Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Improvement

Chapter 3 (Part I)

Feature Improvement

- Making existing features more usable by cleaning and augmenting.
- Developing a better sense of which features are important within our data.
- Improvement tasks:
  - Cleaning data
    - Impute missing values
    - Treatment outliers
  - Augmenting
    - Normalize data
    - Discretize data
    - Transform data value
    - Encode data value

Missing Data Handling

- Data can have missing values due to a variety of reasons.
- Mechanisms of missing data:
  - Structural deficiencies in the data
  - Random occurrences
    - Missing completely at random (MCAR) (independent of the data)
    - Missing at random (MAR) (depends on the observed data but not on the unobserved data)
  - Specific causes (not missing at random: NMAR)
- Applying learning algorithms to data with missing values, most algorithms are not able to cope with missing values. (e.g., SVM and ANN)
- Steps to handle missing data:
  - Identify missing data
  - Get a sense of what the missing values
  - Decide how we want to handle missing values

Dealing with Missing Values

Two major ways for dealing with missing data are:

- Remove rows/columns with missing values in them
- Encode missing values
- Impute missing values

* Each method will have its pros and cons

Dealing with Missing Values

Deleteion of Data

- Remove entire predictor and/or sample that contain missing values.
- Must carefully consider:
  - A number of aspects of the data
  - The intrinsic value of samples as compared to predictors.
  - Removing samples (rows) of the training set is that it might bias the ML model.
- In practice
  - Specific predictors and specific samples contain a majority of the missing information.
  - When it is difficult to obtain samples or the data contain a small number of samples (i.e., rows), then it is not desirable to remove samples from the data.
  - If the data are missing completely at random, removing samples might be viable.

Dealing with Missing Values

Encoding Missingness

- Encode the missing value as a category if the feature is a categorical data in nature.
- Could simply be encoded as “missing” or “unknown”

Dealing with Missing Values

Imputation Methods

- Uses available values in the same feature to estimate to fill the missing value.
  - Mean/Median Imputation
  - End of Tail Imputation
  - Arbitrary Value Imputation
  - Frequent Category Imputation
- Uses information and relationships among the non-missing predictors to estimate to fill in the missing value.
  - K-Nearest Neighbors Imputation
  - Regression Imputation

Dealing with Missing Values

Mean/Median Imputation

- Replacing all occurrences of missing values (NA) within a variable by the mean or median.
- Can only be applied to numerical variables.
- Assumptions:
  - Data is missing completely at random (MCAR)
  - The missing observations, most likely look like the majority of the observations in the variable.
  - If data is MCAR, then it is fair to assume that the missing values are most likely very close to the value of the mean or the median of the distribution.
- Note:
  - If the number of missing data is big (>5%), the variance of the variable will be distorted.
  - Estimates of covariance and correlations with other variables in the dataset may also be affected.
  - If the variable is skewed, the mean is biased by the values at the far end of the distribution.

Dealing with Missing Values

End of Tail Imputation

- Automatically selecting arbitrary values at the end of the variable distributions.
- If the variable is normally distributed, use the mean ± 3 S.D.
- If the variable is skewed, use the IQR proximity rule
  - upper limit = 75^{th} Quantile + 1.5 IQR
  - lower limit = 25^{th} Quantile – 1.5 IQR.
- Assumptions:
  - Data are not missing at random (MNAR)
- Advantages:
  - Captures the importance of “missingness” if there is one.
- Disadvantages
  - Distortion of the original variable distribution and variance
  - Distortion of the covariance with the remaining variables
  - This technique may mask true outliers in the distribution

Dealing with Missing Values

Arbitrary Value Imputation

- Replacing all missing values (NA) within a variable by an arbitrary value.
- Typically used arbitrary values are 0, 999, -999 or -1 (if the distribution is positive).
  - If the variable is a categorical feature, used “Missing” label as the arbitrary value.
- Assumptions:
  - Data is is not missing at random
- Advantages:
  - Captures the importance of “missingness” if there is one.
- Disadvantages
  - Distortion of the original variable distribution and variance
  - Distortion of the covariance with the remaining variables
  - If the arbitrary value is at the end of the distribution it may mask or create outliers

Dealing with Missing Values

Frequent Category Imputation

- Replacing all occurrences of missing values (NA) within a variable by the mode (most frequent category).
- Appropriate for categorical variables.
- Assumptions:
  - Data is missing completely at random
  - The missing observations, most likely look like the majority of the observations in the variable.
- Note:
  - If the number of missing data is big (>5%), lead to an over-representation of the most frequent label.
  - Estimates of covariance and correlations with other variables in the dataset may also be affected.

