Machine Learning Tools for Social Service Providers Funded by the United Way of the Columbia-Willamette

Bruce Marron, Portland State University

Alejandro Queral, United Way of the Columbia-Willamette

May 23, 2016

Advances in computer science, statistics, database design, and learning algorithms have led to a rich selection of new tools for tackling the tough problems of inference in the modern world. Many of these new tools are the direct result of efforts in the subfield of machine learning. Machine learning aims at detecting patterns and regularities in big datasets by using computer programs that can be iteratively optimized (updated) through "training." Machine learning is the process of "programming computers to optimize a performance criterion using example data or past experience"(Alpaydin, 2014). Machine learning also refers to the large suite of powerful computer algorithms for applied data science. Machine learning tools have enjoyed great success in a wide variety of areas including finance (fraud detection, credit evaluation), bioinformatics (facial recognition), medicine (diagnosis, disease risk), signal processing (pattern recognition for speech-to-text processing), language processing (spam detection, handwriting analysis, authorship determination), and general science (outlier detection) (Cruz & Wishart, 2006; Fawcett & Provost, 1996; Alpaydin, 2014). Machine learning tools also have been severely criticized. Precisely because of their immense extractive and predictive power, there are real, ethical concerns about the potential abuse of machine learning tools (Miller, 2015; Datta et al., 2015; Sumner et al., 2012).

This document outlines a pilot project for the development of machine learning tools for use by social service organizations funded by United Way of the Columbia-Willamette (hereafter, United Way). The aim of the pilot project is to extend and diversify the set of tools available to United Way-funded organizations for efficacy assessment of social service provisioning. Specifically, this pilot project proposes to use supervised learning algorithms called Naive Bayes classifiers to assist United Way's partner organizations in identifying combinations of programs and services that can lead to predictable outcomes of success in the *Breaking the Cycle of Childhood Poverty* campaign. As discussed more fully below, the pilot project is committed to the highest levels of ethics and transparency in all aspects of data acquisition, data processing, and inference.

The Vision: Why new tools would be useful

To obtain tangible evidence for the practices and processes that lead to successful outcomes in its *Breaking the Cycle of Childhood Poverty*, United Way must draw valid inferences from many distinct datasets. This is a non-trivial task. Currently United Way is funding over 30 regional organizations all of which collect different types of data (with various metrics and units) often on different populations. These organizations provide a diverse array of services ranging from education supports, mentorship and advocacy in schools, housing and income stability, health promotion and health care, and strengthening of cultural identity of individuals within organizations, to mention a few.

Standardized data collection methods and metrics as well as standardized records disposition into databases would ideally allow for the use of classical survey design and statistical analyses to obtain evidence and make inferences about best-practices. However, until such standardized datasets are available, machine learning classifiers offer a useful and readily-available set of tools for predicting what combinations of practices and processes lead to the best possible outcomes.

The Reality: How machine learning tools work

Machine learning tools are advanced data processing algorithms designed to answer two of the most fundamental questions in any scientific investigation: Are there meaningful patterns in a dataset? If so, is it possible to construct a simple mathematical model from which useful inferences can be drawn? Machine learning is founded on the following logic:

1. There are specific factors that affect any clearly defined, observable outcome. For example, a student's graduation from high school may be the result of factors such as housing stability, access to extra-curricular activities, access to computers and internet supports, etc. Put another way, outcomes and events have discernible, non-random causes. In machine learning, the clearly defined, observable outcomes can be thought of as process outputs and the specific factors can be thought of as process inputs.

2. Data generated by non-random processes contain patterns, even when randomness is introduced. For example, regardless of race, poor nutrition and poor health are detrimental to student success. If we were to examine the graduation outcomes from a large number of cases (students), we would suspect that the factors related to diet and health care are causal; that is, they contribute directly to a successful or an unsuccessful outcome. Likewise, there may be hidden or latent variables that are not directly measurable but nonetheless have a causal relationship to the outcome. For example, the morale and general happiness of a neighborhood may be a latent variable that affects graduation outcomes. The point is that there is a fundamental scientific assumption about non-random data: it is the network of causal chains that produce patterns in non-random process. There are identifiable inputs, often at multiple spatio-temporal scales, that directly affect what we measure as patterns. Put another way, there are historical flows of matter, energy, and information that produce today's events and processes. Some of those historical flows may have happened very recently (loss of a parent) while others may have happened decades ago (segregation). Science, in part, is designed to sort out the random from the non-random. In the current example, race can be categorically measured, but this factor has no causal relationship to the outcome and is therefore a random factor.

