Lecture 9

Data Preparation

Organizations today continue to look for ways to prepare data quickly and more accurately to solve their data challenges and enable machine learning (ML). But before bringing your data into a machine learning model or any other analytics project, it's important to ensure that it is clean, consistent, and accurate. Because much of today's analytics is dependent on the context of the data, the task is best done by those closest to what the data actually represents; the business domain expert who can apply hunches, theories, and business knowledge to the data.

Unfortunately, business users don't usually come equipped with data science skills so bridging that gap can make the difference between gaining value from your data quickly. As result, many are applying data preparation (DP) to help data scientists and ML practitioners rapidly prepare and annotate their enterprise data to extend the value of the data across the enterprise for analytic workloads.

How data collection and preparation are the foundation for trusted ML models

To create a successful machine learning model, it is imperative that an organization has the ability to train, test, and validate them prior to deploying into production. Data preparation technology is being used to create the clean and annotated foundation needed for today's modern machine learning yet good DP historically takes more time than any other part of the machine learning process.
Reducing the time necessary for data preparation has become increasingly important, as it leaves more time to test, tune, and optimize models to create greater value. To prepare data for both analytics and machine learning initiatives teams can accelerate machine learning and data science projects to deliver an immersive business consumer experience that accelerates and automates the data-to-insight pipeline by following six critical steps:

Step 1: Data collection

This is the by far the essential first step as it addresses common challenges, including:

- Automatically determining relevant attributes in a data string stored in a .csv (comma-separated) file
- Parsing highly-nested data structures such as those from XML or JSON files into a tabular form, for easier scanning and pattern detection.
- Searching and identifying relevant data from external repositories.

However, when considering a DP solution, make sure it can combine multiple files into one input, such as when you have a collection of files representing daily transactions, but your machine learning model needs to ingest a year of data. Also, be sure to have a contingency plan in place for overcoming problems associated with sampling and bias in your data set and your machine learning model.

Step 2: Data Exploration and Profiling

Once the data is collected, it's time to assess the condition of it, including looking for trends, outliers (extreme values that fall a long way outside of the other observations), exceptions, incorrect, inconsistent, missing, or skewed information. This is important because your source data will inform all of your model's findings, so it is critical to be sure it does not contain unseen biases. For example, if you are looking at customer behavior nationally, but only pulling in data from a limited sample, you might miss important geographic regions. This is the time to catch any issues that could incorrectly skew your model's findings, on the entire data set, and not just on partial or sample data sets.

Step 3: Formatting data to make it consistent

The next step in great data preparation is to ensure your data is formatted in a way that best fits your machine learning model. If you are aggregating data from different sources, or if your data set has been manually updated by more than one stakeholder, you'll likely discover anomalies in how the data is formatted (e.g. USD5.50 versus $5.50). In the same way, standardizing values in a column, e.g. State names that could be spelled out or abbreviated) will ensure that your data will aggregate correctly. Consistent data formatting takes away these errors so that the entire data set uses the same input formatting protocols.

Step 4: Improving data quality

Here, start by having a strategy for dealing with erroneous data, missing values, extreme values, and outliers in your data. Self-service data preparation tools can help if they have intelligent facilities built in to help match data attributes from disparate datasets to combine them intelligently. For instance, if you have columns for FIRST NAME and LAST NAME in one dataset and another dataset has a column called CUSTOMER that seem to hold a FIRST and LAST NAME combined, intelligent algorithms should be able to determine a way to match these and join the datasets to get a singular view of the customer.

For continuous variables, make sure to use histograms to review the distribution of your data and reduce the skewness. Be sure to examine records outside an accepted range of value. This "outlier" could be an inputting error, or it could be a real and meaningful result that could inform future events as duplicate or similar values could carry the same information and should be eliminated. Similarly, take care before automatically deleting all records with a missing value, as too many deletions could skew your data set to no longer reflect real-world situations.

Step 5: Feature engineering

This step involves the art and science of transforming raw data into features that better represent a pattern to the learning algorithms. For example, data can be decomposed into multiple parts to

capture more specific relationships, such as analyzing sales performance by the day of the week, not only the month or year. In this situation, segregating the day as a separate categorical value from the date (e.g. “Mon; 06.19.2017”) may provide the algorithm with more relevant information.

Step 6: Splitting data into training and evaluation sets

The final step is to split your data into two sets; one for training your algorithm, and another for evaluation purposes. Be sure to select non-overlapping subsets of your data for the training and evaluation sets in order to ensure proper testing. Invest in tools that provide versioning and cataloging of your original source as well as your prepared data for input to machine learning algorithms, and the lineage between them. This way, you can trace the outcome of your predictions back to the input data to refine and optimize your models over time.