Supervised learning

Dirk Valkenborg,[a] Melvin Geubbelmans,[a] Axel-Jan Rousseau,[a] and Tomasz Burzykowski[b]
Hasselt, Belgium, and Bialystok, Poland

As mentioned in a previous article,[1] unsupervised learning involves using datasets without clearly noticing the dependent (response) variable. Thus, in principle, all variables are treated in the same way. In supervised learning, 1 variable receives special focus. It is the response that determines the scope of the learning task variable. The response variable is the dependent variable, outcome, explained variable, output, label, or target. In principle, the response is not any different than the other variables. It is just a variable of interest for predicting its value when the other (explanatory) variables are given.

In general, supervised learning distinguishes 2 types of tasks:

1. Regression: in this task, the response is a continuous variable. For example, we want to predict the duration of the orthodontic treatment on the basis of the amount of crowding.
2. Classification: in this task, the response is a categorical variable. For example, we want to predict canine impaction on the basis of the position or angle of the tooth in the maxilla.

How is it possible that a computer can learn from a dataset? Algorithms mimic the learning process of humans via various mathematical techniques. The main principle here is to teach by example. Consider the following analogy. A person who has never seen a vehicle before is confronted with pictures of trucks, cars, and motorbikes. The labels are provided along with the pictures so the person knows exactly to which class each vehicle belongs. This person must learn rules from these examples to categorize the vehicle. Toward this aim, the person will try to look for characteristics in the pictures that are different between the 3 classes allowing them to understand the definition of a truck, car, and motorbike to explain why a picture belongs to a certain vehicle category. For instance, the number of wheels, the size of the vehicle, the presence of interior space, speed, weight, etc, are good characteristics to categorize these vehicles. However, the first issue with supervised learning tasks is whether the examples are representative enough to cover the learning objective. Consider that, by coincidence, we have provided the person with only pictures of red motorbikes, blue cars, and yellow trucks. From this limited dataset, the person could incorrectly reason that color is a good discriminating variable to classify a vehicle.

A second issue is that the person could overemphasize the examples and memorize all the vehicles presented to them. By memorization, the person cannot anticipate a new situation. Thus, the concepts learned have limited value as the person could only apply the knowledge correctly on this limited "learning" dataset. The latter 2 examples indicate the risk of overfitting. In such a case, the relationship inferred by a person or computer algorithm is not generalizable. Generalizability means that the concepts learned can be applied elsewhere in different situations within a different context. Thus, smart persons are people that can generalize well.

A similar thing can be said about machine learning (ML) models. However, we do not call these models smart but flexible. A flexible model can detect and use subtle patterns present in a dataset. In contrast, a rigid, inflexible model will only capture general trends in a dataset. The concept of flexibility is further explained in Figure 1. Consider a continuous explanatory variable (covariate) x and a response variable y. The dependence of the true value of y on x can be expressed by using the function f(x), denoted by the *line* in Figure 1. We can learn about this relationship via an experiment by controlling the covariate x and measuring the response y. The measurement is affected by uncertainty because of unobserved factors that influence the value of the response or just random errors caused by the measurement process. The term ε denotes this random error and is sometimes regarded as normally distributed. The *dots* surrounding the line in

[a]Data Science Institute and Center for Statistics, Hasselt University, Hasselt, Belgium.
[b]Data Science Institute and Center for Statistics, Hasselt University, Hasselt, Belgium; Department of Biostatistics and Medical Informatics, Medical University of Bialystok, Bialystok, Poland.
All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest, and none were reported.
This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.
Address correspondence to: Dirk Valkenborg, Data Science Institute, Hasselt University, Agoralaan 1, Building D, B-3590 Diepenbeek, Belgium; e-mail, dirk.valkenborg@uhasselt.be.
Am J Orthod Dentofacial Orthop 2023;164:146-9
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2023.04.010

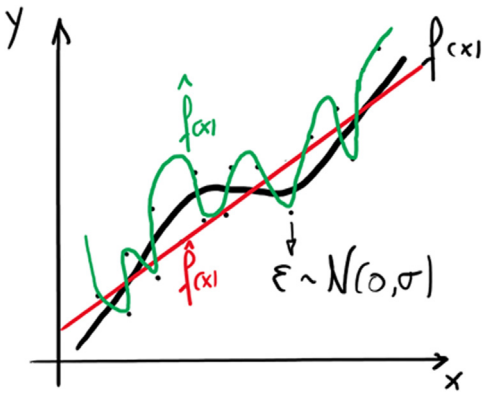

Fig 1. Hypothetical regression example. The *line* indicates the true relation between the covariate x and response y. The *dots* are measurements taken from this system and are influenced by random error. The *red* results from fitting a simple linear model to the *dots*, whereas the *red* results from fitting a flexible model to the data points.

