Causes of doppelgänger effects and areas of presence

Qianruo Zhang February 1st, 2022 1547 words

Executive Summary

Machine learning has a wide range of applications in areas such as data mining, computer vision, biometric recognition, search engines, and medical diagnosis. However, the high similarity of the validation and training sets may still lead to an inflated prediction result. This report will summarise the possible causes of the effects based on machine learning and the characteristics of doppelgänger effects, as well as their representation in protein interactions, finance, and time-series data. Finally, two suggestions for improvement are made from a visualization and time-series perspectives.

1.0 Introduction

Machine learning is the study of how computers can simulate or implement human learning behavior to acquire new knowledge or skills and reorganize existing knowledge structures to continuously improve their performance, and it is the fundamental way to make computers intelligent. However, the reason why a model, after being adapted, does not perform well for test data, is because doppelgängers effects make the validation effect better than the testing effect. From the validation point of view, the aim of the validation set is to verify the accuracy of the training model, so the validation set is hopefully different from the training set. However, due to particular data processing, different data distribution, and the way the data is selected, the validation data and the training data may have high similarities or even be duplicative. Consequently, though the validation results show a good fit of the model, the prediction results of the test data are less accurate. This paper will sort out the possible causes of the doppelgänger effects, then give examples of areas and data types that may have doppelgänger effects, and finally make a few suggestions on how to avoid or check the doppelgänger effects.

2.0 Causes of doppelgänger effects

Data doppelgängers can lead to models that appear to be accurate but become inaccurate when used to make predictions in the real world. This refers to inadvertently putting future information or information with a high degree of similarity to the validation set into the training set during data collection and data processing. When this information is introduced into the training model, the model often performs very well, but the accuracy of prediction results is greatly low.

Combined with machine learning knowledge, and its use in the biomedical field, it is believed that the doppelgänger effects may commonly arise for several reasons, but not limited to the following ones.

2.1 Data processing

The training data is processed in the same way as the validation data, but not in the same way as the test data. Alternatively, when the training set is processed in a way that depends on the results of validation set data, information leakage ("double-dipping") occurs, which can lead to dependencies between examples and interfere with the utility of the test set for evaluating model performance. For example, data is normalized in both the training set and the validation set, but the prediction data is not normalized.

2.2 Distributional differences

There are distributional differences between the training, validation, and test data[1]. For example, the training and validation sets have the same distribution, while the test set has a different distribution.

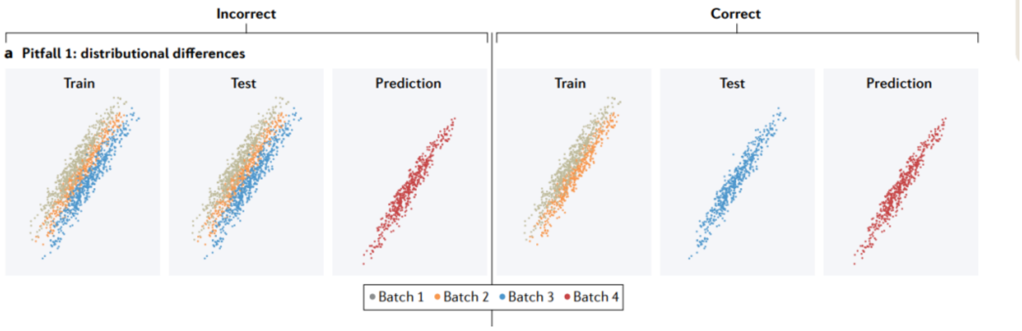

Fig. 1 Data distribution

In the figure 1, Differences in distribution may come from different sources, such as batch effects. If the training and test sets are a mixture of examples from different batches (left), the performance of the test set will be much higher than that of the new batch. To fit a model suitable for use with a new batch, the training and test sets should consist of different batches (right). In this case, better performance should be expected from the cross-validation than from the test set. As the test and validation sets differ in distribution, the relationship between features learning during model fitting and the results may not hold in prediction.

2.3 Confounding

In imaging, it is difficult to identify confounders even though they are in front of our eyes. For example, background scenery confounding predicts animal types or radiographic scanner type confounding predicts hip fracture. Confounding in genetic studies can arise from unmodelled environmental factors, population structure, and other factors. And in biomedicine, a common example is inadvertent confounding of data by sampling or treating samples with different outcomes, such as sick versus healthy or different treatment groups, in different batches.

3.0 Data doppelgängers in different data types

The impact of the doppelgänger effects on biomedical data is more significant and its relative studies are numerous. In addition to the examples cited by Li Rong Wang, Limsoon Wong, there is also a study on the relationship between prediction accuracy of protein interactions and sample reproducibility of datasets by Ji Limin[1]. She compared humans (Human) and yeast (S. cerevisiae) from the DIP database, the BIND database, and the bioGRID database. In the experiments, the authors constructed positive and negative data sets for modeling, where positive data sets refer to protein pairs that interact with each other and negative data sets refer to protein pairs that do not interact with each other. From the experimental results, we can see that the accuracy of the prediction varies with the repetition rate of the protein samples in the positive and negative datasets, and as the repetition rate of the protein samples in the dataset changes from high to low, the corresponding prediction accuracy also changes accordingly.