3. For many processes, uncertainty remains about (a) how many non-random factors are involved; (b) the mechanisms and causal relationships between the non-random factors; and

(c) the weight of importance for any individual factor or any set of factors. For example, what other factors may affect high school graduation success? Which of these additional factors may be coupled to other factors (i.e., not independent)? Which factors are most important?

4. Computer algorithms can be used to "learn" how process outputs might be generated from process inputs. Machine learning is simply another method for building a mathematical model of a process. For example, Ohm's Law ($I = \frac{V}{R}$) is a model of the process by which electricity flows between two points. This simple model says that if the voltage (V) and the resistance (R) between any two points is known, then the amount of current (I) can be predicted. Typically, mathematical models in the social sciences are complex and may not have such simple, closed form solutions. This is to say that the mathematical relationships between inputs are very complicated and are often unknown. Machine learning tools are like sophisticated code readers: they take seemingly chaotic information (the inputs) and derive the mathematical relationships that underlie the observable, empirical patterns (the outputs).

5. Mathematical models can be used to unravel possible causal links and to make predictions. This has been a fundamental tenet of science for hundreds of years. As science investigates evermore complex systems, mathematical models have gradually shifted from being highly deterministic (like Ohm's Law), to being stochastic or probabilistic (like Naive Bayes classifiers).

In many ways, machine learning sits at the intersection of statistics, computer science, pattern recognition analysis, neural network theory, signal processing, communications theory, and artificial intelligence and has become increasingly popular because datasets have become increasingly large and complicated ("big data"). The application of traditional statistical methodologies to such datasets for the purpose of unraveling causal relationships, extracting process knowledge, and making predictions has become increasingly difficult. Machine learning offers a valid alternative. Machine learning exploits the fact that while we lack knowledge, we have lots of data.

The process of "learning" in machine learning means that the numerical values associated with the structural components of a machine learning tool are updated by new data. Typically, a general class of machine learning tools or models is selected because the class has the capacity to define a unique and usable mathematical model given a certain amount and type of data. Here, capacity means the configuration and type of structural components present in the model. For example, regression models can be considered as machine learning tools. Linear regression models are a class of tools that have the capacity to effectively model linear relationships between multiple inputs and an output. The linear regression coefficients can be learned from the data. However, we would expect that the general class of linear regression models would not do well if the underlying rela-

tionships in the data were non-linear. The capacity of linear regression models is thus limited to linear relationships in the data. Once a class of tools is selected, so-called training data are used to define a specific model. The training data permit the machine learning algorithm to "learn" the values for the model's structural components. The structural component placeholders in a model are instantiated by replacement with real values. This process is called parameterization of the model. For example, if a linear regression model ($y = a_1x_1 + a_2x_2$) has two structural components (a_1x_1, a_2x_2), then parameterization of this model means supplying real values for the regression coefficients (a_1, a_2). A machine learning algorithm updates a model's parameter values based on information in the training data.

The terms, *supervised* and *unsupervised* refer to whether or not the training data includes class label values. Class labels are used to separate objects into mutually exclusive and exhaustive sets. For example, a population of students might be grouped into two classes according to whether or not they have graduated from high school. There is no arbitrariness in this classification (mutually exclusive) and every student will be in one class or the other (exhaustive). Class label values for this example could be +1 for graduated and -1 for not yet graduated. Such a classification scheme represents the outcome of a Bernoulli (binary) random variable. Other nominal classification schemes are of course possible (e.g., student has no siblings, student has 1 - 3 siblings, student has greater than 3 siblings). Supervised learning means that the machine learning tool develops a way to predict the class of a new instance or case by training on data where the cases include the class label values. Unsupervised learning means that the machine learning tool first must decide how the data are to be classified or clustered into useful sets before it "learns" the underlying classification model. No class label values or target values are given in unsupervised datasets.

