Section 3: Week 8: Data Mining IoT

Nate Bachmeier

TIM-8130: Data Mining

March 28, 2020

North Central University

Data Mining IoT

With ubiquitous access to high-bandwidth wireless networking and cloud computing, physical devices are evolving business capabilities to both monitor and react to changes across their supply chains and operational footprint. These devices produce enormous volumes of unstructured telemetry that require curation processes to transform raw data into business intelligence, enabling data-driven decisions that make the organization more competitive. During this transformation, data mining strategies extract patterns and statistical inferences through regression, clustering, classification, and rule association algorithms (Barua & Mondal, 2019). Each of these algorithmic categories has distinct objectives for scenario-specific applications. Consider the difference between asking (1) how much a customer will spend versus (2) which customers are most alike. Point of sales (PoS) records can answer either question, though the question structure creates a dependency on which sales information becomes relevant facts into (1) regression versus (2) clustering solutions. Using the wrong approach or not defining the outcome goals upfront always leads to nonsensical results (Snee, 2015). Instead, a formal data mining lifecycle begins with a specific question, then collects relevant facts to derive a conclusion. Next, an evaluation method confirms these conclusions are scientifically sound and not wishful thinking through some statistical variance or cross-validation testing. After constructing a probabilistic model of the scenario, the company needs to deploy it into its production environment and begin collecting a return on investment (ROI). Measuring the amount of return depends on Key Performance Indicators (KPI) that typically align with high-level corporate mandates, such as increasing sales per customer or reducing inventory carry times. Despite alignment challenges across data producers, business questions, relevant facts, conclusions, operationalization, and KPIs—planning and methodical approaches lead to success.

# Section I: Business Structure

## Case Study

Black Bean started as a small ‘mom and pops restaurant,’ and over the last several years, expanded operations to include hundreds of international locations. The company prides itself on delivering consistently high-quality fruit desserts, regionally sourced from local farms. This local approach to supply chain management allows each site to reduce shipping times and ensure the freshest produce. However, decentralization increases the complexity of the corporate offices to gain insights into individual sites and confirm that it meets health and safety standards. The senior leadership addressed these issues by first defining expectation guidelines for middle management. Initially, supervisors would manually collect metrics across the store though this approach was tedious and full of errors. For instance, when the restaurant becomes crowded, the team wholly dedicates to the customer, causing missed readings. Instead, a transition towards automated solutions with IoT devices provides an economical approach to get this data more consistently and reliably. Technologies, such as Arduino and Raspberry PI, allow users to connect specialized sensors to wireless networks for under ten dollars a unit. This capability opens the door for smart restaurants to measure virtually unlimited aspects of the site, such as refrigeration temperatures, humidity sensors, customer counts, and power consumption. After realizing the success of IoT within the eatery, the leadership team began to ask, what about the supply chain? Using Radio Frequency Identifiers (RFID) and Global Positioning System (GPS) sensors allow shipment tracking with fine-grained granularity. Consider scanning an RFID tag on a fruit container as it passes from the farmer, the distribution center, the receiving dock, and finally, the chef—creating an analyzable lifecycle. However, the lifecycle of the desserts does not end here. It continues with the customer scanning their mobile device, creating a point of sales record, and discussing the purchase on social media.

## What types of data artifacts exist

Black Bean relies on a central data lake hosted in the public cloud to store all raw results from the logistical systems, restaurants, mobile apps, and social media impressions. These unstructured artifacts need to go through several iterations of curation as they promote through the corporate data catalog. For instance, the organization manages thousands of temperature sensors that were manufactured by dozens of providers, each with minor differences to the telemetry schema, such as property names, data type encoding, and units (Fahrenheit versus Celsius). After normalizing the sensor values, they can start to connect into semantic models that describe each aspect of the business. These aspects extend beyond inventory management and include scenarios, such as detecting training gaps and safety concerns. Perhaps an employee frequently leaves the meat slicer running unattended or the freezer door ajar—the telemetry contains sufficient information to derive these states. With the capability to predict state comes the ability to remediate, such as killing the meat slicer’s power. The predictive capabilities become even more advanced as IoT devices include video recording, and rely on deep learning to discover domain-specific actions. Consider the benefits of being able to analyze every customer’s cashier interaction, and then evangelize the more successful traits.

