensions – "participation" and "contestation". While we thus abstract away from much of the discussion of measurement that has dominated the recent literature on this topic (e.g., Little and Meng 2023), the basic issue of determining the cut-off point between democracies and nondemocracies remains highly salient both theoretically and empirically (Baltz, Vasselai, and Hicken 2022) and is complicated by small sample size.

Values are calculated by fixing $\eta = 0.05$ and either $\epsilon = 0.06$ or $\delta = 0.01$. In this setting, Bound 4 gives a theoretical bound of 178 observations, which is borne out by the simulation. Notably, this bound is significantly higher than the sample size needed to achieve performance of over 95% on conventional out-of-sample performance metrics, since it involves the more stringent requirement that the probability of high error rates in *any* sample be controlled, and not simply the average misclassification rate. Although this may lead to more conservative conclusions regarding the target sample size, a key advantage is that researchers must explicitly specify the confidence δ with which they hope to achieve the desired out-of-sample accuracy, improving transparency and

replicability. Morevoer, the theoretical bound of 178 cases closely corresponds to the simulation results under simple random sampling, shown in Figure 2, indicating that it is not unduly pessimistic even in this simple application.

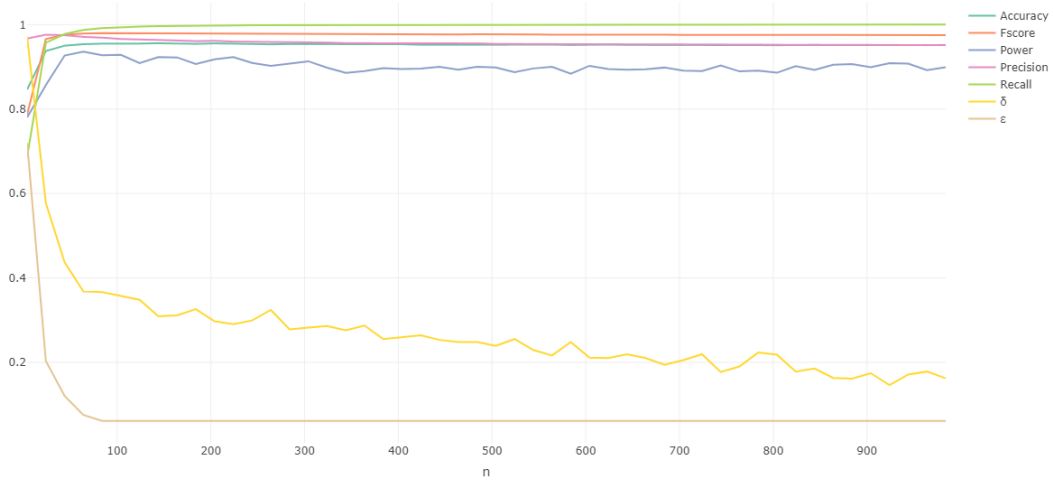

FIGURE 2. Learning to Classify Polyarchies

Figure 2 also shows the results of 1000 thousand simulated experiments attempting to identify the effect of the latent concept – being a polyarchy – on an arbitrary outcome generated based on the true labels. Here, we hold sample size constant and evaluate the ability of the trained model to correctly identify a significant effect ($\alpha = 0.05$) through out-of-sample predictions. The results demonstrate that the sample size needed to achieve the desired level of power in a downstream quasi-experiment again corresponds closely to the theoretical sample complexity bound, further underlining the value of incorporating estimation error at the pre-analysis stage.

4. Application to Predicting Recidivism (Dressel and Farid 2018)

In the appendix, we offer an application of the method to actual noisy data, based on the study implemented by Dressel and Farid (2018). The original authors studied how predictions made by people with little or no criminal justice expertise can be

comparable to machine-based commercial risk assessment software. The application of our method allows us to precisely assess the additional benefit provided by big data in this context, demonstrating how it could be applied by researchers prior to data collection. Our findings highlight the concept formation problem, as big data provides minimal additional benefit due to imprecise specification of the target concept. Further details are provided in Appendix C.

