DLI Accelerated Data Science Teaching Kit

Lecture 18.2 - Data Preprocessing

Data Preprocessing

Data preparation is an essential step to build a machine learning model.

To enhance the model performance, it requires transforming raw data and even select relevant attributes by data processing.

Two reasons for performing data preprocessing before training a machine learning model:

- To improve fitting of learning algorithms on the data;
- To extract or keep only the most relevant data.

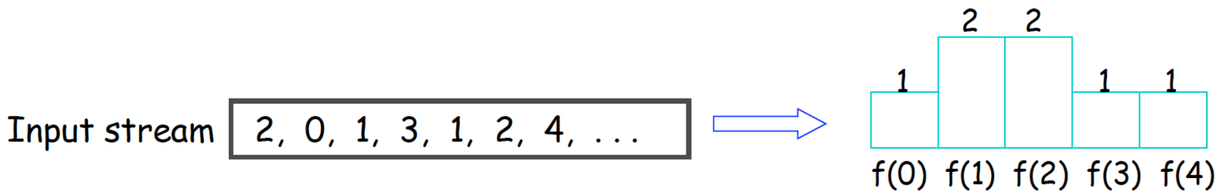
Basic preprocessing methods like normalization will be more complicated in a streaming setting since statistics about the data will be an unknown priori, e.g., the minimum and maximum values of a given feature/attribute.

Feature Transformation

Summarization Sketches

- A ‘sketch’ of the data refers to the summary to avoid storing and maintaining a large amount of data.
- Sketches are probabilistic data structures that summarize streams of data, such that any two sketches of individual streams can be combined into the sketch of the combined stream in a space-efficient way.
- Using sketches requires a number of compromises. For example, sketches must be created by using a constrained amount of memory and processing time. If the sketches are not accurate, then the information derived from them may be misleading.

- COUNT SKETCH ALGORITHM (Charikar et al. 2004)
- Goal
  - k most frequent elements in a stream (for large number N of distinct values)
  - Ex. 100 most frequent IP addresses going through a router



Source: https://www.slideshare.net/rchiky/dds-workshop05122014

Feature Transformation

Feature Scaling

- Feature scaling is to transform the features domain in a way that they are on a similar scale.
- Generally, scaling refers to normalizing, i.e., transform features such that their mean $\bar{x}$ = 0 and standard deviation σ = 1.

$$\textbf{Mean } (\bar{x}) = \sum [x_i * P(x_i)]$$

$$\textbf{Standard Deviation } (\sigma) = \sqrt{\sum (x_i - \bar{x})^2 * P(x_i)}$$

- The two most popular approaches consist of i) centralizing the data by subtracting the mean and diving by the standard deviation; or ii) dividing each value by the range (max - min).
- It is unfeasible to perform feature scale for data streams as aggregate calculations must be estimated throughout the execution.

Source: https://www.educba.com/probability-distribution-formula/

Feature Transformation

Feature Discretization

- It is to divide numeric features into categorical ones using intervals.
- Depending on the application and predictive model, discretization can bring several benefits
  - Speeds up computation time as discrete variables are usually easier to handle compared to numeric ones;
  - Decreases the chances of overfitting since the feature space becomes less complex.
- Similar to feature scaling, the effort on feature discretization should target the provisioning of efficient implementations that integrate with different parts of the streaming process, such as image classification and concept drift detection.

Dimensionality Reduction

Dimensionality reduction is to represent the data with lower space.

Dimensionality reduction techniques that apply transformations to the input data, e.g., Principal Component Analysis (PCA) (Linear method), and Autoencoder (Non-linear method).

Feature Selection

Feature selection targets the identification of which features are relevant to the learning task.

In contrast to dimensionality reduction, feature selection does not apply transformations to the data, and thus, the features can still be interpreted.

Feature selection methods require the entire dataset to determine which features are the most relevant according to some goodness-of-fit criterion. Nevertheless, this is a requirement that does not hold in streaming scenarios, as new data becomes available over time.

Feature selectors are expected to be “stable”, meaning that they should select the same features despite being trained with different subsets of data

DLI Accelerated Data Science Teaching Kit

Thank You