There are technical challenges in using machine learning tools, as there are in using any of the methods for scientific inference. Most prominent among these types of challenges are:

- *The tool selection challenge* – There are many machine learning tools available: the open-source Python-based machine learning library, *sci-kit learn* lists nine algorithms for unsupervised learning and well over 20 algorithms for supervised learning [1]. Similarly there are numerous machine learning packages for the open-source statistical computing environment, R [2]. Tool selection involves first defining the task to be accomplished (clustering, classification, regression) and then evaluating possible tools based on what data are available, the costs of analysis, and the decision-making end use of the analysis. In this pilot project we will evaluate the use of Naive Bayes, supervised learning algorithms. The reasons for selecting this set of machine learning tools are discussed below.

[1] See http://scikit-learn.org

[2] See https://cran.r-project.org/web/views/MachineLearning.html

- *The feature selection challenge* – Big data are nearly always multivariate. That is, the datasets are large matrices (spreadsheet-like) where the rows are the individual cases or instances of observation and the columns are the results of measurement on individual variables. For example, we might have social service data per student that includes (1) age; (2) sex; (3) zip code; and (4) number of years in a given program. The four variables are termed *features* in machine learning and the feature selection challenge is to decide which features best characterize the cases, given the task. There are certainly good reasons to use all of the data, but because of the so-called "curse of dimensionality" and the idea of parsimony (simplicity), it is often desirable to reduce the number of features in order to remove redundant information.

- *The dimensionality reduction challenge* – As the dimensionality (i.e., the number of features) of a dataset increases, the effectiveness of a machine learning algorithm to perform an optimization routine may be compromised because of data sparseness or an exponential increase in computer time. This can become a real issue when datasets exceed hundreds of features per case which is not at all uncommon! The so-called "curse of dimensionality" is not expected to be a problem in this pilot study.

- *The data quality challenge* – Determining the quality of the data is the ultimate challenge. If the task involves supervised learning, important data quality questions include: Do all of the cases have a label? Are the labels correct (i.e., did the human "oracle" provide the correct label)? Have the data been corrupted by transfers or transcriptions where values have been inadvertently altered? Are the classes balanced in the data (i.e., are there are about the same number of cases per class)? Do the data actually have the necessary features to use machine learning for the task? (An open-ended question that can only be addressed by subject matter experts.) Are there enough data to use a machine learning tool?

The Goals: What we can do now and in the future

With this pilot project we can create a set of first-generation, machine learning classifiers designed specifically for use by social service organizations funded by United Way. We can evaluate these tools to determine if, in fact, they would prove useful to such organizations. That is, this pilot project is designed to investigate whether or not machine learning tools could actually help individual social service organizations assess their own system of social service provisioning.

This pilot project has the long-term goal of providing a repository of machine learning tools for use by the social service organizations funded by United Way. Machine learning algorithms are iterative and not surprisingly, the process of building machine learning tools is likewise iterative. Should the outcomes of this pilot project prove acceptable, we hope to move forward to refine the tools, train organizations in their use, and build a machine learning tool repository.

The Pilot Project

Below we present the pilot project as a sequence of project planning steps, introduce Naive Bayes classifiers, outline the technical approach, and discuss our expectations for the pilot project.

Project Planning Steps

Step 1. Construction of a basic, project plan document that details what is to be done, how it will be done, who will be involved, and who will have access to the results. The project plan document will ensure that ethical safeguards, transparency guarantees, and legal frameworks are in place. The project plan document (a) identifies the individuals and organizations involved as well as their roles and responsibilities; (b) identifies the specific tasks to be performed and the expected outcomes from such tasks; (c) identifies legal provisions and authorizations for data acquisition and use; (d) identifies data processing procedures, data analysis procedures, and the software and hardware used to perform them; (e) identifies acceptable limits of statistical inference and well-defined thresholds for project evaluations; (f) identifies the records management scheme, including records storage, disposition, and access; (g) identifies independent project reviewers; and (h) identifies project reporting requirements.