Dealing with Missing Values

K-nearest neighbors

- Procedure
  1. Identifies a sample with one or more missing values.
  2. Identifies the *K* most similar samples in the training data that are complete.
  3. Calculate the average value (mean/median/mode) of the predictor of interest.
  4. Replace the missing value of the sample with the average value.

Age	Sex	Salary	Province
25	M	25,000	CM
26	M	23,500	CR
26	F	24,000	CM
25	M	26,000	CM
Average salary		24,625	

4 similar samples

26	F	24,625	CM

Missing data

Advantages:
- Can be applied when the training set is small or moderate in size.

Dealing with Missing Values

Regression Imputation

- Use complete features as predictor and missing variable as the target.
- Use complete samples as training data to built a regression model.
- Use the trained model to estimate the missing value.
- Assumptions:
  - When you use a linear model, complete predictors should show strong linear relationship with a predictor that requires imputation.
  - Linear regression can be used for a numeric predictor that requires imputation.
  - Logistic regression is appropriate for a categorical predictor that requires imputation.

Dealing with Missing Values

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	78	7.529
$\mathbf{x}_2$	84	8.267
$\mathbf{x}_3$	84	8.267
$\mathbf{x}_4$	85	8.390
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	12

$$JP = 0.123(78) + (-2.065) = 7.529$$
$$JP = 0.123(84) + (-2.065) = 8.267$$
$$JP = 0.123(84) + (-2.065) = 8.267$$
$$JP = 0.123(85) + (-2.065) = 8.390$$

$$JP = \beta_1(IQ) + \beta_0 = 0.123(IQ) + (-2.065)$$

incomplete variables

complete variables

Example of Regression Imputation

1. Estimate a set of regression equations
2. Generate predicted values for the incomplete variables
3. Fill in the missing values

Outlier Detection and Treatment

Outlier is a data point that is significantly different from the remaining data.

Effects of outliers

- The mean and variance are sensitive to outliers
- The performance of some machine learning models (linear regression or AdaBoost)

Dealing with outliers

- Identify outliers
  - Z-score
  - Boxplot
- Handle outliers
  - Perform variable discretization
  - Treat outliers as missing data and carry out any of the missing imputation techniques
  - Remove samples with outliers
  - Perform winsorization

Outlier Detection Techniques

Z-score

- Consider on each individual feature of the dataset.
- How far the value of the data point/sample is from its mean for a specific feature.

