Sample Complexity for Experimental Studies

Examining Its Validity For Open-Ended Survey Responses

Perry Carter & Dahyun Choi & Narrelle Gilchrist
Ph.D. Candidates
Princeton University
Correspondence: dahyunc@princeton.edu

PRINCETON UNIVERSITY

Overview

- **Question**: What constitutes "good enough" data for experiments?
- **Method:** Applying *sample complexity bound* at the design stage in a high-cost research setting
- **Empirical Setting:** Open-ended online survey in Nigeria
- **Goal of Experiment:** Measuring "polarization" inferred from historical narratives among students
- **Contribution:** Demonstrating the validity of *sample complexity* for resource-intensive measurement tasks

Probably Approximately Correct (PAC) Model

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

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

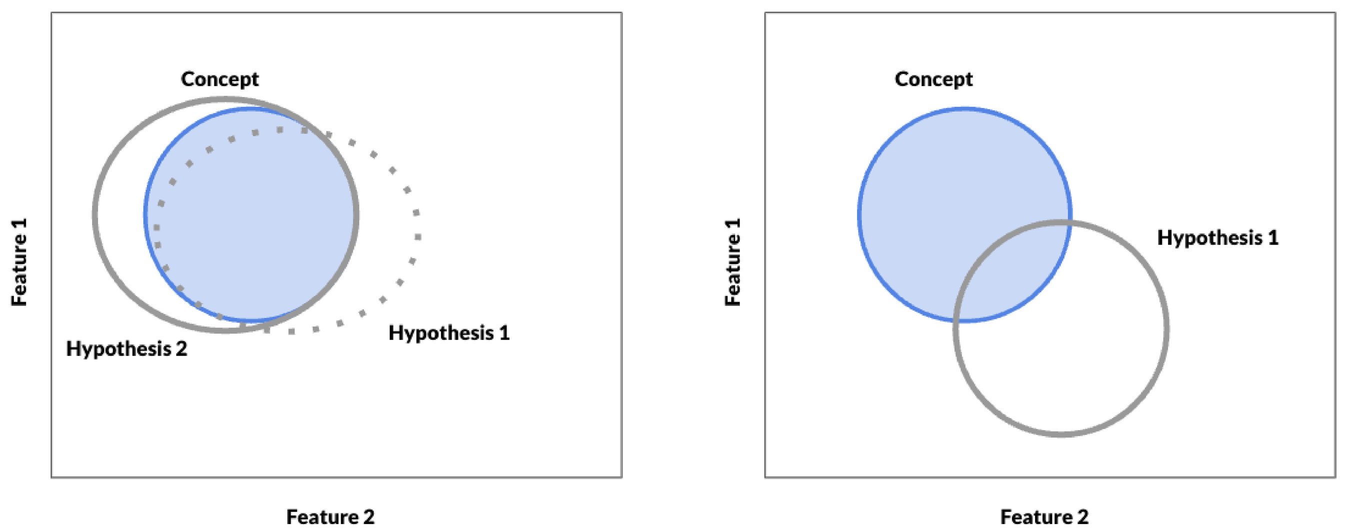
- **Empirical Error:** Given a sample set S, the empirical error of a concept h with respect to S is the fraction of instances in S that are incorrectly labeled by h.

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

- Given a hypothesis class, H, the learner evaluates the risk, $|R(h) - \hat{R}_m(h)|$, of each h in H on the given sample and outputs a member of H that minimizes the empirical risk.

$$\forall h \in H, |R(h) - \hat{R}_m(h)| < \epsilon \tag{3}$$

Figure 1: Schematic Illustration of PAC Learning



- The accuracy parameter ϵ determines how close the output can be to the optimum.
- The confidence parameter δ indicates the likelihood that the classifier will meet the accuracy requirement.
- Researchers seek to achieve $P_A(e > \epsilon) \leq \delta$, for an algorithm A producing hypotheses h with error rate $e = |R(h) - \hat{R}_m(h)|$.

Sample Complexity Bounds (SCB)

- The smallest size necessary to achieve PAC-Learning for all distributions and target concepts, given noisy labels with probability $\eta < 1/2$.
- Combining [7] with [1], a general lower bound on sample complexity (SCB) is given by

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

where $VC(\mathbf{C})$ indicates the Vapnik–Chervonenkis dimension, which measures the underlying complexity of the target concept.

- Calculating VCD analytically is challenging for most concepts [5].
- Solution: estimate empirically based on known relationship between worst-case generalization error and $VC(\mathbf{C}) = d$:

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

Simulation-based Approach of Carter and Choi (2024)

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

Key advantages of Carter and Choi (2024)

- **A more precise alternative to the assumption that the sample size is "large enough" for asymptotic approximations to hold**
- Considering the role played by labeling error and concept definition on model performance, a factor that has generally been overlooked in applied work
- `scR` Package [3] provides computationally efficient way to implement the proposed methods.