## What business goals use these artifacts

The first and most critical step in any data mining exercise is to determine the question and then discover supporting evidence. Until this action occurs, the business is unlikely to have a successful deliverable and will spend excessive resources investigating irrelevant materials. After clearly articulating the business value, the engineer teams can perform broad filtration of data sources based on their ability to address those questions. During filtration, having a logical framework can improve the search process through partition pruning of the relevant data stores. For instance, if the business operates in Michigan, there is minimal value in exploring Texas-specific data. 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). However, they are unlikely to encounter the global maxima. If instead, the company explicitly 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 target physical advertisements.

Businesses with perishable supply chains hold risk in their carrying inventory. If there is too much inventory, then capital is unavailable for other activities, and too little will miss sales opportunities (Viale, 1996). The organization needs to define this optimization objective Specific, Measurable, Attainable, Realistic, and Timely (SMART). For instance, using the inventory tracking and point of sales information can funnel into a forecasting model that predicts how many apples the business will need for the next two weeks. As the precision of this model increases the amount of excess carry can decrease, and improve sales margins. Social media intelligence is a crystal ball into customer preferences, and provides should be a goal of all businesses (Gioti, Ponis, & Panayiotou, 2018). Black Bean could look to expand its loyalty rewards members through Facebook or Twitter advertising.

# Section II: Collecting and Enhancement

## Preprocessing Data Resources

Countless failed data mining experiments begin with the assumption that pairing large quantities of data with advanced statistical algorithms produces quality insights. The reason this approach does not work is that the algorithms can only apply statistical inference based on the provided data, and cannot derive domain-specific rules. For instance, a missing value for a temperature sensor might indicate a network failure, versus a smart outlet means the associated device is disabled (Rawal et al., 2017). Depending on the prediction scenario, these missing values could be the most or least important detail, since the algorithm cannot magically discover these internal rules, training data must account for these situations.

Before analysis can begin, the data analyst needs to normalize the incoming data through an extract-transform-load (ETL) process. This process needs to perform column renaming and reordering, adjusting quantity units, filtering erroneous values, populating missing values, and similar cleanup actions. Many analysts reportedly spend nearly 70% of the experimentation time cleaning and preparing data (Gibert, Sanchez-Marre, & Izquierdo, 2016). Another common challenge comes from incoming data arriving in different shapes and sizes, and normalization processes need to standardize those values. For instance, inventory reports can contain units in both pounds and kilograms, making direct comparisons impossible, and other numerical values might have entirely different range scales that require nonparametric conversions. After normalization, there can be multiple rounds of using descriptive statistics to discover anomalies and unexpected outliers. When analysis does not handle these aspects upfront, it creates garbage-in/garbage-out scenarios. These scenarios can cascade of problems, such as (1) an erroneous manual entry of 555 degrees instead of 55 degrees, (2) becomes part of a feature mean calculation, (3) that becomes the default for missing values, and (4) skews training that relies on this feature. During this cleaning process, it is critical to record the applied actions so that the results are reproducible (Zambetti, Pinto, & Pezzotta, 2019). One solution is to have automation scripts that register each transformation with a data catalog and archive into the data lake.

