A mathematical example of identifying partially redundant features using trend-constrained DNN

By Leixin Ma, MIT PhD 2021

We first verify that the proposed machine learning architecture is able to differentiate the importance of partially correlated (redundant) features using a designed mathematical example.

We consider a nonlinear relationship between output variable *Y* and input feature (X_1, X_2) as,

$$Y = \begin{cases} \sin\left(\pi X_1\right) + 1 + 0.01\varepsilon & X_1 \le 0.1 \\ \sin\left(\pi X_2\right) + 1 + 0.01\varepsilon & X_2 > 0.1 \end{cases} \tag{1}$$

Where X_1 is uniformly sampled from 0 to 0.4, i.e., $X_1 = U(0, 0.4)$, and $X_2 = X_1 \ (X_1 > 0.1), \ X_2 = 0 \ (X_1 \le 0.1)$. It can be noticed that the input feature set (X_1, X_2) are partially correlated. A feature selection technique without prior knowledge may not be able to distinguish the difference between X_1 and X_2.

However, if prior knowledge about the effect of X_1 on *Y* is known, such as X_1 can only contribute to the increase of *Y*, then the X_2 is found to be important, in that it contributes to the decreasing of output *Y*. To embed the prior knowledge of increasing trend between X_1 and *Y* in the neural network, the optimization objective is modified as,

$$w* = \arg\min\left\{\left(1-\theta\right)E_{pred} + n\theta E_{trend}\right\} \tag{2}$$

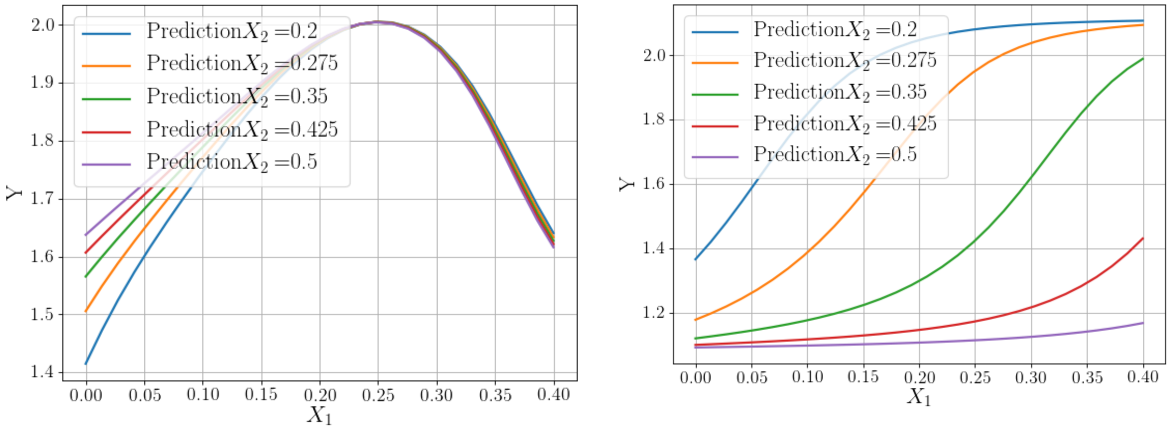
Where $E_{pred} = MAPE = \frac{100\%}{T}\sum_{i=1}^{T}\left|\frac{y_i\left(X_o, X_t\right) - \hat{y}_i\left(X_o, X_t\right)}{y_i\left(X_o, X_t\right)}\right|$,

$$E_{trend} = \sum_{i=1}^{T} Sig\left[\frac{y_i\left(X_o, X_t\right) - \hat{y}_i\left(X_o, X_t\right)}{Ty_i\left(X_o, X_t\right)/1000}\right]$$

Where X_o and X_t represents the unperturbed and perturbed feature, respectively. In this case, $X_t = X_1$. While each input feature X_1 is being randomly increased by $0.1 + 0.1\varepsilon$, where is a uniformly random variable between 0 and 1. *T* is the size of the original training data, which is 320. In this paper, the Reynolds number and the damping coefficient of each measured data points were perturbed four times, hence n was set to 4. The DNN predictions of each perturbed data $\hat{y}_i\left(X_o, X_t + \Delta X_t\right)$ are compared with the measured output $y_i\left(X_o, X_t\right)$ before perturbation.