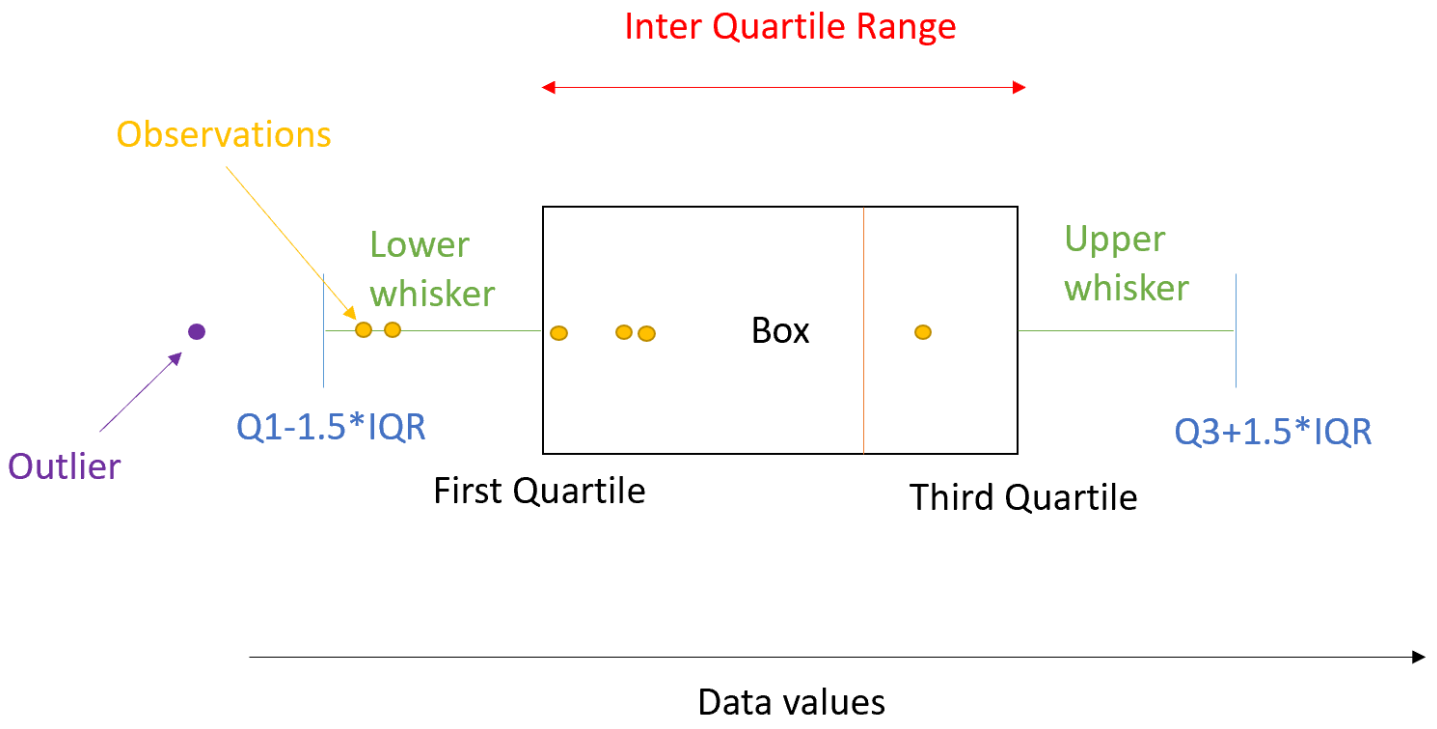
$$Z - score = \frac{x_i - \bar{x}}{\sigma}$$

- A Z-score of 1 means the sample point is 1 standard deviation away from its mean.
- Z-score values greater than or less than + 3 or − 3, respectively, are considered outliers.
- Note that this method assumes that the features have normal distributions.

Outlier Detection Techniques

Boxplot

The values that are not in range [Q_1-1.5IQR,Q_3+1.5IQR] are considered as extreme values.



Source: https://medium.com/analytics-vidhya/introduction-to-box-plots-and-how-to-interpret-them-22464acbcba7

Winsorization: Outlier Treatment Technique

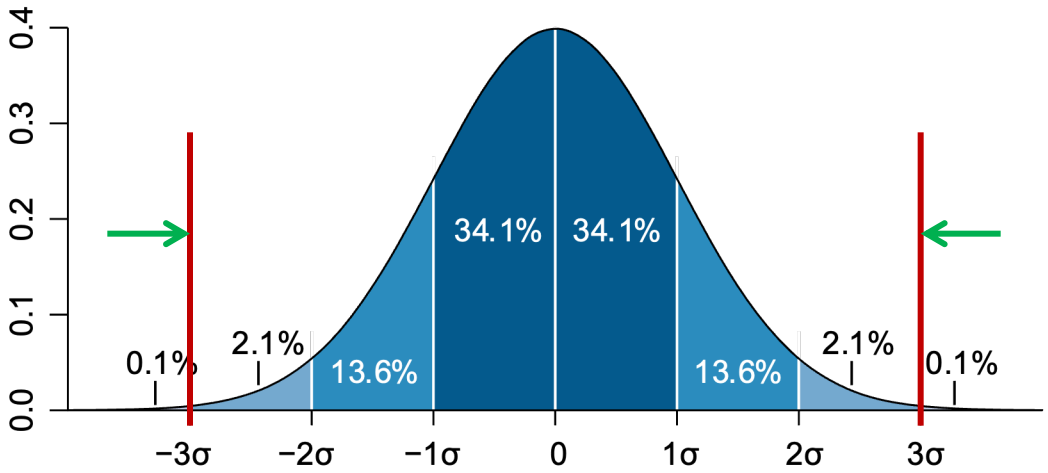
Winsorization

- The process of transforming the data by limiting the extreme values to a certain arbitrary value.
- The extreme values are replaced by:
  - Maximum and minimum values
    - Gaussian approximation
    - Inter-quantile range proximity rule
    - Percentiles
  - Zero-coding

Winsorization: Outlier Treatment Technique

Gaussian approximation

- Related to Z-score, the values that higher/lower than mean±3S.D. are considered as outliers.
- The values that higher than mean+3S.D. are replaced by mean+3S.D.
- The values that lower than mean-3S.D. are replaced by mean-3S.D.
- Note that this method assumes that the features have normal distributions.