Table 1: Comparing Sample Complexity with Power Analysis

	Power Analysis	Sample Complexity
Purpose	Probability of detecting an effect	Expected degree of predictive accuracy
Setting	Distributional features	All distributions & All target concepts
Researcher-specific Parameters	Effect size, significance level, power	Accuracy, confidence, misclassification parameters

Open-ended Online Survey in Nigeria

- In experimental settings, the high cost of data acquisition motivates researchers to use the smallest sample size necessary for reliable statistical inference.
- **Limitation of power analysis:** target sample size on a power analysis does not account for the additional sampling demands of upstream measurement tasks.
- Most readily available datasets have predictable structures, making them a weak test of SCB.
- The design stage in a high-cost research setting, where the impact of misjudging sample size is significant and the sampling distribution is unpredictable
- Estimating the historical narratives about Nigeria's civil war among Nigerian students
- Open-ended responses that have traditionally been considered more difficult to analyze [6]
- Measuring a latent concept (= "Polarization") using topic modeling and random forest

Figure 2: Example Settings for Online Surveys

The survey is conducted online, with hybrid recruitment. Enumerators recruit participants face-to-face in randomly selected secondary schools (stratified by ethnic composition and neighborhood income), across 8 cities (selected based on ethnic composition and level of wartime violence). Once students provide contact information, invitations to participate in the survey will be sent to a random selection of respondents via WhatsApp or email. This method follows the practices of [4], which were highly effective in West Africa.

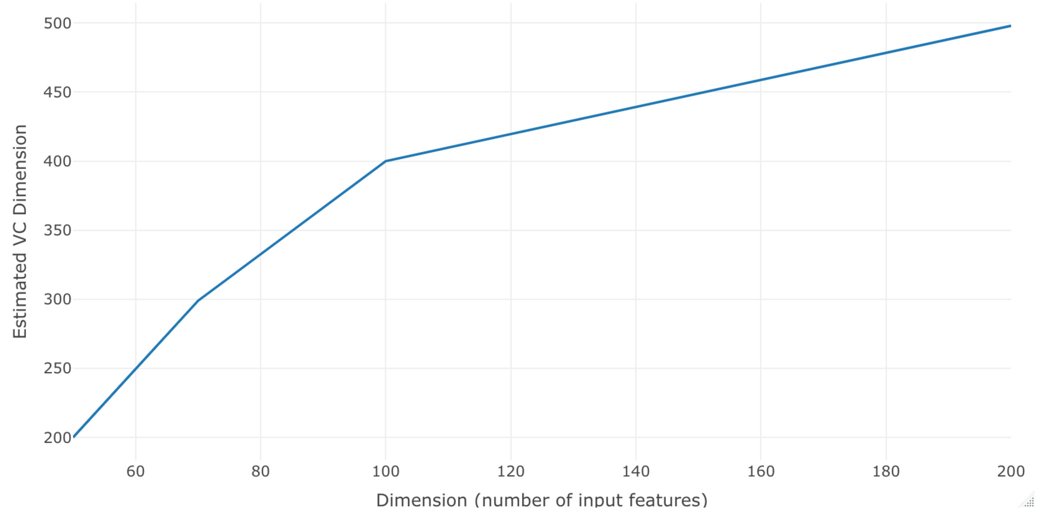
Application of Sample Complexity: Random Forest

- Minimum description length principle: Trading off empirical risk for saving description length
- We define a tree with n nodes, described in n +1 blocks, each of size $log_2(d+3)$ bits.
- We aim to find a tree with both low empirical risk and a number of nodes n not too high.
- While trees of arbitrary size have infinite VC dimension, we can restrict the tree and construct an ensemble of trees [2], thereby reducing the danger of overfitting.

$$R(h) \leq \hat{R}_m(h) + \sqrt{\frac{(n+1)log_2(d+3) + log(2/\delta)}{2m}} \tag{5}$$

- Smallest sample size m that satisfies the condition 5 with a probability of at least $1-\delta$ for every n and every tree h $\in H$ with n nodes.
- Estimating the empirical VC dimension of random forest using the `scR` package [3]

Figure 3: Estimating VC Dimension of random forest



- The SCB under researcher-set parameters with $\epsilon = \delta = \eta = .1$ is 4708 (assuming 100 features).
  $\rightarrow$ The necessary minimum sample size to achieve 90% accuracy achieved with 90% confidence and a noisy rate of 10%
- Actual survey planned for August 19 - September 9
- Following data collection, predicted accuracy will be evaluated against the observed results through cross-fold sample splitting.

Acknowledgement

We would like to thank Brandon Stewart for his discussions at various stages of this project.

References

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

[2] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[3] Perry Carter and Dahyun Choi. Learning from noise: Applying sample complexity for social science research. *Working Paper*, 2024.

[4] Arnim Langer, Bart Meuleman, Abdul-Gafar Tobi Oshodi, and Maarten Schroyens. Can student populations in developing countries be reached by online surveys? the case of the national service scheme survey (n3s) in ghana. *Field Methods*, 29(2):154–170, 2017.

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

[6] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082, 2014.

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