Figure 1 are the observations from the experiment and are affected by this random error. Mathematically, the dependence of the observed response value y on x can then be presented as follows:

$$y = f(x) + \varepsilon$$

The main assumption in ML is that the unknown function f(x) can be approximated by presenting examples to the machine learner. The obtained approximation (estimate) of the unknown function is usually denoted by adding a caret to the function $\hat{f}(x)$. Ideally, the estimate is so good that we can use $\hat{f}(x)$ to predict the value of the response variable $\hat{y}$ for a particular value of x. Figure 1 presents 2 different ML models that are fitted to the data. The *red* presents predictions from an inflexible model, whereas the *green* presents a flexible one. The *green* model can follow the *dots* in Figure 1 accurately. In this case, the difference $(y - \hat{y})$ between the observed and the predicted response value is very small. Minimizing this difference (prediction error) is exactly the objective of the model fitting procedure. Note that this error should be minimized for all subjects in the dataset, and often this difference is squared such that underpredictions and overpredictions are not canceled out by differences in the sign. As a result, we arrive at the sum of squared errors by taking the sum of $(y - \hat{y})^2$ across all subjects, also known as a residual sum of squares.

In contrast, the *red* cannot accurately predict the response values, yielding a larger residual sum of squares. Based on this criterion, the *green* (flexible model) is preferred over the *red* (inflexible model). However, one could argue that, although the *red* is off most of the time, it can capture the global trend of the data better than the *green* without being unduly influenced by individual observations. Put differently, the *green*, corresponding to the more flexible model, is subject to the risk of overfitting.[2] Informally, the flexible model memorizes the example dataset, which may lead to bad generalizations. Be aware of a seemingly contradictory statement here, although the flexible model can accurately learn the presented data, this does not mean that the model is also flexible to translate the learned concepts to different situations. This lack of generalization and measures to prevent overfitting will be discussed in more detail in the next article in this series on ML.

Let us consider a clinical orthodontic example. The dataset of Konstantonis et al.[3] has been collected to support the orthodontist in their decision-making process on the basis of the measured explanatory variables (26 cephalometric, 6 models, and 2 demographics).

Ignoring, for now, the extraction or nonextraction treatment, it could be of interest to model the relationship between the Frankfort plane to mandibular incisor angle (FMIA), which can be treated as the response, and the covariate Incisor mandibular plane angle (IMPA). Do note that by selecting these 2 variables, we can expect a dependency as both angles are computed with respect to a common line. Therefore, the orthodontic applicability might be limited, but the covariates are well suited to explain regression as a toy example.

Many machine learners could be considered for this regression task. Here, we consider simple polynomial regression models, which try to describe the relationship between a continuous response variable and a covariate in an additive way by incorporating different powers of the covariate. In this case, a first-order polynomial offers optimal model flexibility using cross-validation, which will be introduced in the next article. The resulting estimated linear regression model has the following form:

$$\widehat{\text{FMIA}} = 124.99 - 0.68656 \times \text{IMPA}.$$

Figure 2 presents a scatter plot of the observed values of the response and the covariate along with the regression line (*black*) corresponding to the linear regression model.

The linear model may seem to be inflexible. However, using a more flexible, higher-polynomial model might lead to overfitting. To illustrate the issue, Figure 2 includes the result (*red*) of a model that uses the 10th-order polynomial of the covariate. Note that the *red* behaves unpredictably at the boundaries of the x-axis in that it follows the few data points near the boundaries too closely. The predictions provided by the model at the boundaries might