5. Discussion

In this paper, we consider the question of what constitutes "good enough" data, both in terms of sample size and labelling accuracy. To address this persistent question in applied research, we propose a novel application of the PAC model including both a theoretical and simulation-based approach for estimating sample complexity bounds. This is a general-purpose tool for quantitative research in political science that permits researchers to make rigorous predictions about what can be "learned" from data before investing in costly collection efforts.

A key advantage of our approach is that it provides a more precise alternative to the assumption that the sample size is "large enough" for asymptotic approximations to hold. Furthermore, we directly consider the role played by labeling error and concept definition on model performance, a factor that has generally been overlooked in applied work, with hand-labeled data assumed to represent ground truth. Our approach thus provides a practical toolkit for researchers to guarantee adequate performance at the design stage, improving the replicability and external validity of studies reliant on machine learning classifiers.

An important consideration is that our application of PAC learning focuses on binary-valued measures. While researchers can collapse continuous measures to binary by setting thresholds, a valuable extension would be to apply the framework to contin-

uous concepts. Next, although our validation exercises suggest that the theoretical bounds hold well in practice, one might still be concerned that the bounds may be too pessimistic. Related to this issue, one of our future tasks would be to update our simulation-based approach to allow researchers to parametrically pre-specify their "optimism" regarding the sampling procedure, generating an interval of progressively looser bounds.

Our goal is not to provide the definitive answer to the methodological question of what constitutes "good enough" data. Instead, we aim to initiate a dialogue. Social scientists have made significant strides in developing rigorous statistical models. However, the need to assess the quality of input has often been overlooked when performing downstream statistical inferences. We hope this paper contributes to rekindling attention to the importance of both data quantity and quality in applying statistical learning to social sciences.

6. Acknowledgements

We thank Brandon Stewart, Justin Grimmer, and P. M. Aronow for their discussions at various stages of this project. The authors thank the anonymous reviewers in advance for their suggestions.

7. Conflict of Interest

The authors are not aware of any conflicts of interest.

References

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A Robinson. 2019. Democracy does cause growth. *Journal of political economy* 127 (1): 47–100.

Aslam, Javed A, and Scott E Decatur. 1996. On the sample complexity of noise-tolerant learning. *Information Processing Letters* 57 (4): 189–195.

Baltz, Samuel, Fabricio Vasselai, and Allen Hicken. 2022. An unexpected consensus among diverse ways to measure democracy. *Democratization* 29 (5): 814–837.

Bansak, Kirk. 2019. Can nonexperts really emulate statistical learning methods? a comment on "the accuracy, fairness, and limits of predicting recidivism". *Political Analysis* 27 (3): 370–380.

Barberá, Pablo, Amber E Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. Automated text classification of news articles: a practical guide. *Political Analysis* 29 (1): 19–42.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review* 110 (2): 278–295.

Carlson, David, and Jacob M Montgomery. 2017. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review* 111 (4): 835–843.

Collier, David, and James E Mahon. 1993. Conceptual "stretching" revisited: adapting categories in comparative analysis. *American Political Science Review* 87 (4): 845–855.

Dahl, Robert A. 2008. *Polyarchy: participation and opposition.* Yale university press.

Dressel, Julia, and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4 (1): eaao5580.

Fong, Christian, and Matthew Tyler. 2021. Machine learning predictions as regression covariates. *Political Analysis* 29 (4): 467–484.

Gross, Samuel R, Barbara O'brien, Chen Hu, and Edward H Kennedy. 2014. Rate of false conviction of criminal defendants who are sentenced to death. *Proceedings of the National Academy of Sciences* 111 (20): 7230–7235.

Laird, Philip D. 2012. *Learning from good and bad data.* Vol. 47. Springer Science & Business Media.

