Individual Assignment

TECHNOLOGY PARK MALAYSIA

CT051-3-M-DM

Data Management

APUMF2204DSBA(DE)(PR)

Assignment Part-1

Student's TP: TP067696

Student's Name: Mr. Muhammad Arif Bin Jamaluddin

Lecturer's Name: Dr. MURUGANANTHAN VELAYUTHAM

Table of contents

Introduction to Feature Engineering

Feature engineering is a part of machine learning approach whereby it extorts data to produce new variables that does not exist in the training set. The goal is to simplify and stepping up the data transformations and at the same time improve model accuracy. Other purposes it also can create new features for supervised and unsupervised machine learning. Feature engineering can be categorized into two parts which are variable transformation and variable creation and also is a mathematical depiction of unprocessed data and hence it can be defined as the process or method that formulates the best appearances provided the information, model and assignment. Feature engineering is also a mechanism that support to create new features which then can be reused again in another models.

Data scientists used feature engineering to conduct model evaluation and hence to achieve automated feature engineering, there are several current ways adopt by using assisted search in feature space by using feature quality measures such as knowledge gain and alternative measures of conduct (Markovitch and Rosenstein, 2002; Fan et al., 2010). On the other hand, others used distinct approach for example feature assembly and choice established on model assessment (Dor and Reich, 2012; Khurana et al., 2016).

Another method used by other researchers whereby Kanter et al states that Data Science Machine (DSM) technique which review at feature engineering problems as quality selection on area of novel trademark and this method depends on all available features from the dataset, given by the progression stemming from the set of transformations and after that it will perform feature selection on the magnified dataset (Kanter and Veeramachaneni, 2015). Raw data need to be pre-processed due to its also affects the performance of classifiers before feature engineering takes place.

Variable/ Feature transformation

Variable or feature transformation is to transform the data to make it works better in the model and data variables can be divided into two types such as numeric and categorical variables where each transformation would have distinct approaches. Numeric variable transformation is whereby converting numeric variable to another numeric variable and its usually designed

to modify the range of values and to tune the data distribution into gaussian-like features. Next is categorical variable transformation where it converts categorical variable into numeric whereby this transformation is very crucial as most of machine learning model works when the input is in numerical values. Another name of this is encoding or in text mining it is called embedding which to handle similar situation but embedding function is to return numerical values that have semantics of original data. The transformation of categorical variable is very crucial for all models and its selection very important for most model performance.

There are two types of types of variable transformations which are simple functions and normalization. Simple function in statistics like sqrt, log, and 1/x are applied so that it can convert record that do not have gaussian distribution into information that does. Thus, in this case the data can be alter into log transformation but it needs to be applied under caution because it easily alters the nature of the data. The other type of feature transformation is standardization or normalization where it is known that after the alteration the new variable contain mean of zero and standard deviation of one. These are some generic examples of classification task in feature engineering.

1. Primary variable transformation such as counting new feature value from current feature value.
2. All of the features undergo normalization and thus feature scaling I used
3. Standardization
4. Box-cox transformation is used to normalize feature values.
5. Min-Max normalization:
6. The log-odds(logit) feature transformation
7. The density of occurrence for every level of categorical values are captured under high cardinality features
8. Discretization

Variable/ Feature creation

Variable creation is a way creating new features depends upon existing attributes for example if input variable in the dataset is date(dd-mm-yy) the output variables created is day, month, year, week, weekday. The example below shows the method to create new features such as

1. For example, analyzing missing information and add new features to find the disappeared values and enhanced the classifiers performance by evaluation errors created by classifiers (S. Bird, E. Klein, E. Loper, 2009).
2. Patel et al. suggested the system of discriminating features by measuring the labeled data with classifiers “confused region” with the differing class.
3. Arrangement of feature weights in every iteration (H. Raghavan, O. Madani, R. Jones, 2005)
4. Crowd generating features (J. Cheng and M. S. Bernstein, 2015)
5. Interactively debugging features (F. Heimerl, C. Jochim, S. Koch, T. Ertl, 2012)
6. Hsiang-Fu Yu et. al. use two-levels method where binarization and discretization is used to produce sparse feature and statistical approach implemented to the data to produce compact features. Hence, this new feature can be input to the classifiers to study and enhance the veracity of undiscovered data.

Conclusion

Thus, it can be concluded feature engineering is one of the elements of Machine Learning algorithms that are designed to work with qualitative or quantitative data and very few algorithms support mixed data. Hence, various feature transformation methods act differently in the environment where it’s been experimented and checked, thus each transformation is usually gained by analyzing several functions from the original features.

References

[Markovitch and Rosenstein, 2002] Shaul Markovitch and Dan Rosenstein. Feature generation using general constructor functions. Machine Learning, 2002.

Fan et al., 2010] Wei Fan, Erheng Zhong, Jing Peng, Olivier Verscheure, Kun Zhang, Jiangtao Ren, Rong Yan, and Qiang Yang. Generalized and heuristic-free feature construction for improved accuracy. SDM, 2010.

[Dor and Reich, 2012] Ofer Dor and Yoram Reich. Strengthening learning algorithms by feature discovery. Information Sciences, 2012.

[Kanter and Veeramachaneni, 2015] James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. DSAA, 2015.

S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. O'Reilly Media, Inc., 2009.

H. Raghavan, O. Madani, and R. Jones, "Inter Active Feature Selection," in Proc. IJCAI 2005, 2005, vol. 5

J. Cheng and M. S. Bernstein, "Flock: Hybrid Crowd-Machine Learning Classifiers," 2015, pp. 600–611

F. Heimerl, C. Jochim, S. Koch, and T. Ertl, "Feature Forge: A Novel Tool for Visually Supported Feature Engineering and Corpus Revision," in Proceedings of COLING 2012: Posters, Mumbai, India, 2012, pp. 461–470.