Question 3

0.1 Introduction

Machine learning(ML) is a subset of Artificial intelligence and computer science, which have been widely adopted in statistics, finance analysis, bio-informatics, etc. Cross-validation is typically used to assess these ML models' performance and generalizability. However, more and more scholars have realized that, especially in bioinformatics, these internal validation techniques sometimes cannot objectively evaluate the ML models due to the potentially biased training data. That is the data doppelgängers.

For example, when the training and validation data are very similar because of coincidence or other reasons, the ML model trained based on this data set is likely to perform well regardless of its actual quality. These doppelgängers effects are hard to detect and verify during the entire model construction and training process, even from the final results.

0.2 Widespread Presence of Data Doppelgängers and Causes

As genomic databases grow and the specificity of available data, the chance researchers encounter data doppelgängers in the biomedical field increases. Waldron et al. studied various cancers databases assessed there accuracy via further manual inspection of expression data, clinical annotations, and sample index[1]. Among all databases, they confirmed that more than 50% of them exist data doppelgängers. The study of Cao and Fullwood also pointed out these doppelgängers effects[2]. Cao and Fullwood detailed evaluate the performance of an Ml model called TargetFinder[3]. They obtained very close F_1 scores between real and shuffled motifs. After examining the original functional genomic features of the TargetFinder data sets, they concluded that the high sample similarity is likely to influence the cross-validation results, thereby overstated the model performance. Besides, in the Alzheimer's Disease Neuroimaging Initiative (ADNI), a data set which contains a large number of patient measurements[4]. Huckvale et al. noted that more than 90% features (e.g., biomarkers, gene expression, and magnetic resonance imaging) were strongly correlated with at least one other feature

in data set[5].

These doppelgängers effects are not unique to biomedical data. Almost all ML models constructed on real world data sets may face this problem. For example, in some models that use convolutional neural networks for image recognition and classification. Due to unavoidable factors at the beginning of the data collection, such as limited images sources and categories, there is a high degree of similarity in the images. This occurs more frequently in data sets with smaller sample sizes. Meanwhile, data doppelgängers are present in the fields of natural language processing. When classify tweets into positive and negative classes, tweets with similar texts are inferred to belong in the same attitude. However, because different people express their views with different degrees of severity, with different levels of education, or due to the 'subtexts' that exist in different language systems, sometimes tweets that use entirely different words hold the same attitude. This scenario also brings the doppelgängers effects on NLP models.

There are several possible reasons for data doppelgängers.

First, the biased or potentially erroneous data sets may bring data doppelgängers. The percentage of duplicate medical records is growing rapidly in the medical field, as the American Health Information Management Association research. Because at the initial stage of data collection, repeated made case notes, ununified forms, and index and other programmatic errors resulted in the presence of data doppelgängers. The impact of such subjective manipulation is difficult to detect. This is especially the case for medical health data with a large volume.

Second, this may be due to the characteristics of the data itself. For biological data such as genome sequence and protein sequence, the inherent nature of the data dictates that there is a large amount of replication in different samples.

0.3 Measures

Nowadays, as machine learning models are increasingly employed in more and broader fields, identifying the presence of data doppelgängers in advance and minimizing its

potential side-effects on the model is a crucial step that cannot be overlooked in research.

0.3.1 Identification of Data Doppelgängers

In terms of the identification of data doppelgängers, both Wang and Waldron verified that the Pearson's correlation coefficients (PCCs) are a practical feature to detect the presence of data doppelgängers[1, 6]. The purpose of PCCs is to measure the association between two variables and the relationship strength between them. Therefore, from this point of view, we may try other methodologies which could valuate the similarity between samples—for example, the Kullback-Leibler Divergence[7]. The KL divergence measures the relative difference between two probability distributions. If we could obtain the distribution of information or data features, the value of KL divergence may help us identify the data duplication. Besides, other similarity metrics such as Bhattacharyya Distance, Hellinger Distance, and Kolmogorov Distance may also be helpful. From other point of view, we could regard each sample as a system. Such metrics as entropy can be used to measure the amount of information in a system. Thus, comparing the entropy and other measurements based on entropy or information theory of each data sample may also be effective in detecting the presence of data doppelgängers.

0.3.2 Reducing the Doppelgängers Effect

When we realize that there exist data doppelgängers, we should take steps to minimize the doppelgängers effect when using data, building models and evaluating the model.

In the field of machine learning, there is an expression called *Data augmentation*. It is widely used with image recognition and natural language processing. Data augmentation performs operations on the sample such as rotation, shearing, zooming, and flipping. These operations could reduces data doppelgängers by expanding the sample richness and information content.

In addition to finding ways to reduce the similarity between samples at the data processing stage, we can also use other, more objective methods to develop the ML model.

To some extent, the overstated model was caused by data doppelgängers due to the over-fitting when training the model. High similarity samples in training and validation sets are likely to over-weighted some parameters, leading to a biased model. Therefore, we could applied *dropout* method in the model construction process to avoid this and enhance model generalizability. Dropout refers to some number of layer outputs are randomly ignored during model training.

Moreover, in the model validation phase we can take some measures to reduce the doppelgängers effect. Splitting a data set that is data doppelgängers into the training set and validation set will undoubtedly result in an evaluation that reinforces the bias. The use of bootstrap and cross-validation methods such as testing on the same input data set is called internal validation. In contrast, external validation refers to using independently derived data sets to validate the performance of a model trained on the initial input data. If we can choose external validation rather than internal validation or combine them, we could get a more objective model assessment. Ho et al. proposed two extensions of the external validation and discussed its application[8].

Finally, suppose we have already constructed a model based on doppelgängers data and are hard to modify. In that case, we can try to locate the extent to which these data are doppelgängers, thus reducing this model's application to that level. Because sometimes, it is not virtuous to go overboard with the model generalizability.

0.4 Conclusion

In practice, any data we research is effectively a sample and may not be an accurate representation of the population under study. Before building the ML model or applying other data analysis methods, it is often good to check for data doppelgängers and other errors such as sample size, bias, noise, etc. Because we can never guarantee that all input data are error-free and unbiased, we need to remain objective and cautious when constructing and applying the ML models.

References

[1] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. Birrer, "The doppelgänger effect: hidden duplicates in databases of transcriptome profiles," *JNCI: Journal of the National Cancer Institute*, vol. 108, no. 11, 2016.

[2] F. Cao and M. J. Fullwood, "Inflated performance measures in enhancer–promoter interaction-prediction methods," *Nature genetics*, vol. 51, no. 8, pp. 1196–1198, 2019.

[3] S. Whalen, R. M. Truty, and K. S. Pollard, "Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin," *Nature genetics*, vol. 48, no. 5, pp. 488–496, 2016.

[4] [Online]. Available: http://adni.loni.usc.edu/

[5] E. D. Huckvale, M. W. Hodgman, B. B. Greenwood, D. O. Stucki, K. M. Ward, M. T. Ebbert, J. S. Kauwe, J. B. Miller, A. D. N. Initiative *et al.*, "Pairwise correlation analysis of the alzheimer's disease neuroimaging initiative (adni) dataset reveals significant feature correlation," *Genes*, vol. 12, no. 11, p. 1661, 2021.

[6] L. R. Wang, L. Wong, and W. W. B. Goh, "How doppelgänger effects in biomedical data confound machine learning," *Drug discovery today*, 2021.

[7] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[8] S. Y. Ho, K. Phua, L. Wong, and W. W. Bin Goh, "Extensions of the external validation for checking learned model interpretability and generalizability," *Patterns*, vol. 1, no. 8, p. 100129, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389920301707