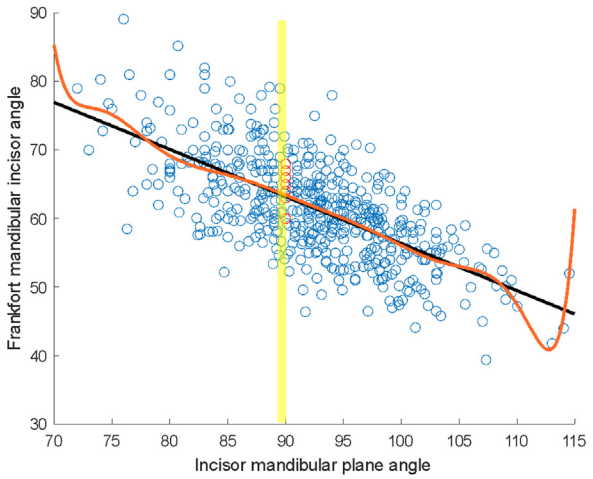

Fig 2. Relationship between the IMPA and FMIA of the subjects in the dataset of Konstantonis et al.[3] Each *circle* indicates a patient. The *line* is a simple model with only intercept and slope able to capture the overall trend in the dataset. The *red* is a 10th-order polynomial model behaving unpredictably near the boundaries of the x-axis. The highlighted bar indicates the 10-nearest data points around an IMPA of 90. This selection of 10 points will be used to calculate their FMIA values, resulting in an average FMIA of 64.8. This result of the KNN model can be contrasted with the result of the simple linear model, which yields a prediction for the FMIA of 63.2.

be problematic if the model is applied to a new set of observations. Again, this example illustrates that model flexibility is the capability of a machine learner to bend into some of the presented data points, and it is not about the capability of adjusting to different situations.

Thus, a flexible model can fit a particular dataset better than an inflexible one (rigid). However, one must evaluate which model can adequately generalize the learned concept to new, unseen situations.

Note that the dataset of Konstantonis et al[3] included, in fact, a binary response variable (ie, extraction or non-extraction treatment). It was interesting to build a model capable of predicting treatment on the basis of a subset of the measured explanatory variables for a new patient. This was a classification task. The principles discussed for regression also hold for a classification task. The major difference is that the calculation of the residual sum of squares or, equivalently, the error is replaced by the misclassification error. The misclassification error denotes the percentage of misclassified subjects in a dataset. Another term to indicate the performance of a classification model is accuracy which denotes the percentage of correctly classified subjects. A well-known classification model is, for instance, a logistic regression model. The model is an example of a generalized linear model. Generalized linear models will be discussed in more detail in one of the next articles in this series on ML. The misclassification error or accuracy is not always the best metric to describe the performance of the classification. Other metrics, such as sensitivity and specificity, can be more informative and will be explained later in this series on ML.

In the regression and classification examples discussed previously, the learning effort is materialized in a model that describes the underlying relationships in a compact mathematical manner. Such ML models are called eager learners as they replace the data set with a mathematical formulation. Another choice for a model could be a lazy learner. For example, a typical lazy learner is the K-nearest-neighbor (KNN) model.[4] A KNN model considers nearby subjects regarding their covariates and a particular distance measure. When a user-specified number of KNN subjects are found, the average of their responses is taken in the case of a regression task, or a majority vote is conducted in the case of a classification task. The flexibility of the model is controlled by K, the number of the nearest neighbors to be considered. KNN is a lazy learner because no formal description exists for the underlying relationship. Therefore, for each new prediction, we need to consult the data again, like a lazy student who needs to consult his textbook every time a question needs to be answered. Although the KNN model does not learn anything and adheres to very simple principles, it is often worth considering as it can be used as a baseline model compared with more advanced supervised methods. In Figure 2, the highlighted data points indicate the 10 nearest neighbors to an IMPA of 90. Averaging their FMIA values yields a value of 64.8. This result of the KNN model can be contrasted with the simple linear model, which yields a prediction for the FMIA of 63.2. The KNN method is also an intuitive ML model to exemplify the so-called curse of dimensionality when the data is of high dimensionality (ie, many covariates). These phenomena will be further elaborated on in an upcoming article in this series.

In contrast, instead of lazy, some ML models are very eager (eg, artificial neural networks, which will be discussed in more detail in one of the next articles in this series on ML) and include many explanatory variables or coefficients. As a result, many coefficients in the model formally describe the underlying relationship between covariates and the response. Such ML models are extremely flexible, prone to overfitting and may appear as black boxes, as no meaningful interpretation can be attached to their coefficients. Therefore, specialized methods are needed to explain those black boxes. This research domain is called explainable artificial

intelligence. It will be discussed in more detail in one of the next articles in this series on ML.

REFERENCES

1. Valkenborg D, Rousseau A-J, Geubbelmans M, Burzykowski T. Unsupervised learning. Am J Orthod Dentofacial Orthop 2023;163:877-82.
2. Burzykowski T, Rousseau A-J, Geubbelmans M, Valkenborg D. Introduction to machine learning. Am J Orthod Dentofacial Orthop 2023;163:732–4.
3. Konstantonis D, Anthopoulou C, Makou M. Extraction decision and identification of treatment predictors in Class I malocclusions. Prog Orthod 2013;14:47.
4. Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992;46:175-85.