1. Hi every one, I am Yun-Hao yang, form the Institute of Industrial Engineering of NTU. I am glad I can be here to share the topic, which is about variable encoding for binary features.
2. while doing Data pre-processing, variable encodings help us transform the categorical data into numerical, most of them assigning numerical value to each type in the feature, like ordinal, frequency, and target encoding.

   Among all, One Hot encoding is also a commonly used methods, for its convenience and efficiency. Unlike other methods, One Hot encoding creates dummy columns for each type in the categorical feature. And represent original data into binary
3. Unfortunately, If the categorical feature has numerous types, one hot encoding will largely expand the dimension of the data, and might trigger the curse of the dimensionality, causing ML model perform poorly.

   The reason why is mainly due to too many features encumber the classification model in making decision
4. The common way to prevent the problem is simply encode with others methods, instead of presenting categorical data in numerous binary columns. Imagine we receive one hot encoded data as original input, with many grouped binary features, we can reverse the data back to categorical, then choose another method to solve the dimensionality problem.

5. But, what if the data we got in first hand is not grouped binary? What if there are too many binary features to deal with? There is no way to revered this non-grouped binary data to categorical like the previous slide. Is there a better way to describe, or encode this kind data?
6. As the traditional encoding is only for categorical part. In the research, we try to develop a variable encoding method built for binary data, which is able to transform non-grouped binary into numerical, within these 3 steps: column Grouping, Sequencing, and binary-coded decimal. The following animation will show how these steps works.
7. As we gather the non-grouped binary data, there might be some relationship between columns, yet we don't know about it.
8. In grouping phase, hoping by the feature selection method can find this kind of relationship. In this case, we successfully group the columns by their physical meaning, which is animal types, colors, and sizes.
9. Secondly in the sequencing phase, for generating a more competitive numerical value, we need to rearrange the sequence of columns in groups. The detail and the reason why will be explained later on
10. F
11. In the last step, with the binary – code decimal, we can represent columns into numerical values. And that's it, we have generated a both understandable and tidy dataset with numerical value form the non-grouped binary data.

12. Here is a little example. Assume you are a zookeeper who have to check all animal's health status, and here is the checklist, showing the animal ’s type, size, color, and health status. Each checkbox in same group is exclusive to each other. After collecting the data of all 300 animals in your zoo, the result is at the right, presenting how many times each checkbox was checked, and what kind of health status the checked animal was.

    Like, in the size: mid checkbox, 100 out of 300 samples are checked, and half of them are in the good status, but the other half of them don’t.

    Our target is using these 18 checkboxes as binary feature data, on predicting the animal’s health status.

13. First, we try to find the relationship of these 18 binary features. for now, we used Principle Components Analysis in grouping the columns.

    In here, PCA helped us makes 3 different columns groups, via PC1 to PC3.

    ~~The main purpose using PCA here is by flittering out the less importance features, we can create groups that contain the most importance ones.~~

14. After grouping the features by PCA, we now can rearrange the columns. We had been trying using column sum, impurity, and Feature importance as our sequencing rule.

15. In the right-hand side, we can see the columns have been rearrange with its columns sum decreasingly.

    But why bother to rearrange the sequence of the columns? Will this matter?

16. In fact, sequencing is decisively importance to the new encoded numerical value. These figures show how the new data will distributed in numerical feature by the sequence at the left, if a certain sequence can make these two types be clearly separated in numerical value, will makes classifier easily to find the cut point in this new feature.

    In a nut shell, a good sequence can make our encoded numerical data easier to be classified.

17. But how to transform multiple binary value into numerical by BCD? In the right is the how the BCD works. It basically is by decimalize a sample's grouped columns as binary values. Like the sample here, 2 power 5 plus 2 power 0 is 33 in total.

    As you can imagine, if we can put a more Informative column at front, it can determine the BCD value more than other columns do.

    Like the color: blue checkbox only checked by the good animals, and putting it at front will largen both their BCD value and their gap to the bad animals on the new numerical feature.

    ~~Hence putting informative column at front can help us in output a better numerical value in separating two types.~~

18. That's the main steps of whole process, the goal of the research is focus on how to find a better grouping and sequencing methods, in order to generate a more easily classified data.

19. In case study we will go through both a simulated, and a Kaggle dataset. To demonstrate our method, and compare with the traditional encoding in the same time.

20. These datasets were generated procedurally, with around 3 thousand samples and ten to one ratio in two types, and several noises in the majority. We plan to use its 3 dimensions as our inputs.

21. First, we binarize each axis from the dataset, and use both our method and the traditional variable encoding to numericalize the data. Then compare the classification result of the those data.

    Here we can see the 3 axes of dataset 1, and type was showing by the color. And the right shows after binarize each axis has 10 categories, makes 30 binary features in total.

22. These is a demonstration of PCA grouping before sequenced by column sum. in each PC we can see columns from different axis, and also, compare PC1 to PC5, PC5 only has small size binary features. Meaning binary features in PC1 has more samples can explain more variance to the whole data.

    The BCD result is reasonable too, the informative seems dwindling by PCs.

23. The method we develop can also use as a way of dimension reduction, visualize the binary data in to the dimension we wanted to.

    We can also see a huge difference of sequencing.

It’s not hard to image what kind of data will be benefit whiling training the classification model.

24. The line chart shows the k-folds validation average f1 score under different binary features we binarize. We can observe the variance of score became wider as the binary features increase.

25. Bar chart shows the average f1 score of the previous line chart. We can see the score of sequenced by impurity and the feature importance is much higher than the random in all groups, and even or above the most variable encoding method under certain combination. Our conclusion here is different grouping and sequencing will definitely effect on classification result.

The numbers below show the data’s number of features. number of features of One hot and binary encoding variate with how we binarize the continuous data.

26. In the Kaggle dataset. We selected some of features as the read-in data. And it contains multiple data types, includes numerical, categorical, and binary. After one hot encoded, it will generate 75 dummy features.

27. Although two the grouping method cannot outperform the default group or some traditional encoding methods,

28. But, in facing non-grouped binary data, those method can’t be used anymore. And there we can see the advantage. By compress the 83 binary features into 5 numerical, and maintain a same or slightly better classification result.

29. In conclusion, we formulate a way to encode non-grouped binary data into numerical, and can preserve the classification result with certain sequencing method. Or the method can be used as a dimension reduction method for binary data.

That’s all, thank you.