Universidad Pólitecnica de Yucatán

Subject: Machine Learning

Grade and group: 9°B

Teacher´s name: Victor Alejandro Ortiz Santiago

Student´s name: Hernando Enrique Te Bencomo

Date: 15/09/ 2023

- Define the concepts of: Overfitting & Underfitting.

Overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models.

Underfitting when a model is too simple to capture data complexities. It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data. In simple terms, an underfit model's are inaccurate, especially when applied to new, unseen examples.

- Define and distinguish the characteristics of outliers.

Outliers are data points that significantly differ from the majority of the data in a dataset. They can be identified by their unusual values or patterns and can have a significant impact on statistical analyses and machine learning models. Outliers can arise for various reasons, such as data entry errors, measurement errors, or rare events. Here are some key characteristics that help define and distinguish outliers:

1. Unusual Value:
    - Outliers typically have values that are notably different from the central tendency of the data. They can be much larger or smaller than the majority of the data points.
2. Deviation from the Mean:
    - One common way to identify outliers is by examining how far a data point deviates from the mean (average) of the dataset. Outliers have a large distance from the mean in comparison to other data points.
3. Impact on Statistical Measures:
    - Outliers can significantly affect statistical measures like the mean, median, and standard deviation. For example, a single extremely high value can inflate the mean.
4. Skewed Distribution:
    - In the presence of outliers, the distribution of the data may be skewed. The tail of the distribution may be elongated in the direction of the outliers.
5. Influence on Relationships:
    - Outliers can distort relationships and patterns in data. In regression analysis, for instance, an outlier can disproportionately influence the slope and intercept of the regression line.
6. Context Matters:

- Whether a data point is an outlier may depend on the context of the analysis. In some cases, what appears to be an outlier in one context may not be considered an outlier in another.

7. Data Context:
    - Understanding the domain and the data collection process is crucial for identifying outliers. Some values that seem like outliers may actually be valid data points in certain contexts.
8. Visualization:
    - Data visualization tools, such as box plots, scatter plots, and histograms, can help identify outliers by providing a visual representation of the data's distribution and any unusual data points.
9. Statistical Tests:
    - Various statistical methods and tests, such as the Z-score, IQR (Interquartile Range), and Grubbs' test, can be used to quantitatively identify outliers based on their deviation from the norm.
10. Impact on Analysis:
    - Outliers can have a significant impact on the results of data analysis and modeling. It's essential to consider their presence and decide whether to remove them, transform the data, or use robust statistical methods that are less sensitive to outliers.

In summary, outliers are data points that stand out from the rest of the dataset due to their extreme values or unusual patterns. Identifying and handling outliers appropriately is essential to ensure the accuracy and reliability of data analysis and modeling efforts.

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

Underfiting:

- Increase model complexity.
- Increase the number of features, performing feature engineering.
- Remove noise from the data.
- Increase the number of epochs or increase the duration of training to get better results.

Overfiting:

- Increase training data.
- Reduce model complexity.
- Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
- Ridge Regularization and Lasso Regularization.
- Use dropout for neural networks to tackle overfitting.

- Describe the dimensionality problem.

Dimensionality reduction is the process of reducing the number of random variables in the data set under consideration by obtaining a set of principal variables.

Consider this scenario where you need many indicator variables in the data set to reach a more accurate result from the Machine Learning model, then tend to add as many features as possible at the beginning. However, after a certain point, the performance of the model will decrease with the increasing number of elements. This phenomenon is known as “the curse of dimensionality.” The curse of dimensionality occurs because sample density decreases exponentially with increasing dimensionality. When we keep adding features without increasing the number of training samples, the dimensionality of the feature space grows and becomes more and more sparse. Due to this sparsity, it becomes much easier to find a perfect solution for the Machine Learning model, which most likely leads to overfitting.. An oversized model would perform too well on the training data set to fail on future data and make the prediction unreliable.

- Describe the dimensionality reduction process.

Generally speaking, dimensionality reduction has two kinds: feature removal and feature extraction.

