Section 2: Week 4: Techniques to Discover Logic

Nate Bachmeier

TIM-8130: Data Mining

March 1, 2020

North Central University

Techniques to Discover Logic

# Section I: Determining Relevance

## Identify Business Value

It is not possible to answer a question if either the question or the necessary facts are not known. Consider the scenario where the organization wants to execute the most efficient marketing campaign using the least amount of resources. Without proper planning, the business might stumble upon an acceptable deliverable (local maxima) though they are unlikely to encounter the global maxima. If instead, the company specifically defined the objective as increase awareness of their product to minorities and underserved rural populations, then it becomes possible to rate the quality of supporting evidence. Now that a logical base case exists, the company can review public and private data providers and perform an initial inclusion filter. For instance, governmental census information contains population statistics that describe high-value segments to place physical advertisements. Though even within this vast dataset, only a subset will be useful today or even tomorrow. Perhaps the organization has a strong presence in the southwest, and the business model does not support expanding into the northeast (e.g., licensing, political, or transportation concerns). These limitations remove the need to have either humans or machines mine those areas (partitions), as there is no potential value.

## Identify Completeness

After repeating this filtration process multiple times, the conversation can transition the focus to the shape and volume of these facts. If the analyst is attempting to build a statistical model across five thousand features, they will need a lot more data than an alternative that contains five features. Wall and Toit (2011) suggest that a minimum of ten examples need to exist for every parameter. That logically makes sense as a deficit of information leads to speculation and bias. Imagine asking two random people at an NYC Metro bus stop their income, and then predicting the median for the region. Alternatively, a person that samples from different neighborhoods along the Metro is more likely to estimate this value accurately. Hsu et al. (2017) caution that the absence of evidence is not evidence of absence, and that under-generalizations frequently occur. For instance, if the collection of these samples takes place during the afternoon, it is likely to miss high earning professionals that commute during the mornings and evenings. Data miners can detect some of these gaps by looking at descriptive statistical and broad aggregates. Consider how a pivot against time or industry (e.g., financial, technology, hospitality, etc.) would highlight missing examples from the metro scenario. When the shape of these pivots does not align with expectations, then there is a problem.

## Identify Shape

Understanding the business-value and determining a complete collection of data enables the organization to begin testing their hypothesis. Snee (2015) directs research towards a graphical investigation first. For instance, plotting a simple line chart of market segment growth over time makes specific cyclical trends visually discoverable, versus a detailed analysis of rows in a table. Another critical strength of graphically approaching the data’s shape comes with the relative ease of explaining rational to other parts of the institution. Not all team members are ‘deeply in the trenches’ and need some mechanism to connect the discoveries with their external perspectives. Consider the distinction between presenting *only* a polynomial equation to predict housing prices versus the inclusion of a scatter plot and estimated curve. Aside from data visualizations, the input shape is an important characteristic of mining algorithms. If the algorithm learns on individual instances, then aggregate counts are not useful regardless of its relevance to the problem. Similar requirements are present concerning time components and multi-dimensional data (e.g., images). These constraints can force the researchers to search for more supporting evidence or alternative expressions of the problem. However, specific scenarios are inherently spare, or it’s prohibitively expensive to acquire more data, such as collecting project bids from contractors. Instead of mapping these bids to precise quantitative estimates through regression (e.g., $152,593.22)—the company could use a qualitative bucketing strategy (e.g., cheap versus expensive). While the qualitative approaches are less specific, they provide a mechanism to reduce the search space further and can ensemble with other algorithms to form a unified collection of signals.

# Section II: Compare Predictive Techniques