A DNN with three hidden layers was constructed. Each hidden layer was made up of 15 neurons with a sigmoid activation function. 80% of the data were used as the training data, while the rest were used as the test data. The neural network optimization was conducted to minimize the regularized total error of the training data via the mini-batch Adaptive Moment Estimation (Adam), which is a commonly used variant of stochastic gradient descent. During neural network training, the batch size was 32 and learning rate was set as 0.005. A total number of 10000 iterations were conducted to optimize the unknown weights, w, connecting the neurons.

Figure 1 compares the predicted relationship between X_1 and Y for several fixed values of X_2 when both features X_1 and X_2 are input to the neural network. Without imposing the prior trend constraint, the relationship between X_1 and Y is very nonlinear, while after a certain level of trend constraint is imposed, the relationship between X_1 and Y becomes monotonic.



(a) Without prior trend constraint ($\theta = 0$) (b) With prior trend constraint ($\theta = 0.3$)

Figure 1 The machine learning predicted relationship between X_1 and Y for fixed X_1

Next, we assess the importance of feature X_2 by comparing the prediction error (MAPE) using either feature X_1 only or both features X_1 and X_2 for the test data, as shown in Figure 2. When the constraint is weak, the MAPE with and without feature X_2 are almost identical, which suggests that X_2 is a redundant parameter. However, if the prior physical constraint is strong, the prediction with only X_1 is no longer able to give accurate predictions. The comparison shows the effectiveness of trend constrained feature selection in distinguishing the effect of partially correlated input features on the target output.

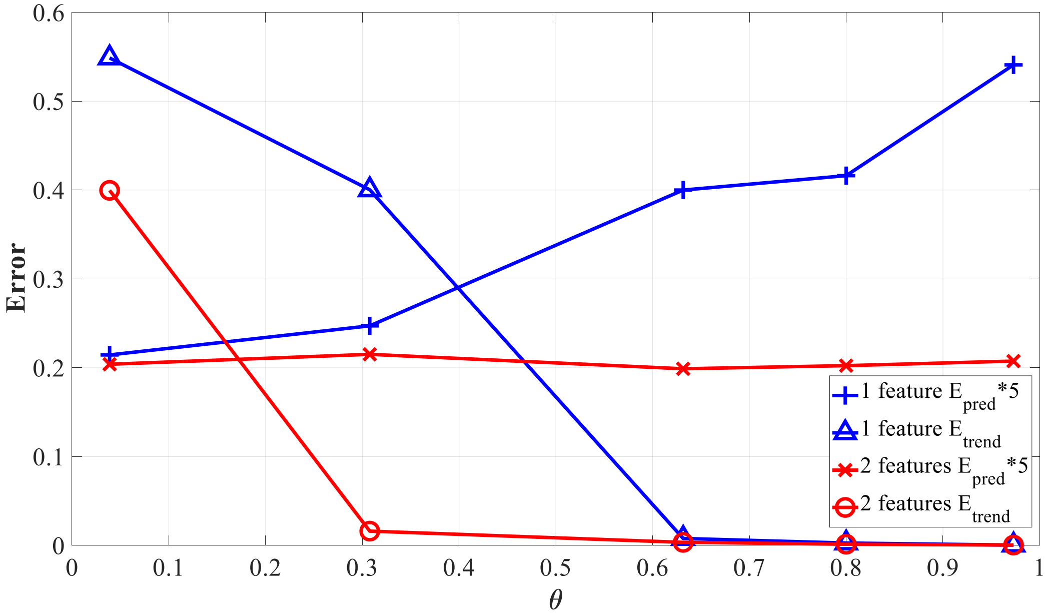


Figure 2 The variation of MAPE and the loss for trend constraints with different levels of hyper-parameters θ (the errors corresponds to the test data)