Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Improvement

Chapter 3 (Part II)

Feature Discretization

Reducing the number of possible values a feature can take

- Continuous feature → a discrete feature (integer).
- Continuous feature → an (ordered) categorical feature.

Effects of discretization

- Discretization error
- Result in fewer parameters for ML models that can take categorical feature as input.
- Increase the number of parameters for ML models that cannot accommodate categorical features directly.

Usefulness

- Improve generalization
- Error analysis and understanding the behavior of the system

Feature Discretization

Feature discretization methods:

- **Unsupervised discretization**: perform on over the feature values alone, in isolation from the target class values.
    - Equal interval width discretization
    - Equal frequency Interval discretization
    - K-mean Clustering
- **Supervised discretization**: done relative to the target class.
    - ChiMerge discretization
    - Decision trees

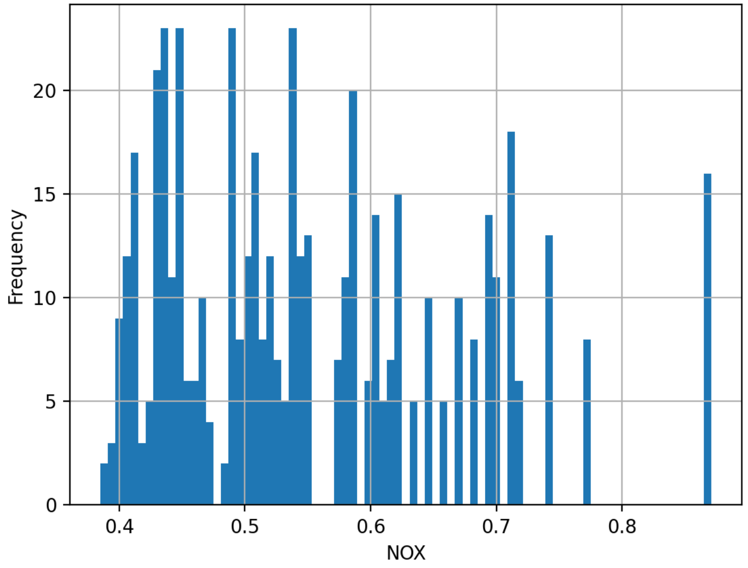
Feature Discretization

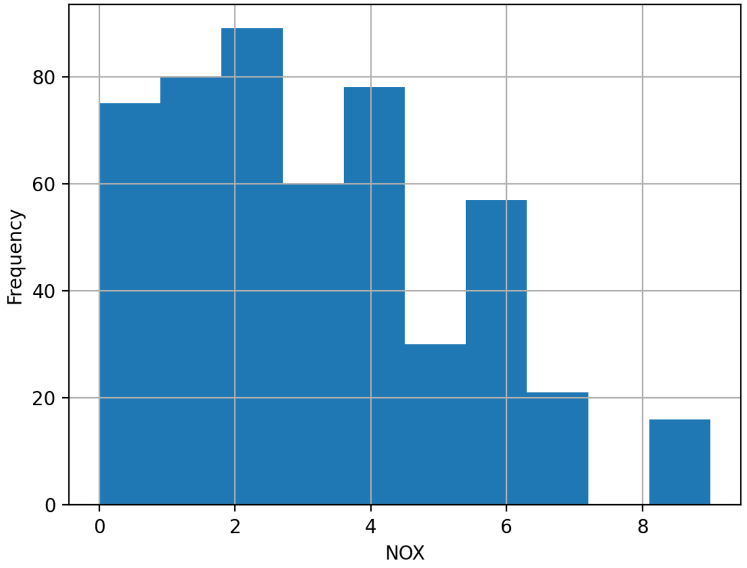
<u>Equal interval width discretization</u>

- Get the range of variable: $l = \max(X) - \min(X)$
- Divide the range l into k equal region: $w = L/k$
- Obtain the boundaries of each bin

$$([\min(X), \min(X) + w), [\min(X) + w, \min(X) + 2w), [\min(X) + 2w, \min(X) + 3w), \ldots, [\min(X) + (k-1)w, \max(X)])$$

Note that this method is very sensitive to outliers, so either remove them first or do not use them if you have many outliers.

