Question 3

In this paper, the authors discuss how the doppelganger effect appears in biomedical data and analyze the implications of this problem for machine learning. The authors suggest that the doppelganger effect may affect the accuracy and reliability of machine learning models, and suggest using more diverse training data, cross-validation and data enhancement techniques to avoid the doppelganger effect. In this paper, the authors discuss how the doppelganger effect appears in biomedical data and analyze the implications of this problem for machine learning. The authors suggest that the doppelganger effect may affect the accuracy and reliability of machine learning models, and suggest using more diverse training data, cross-validation and data enhancement techniques to avoid the doppelganger effect. Through this article, I propose some of my own understanding of the doppelganger effect.

The doppelganger effect was first used to refer to a psychological phenomenon in which people feel that a shadow, reflection, or alter ego of themselves appears in front of them in certain situations. In machine learning models, the doppelganger effect may refer to the fact that the model produces the same or similar output for some input data, even though the input data varies greatly in content. Suppose a machine learning model is trained to classify images, some of which contain dogs and others contain cats. If the model produces the same output "cat" for two very different images, there may be a doppelganger effect in the model. This situation may be caused by the lack of sufficient diversity in the training data of the model or some problems in the model itself.

The doppelganger effect is not unique to biomedical data. The same is likely to happen in other areas of data. For example, when predicting consumer interest in a product, a machine learning model may also have a doppelganger effect if the model produces the same output for consumers of different ages, genders, and regions. In summary, the doppelganger effect is a phenomenon that can be seen in machine learning models and may be seen in various types of data, including biomedical data.

When developing machine learning models related to health and medicine, the doppelganger effect can be avoided in the following ways:

1. Use training data with sufficient diversity: To avoid the model learning only

part of the characteristics of the training data, the training data should be as diverse as possible. This helps the model learn more information to better predict unknown data.

2. Error of analysis model. When the prediction results of the model do not match expectations, the model should be tried to analyze the error. This can help you understand what is wrong with your model so you can better tune it.
3. Use multiple metrics. In order to better evaluate the performance of the model, a number of different evaluation metrics should be used. This can help you get a more complete picture of how the model is performing, so you can better determine if you need to tweak the model further.
4. Make adjustments to the model. If you find some problems with the model, try adjusting the parameters or structure of the model to improve its accuracy and reliability.
5. Use professional data cleaning tools. When training models with real-world data, the data often has dirty data or missing values. Using professional data cleaning tools can help you deal with these issues, making training of your models much smoother.
6. Use cross validation. Cross-validation is a common model evaluation method that helps you better evaluate the performance of your models. With cross-validation, you can better understand how the model behaves on different data sets, so you can better determine if you need to adjust the model.
7. Use data enhancement techniques. Data enhancement is a technique that generates new data by transforming training data. Using data enhancement can help you increase the diversity of your training data, thereby improving the accuracy and reliability of your models.
8. Use the right model type: When choosing a machine learning model, you should consider the scope and performance of the model. For example, tasks

The doppelganger effect occurs in other domains as well. For example, in the fields of gene sequencing and metabolomics, doppelganger effect still occurs in machine learning models, but it is essentially caused by the lack of sufficient diversity in the training data of the models or some problems in the models themselves.

In quantitative terms, suppose a machine learning model is trained to predict some target variable. If there is not enough diversity in the training data of the model, the model may only learn some features in the training data and ignore other features. In this case, if the model encounters a new data point during prediction that has very similar characteristics to some data points in the training data, but whose target variables are completely different, then the model may have a doppelganger effect. This means that the models may produce very different predictions for similar data points, which affects the accuracy and reliability of the models. In addition, the doppelganger effect can be caused by some problems with the model itself. For example, if there is underfitting or overfitting in the training process of the model, or the complexity of the model is too high or too low, then the model may also have a doppelganger effect when predicting new data points.

Reference

Wang LR, Wong L, Goh WWB. How doppelgänger effects in biomedical data confound machine learning. Drug Discov Today. 2022 Mar;27(3):678-685. doi: 10.1016/j.drudis.2021.10.017. Epub 2021 Oct 28. PMID: 34743902.

Wang LR, Choy XY, Goh WWB. Doppelgänger spotting in biomedical gene expression data. iScience. 2022 Jul 19;25(8):104788. doi: 10.1016/j.isci.2022.104788. PMID: 35992056; PMCID: PMC9382272.

Wang LR, Fan X, Goh WWB. Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier. STAR Protoc. 2022 Oct 26;3(4):101783. doi: 10.1016/j.xpro.2022.101783. PMID: 36317174; PMCID: PMC9617193.