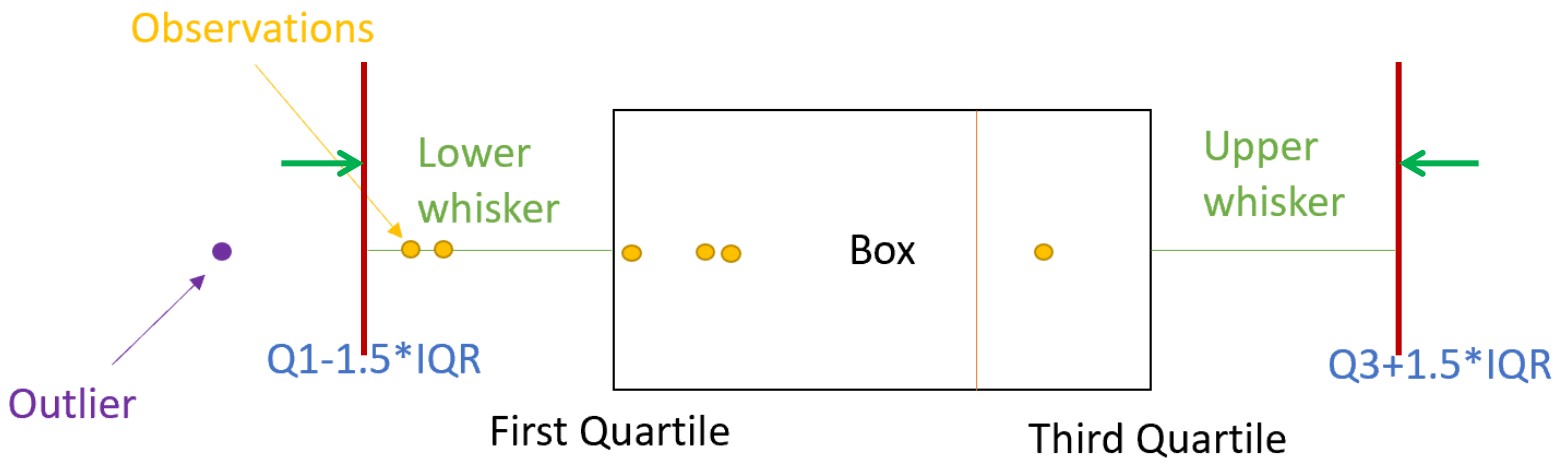


Source: https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Winsorization: Outlier Treatment Technique

Inter-quantile range proximity rule

- The values that higher than Q_3+d*IQR are replaced by Q_3+d*IQR.
- The values that lower than Q_1-d*IQR are replaced by Q_1-d*IQR.
- The parameter d can be 1.5 or 3.



Winsorization: Outlier Treatment Technique

Percentiles

- The values that higher than 5^{th} percentile are replaced by 5^{th} percentile .
- The values that lower than 95^{th} percentile are replaced by 95^{th} percentile .

Winsorization: Outlier Treatment Technique

Zero-coding

- A variant of bottom-coding and refers to the process of capping, usually the lower value of the variable, at zero.
- It is commonly used for variables that cannot take negative values, such as age or income.
- The values that lower than o are replaced by o.

References & Study Resources

- Sinan Ozdemir and Divya Susarla. (2018). *Feature Engineering Made Easy*. Packt Publishing.
- Sinan Ozdemir. (2022). *Feature Engineering Bookcamp*. Manning Publications Co.
- Max Kuhn and Kjell Johnson. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.
- Soledad Galli. (2020). *Python Feature Engineering Cookbook*. Packt Publishing.
- https://medium.com/analytics-vidhya/feature-engineering-part-1-mean-median-imputation-761043b95379
- https://medium.com/analytics-vidhya/feature-engineering-part-1-end-of-tail-imputation-c5069a41869a
- https://medium.com/analytics-vidhya/feature-engineering-part-1-arbitrary-value-imputation-e81444bd79b2
- https://medium.com/geekculture/frequent-category-imputation-missing-data-imputation-technique-4d7e2b33daf7
- https://medium.com/analytics-vidhya/introduction-to-box-plots-and-how-to-interpret-them-22464acbcba7
- https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg