Data Preprocessing

COSC757: Data Mining

Devere Anthony Weaver

1 Why Do We Need to Preprocess the Data?

Much of the raw data contained in databases is unprocessed, incomplete, and noisy. E.g.

- Fields are obsolete or redundant
- Missing Values
- Outliers
- Data in a form not suitable for the data mining models
- Values not consistent with policy or common sense

We want to minimize *GIGO* (Garbage In Garbage Out).

2 Handling Missing Data

A common methods of handling missing values is simply to omit the records or fields with missing values from the analysis. However, this may be dangerous since the pattern of missing values may in fact be systematic, and simply deleting the records with missing values would lead to a biased subset of the data.

Some common criteria for choosing replacement values for missing data include:

1. Replace the missing value with some constant, specified by the analyst.
    - Choosing the field mean as a substitute for whatever value would have been there may sometimes work, but the end-user needs to be informed that this process has taken place.
2. Replace the missing value with the field mean (numeric) or the mode (categorical).
3. Replace the missing values with a value generated at random from the observed distribution of the variable.
4. Replace the missing values with imputed values based on other characteristics of the record.

Data imputation methods can take advantage or previous knowledge and often ask "What would be the most likely value for this missing value, given all the other attributes for a particular record?"

3 Graphical Methods for Identifying Outliers

Outliers are extreme values that go against the trend of the remaining data. They may represent errors in data entry or may be legitimate values.

Identification of these observations is important since certain statistical methods and data mining algorithms are sensitive to the presence of outliers resulting in unreliable results.

Two of the most common graphical methods for identifying outliers are histograms (univariate) and scatter plots (bivariate). One thing to note about the bivariate scatter plot is that a record may be an outlier in a particular dimension but not in another.

4 Measures of Center and Spread

You know the drill with these.

5 Data Transformation

For some data mining algorithms, differences in the ranges of different variables will lead to a tendency for the variable with greater range to have undue influence on the results. That is, the greater variability in one variable will dominate the lesser variability in another, depending on the algorithm used.

Thus, data miners should *normalize* their numeric variables to standardize the scale of effect each one has on the results. For example neural networks and other algorithms that make use of distance measures benefit greatly from data normalization.

6 Data Normalization

There are numerous ways to normalize data, but in these notes, we're only going to cover three prevalent methods.

For the following subsection, let X refer to the original value and X^* refer to the normalized field value.

6.1 Min-Max Normalization

Min-max normalization works by seeing how much greater the field value is than the minimum value $min(X)$, and scaling this difference by the range.

$$X^* = \frac{X - min(X)}{range(X)} = \frac{X - min(X)}{max(X) - min(X)} \tag{1}$$

The data values that represent the minimum for the variable will have a min-max normalization value of zero. The data values that represent the midrange data value has a min-max normalization of 0.5. The data values representing the field maximum will have a min-max normalization of 1.

Thus, the min-max normalization will range from 0 to 1.

6.2 Z-Score Standardization

Z-score standardization works by taking the difference between the field value and the field mean value, and scaling this difference by the standard deviation of the field values.

$$Z - score = \frac{X - \mu_X}{\sigma_X} \tag{2}$$

The data values that lie below the mean will have a negative Z-score standardization. The values falling exactly on the mean will have a Z-score standardization of zero. Data values that lie above the mean will have a positive Z-score standardization.

6.3 Decimal Scaling

Decimal scaling ensures that every normalized values lies between -1 and 1.

$$X^*_{decimal} = \frac{X}{10^d} \tag{3}$$

where d represents the number of digits in the data value with the largest absolute value.

7 Transformations to Achieve Normality

Find a more in-depth resource for this as well as just get hands-on practice.

---

8 Numerical Methods for Identifying Outliers

Recall, the Z-score method for identifying outliers states that a data value is an outlier if it has a Z-score that is either less than -3 or greater than 3. A potential issue with this is that the Z-score standardization is computed using the standard deviation and the mean which are both sensitive to outliers.

A more robust method is to use the *interquartile range (IQR)*. Using the IQR, a data point is considered an outlier if

- It is located 1.5(IQR) or more below Q1 or
- It is located 1.5(IQR) or more above Q3.

9 Flag Variables

Flag variables are also known as dummy variables. They take on either 0 or 1 as values.

When $k \geq 3$, we define $k - 1$ dummy variables and use the unassigned category as the reference category.

---

10 Binning Numerical Variables

Some algorithms prefer categorical rather than continuous predictors. For these, we'll need to partition any numerical predictors into bins or bands.

The four common methods for binning numerical predictors are:

1. Equal width binning - dividing the numerical predictor into k categories of equal width, where k is chosen by the client or analyst.
2. Equal frequency binning - dividing the numerical predictor into k categories each having n/k records, where n is the total number of records.
3. Binning by clustering - uses a clustering algorithm to automatically calculate the optimal partitioning.
4. Binning based on predictive value - binning based on predictive value partitions the numerical predictor based on the effect each partition has on the value of the target variable.

Pro tip, try to avoid the first two methods.

---

11 Reclassifying Categorical Variables

Reclassifying categorical variables is the categorical equivalent of binning numerical variables.

Some data mining methods will perform sub-optimally when confronted with predictors containing too many field values. In such a case, the analyst should reclassify the field values.