# Report

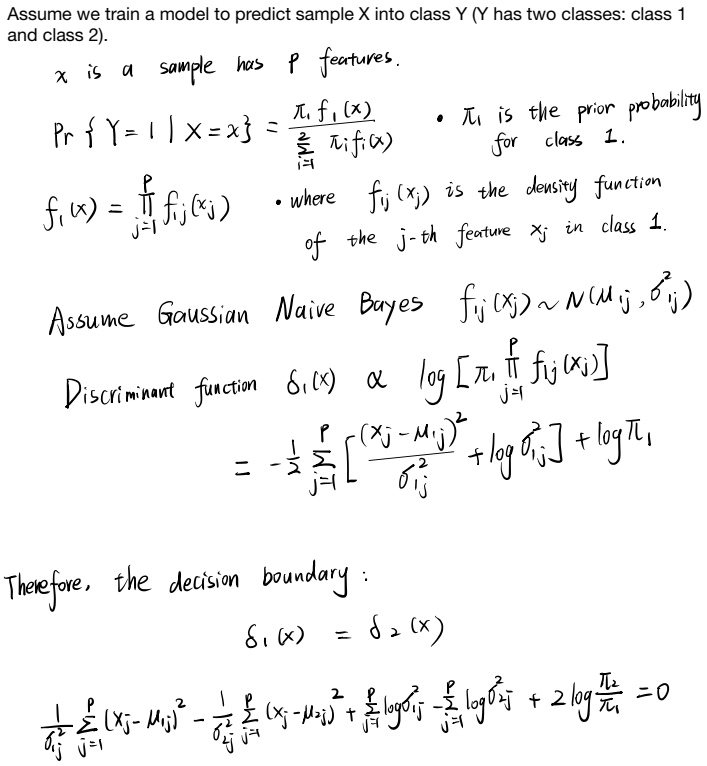
Yuqi Huang

Doppelganger effects are not unique to biomedical data. Based on the definition in the paper, if the train and validation set are highly similar during the cross-validation process, the model might perform well regardless of the quality of training, which is the doppelgänger effect (Wang, Wong, & Goh, 2021). Besides biomedical data, doppelganger effects may also occur in other fields, including finance, marketing, business, transportation, action, etc. For example, when collecting the data of customer behaviors to predict how the customers and prospects will interact with a product or service, doppelganger effects may be encountered when training a model in the situation as follows. Two groups of behavior data are derived independently, but they are very similar to each other because two groups of people have similar sociodemographic characteristics. Hence, if these two groups of data are separated to training set and validation set respectively, then they are data doppelgangers, and if they are functional doppelgangers, they may cause doppelganger effects when training a machine learning model.

It is especially common to have doppelganger effects in biomedical data. In addition to the RCC proteomics data mentioned in the paper, there are many other examples of doppelganger effects. For instance, when training a ML model to predict the type (or stage) of a disease using the 16S ribosomal RNA sequences of gut microbiome in patient samples, samples derived from the same class but from different patients may share similar gut microbiome composition pattern, including the species diversity among population and within population. These samples may present an impression of high accuracy of prediction, but when the model is applied to less similar sequence samples, the result is not satisfying. Another example can be shown in the protein

structural prediction mentioned in the paper. Protein pairs with a sequence identity larger than 30% are found to be structurally similar, and thus are similar in cellular function, which may lead to doppelganger effects. Similarly, in the prediction problem using gene expression profiles, doppelganger effects will happen if two genes are homolog, which means they related to each other by descent from a common ancestral gene by speciation and therefore will have similar function and expression profiles.

The doppelganger effects emerge due to the inaccuracy when evaluating the model using the validation set. In the paper, the authors argue that naïve bayes models have a clearer linear relationship between performance inflation and doppelganger dosage (Wang, Wong, & Goh, 2021). Hence, I take a simple naïve bayes model as an example to show the quantitative angle of this effect. The following picture shows the decision boundary derived in naïve bayes model.




The decision boundary is derived using the training data set, which estimates the prior probability, the mean and variance of each class in the decision

boundary formula. When the validation set have doppelganger data, which means that validation samples have similar features($x_j$ in the equation) as the samples in training dataset(predictor x in the model). Therefore, the evaluation result will have low error rate. However, when using this trained naïve bayes model to predict other samples, which have dissimilar features with the samples in training dataset, the decision boundary will not have accurate predictive effects on these samples, and thus will show poor prediction result.

The paper has proposed several approaches to avoid doppelganger effects of machine learning models. The overall process is to check the similarity of functional doppelganger and the proportion of functional doppelganger in the validation set and then reduce the negative effects of functional doppelganger on the model. The following methods are presented by the paper (Wang, Wong, & Goh, 2021).

In terms of identifying the functional doppelgangers, there are three methods. Firstly, we could calculate pairwise Pearson's correlation coefficient(PPCC) and then check the sample pairs with high PPCC values. The paper has concluded that PPCC data doppelganger acts as functional doppelganger in ML models, so we could use this conclusion. Secondly, we could perform cross-checks using metadata as a guide. The paper has given an example that from RCC data, samples derived from same class but different patients or technical replicates arising from the same sample should be dealt with. Hence, we could check the metadata to find the relationship between samples, which may indicate the data doppelganger. Finally, we could identify functional doppelganger directly by looking for subsets of a validation set that are predicted correctly without using ML models.

After identifying the function doppelganger, we should consider how to eliminate the doppelganger effects. There are three methods proposed in the

paper (Wang, Wong, & Goh, 2021). Firstly, after identifying the data doppelgangers, we could place them together in the training dataset, then the doppelganger effect is eliminated. This is a suboptimal solution, because it may lead to models that are not generalized well for lack of knowledge. Secondly, we could split the training and validation data based on individual chromosomes, or could use different cell types to generate the dataset. However, this method is difficult in practice. Thirdly, the data doppelganger could be removed directly to mitigate the effects. However, if the datasets contain a large proportion of data doppelgangers, then direct removal will reduce the data to an unusable size. Finally, we could trim the data by removing variables contributing strongly to doppelganger effects. This method may not be effective because the doppelganger effects are complicated that may not be determined by only part of the variables. The paper also argues that the most robust method is to check the independent validation(divergent validation) by involving as many datasets as possible.

In conclusion, each method listed above is not perfect, and functional doppelgangers are sometimes ignored during data analysis process, especially when dealing with a large amount of dataset in health and medical science. Therefore, it is better to combine several methods together as an essential procedure of data preprocessing, in order to avoid the doppelganger effects when conducting cross-validation in ML training.

## References

Wang, L. R., Wong, L., & Goh, W. W. B. (2021). How doppelgänger effects in biomedical data confound machine learning. *Drug Discovery Today*. doi:https://doi.org/10.1016/j.drudis.2021.10.017