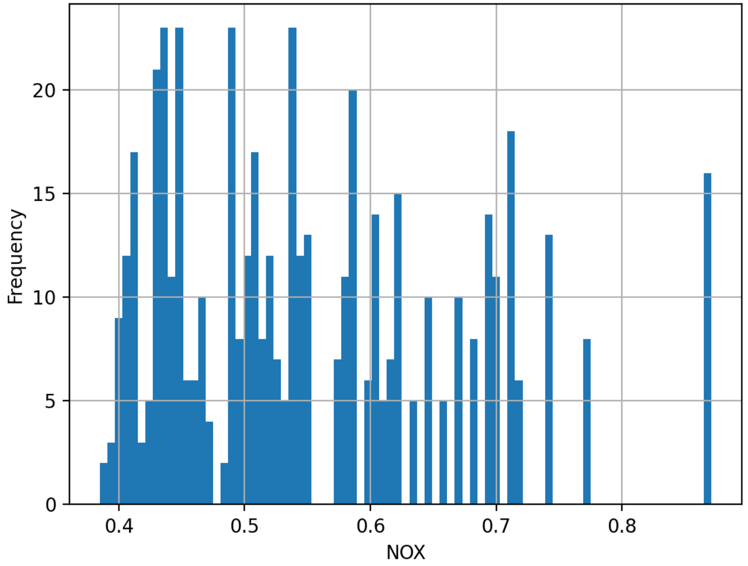


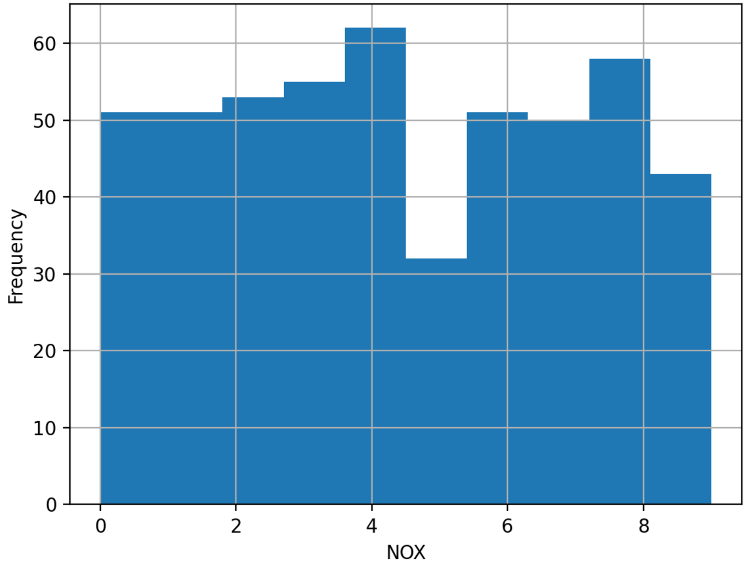


Feature Discretization

<u>Equal frequency Interval discretization</u>

- Get the number of instances m
- Divide m instances into k groups: $f = m/k$ (all groug contain the same number of numerical values.)
- Sort instances by their values and pick the boundary items
- This methos helps when you have different levels of data density in different regions of the possible values.





Feature Discretization

<u>K-mean Clustering</u>

- Perform k-mean clustering on the feature
- Use the number of the cluster as the feature category

Feature Discretization

ChiMerge discretization

- Applies the Chi Square method to determine the probability of similarity of data between two intervals.
- Procedure:
    1. Sort the feature values
    2. Consider each feature value as a separate internal. The boundary between two intervals is $\frac{x_i+x_{i+2}}{2}$
    3. Replete until there are no interval can be merged:
        1. For each interval and its neighbors, calculate chi-square test over the values of the target class.
        2. Merge an interval with its neighbors if the chi-square test cannot reject the null hypothesis.

Feature Discretization

<u>Decision trees</u>

- Use a decision tree to identify the optimal splitting points that determine the bins or contiguous intervals.
- Procedure:
    1. Train a decision tree of limited depth (e.g., 2, 3 or 4) using the variable we want to discretize to predict the target.
    2. Replace the original value by the prediction returned by the tree.

Note that this method may cause over-fitting.

Categorical Feature Encoding

Turn the nonnumeric categories into numbers.

- Some ML models work with numeric data only.

Categorical feature encoding methods:

- Ordinal Encoding
- One-hot Encoding
- Dummy Coding
- Effect Coding
- Feature hashing

Categorical Feature Encoding

Ordinal Encoding (Label Encoding)

- The encoding of variables retains the ordinal nature of the variable
- Each category is assigned a value from 1 through the number of possible values by considering the order of values.

Temperature
Cold
Warm
Hot
Very Hot

Temperature
1
2
3
4

Categorical Feature Encoding

One-hot Encoding

- Use *k* bite to represent *k* possible categories.
- Map each category to a vector that contains 1 and 0
  - 1 - presence of the feature
  - 0 - absence of the feature

Gender
Male
Female
Other

isMale	isFemale	isOther
1	0	0
0	1	0
0	0	1

Note that this method it uses one more bit than is strictly necessary.

Categorical Feature Encoding

Dummy Coding

- Dummy coding removes the extra degree of freedom by using only k–1 features in the representation.
- A category, called referent category, is represented as a vector of all zero.