Little, Andrew, and Anne Meng. 2023. Subjective and objective measurement of democratic backsliding. *Available at SSRN 4327307.*

Long, Philip M. 1995. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks* 6 (6): 1556–1559.

McDonald, Daniel J, Cosma Rohilla Shalizi, and Mark Schervish. 2011. Estimated vc dimension for risk bounds. *arXiv preprint arXiv:1111.3404.*

Miller, Blake, Fridolin Linder, and Walter R Mebane. 2018. Active learning approaches for labeling text. *Political Analysis.*

Simon, Hans Ulrich. 1993. General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the sixth annual conference on computational learning theory,* 402–411.

Tian, Tian, and Jun Zhu. 2015. Max-margin majority voting for learning from crowds. *Advances in neural information processing systems* 28.

Vapnik, Vladimir, Esther Levin, and Yann Le Cun. 1994. Measuring the vc-dimension of a learning machine. *Neural computation* 6 (5): 851–876.

Ying, Luwei, Jacob M Montgomery, and Brandon M Stewart. 2022. Topics, concepts, and measurement: a crowdsourced procedure for validating topics as measures. *Political Analysis* 30 (4): 570–589.

Supporting Information for

Learning from Noise: Applying Sample Complexity for Political Science Research

Perry Carter Dahyun Choi

November 2023

*Ph.D. Candidate, Department of Politics, Princeton University.Email: pjcarter@princeton.edu
†Ph.D. Candidate, Department of Politics, Princeton University. Email: dahyunc@princeton.edu

A. Identification from Noisy Cases

Here we provide further elaboration on Section 2.1 concerning the sample complexity bounds for applied research. This section provides a more detailed overview of the setup in Laird (2012) and Simon (1993) on which our approach is based. Suppose we have a rule e with error p. For example, the rule fails or disagrees with an example on average pm times in m examples. With the addition of noise, it may fail more often or less. Then the expected failure rate p_η is as follows:

$$p_n = (1 - \eta)\, p + \eta(1 - p) \tag{5}$$

The first indicates that if no classification errors occurs with probability $1 - \eta$, the probability of failure is p. The second term suggests that if a classification error does not strike with probability η, the rule will fail only if it would not have failed without the error, with probability $1 - p$. Note the cases below.

- When p =0 (zero error). its failure rate increases to η with noise.
- When $p \geq \epsilon$, its failure rate is at least $\eta + \epsilon(1 - 2\eta)$; and since $(1 - 2\eta) > 0$, this failure rate is greater than that of any correct one with zero error. [1]

We refer to rules with error greater than ϵ as ϵ -bad. Rules that are not ϵ -bad are ϵ -good. Rules with zero error are described simply as good. On average, ϵ -bad rules have a failure rate that is greater than that of good rules by at least $\epsilon(1 - 2\eta)$.

By the Law of Large Numbers, as $m \to \infty$

$$Pr[|\, p - \hat{p}| > \epsilon] \to 0$$

[1] When η is $\frac{1}{2}$, all of the information in the example become obliterated. For noise being $\frac{1}{2}$, we cannot see how PAC-identification could be achieved with much noise.

that is, $\hat{p}$ will be arbitrarily close to p for sufficiently many tests of an event whose probability of occurring is p, for all $\epsilon > 0$. If a correct hypothesis fails on average at the rate η, then with enough examples m, we will measure a failure rate closest to η with high probability. Similarly, an ϵ- bad rule will fail at nearly its expected rate, $\eta + s$, where $s \geq \epsilon(1 - 2\eta) > 0$.

LEMMA 1 (Hoeffding's inequality). *Consider a Bernoulli random variable with probability p of having the value 1 and $1 - p$ of having value 0. Let $GE(p, m, r)$ be the probability of at least $\lceil rm \rceil$ successes in m independent trials, and $LE(p, m, r)$ be the probability of at most $\lfloor rm \rfloor$ successes. If $0 \leq p \leq 1$, $0 \leq s \leq 1$, and m is any positive integer then*

$$LE(p, m, p - s) \leq e^{-2s^2 m}$$

and

$$GE(p, m, p + s) \leq e^{-2s^2 m}$$

This lemma bounds the probability that r, the empirical rate of success, is very different from p.

Suppose $\epsilon > 0$, $\delta \leq \frac{1}{2}$ and $0 \leq \eta \leq \eta_b < \frac{1}{2}$. Let success for a rule e_i refer to the event of disagreeing with a random sample. In m examples, F_i is the number of successes, and $\frac{F_i}{m}$ is the empirical rate of disagreement.

THEOREM 1. *When*

$$m \geq \frac{2}{\epsilon^2(1 - 2\eta_b)^2} ln\frac{2N}{\delta}$$

, the algorithm pac-identifies

Or a tighter bound:

THEOREM 2. *Let* $\eta < 1/2$ *be the rate of classification noise and* N *the number of rules in the class* $\mathcal{E}$. *Assume* $0 < \epsilon, \delta < \frac{1}{2}$. *Then the number* m *of examples required is at least*

$$m \geq \max\left[\frac{\ln(1/2\delta)}{\ln[1-\epsilon(1-2\eta)]^{-1}}, \log_2 N(1-2\epsilon(1-\delta)+2\delta)\right] \tag{6}$$

and at most

$$m \leq \frac{\ln(N/\delta)}{\epsilon\left[1-\exp[-\frac{1}{2}(1-2\eta)^2]\right]} \tag{7}$$

Proofs

The probability that a good rule fails on an example is η, while the probability that an ϵ-bad rule fails is at least $\eta + \epsilon(1-\epsilon)$. The difference between these two rates is at least $\epsilon(1-\eta) \geq \epsilon(1-2\eta_b) = s$. The algorithm is poor if some ϵ-bad rule happens to fail less often than all acceptable rules. Consider such an ϵ -bad rule, e_i, and let e_t be a correct rule. Let F_i and F_t be their respective failure statistics. By Hoeffding's inequality above, the probability of the first of these is at most,

$$\begin{aligned} LE(\eta + s, m, \eta + s - \frac{s}{2}) &\leq e^{-2(s/2)^2 m} \\ &\leq \frac{\delta}{2N} \end{aligned}$$

Similarly, the latter is

$$GE(\eta, m, \eta + \frac{s}{2}) \leq \frac{\delta}{2N}$$

Thus, the probability that any ϵ-bad rule e_i fails less than e_t is at most $\frac{\delta}{N}$. Since there

are fewer than N bad rules, the probability that one of them minimizes the number of failures by the algorithm is less than δ.

Simon 1993's formalized proofs (Corollary 3.13)

THEOREM 3. *Let C be a class of concepts and* $VC(C) \geq 2$ *its Vapnik-Chveronenkis dimension. Any algorithm that learns C with regard to classification noise rate* η *needs* $\Omega(\frac{VC(C)}{\epsilon(1-2\eta)^2})$ *observations.*

Proof The sample for f is defined as the sample of a p-concept f_η, where $f_\eta(x) = \eta$ if f(x)=0, and $f_\eta(x) = 1 - \eta$ otherwise. Let $C_\eta = \{f_\eta | f \in C\}$. Let $d = d(C)$ and S be the sequence of size d which is shattered by C. S is γ-shattered by C_η for $2\gamma = 1 - 2\eta$. Let A be an algorithm which learns C under classification noise rate η. The domain distribution D can be chosen as in the proof of Theorem 3.1. If A's output $\bar{h}$ is a hypothesis for target concept f whose error is bounded by ϵ, then h is an $(\epsilon, 0)$-good model for f_η on S.

B. Estimated VC dimension for risk bounds

McDonald, Shalizi, and Schervish (2011) propose a simulation-based method to estimate the VC dimension, which measures the generalization capacity of learning algorithms. They prove two main results: first, that the estimated VC dimension will concentrate around the true dimension with high probability, and second, that using the estimated VC dimension allows for the recovery of accurate bounds on generalization error.

