OVERFITTING

Akalanka Galappaththi
Board of Study in Statistics and Computer Science

Using models to capture relationships in data is trending in many disciplines. Samples were used to identify the hidden characteristics in a population. Having few observations in a sample and many features describes patterns that are only available in the sample itself, rather in population. This is a situation a model has been overfitted to the sample data (training data).

Overfitting is a common problem occurred in creating models for training data. This may be due to using a large number of features or applying an unnecessary large polynomial model. Overfitting can be a serious problem since it seems to have a good fit on training data and result erroneous output for new data [1].

For example in a regression model that identifies the relationship between a dependant variable and predicting variables, it is possible to find a model that best fit to the given dataset (the training dataset) and then predict the value for the dependant variable for a given set of independent variables. Let's have a look at three different models to understand the concept.

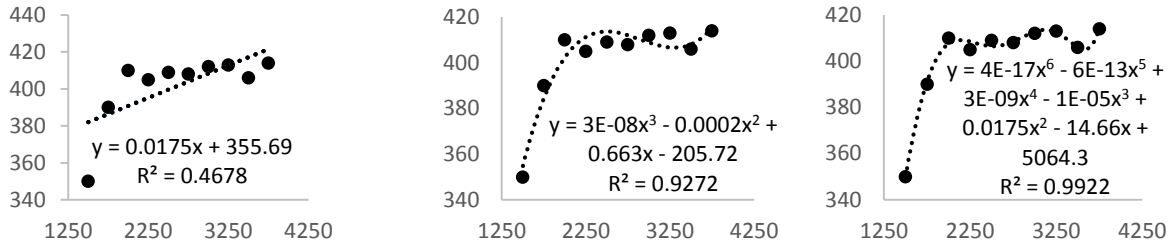

Figure 1: Three regression models for same data. (a) Linear model doesn't have a good fit for the data. (b) Third order polynomial model has a good fit for the data. (c) Sixth order polynomial model with extremely good fit compare to (b).

There are three different models (a linear, a third order polynomial and a sixth order polynomial respectively) has been fit to the same dataset. The equation of the model and R^2 (coefficient of determination) are also displayed on each plot. By considering the R^2 we can see that the sixth order polynomial model has the best fit to the data. It can be seen that linear model is underfitted to the dataset. However, the third order polynomial model also has a better fit to the dataset than linear fit. Even though the measure of R^2 provides the goodness of fit of the model to the data at present it doesn't provides how well the model will perform on data not presented [2].

Assume a classification problem that classifies two different classes of data. The following images show three different classifiers for the same data set.

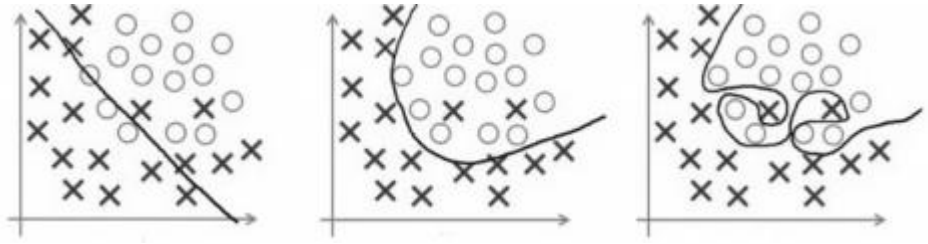

Figure 2: Classification models for a two classes of data (courtesy of https://affineanalytics.wordpress.com). (a) Classifier is a linear model. However it doesn't seems to classify well. (b) Classifier is a non-linear model. Even though two data points have been incorrectly classified the model has good performance. (c) Classifier is a higher order polynomial model. The model classify all data points correctly. However this may not perform well on a new data point.

In the problem of overfitting, model may fit to the data at present but fail to generalize the model to new data. Therefore model validation algorithms can be used to find the best model explains the data.

Having multiple features may cause overfitting. Therefore, it is recommended to drop the features that don't provide much information. Using a small sample tends to return biased model which can't generalize information of population [3]. In the case of having large number of features for a dataset, it is possible to reduce number of features by selecting features manually for the model. Or, it is possible to use regularization techniques so that all features are included to the model and the magnitude of them is reduced. Therefore, each feature contributes a bit to the model. So the effect of certain features to the dependant variable is minimized [3]. This is done by adding a penalty for parameters (except the interception). But regularization can cause underfitting when penalty for parameters are extremely large. In extreme cases, the effect of all features equal to zero, so that, only the interception will contribute in the model [4]. If we consider the linear regression example and try to penalize the parameters of the sixth order polynomial model with a high value, end result will be a horizontal line (y = 5064.3).

References

1. Friedman, J., & Popescu, B. E. (2003). Gradient directed regularization for linear regression and classification. Technical Report, Statistics Department, Stanford University.
2. Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water resources research, 35(1), 233-241.
3. Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine*, *66*(3), 411-421.
4. Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, *44*(1), 1-12.