For the pairwise Pearson's correlation coefficient (PPCC) data, different variables in the training and validation sets that are highly correlated may also lead to data doppelgängers. Variables that are highly correlated with each other give the same information to the model, and hence, it becomes unnecessary to include all of them for our analysis. For example, in the dataset of housing price in Kaggle[3], the training set contains a feature "GarageYrBit" and the validation set contains "YearBuilt ", their correlation coefficient is 0.84, then we can imagine that these two variables will be correlated to some extent, and we would see this high correlation even if we pick up an unbiased sample of the data. By mapping a correlation heat map of the features in the training and validation sets, we can see that there are many highly correlated features in the two datasets, and these highly correlated data may lead to incorrectly assuming that the model performs well. These highly correlated features may lead us to believe that the model is predicting well, resulting in that the model adapted by the validation set will have higher prediction results than the true results.

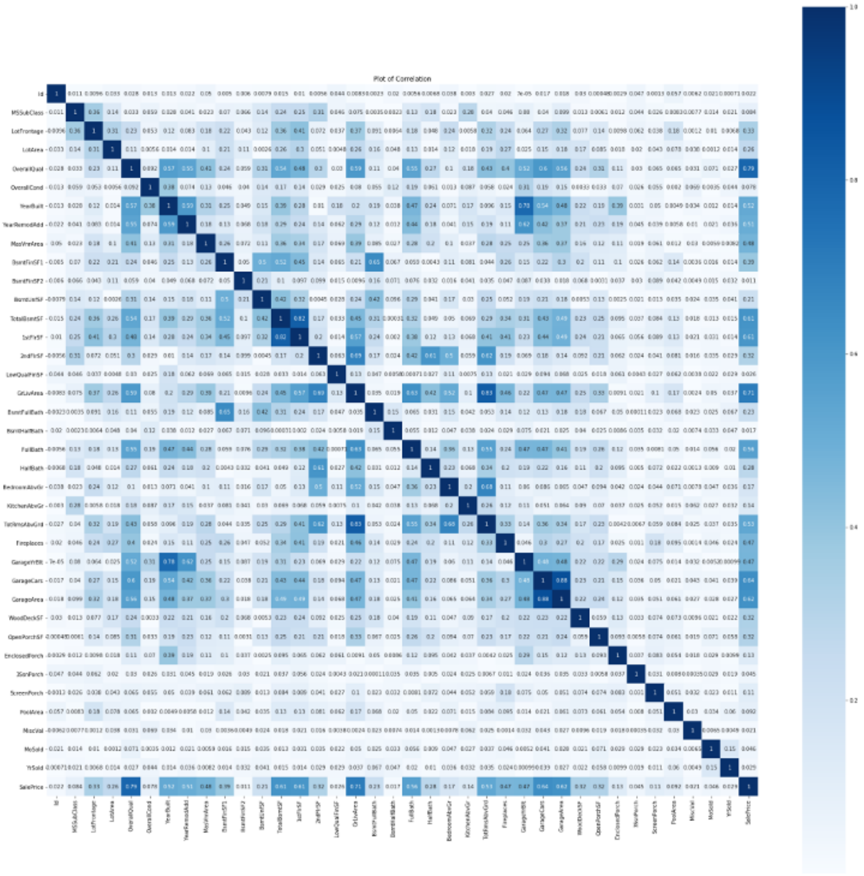

Fig. 2 Heat map of data correlation

In addition, the doppelgänger effects may be more pronounced for data with time-series properties, such as sales, stock, temperature, or credit data. For example, data on the sales volume of different goods provided by 1C Company, the largest software company in Russia, contains data on the sales of different categories of goods in different shops in Russia from January 1, 2013, to October 31, 2015[4]. Therefore, when dividing the training and validation sets, the data for the same sales cycle may be extremely similar, and we may also treat this PPCC data simply as a periodic feature when training the model. If the root mean squared error (RMSE) is used as the evaluation criterion for the results, the RMSE of the test set is predicted using the machine learning model adapted according to the validation set, and the RMSE of the prediction results is shown in table 1. It can be found that because of the similarity at the temporal level between the data in the validation and training sets, the RMSE of the test data prediction is larger.

Table. 1 RMSE of prediction results

Method	RMSE
Ramdom Forest Regression	0.5814
KNN Regression	0.6032

4.0 Recommendations

As for distributional differences, ideally, the marginal distribution of results and features should be checked. However, in a prediction setup, the outcome is usually unknown, and therefore we can solely evaluate the feature distribution. Visualization is a simple way, we can draw a scatter plot, or compare histograms of feature values. More sophisticated methods are using statistical test methods to detect differences in distributions: for example, the binomial test for binary features, the Kolmogorov-Smirnov test for univariate continuous features, or the Maximum Mean Discrepancy for multivariate continuous features[2].

For data with time-series characteristics, both exploratory analyses of the data and correlation analysis can be used to check for possible doppelgänger effects. Also, there are several ways to avoid doppelgänger effects. Firstly, relevant processing methods such as data information extractions (i.e. feature selection, outlier removal, coding, feature scaling, dimensionality reduction, etc.) are only performed on the training dataset or the training set for cross-validation, while independently within each cycle for cross-validation. Secondly, doppelgänger data is removed as appropriate and not used when there is a suspicion that a particular validation set data has a high likelihood of being similar to the training data.

5.0 References

[1]纪丽敏. 基于机器学习的蛋白质相互作用预测精度与数据集关系的研究[D]. 华南理工大学, 2013.

[2]Navigating the pitfalls of applying machine learning in genomics[J]. Nature Reviews Genetics. https://blog.csdn.net/weixin_45822007/article/details/121804923

[3]https://www.kaggle.com/anupahuje1/housing

[4]https://www.kaggle.com/c/competitive-data-science-predict-future-sales