Course: Machine Learning
Unit 1, Portfolio Evidence 2: Solution to most common problems in ML

Name: Victor Alejandro Moo Quintal
Group: IRC 9B
e-mail: 2009098@upy.edu.mx

Define the concepts of: Overfitting & Underfitting.

Overfitting

Overfitting is a recurrent challenge in training neural networks and is characterized by the model's propensity to fit the training data excessively. This phenomenon occurs when a neural network does not refine its problem-solving abilities during the training period but instead begins to learn random regularities present in the training dataset. As a result, overfitting is corresponding to the empirical observation that the error on the test set reaches a minimum, indicating the network's optimal generalization ability, before this error starts to increase again. Consequently, overfitted models describe random error or noise rather than the underlying data relationship, as seen in *figure 1*.

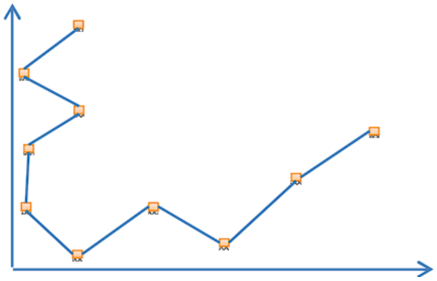

FIgure 1. Overfitting graph example. Source: Adapted from Allamy (2014).

Overfitting is typically associated with low bias and high variance estimators. This means that while such models may demonstrate high accuracy on training datasets, they may not generalize well to new, unseen data (Allamy, 2014).

Underfitting

Underfitting is the opposite of overfitting. It arises when the model fails to capture the inherent variability and patterns in the data. An example illustrating underfitting is when one attempts to train a linear classifier (represented by the equation $y = ax + b$) on data that follows a parabolic trend. Such classifiers, marked by high bias and low variance, lack predictive power and cannot map the training data accurately.

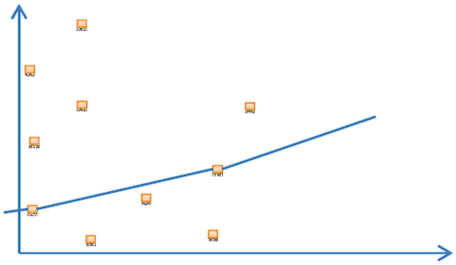

FIgure 2. Underfitting graph example. Source: Adapted from Allamy (2014).

This results from the attempt to utilize a model that is overly simplistic for the dataset in question, and therefore, the model will generally misrepresent the data, as seen in *figure 2* (Allamy, 2014).

Differences between overfitting and underfitting:

An overfitted model possesses an excess of parameters, allowing it to perform exceptionally well on the training data. However, this added complexity restricts the model's capability to predict future points accurately. The underlying issue is that the model becomes attuned to noise, hidden factors, and complex relations in the training data, rendering it less effective for future predictions.

In contrast, underfitting occurs when the model is too rudimentary, not capturing the data's inherent variability. A classic instance of underfitting is evident when a linear classifier is applied to a dataset that intrinsically follows a parabolic curve. Such models don't have predictive capabilities and misrepresent the training data (Allamy, 2014).

Define and distinguish the characteristics of outliers.

Outliers can be understood as the responses that deviate from what is usually expected. These deviations can sometimes resemble appropriate responses, making them difficult to detect (Cousineau & Chartier, 2010). Outliers can present themselves in both low and high extremes of a given measure. Specifically, when a measure is one-dimensional, such as IQ or response time, outlier responses can either be suspiciously small or large.

The responses on the left of the scale are termed as low-outliers while those on the right are called high-outliers. However, it's significant to note that many of these problematic responses might be intertwined with appropriate ones, making their detection challenging without a clear framework. This necessitates strategies either to accept outliers, mitigate their effects, or design experiments in a way that minimizes their occurrences. It's crucial to give due attention during the design phase to avoid accumulating excessive outlier responses, as designs cannot be rectified post by any analysis (Cousineau & Chartier, 2010).

The presence of outliers can notably affect statistical inferences, especially when the tests are based on means. The standard deviation, which gauges the typical fluctuation of examined processes, can be disproportionately influenced by outliers. For instance, in parametric tests, even a few high or low outliers can skew the mean response. This can further heighten the probability of a Type-I error if the outliers are not evenly distributed across different conditions. Conversely, the simultaneous presence of both low and high outliers can elevate the standard deviation, leading to an augmented likelihood of a Type-II error (Cousineau & Chartier, 2010).

Discuss the most common solutions for overfitting, underfitting and presence of outliers in datasets.

Overfitting and Underfitting solutions:

Penalty Methods:

Penalty methods apply additional constraints to the model in the form of a penalty to avoid overfitting, thereby controlling the complexity of the model.

Methods under Penalty Techniques:

1. Map Penalty: This imposes a penalty based on $P(H)$, a prior probability.

2. Minimum Description Length (MDL) Principle: It balances model fit with model complexity.

3. Structural Risk Minimization: It seeks to balance the empirical risk with model complexity.

4. Generalization Cross-Validation: A method to determine the performance of the model on new, unseen data.

5. Hold and Cross-Validation: A type of cross-validation where some data is held out for testing.

- Let E_{train} be our training set error and E_{test} be our test error. The real objective is to find a hypothesis h that minimizes E_{test}. However, E_{test} is not directly measurable. Instead, penalty methods find a penalty such that $E_{test} = E_{train} + Penalty$.

