Question 3

In machine learning approach, the training data set is the general term for the samples used to create the model, while the test or validation data set is used to qualify performance. The danger of the Doppelganger effect, known as the high similarity between test data and training data is that the model might be changed (e.g., adding another layer, and/or adding more units to an existing layer) because it provides a better result on test data. When altering the model in response to observations of the test error, it risks overfitting to data.

Besides the reported doppelganger effect observed in applying machine learning models in drug development [1], other field in modern bioinformatics like metabonomics also report similar problem. In one notable case, Tusharkanti, Weiming, Debashis and Katerina performed a detailed evaluation of predictive modeling for metabolomics data [2]. Their work revealed that the performance of these systems might be overstated due to the reported model validation. In specific, model validation presented that the top 5 metabolites for all classifiers using the test dataset were exactly the same as the training dataset. This would therefore weaken the strengthen the corresponding model in its performance in other data and in future predictions.

Doppelganger effect may also be related to the problem of overfitting, which the trained model performs well on the training set but performs poorly on the test set. Overfitting the training data without considering the generalization ability. Over-fitting of the model may be caused by: (1) the data is noisy (2) the training data is insufficient, and the limited training data (3) the model is very complicated due to the excessive training of the model. Resolving these causes can ultimately avoid the problem of doppelganger effect during the development of machine learning models for health and medical science.

First, noisy data. Machine learning approach is for finding a hypothesis space, which is to search for a set of parameters in the model parameter space to minimize the loss function, that is, to keep close to the real hypothesis model, and the real model can only be obtained by knowing all the data distribution. Often it is to find the optimal model that minimizes the loss function when the training data is limited, and then generalize the model to all other parts of the data. This is the essence of machine learning.

Assume that the overall data is as shown in Figure 1 below: (Assume overall data distribution satisfies a linear model y = kx+b)

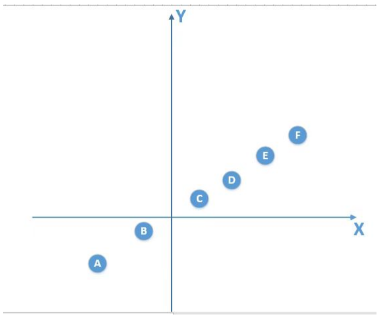


Figure 1

Part of the data obtained with noise, as shown in Figure 2 below: (noise – red data points)

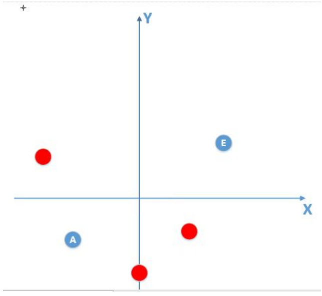


Figure 2

The above trained data points do not form a linear model (a standard model that is satisfied under the overall data distribution). For example, the trained model is as Figure 3 below:

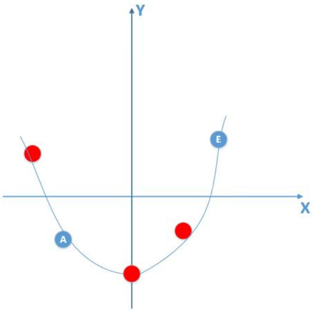


Figure 3

Then this noisy trained model can achieve a loss function value of 0 through continuous training on the training set but holding this model to the real overall data distribution (satisfying the linear model) de-generalization, the effect will be very poor because a nonlinear model is used to predict the true distribution of a linear model, the obvious and easy-to-obtain effect is very poor, and then over-fitting occurs.

Second, when the training data is insufficient, even if the obtained training data has no noise, the trained model may have over-fitting phenomenon

Suppose the overall data distribution is as the left and the obtained trained data is as the right as Figure 4: (Training data only got A and B)

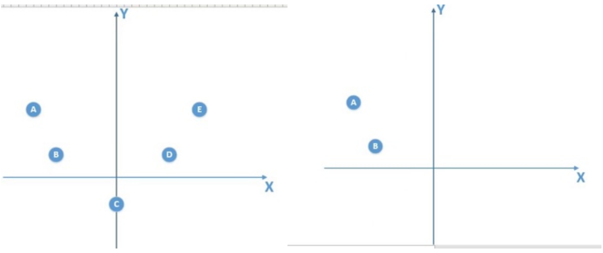


Figure 4

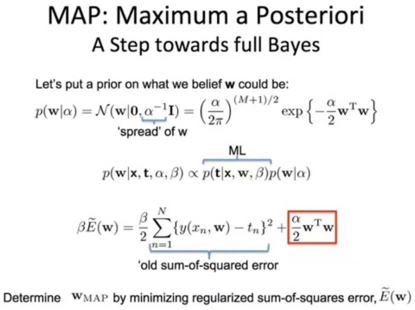
Then from this training data, the model obtained is a linear model. By training more times, a linear model with a loss function of 0 can be obtained in the training data. Use this model to generalize the real overall distribution data (in fact, it satisfies the quadratic function model). Obviously, the generalization ability is very poor, and over-fitting occurs.

Third, excessive training of the model leads to a very complicated model. Excessive training of the model makes the model very complex, and it can also lead to overfitting.

When training the training data, if the training data is over-trained and the training data is completely fitted, the resulting model may not be reliable.

For example, in the noisy training data, if the training is over-trained, the model will learn the characteristics of the noise, which will undoubtedly cause the accuracy of the real test set without noise to decrease.

To solve overfitting together with Doppelganger effect, regularization can be utilized to add a priori to the model parameters, so that the model complexity is small, and the input disturbance to noise and outliers is relatively small. Take the regularization term and loss function as l_2 norm as an example, as Figure 5 below:



MAP: Maximum a Posteriori

A Step towards full Bayes

Let's put a prior on what we belief **w** could be:

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \underbrace{\alpha^{-1}}_{\text{'spread' of w}}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto \overbrace{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)}^{\text{ML}}p(\mathbf{w}|\alpha)$$

$$\beta\widetilde{E}(\mathbf{w}) = \underbrace{\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2}_{\text{'old sum-of-squared error}} + \boxed{\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}}$$

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$

Figure 5

It is equivalent to add a zero-mean Gaussian distribution prior with a covariance of 1/alpha to the model parameter w. For alpha =0, that is, without adding regularization constraints, it is equivalent to the Gaussian prior distribution of the parameters has infinite covariance, then this prior constraint will be very weak, in order to fit all the training data, w can be changed Be arbitrarily unstable. The larger the alpha, the smaller the prior Gaussian covariance, the more stable the model, and the smaller the relative variance. Joining regularization is to make a tradeoff between bias and variance.

Another way to address the doppelganger effect is to create a default setting system which resembles the idea of positive and negative control in the field of biochemistry. To elaborate, dataset based on a drug with well-known mechanism with its target can be used for building up positive control and negative controls for training and validation data sets. Dataset for drug testing on sensitive cell lines can be considered as positive control and dataset for drug testing on insensitive cell lines can be considered as negative control. These two controls can then be used as the default setting system and implemented into machine learning model for the purpose checking of doppelganger effect and improving the model.

Reference List:

1. Wang, L. R., Wong, L. & Goh, W. W. B. (2021) How doppelganger effects in biomedical data confound machine learning, Drug Discov Today.

2. Ghosh, T., Zhang, W., Ghosh, D. & Kechris, K. (2020) Predictive Modeling for Metabolomics Data, Methods Mol Biol. 2104, 313-336.