Section 1: Week 3: Applying Machine Learning

Nate Bachmeier

TIM-8130: Data Mining

February 22nd, 2020

North Central University

Applying Machine Learning

# Section II: Application of Machine Learning to Organizations

Many organizations are collecting vast quantities of data with the hopes of unlocking the potential by transforming it into business intelligence. A common misconception arises when these businesses assume that adding machine learning will magically create insights. In reality, data cleaning and curation accounts for roughly 70% of the effort (Gibert, Sanchez-Marre, & Izquierdo, 2016) and meta-learning analysis accounting for the remainder (Tripathy & Panda, 2017). The No Free Lunch Theorem explains the disconnect by stating that no universal algorithm exists that is equally efficient for all scenarios. That statement makes sense, given that a machine learning algorithm produces a scenario-specific statistical model. For instance, determining association rules requires Apriori-like algorithms that focus on groupings of frequent itemsets. Both the approach and end goal are significantly different than logistical regression, which maps features to a continuous range. Even similar objective strategies, such as classification versus clustering, are dissimilar because of the unique mapping characteristics between labeled and unlabeled buckets. These distinctions suggest that a successful application of machine learning requires (1) a concise business question, (2) relevant facts about the issue, and (3) an algorithm that transforms the evidence into answers.

Having a specific business question is the first and most critical step in the process. Without a clear objective, the project is unlikely to be successful. Consider the impact of attempting to discover cures to heart diseases with banana cells. The engineering team is immediately at a disadvantage against a competitor that is using heart cells. As the fruit cell data set continues refinement, these facts remain irrelevant to their goal. No matter the algorithm, it is improbable that a revelation arises. However, if the specific question asks for similarities between fruit cell and heart cell decay, then these facts are now relevant. Other examples are less black and white, such as predicting website click-streams or product recommendations. These scenarios can be more nebulous, and only through a refinement of the target objective does it become clear whether evidence is significant.

After determining that a collection of facts address the business question, the next step is to curate them into an efficient format for analysis. Typically the evidence will come from multiple sources such as customer reviews, order histories, search histories, and third party providers. The heterogeneous sources use different identifiers, categorical labels, and value scales (e.g., inches versus centimeters) that need to become standardized before processing the results. Along with standardization, the data sets need to be joined, aggregated, deduplicated, and filtered for erroneous examples. When these actions do not take place, then the statistical predictions will be skewed and contain bias. Some of these incorrect values are easy to detect, such as a value that is several deviations from the mean. Other scenarios, like miscalibration, are more difficult to detect because they are still within statistical norms. As the volume of data increases, so will the frequencies of these challenges. While it is economically prohibitive to detect and correct every instance, a best-effort needs to occur to avoid a “garbage-in/garbage-out” scenario. Though at the same time, it is critical to understand why the removal of examples from the training sets is necessary. For instance, a list of security transactions could enumerate the date that a customer purchased and sold the stock. If an empty sales date represents the asset is still held, then blinding dropping rows with any empty column would discard the entire user-scenario. Depending on the parameters of the question, this invalid action could result in a model that provides excessive confidence through overfitting and does not make reliable predictions in practice.