Step 2. Assignment of reviewers for the project. Reviewers are to ensure that all of the requirements for the project's ethical safeguards, transparency guarantees, and data processing are met. Requirements are stated in the project plan document and include (a) the exclusive use of data purged of all personal identification information; (b) the use of open-source, readily available software; (c) independent review of data processing and analysis; and (d) the public disposition of all findings.

Step 3. Construction and evaluation of a small set of first-generation, Naive Bayes classifier templates designed for social service applications.

Step 4. Public presentation of the pilot project's results to community cohorts.

Once internal reviews are complete this document as well as the project plan document will be available from Alejandro Queral (Director, Systems Planning and Performance at United Way). It is expected that United Way's *SF 2020* partners will provide an equity lens critique and that Metropolitan Family Service will provide the data for building the first-generation, Naive Bayes classifier templates.

Naive Bayes Classifiers

Naive Bayes classifiers are supervised learning algorithms that apply Bayes' theorem with the "naive" assumption of independence between every pair of features. Bayes' Theorem, from probability theory, states that the joint probability of two or more propositions or events, say A and B, can be factored into various conditional probabilities,

$$P(A/B)P(B) = P(A, B) = P(B/A)P(A)$$

The power of Bayes' Theorem lies in its use in drawing scientific inferences from data through the direct application of probability theory as logic (Jaynes & Bretthorst, 2003). Naive Bayes works like this. Say you have many different social service measurements about a student. Additionally, you have some binary outcome data such as the student did or did not pass eighth-grade benchmarks and the student did or did not graduate from high school. As discussed above, the social service data are termed *features* and if listed together would form a *feature vector*. The binary outcome data are termed *targets* and are used to separate students into distinctive groups or classes. For example, the class of those who passed eighth-grade benchmarks AND graduated from high school, the class of those who did not pass eighth-grade benchmarks AND did not graduate from high school, the class of those who passed eighth-grade benchmarks AND did not graduate from high school, and lastly, the class of those who did not pass eighth-grade benchmarks AND did graduate from high school. These are the only possibilities given the example binary outcome data above. If we let y stand for the class variable (which has four possible states) and x_1 through x_n stand for the feature (which is class dependent), then Bayes' Theorem says,

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Namely, there is a probability that the student is in class y given the data $x_1, \ldots, x_n$. This probability ($P(y \mid x_1, \ldots, x_n)$, termed the *posterior*) can be determined if the probability of the class ($P(y)$, termed the *prior*) as well as the *likelihood* ($P(x_1, \ldots x_n \mid y)$) are known. In Naive Bayes, the *prior* and the *likelihood* can be "learned" from the data. To do so, Naive Bayes says that each of the measurements in the feature vector is independent of every other measurement. Clearly this is a big assumption, but it allows the following,

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y \left[P(y) \prod_{i=1}^{n} P(x_i \mid y)\right]$$

Where, ultimately, we can predict the class assignment $\hat{y}$ of a new student as the class with the highest probability ($\arg\max_y$) given the data.

Naive Bayes classifiers can be used with categorical data (e.g., zip codes), continuous data (e.g., time spent in program), or mixtures of both data types. Naive Bayes classifiers are fast, do not require much training data, and can be used as a baseline or first-step classifier. In spite of the overly-simplified assumptions, they perform surprisingly well in classifying complexity in cases such as breast cancer and in identifying spam email (Hand & Yu, 2001; Kotsiantis, 2007).

Technical Approach

The technical approach below outlines how social service data will be processed and how a Naive Bayes classifier will be built. The technical approach is presented as a sequence of numbered action items and for brevity, technical terms are used. The pilot project will use a single dataset provided by Metropolitan Family Service.

1. Identify up to three Bernoulli random variables that can be used to construct project-defined, success classes. Each of these Bernoulli variables must have an unequivocal outcome (yes/no) and must be considered by United Way as an indicator of program success. For example, "passed 3rd grade benchmarks" and "graduated from high school" are possible Bernoulli random variables that could be used to build success classes.

