Preserving Validity in Adaptive Data Analysis

Vitaly Feldman

From discovering new particles and clinical studies to election results prediction and credit score evaluation, scientific research and industrial applications rely heavily on statistical data analysis. The goal of statistical data analysis is to enable an analyst to discover the properties of a process or phenomenon by analyzing data samples generated by the process. Fortunately, data samples reflect many properties of the process that generated them: if smoking increases the risk of lung cancer, then we should expect to see a correlation between smoking and lung cancer in samples of medical records. However, data will also exhibit idiosyncrasies that result from the randomness in the process of data sampling and do not say anything about the process that generated them – these idiosyncrasies will disappear if we re-sample new data from the process. Teasing out the true properties of the process from these idiosyncrasies is a notoriously hard and error-prone task. Problems stemming from such errors can be very costly and have contributed to a wider concern about the reproducibility of research findings, most notably in medical research (Ioannidis, 2005).

Statisticians have long established a number of ways to measure the confidence in a result of analysis, most famously p-values and confidence intervals. The guarantees that a confidence interval or p-value provide have a critical caveat however: they apply only if the analysis procedure was chosen without examining the data to which the procedure is applied.

A simple and well-recognized misuse of this guarantee happens when an analyst performs multiple analyses but reports only the most favorable result (for example having the lowest p-value). It is known by many names including the multiple comparisons problem, multiple testing, p-hacking and data dredging. As a result of such cherry-picking, the reported analysis depends on the data, its stated p-value is incorrect and conclusions often invalid. A number of techniques have been developed to address multiple comparisons when the set of analyses to be performed is known before the data are gathered. At the same time the practice of data analysis goes well beyond picking the best outcome from a fixed collection of analyses. Data exploration inspires hypothesis generation; results from one test determine which analyses are performed next; one study on a large corpus determines the next study on the same corpus. In short, data analysis in practice is inherently an *adaptive* process.

While very useful, reusing data in adaptive analysis can greatly increase the risk of spurious discoveries. Adaptive choices in analysis can lead to an exponential growth in the number of procedures that would have been performed had the analyst received different data samples. In other words adapting the analysis to data results in an implicit and potentially very large multiple comparisons problem.

Adaptive data analysis presents a similar challenge in machine learning. An important goal in machine learning is to obtain a predictive model that generalizes well, that is a model whose accuracy on the data is representative of its accuracy on future data generated by the same process. The accuracy of a predictor is usually estimated using a testing (or holdout) part of the given dataset that has not been used for training. If the predictive model is chosen independently of the holdout dataset, such an estimate is a valid estimate of the true prediction accuracy. However, in practice the holdout dataset is rarely used only once. One prominent example in which a holdout set is often adaptively reused is hyper-parameter tuning. Similarly, the holdout set in a machine learning competition, such as the famous ImageNet competition, is typically reused many times adaptively.

Prior literature recognizes the risks and proposes solutions in a number of special cases of adaptive data analysis. Most of them address a single round of adaptivity such as variable selection followed by regression on selected variables. Yet there is no prior work giving a general methodology for addressing the risks of adaptive data reuse over many rounds of adaptivity and without restricting the type of procedures that are performed. We describe such a methodology, based on techniques developed in the context of privacy-preserving data analysis, together with a concrete application we call the *reusable holdout*.

The key technique we use is referred to as *differential privacy* (Dwork et al., 2006) ensures that the probability of observing any outcome from an analysis is "essentially unchanged" by modifying any single dataset element (the probability distribution is over the randomness introduced by the algorithm). The central insight of the differentially private data analysis is that it is possible to learn statistical properties of a dataset while controlling the amount of information revealed about any dataset element. Our approach is based on the same view of the adaptive data reuse problem: the analyst can be prevented from overfitting to the data if the amount of information about the data revealed to the analyst is limited. To ensure that information leakage is limited, the algorithm needs to control the access of the analyst to the data. We show that this view can be made formal by introducing the notion of maximum information between two random variables. This notion allows us to bound the factor by which uncertainty about the dataset is reduced given the output of the algorithm on this dataset.

Our main technical result demonstrates that any analysis that is carried out in a differentially private manner must lead to a conclusion that generalizes to the underlying distribution. This theorem allows us to draw on a rich body of results in differential privacy and to obtain corresponding results for our problem of guaranteeing generalization in adaptive data analysis.

We show how this theory can be used to give algorithms that allow adaptive reuse of the holdout set. In this application the analyst splits the dataset into a training set and a holdout set. The analyst can then perform any analysis on the training dataset, but can only access the holdout set via queries to our reusable holdout algorithm. The reusable holdout algorithm allows the analyst to validate her models and statistics against the holdout set. Crucially, the analyst is allowed to choose statistics that depend on the results of previous queries to the holdout set. We describe several implementations of the reusable holdout and discuss their applications.

This talk is based primarily on the following publications (all authored by Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth).

1. *The reusable holdout: Preserving validity in adaptive data analysis*, Science, 349 (2015), pp. 636–638.

2. *Generalization in adaptive data analysis and holdout reuse*, Proceedings of ACM Symposium on the Theory of Computing (2015), pp. 117-126

3. *Generalization in adaptive data analysis and holdout reuse.* Proceedings of Neural Information Processing Systems, 2015.