Building on Vapnik, Levin, and Le Cun (1994), which shows that the expected maximum deviation between the empirical risks of a classifier on two datasets can be bounded by a function that depends only on the VC dimension of the classifier, the authors provide the following function of n and parametrized by h:

$$VC(F) = \begin{cases} 1 & n < \frac{h}{2} \\ a\frac{log\frac{2n}{h}+1}{\frac{h}{h}-a''}(\sqrt{1+\frac{a'(\frac{n}{h}-a'')}{log\frac{2n}{h}+1}}+1) & \text{else} \end{cases}$$

Following Vapnik, Levin, and Le Cun (1994) ,the constants were chosen as follows: a = 0.16, $a^{'}$ = 1.2 and $a^{''}$ = 0.14927 so that ϕ (.5) = 1. These values are tuned to be optimal for linear discriminant classifiers, for which the VCD is known theoretically. A key assumption is therefore that the same values can be used for the chosen algorithm without introducing bias.

As we have imperfect knowledge, we generate many observations

$$\hat{\xi}(n) = \Phi_h(n) + \epsilon(n)$$

along a fine grid of design points n. Here ϵ is centered on mean zero as the bound is tight, having an unknown distribution with support on [0,1]. We then estimate the true

VC dimension h^* using nonlinear least squares. Of course, generating $\hat{\xi}(n_l)$ is nontrivial. Vapnik, Levin, and Le Cun (1994) provides an algorithm for generating the appropriate observations. At each fixed design point $n_l : l \in \{1, ..., k\}$, we simulate m data points for i = 1,....,m, so as to approximate $\xi(n_l)$ as defined. Vapnik, Levin, and Le Cun (1994) shows that this approach works well in practice, recovering the known VC dimension of linear classifiers and demonstrates that the method for generating the dataset does not affect the algorithm's performance, since for any data structure it is sufficient to flip labels to ensure the most inaccurate possible algorithm is trained.

Below is the procedure for generating $\hat{\xi}(n_l)$, discussed in Vapnik, Levin, and Le Cun (1994), which we apply here.

Algorithm generating $\hat{\xi}(n_l)$

Given a collection of possible classifier F and a grid of design points, repeat the procedure at each design point, n_l, m times

1. Generate a dataset from the same space $y \times X$ as the training sample that is independent of the training sample. The generated set should be of size $2n_l$.
2. Split the data set into two equal sets W and W'
3. Flip the labels (y values) of W'
4. Merge the two sets and train the classifier simultaneously on the entire set: W with the "correct" labels and W' with the "wrong" labels.
5. Calculate the training error of the estimated classifier $\hat{f}$ on W with the correct labels and on W' with the wrong labels.
6. Set $\hat{\xi}(n_l) = |\hat{R}_{n_l}(\hat{f}, W) - \hat{R}_{n_l}(\hat{f}, W')|$.
7. Set $\hat{\xi}(n_l) = \frac{1}{m}\sum_{i=1}^{m} \hat{\xi}(n_l)$

C. Application to Predicting Recidivism (Dressel and Farid 2018)

Dressel and Farid (2018) compares the overall accuracy and bias in human assessment with the algorithmic assessment of COMPAS. The authors hired 20 human coders recruited through Amazon's Mechanical Turk and used seven Features (e.g., age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior crimes, crime degree, and crime charge) for the analysis. We access the dataset used by Bansak (2019) to replicate Dressel and Farid (2018) and implement linear discriminant analysis as in the original paper. The model is trained on a random 80% of training and 20% testing split, with a VC dimension of 8. Note that in this case, we do not need to simulate the VCD dimension since it is known theoretically; however, it is straightforward to verify that our procedure produces the same value with a sufficiently large number of simulations.