2. Create the project-defined, success classes as n-tuples from the Cartesian product of all Bernoulli variables. For example, if three Bernoulli random variables are to be used, then the possible outcomes (as generated by a Cartesian cross) are $\{000, 001, 010, 011, 100, 101, 110, 111\}$. There are thus eight, project-defined success classes each of which is defined by a unique triple (a 3-tuple).

3. Identify a sufficiently-sized dataset from Metropolitan Family Service with case data (i.e., student data) that includes the Bernoulli random variables of interest. The dataset must have, at a minimum, 20 cases per project-defined success class. For the Cartesian cross example above this would mean a minimum of 8 x 20 = 160 cases. The dataset should have at least 10 features per case. The matrix size of the minimum dataset would be 80 x 10 for four, project defined success classes and 160 x 10 for eight, project-defined success classes.

4. Pre-process the dataset before its delivery to the project. Dataset pre-processing includes: (a) removal of all personal demographic field data (name, sex, age, address, family members, school, religious affiliation, etc.); (b) addition of headers (column names) that define, per feature, the metric and its units; and (c) formatting to a .csv file.

5. Construct a single, Multinomial Naive Bayes classifier for the dataset using k-fold cross-validation for parameter estimates.

6. Evaluate the performance of the classifier using the standard metrics of accuracy, true positive rate, and false positive rate. Classifier performance will be summarized in a confusion matrix.

7. Predict the project-defined success classes for new cases. New cases should be real as well as hypothetical.

8. Explore the effects of features on the project-defined success classes. Exploration should include a sensitivity analysis. Once this is performed, various feature combination effects can be explored.

Expected Outcomes

Two main outcomes are expected from the pilot project. First, that an easily constructed, relatively simple Naive Bayes classifier will prove useful in prediction of success classes given social services data typical of organizations funded by United Way. Second, that the general methodology developed will be portable and applicable to the unique datasets maintained by different social service organizations. Should these expected outcomes be realized, a new project may be initiated to fulfill the long-term goal of providing a United Way-sponsored repository of machine learning tools. This pilot project is expected to be completed by June 2016 with public presentation of results in September 2016.

Final Remarks

Machine learning tools are at the forefront of inferential processing in "big data." As with any statistical tool for inference, care must be taken in constructing the tool and in drawing conclusions from its use. Specifically, this means (a) being explicit in all of the assumptions, components, and processes used to build the tool which is, after all, only a model; (b) being explicit in the model's limitations and range of expected utility; and (c) being explicit in the intentions for the model's end use. Given these safeguards, machine learning tools may prove of great use to social service organizations in helping them fine-tune their systems to deliver more effective programs and services.

References

Alpaydin, E. (2014). *Introduction to machine learning* (Third edition ed.). Cambridge, Massachusetts: The MIT Press.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*. Retrieved 2016-04-15, from `http://search.proquest.com.proxy.lib.pdx.edu/openview/2d3564860fc6664a5f40ec196594c0f5/1?pq-origsite=gscholar`

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on Ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112.

Fawcett, T., & Provost, F. J. (1996). Combining Data Mining and Machine Learning for Effective User Profiling. In *KDD* (pp. 8–13). Retrieved 2016-04-13, from `http://www.aaai.org.proxy.lib.pdx.edu/Papers/KDD/1996/KDD96-002.pdf`

Hand, D. J., & Yu, K. (2001). Idiot's Bayesnot so stupid after all? *International statistical review*, *69*(3), 385–398.

Jaynes, E. T., & Bretthorst, G. L. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK ; New York, NY: Cambridge University Press.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. Retrieved 2016-04-15, from `http://ieee.scripts.mit.edu.proxy.lib.pdx.edu/urgewiki/images/5/52/Dinner1-supervisedlearning.pdf`

Miller, C. C. (2015, July). When Algorithms Discriminate. *The New York Times*. Retrieved 2016-04-30, from `http://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html`

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012, December). Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. In *2012 11th International Conference on Machine Learning and Applications (ICMLA)* (Vol. 2, pp. 386–393). doi: 10.1109/ICMLA.2012.218