* After cleaning and schematizing the incoming data, the next analyst needs to determine which aspects are relevant for their data mining objective. Having large amounts of unrelated information does not improve results, and for many scenarios, it only slows down model training times.
* The cleaned data set might need additional enhancements by combining across related information. For instance, the marketing team can use Point Of Sales + Mobile App + seasonal trends to create targeted marketing campaigns

## Required Collection Resources

Tracking inventory across the supply chain has several integration points that blend between cyber and physical systems. Suppliers need to place RFID passive tags on each container and then scan them for the invoice statements. Trucks load these containers and report their progress through GPS technologies en route to the distribution warehouses. Warehouse employees move the containers into refrigeration units that report both atmospheric conditions and asset tracking within the distribution center. When those items are ready for restaurant consumption, warehouse employees record the specific containers that become shipped. On the receiving end, managers and staff need to unload the ingredients and begin local accounting of health and safety aspects (e.g., temperature controls). Orchestrating this journey requires commodity hardware (e.g., WiFi routers), specialized systems (e.g., IoT sensors), and manual data entry (e.g., pen and paper). Automated steps that rely on hardware are inherently more accurate reliable, as there is a greater chance of the action occurring. For instance, as inventory moves between rooms, RFID technology can automatically update the Enterprise Resource Management (ERM) systems, thus reducing the chances that containers become lost (Zhang et al., 2018). However, having vast quantities of dark data does not provide any value, as new insights are not forming. Instead, cloud analytical solutions need to create rules that discover the container is in the wrong section of the warehouse. Croson and Donohue (2003) propose that these insights need to be bidirectional, and sharing forecast predictions with suppliers can reduce the bullwhip effect. Supporting this symbiotic relationship requires collection resources also to publish aggregate partner feeds.

Resource collection expands beyond the inventory supply chain and into the restaurant, where IoT technologies can improve numerous aspects of the customer experience. Surveillance technologies can detect and measure how long customers hang out and table cleaning cadences. Fong et al. (2016) describe a system for mining custom gestures from video feeds to determine what actions are occurring. Using a similar technology would allow safety systems to detect erroneous behavior and operate as a circuit breaker. For example, if an employee is intoxicated, they should not be permitted to operate heavy machinery, and detecting this state could prevent an accident. These futuristic scenarios are becoming mainstream because of the ubiquitous access to public cloud platforms, such as Amazon SageMaker and Azure Cognitive Services, with support for deep video learning.

This paradigm shift towards video-centric technologies provides higher contextualized data, but also requires substantially more bandwidth. It will become prohibitively expensive to upload these video streams, and that creates a need for edge computing that performs an initial curation and filtration process (Basanta-Val et al., 2017). For instance, a local stream processor might classify simple actions within the feed justifies either higher or lower sampling rates. There will also be specific behaviors that need to stay onsite for auditory or compliance purposes.

## What logical components or assumptions exist

The semantic model represents a logical component of the distribution, food preparation, or customer satisfaction systems. Each model instance represents a state in time snapshot of various interconnected sensors that collectively describe the broader state. For instance, the state of an individual restaurant site might record hourly the number of guests and workers, weather conditions, inventory levels, storage temperatures, and inner atmospheric levels. Independently these data points can only tell part of the story, versus their correlated aggregate structure communicates the detail. Consider knowing that the number of guests was low for a particular hour, and asking is that a concern? If the worker count is also low or there is a hurricane, then perhaps it is expected. Are some of the guests likely to get food poison? If the model knows how many employees are on sick and the storage temperature variability, then a calculatable probability exists. Similarly, the inventory lifecycle is a temporal model where each checkpoint represents a point in time. At each checkpoint, collecting per container features, such as contents, GPS coordinates, time stamps, and storage conditions, allows the business to predict the decay of that produce at higher precision.

Many assumptions come into these models, such they are complete and not malicious. If an employee was negligent with the storage of inventory, they might seek to cover their tracks to avoid retribution from management (e.g., reporting erroneous values). Concept drift occurs when the features of the model no longer align with predictions. This scenario could happen as preparation items move between stations, and relabeling does not occur, causing telemetry signals with the meat-slicer reporting as the toaster. For specific prediction solutions, this inaccurate state could prevent detection and safety alerting, among other undesirable behaviors. Partial mitigation of these issues can occur during the data cleaning phase by including a curation phase to classify and predict the telemetry source. Imagine a smart outlet is reporting the power consumption—a mixer and toaster will have different time graph distributions and thus different signatures (e.g., Jenson-Shannon Divergence).