Gender
Male
Female
Other

isMale	isFemale
1	0
0	1
0	0

Categorical Feature Encoding

Effect Coding

- Similar to dummy coding
- But the reference category is now represented by the vector of all –1's.

Gender
Male
Female
Other

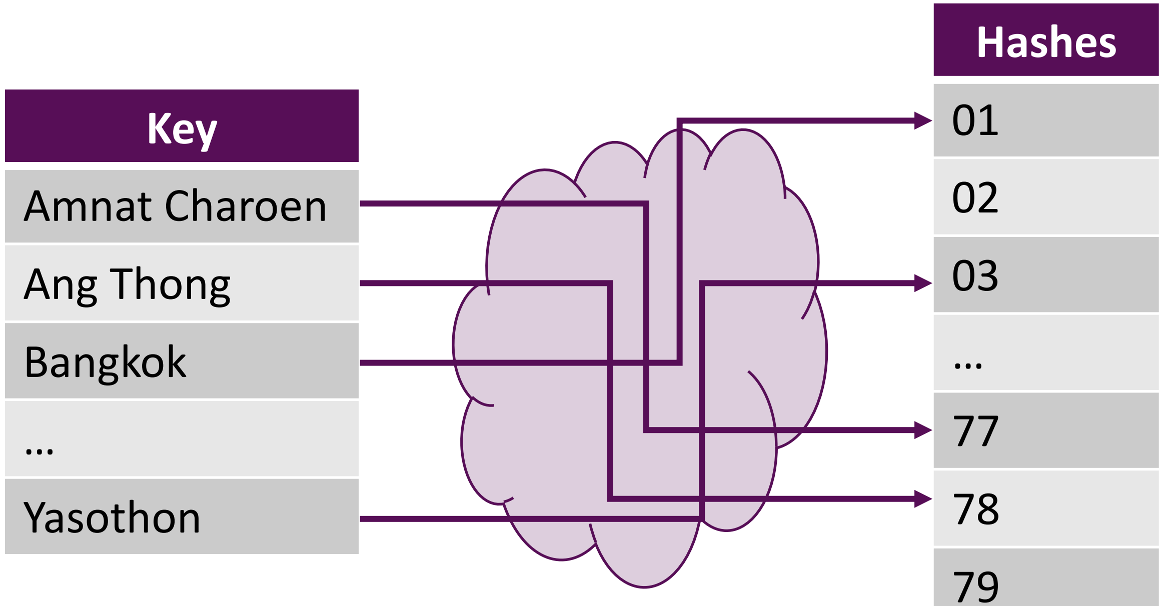
isMale	isFemale
1	0
0	1
-1	-1

Note that one-hot encoding, dummy coding and effect coding break down when the number of categories becomes very large.

Categorical Feature Encoding

Feature Hashing

- **Hash function** is a deterministic function that maps a potentially unbounded integer to a finite integer range.



A hash function

Categorical Feature Encoding

Feature Hashing (Cont.)

- Hash encoding represents the categorical data into numerical value by the hashing function.
- Hashing schemes work on strings, numbers and other structures like vectors.
- Hashed outputs as a finite set of ***b*** bins
  - The same categories are assigned to the same bin (or subset of bins) out of the ***b*** bins based on the hash value.
  - The number of bins ***b*** can be pre-defined.

Note that a high number of categorical values are represented into a smaller number of features, different categorical values could be represented by the same Hash values – this is called a collision.

Categorical Feature Encoding

	Name	Genre	0	1	2	3	4	5
1	Super Mario Bros.	Platform	0.0	2.0	2.0	-1.0	1.0	0.0
2	Mario Kart Wii	Racing	-1.0	0.0	0.0	0.0	0.0	-1.0
3	Wii Sports Resort	Sports	-2.0	2.0	0.0	-2.0	0.0	0.0
4	Pokemon Red/Pokemon Blue	Role-Playing	-1.0	1.0	2.0	0.0	1.0	-1.0
5	Tetris	Puzzle	0.0	1.0	1.0	-2.0	1.0	-1.0
6	New Super Mario Bros.	Platform	0.0	2.0	2.0	-1.0	1.0	0.0

Feature Hashing on the Genre attribute. A signed 32-bit version of the *Murmurhash3* hash function was used.

Source: https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63

References & Study Resources

- Pablo Duboue. (2020). The Art of Feature Engineering: Essentials for Machine Learning. Cambridge University Press.
- Alice Zheng and Amanda Casari. (2018). *Feature Engineering for Machine Learning*. O'Reilly Media, Inc.
- Niculescu-Mizil, et al. (2009). Winning the KDD Cup Orange Challenge with Ensemble Selection. JMLR: Workshop and Conference Proceedings 7: 23-34. KDD 2009.
- https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63