- The error function can be represented as: $E_{test} = E_{train} + \lambda\,(model\,complexity)$ where λ plays a crucial role in managing the bias-variance trade-off (Allamy, 2014).

Advantages:

-Training and pruning are performed simultaneously.

-Can help in reducing the number of free parameters in the network.

Challenges:

-Choosing the appropriate penalty factor λ can be tricky.

-Setting multiple constants for controlling the penalty, like λ_1 and λ_2 may need careful considerations based on the data's nature.

Early stopping:

Early stopping is a regularization method where training is halted before the model starts to overfit the training data. This method requires three datasets: training, validation, and test datasets (Allamy, 2014).

-Monitoring the error on the validation set is essential. Initially, both training and validation errors will decrease. When the model starts to overfit, the validation error will start increasing. Training should be halted at this point.

-Early stopping ensures the model remains linear and does not over-complicate, leading to overfitting. As weights increase, the capacity of the model grows (Allamy, 2014).

Advantages:

-Only requires training the network once.

-Direct performance measure applicable to the actual network parameters.

Challenges:

-Depends on the optimization method specifics.

-Some training data is only used for determining when to stop, which may seem wasteful.

Outliers treatments:

-Transformation approach: One solution is to make the data symmetrical through a non-linear transformation. Techniques such as Log-transform, square-root transform, and $arcsin$ transform are commonly used. However, it is the modified square root transformation that stands out for locating outliers in response time data (Cousineau & Chartier, 2010).

-Recursive and non-recursive approaches using adaptive criterion: Van Selst and Jolicoeur proposed a different approach by developing adaptive criteria based on sample sizes to address biases in reaction time data.

-Multiple regression: In multiple regressions, the relationship between predictor variables (X) and a dependent variable (Y) is considered. Outliers can be extreme in X, Y, or both. Detection becomes complex with more predictors. Multiple regression requires differentiating between outliers affecting the dependent variable (Y) and the predictors (X). Procedures for assessing outliers typically involve their removal to see how target estimates change, followed by an evaluation of their influence to decide on further action (Cousineau & Chartier, 2010).

-Nonlinear regression: Considering non-linear regressions, traditional outlier detection methods might not apply or could be vastly different from linear regression techniques. This divergence is especially evident in logistic and Poisson regressions (Cousineau & Chartier, 2010).

-Multivariate multiple regression: In situations with multiple dependent variables (Y) and no predictor or with multiple predictors and dependent variables, multivariate multiple regression is utilized. The assumption here is that the population is multinormal (Cousineau & Chartier, 2010).

Describe the dimensionality problem.

High-dimensional data is a common challenge in areas such as pattern recognition, data mining, and various data analysis applications. Such high dimensionality introduces complexities because, while more data often suggests a better representation of the problem, it does not always translate to more informative or discriminative data. In fact, in certain cases, increased dimensions may lead to redundant or even irrelevant information. Training with high-dimensional data without domain-specific background or meaning becomes particularly challenging. The absence of explicit meaning or background makes the application of domain knowledge hard, increasing the demand for data-driven dimensionality reduction techniques. Such high dimensionality, if not addressed, can impact negatively the efficiency and performance of learning algorithms, making dimensionality reduction a crucial process (Chao, 2011).

Describe the dimensionality reduction process.

-Redundancy Reduction and Intrinsic Structure Discovery: Multimedia research deals with naturally high-dimensional data, such as digital signals and videos. This data often contains redundancies features that don't necessarily add new information. Dimensionality reduction can help in removing such redundancies, thereby saving memory and transmission costs (Chao, 2011).

-Removal of Irrelevant and Noisy Features: After feature extraction, some features may not directly correlate with the task at hand. For example, a specific feature useful in face detection might be irrelevant in face recognition. DR techniques can help in eliminating such irrelevant features, ensuring only the most relevant ones are preserved (Chao, 2011).

-Feature Extraction: In tasks such as object recognition, DR plays a role in the feature extraction process. The goal is to represent information from the original high-dimensional data more compactly (Chao, 2011).

-Visualization: Visualizing high-dimensional data is challenging. Dimensionality reduction techniques, such as Isomap and LLE, can convert this data into 2D or 3D representations, aiding in understanding and interpretation (Chao, 2011).

-Computation and Machine Learning Perspective: For real-time applications or when computational resources are limited, DR becomes essential. Limiting feature dimensionality also ensures better generalization performance, especially with limited training samples (Chao, 2011).

Explain the bias-variance trade-off.

The bias-variance trade-off is a fundamental concept, it pertains to the dilemma faced when attempting to simultaneously minimize two sources of error: bias and variance. While bias arises from erroneous assumptions in the learning algorithm, variance results from the model's sensitivity to small fluctuations in the training set. The regularization coefficient, λ, in the penalty methods plays a significant role in controlling this trade-off, ensuring that the model doesn't lean too much towards either end of the spectrum (Allamy, 2014).

References:

Allamy, Haider. (2014). Methods To Avoid Over-Fitting And Under-Fitting In Supervised Machine Learning (Comparative Study).

Chao, W.-L. (2011). Dimensionality Reduction. Graduate Institute of Communication Engineering, National Taiwan University. Retrieved from http://disp.ee.ntu.edu.tw/~pujols/Dimensionality%20Reduction.pdf

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research, 3*. https://doi.org/10.21500/20112084.844.