# Section III: Evaluation Procedures

## What statistical techniques can measure process ROI

Successful projects need to expand on business goals and provide measurable value towards the organizational mission. Mature firms can explain their mission in terms of high-level objectives and Key Performance Indicators (KPI) that convey the aggregate performance of that aspect (Gonzales & Wareham, 2019). For instance, Black Bean has a focus on increasing repeat customer traffic, so they track through the loyalty rewards programs and PoS return rates. After launching a marketing campaign, this metric helps to evaluate its performance, as it communicates the impact and value. Snee (2015) warns about multicollinearity and how given enough random variables, some will naturally correlate. External elements, like diet fads and economic recessions, can also dampen the customer’s appetite for discretionary desserts, irrespective of the marketing campaign. Liyanage et al. (2018) propose that KPIs should be decomposable with child-KPIs acting as supporting evidence. Unlike the external factors, monitoring order latencies, recommendation accuracies, and similar interactions with the customer are within the control of the site. When the internal and broad metrics are decaying in unison, it confirms the correlation exists and likely related to internal quality control. Outside correlation, variance analysis can often describe the likelihood of a problem existing. Assume serving a typical order takes five minutes with a couple of minutes variability. Understanding these norms and their standard distribution provides a numeric score to the extent an observation is an outlier and requires further investigation. Third, cross-validation with test data can also measure the accuracy of a system by assessing its performance against real data. If a prediction system can make reliable forecasts towards a business objective, then it has the potential to generate value.

# Section IV: Future Applications

## What data mining techniques can be applied

The four major categories of data mining are regression, classification, association rules, and clustering techniques (Barua & Mondal, 2019). Association rules are patterns that take the form of ‘if X then Y,’ such as a person that buys bread is likely also to purchase butter. Clustering and classification are related strategies that both attempt to group similar items into buckets. The critical difference is that classification knows the bucket labels ahead of time (supervised) while clustering does not (unsupervised). Regression modeling tries to find a mathematical equation that explains the observations. Numerous implementation-specific permutations of these categories exist though they all follow the same general structure. Forecasting expected guests on a future date through Gradient Decent Regression (GDR) or MultiLayer Perceptron Regression (MLPR), both take feature sets and attempt to map them those values to a continuous numeric output range. Even distinctly different solutions such as Decision Tree Regression still support this consistent input and output format. This consistency enables researchers to follow general scenario recipes to produce scenario-specific results, provided sufficient preparation occurs to represent the data domain entirely. These same generalities apply to IoT metric streams, though a few additional curation actions might need to necessary. Harper (2019) warns of predictive analytics on high-volume/low- quality sources as they tend to be noisy and more costly to compute. Instead, stream processing at the edge can perform aggregations and report moving averages and related statistics. After transforming into low-volume/high-quality data points, emit those metrics to central processing systems. When dealing with time-series data, algorithms like Long Short Term Memory (LSTM), assume ordering and relationships between sequences of examples (Talei & Benhaddou, 2018). This characteristic allows the optimizer to consider recent values as features into future predictions. Consider the contrived sequence of (100,200,300, blank) and how your mind assumes the next value is 400, very similar to the trick LSTM exploits.

* Ensemble methods like Netflix (Bell, Koren, & Volinsky, 2007)

# Conclusions

1. Black Bean is a growing multinational organization
   * They have embraced the need for IoT to automate many internal processes
   * Health and Safety, Customer Satisfaction, Inventory Management
   * Collecting for analysis these data points is complex due to the unstructured nature and heterogeneous formats.
2. The organization needs to identify specific KPIs of interest and then determine which relevant facts support their hypothesis. These facts will need to be cleaned, this is 70% of the total data mining. Garbage in/out.
   * After cleaning the data needs to be segmented and structured in such a way that it supports model training. The model produced needs to be scientificially sound and explainable, ideally through a graphical process.
   * Collecting and recording these figures requires numerous personal and hardware resources. Care needs to take place that it does not become corrupt through errorous data entry and similar situations