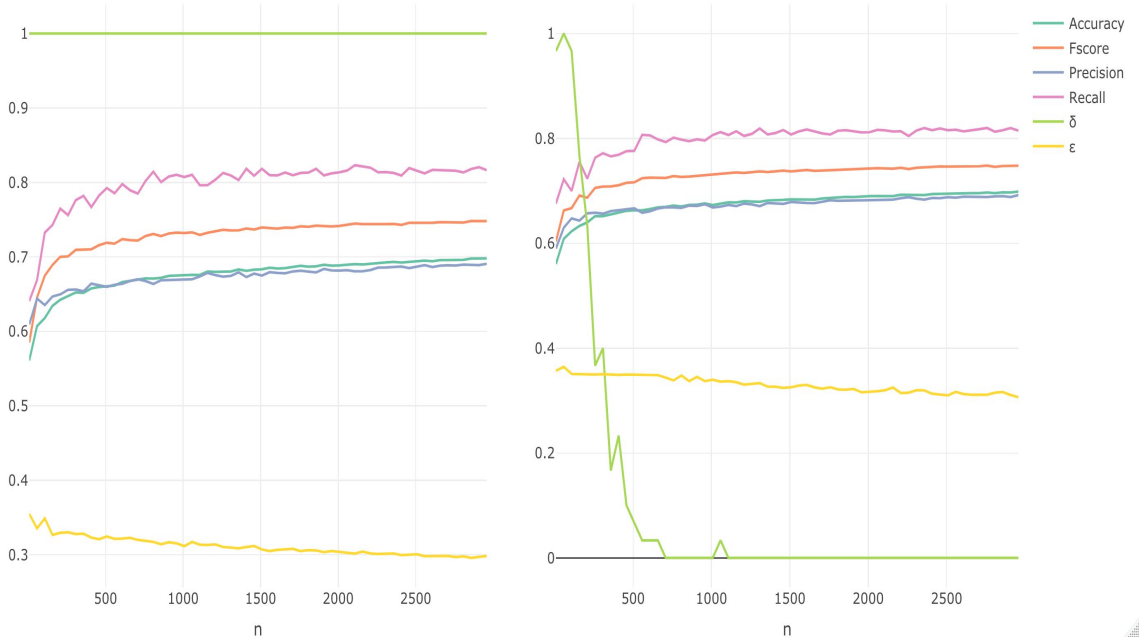

FIGURE C.1. Simulation Analysis When $\epsilon = 0.05$ (Left) & $\epsilon = 0.35$ (Right)

Figure C.1 shows the results of simulations on the observed data using $\delta, \eta = 0.05$

and values of 0.05 (left panel) and 0.35 (right panel) for ϵ. Note that in this case η, or the prevalence of labelling error, essentially corresponds to the false conviction rate since the data are produced directly from the criminal justice system. We use a value of 5%, since available evidence suggests a false conviction rate of between four and six percent (Gross et al. 2014).

The theoretical bound gives a target sample size of 272 observations for these parameters, which is again borne out by the simulation results in the left panel. In practice, however, the results show that the best achievable accuracy with high confidence is approximately 35%, but the additional benefit of sample size above 500 is minimal. This application therefore highlights the central role played by the concept formation problem: the advantages of big data depend on precise specification of the target concept and selection of a corresponding learning algorithm.

As such, this application clearly demonstrates the advantage of the proposed method when used *prior to* data collection. While the original paper employs a dataset of almost 8000 observations – quite costly for many social science applications – Figure C.1 demonstrates that the additional value of this large sample size is negligible given the methods employed. Although an alternative learning algorithm better suited to the target concept may have been able to take advantage of the additional observations, the combination of conceptual mismatch and algorithm choice ensured that most of the sample was essentially wasted. As with power analysis, the calculation of sample complexity bounds prior to undertaking research would prevent such situations from arising, ensuring that scarce research resources are allocated appropriately.

D. R Package: `scR`

The `R` package, `scR`, provides a computationally efficient way of calculating the Sample Complexity Bounds, suggested by Carter and Choi (2024). Please visit the repository:

https://github.com/pjesscarter/scR for more details.