Feature Removal:

It is the removal of some variables completely if they are redundant with some other variable or if they are not providing any new information about the data set. The advantage of feature removal is that it is easy to implement and keeps our data set small, including only the variables we are interested in. But as a disadvantage, we could lose some information about the variables that we stopped evaluating.

Variable extraction:

it is the formation of new variables from old ones. Let's say you have 29 variables in a data set, then the feature extraction technique will create 29 new variables which are combinations of 29 old variables. PCA is the example of one such feature extraction methods.

Feature Selection

- Missing value ratio

Columns of data with too many missing values are unlikely to contain much useful information. This allows data columns with a ratio of missing values greater than a certain threshold to be removed. The higher the threshold, the more aggressive the reduction.

- Low variance filter

Like the previous technique, data columns with few data changes contain little information. In this way, all data columns with a deviation less than a certain threshold can be removed. Note that the variance depends on the range of columns and therefore needs to be normalized before applying this technique.

- High correlation filter

Columns of data with very similar trends are likely to also contain very similar information, and only one of them will be sufficient for classification. Here we calculated the Pearson correlation coefficient between numerical columns and the Pearson chi-square value between nominal columns. For the final classification, we only retain one column from each pair of columns whose pairwise correlation exceeds a given threshold. Note that the correlation depends on the range of columns and, therefore, it is necessary to normalize it before applying this technique.

- Random forests

Random forests are useful for column selection, in addition to being effective classifiers. Here we generate a large, carefully constructed set of trees to predict the target classes and then use the usage statistics of each column to find the most informative subset of columns. We generate a large set of very shallow trees, and each tree is trained on a small fraction of the total number of columns. If a column is often selected as the best split, it is very unlikely that it is an informative column that we should keep. For all columns, we calculate a score as the number of times the column was selected for splitting, divided by the number of times it was a candidate. The most predictive columns are those with the highest scores.

- Removing backward features

In this technique, in a given iteration, the selected classification algorithm is trained on n input columns. We then remove one input column at a time and train the same model on columns n-1. The input column whose removal has produced the smallest increase in the error rate is removed, leaving us with n-1 input columns. The sorting is then repeated using n-2 columns, and so on. Each iteration k produces a model trained on n-k columns and an error rate e(k). By selecting the maximum tolerable error rate, we define the smallest number of columns necessary to achieve that classification performance with the selected Machine Learning algorithm.

- Forward Sequential Feature Construction

This is the reverse process of backward feature removal. We start with a single column, progressively adding one column at a time, that is, the column that produces the largest increase in performance. Both backward feature elimination and this algorithm are quite expensive in terms of time and computation. They are only practical when applied to a data set with a relatively low number of input columns.

Linear dimensionality reduction methods

The most common and well-known dimensionality reduction methods are those that apply linear transformations, such as the following.

- Factorial analysis

This technique is used to reduce a large number of variables to a smaller number of factors. The observed data values are expressed as functions of several possible causes to find the most important ones. The observations are assumed to be caused by a linear transformation of the lower-dimensional latent factors and by added Gaussian noise.

- Principal Component Analysis (PCA)

It is a statistical procedure that orthogonally transforms the original n numerical dimensions of a data set into a new set of n dimensions called principal components. As a result of the transformation, the first principal component has the greatest possible variance. Each subsequent principal component has the greatest possible variance under the constraint that it is orthogonal to the preceding principal components, that is, it is not correlated with them. Keeping only the first m < n principal components reduces the dimensionality of the data, while preserving most of the information in the data, that is, the variation in the data.

- Linear Discriminant Analysis (LDA)

Project the data in a way that maximizes class separability. Examples of the same class are placed close together in the projection. Examples of different classes are placed very far away by the projection.

Nonlinear dimensionality reduction methods

Nonlinear transformation methods or manifold learning methods are used when the data is not in a linear space. It is based on the hypothesis that, in a high-dimensional structure, the most relevant information is concentrated in a small number of low-dimensional manifolds. If a linear subspace is a flat sheet of paper, then a rolled sheet of paper is a simple example of a nonlinear manifold. Some of the most popular learning methods are as follows.

- Multidimensional Scale (MDS)

A technique used to analyze the similarity or dissimilarity of data as distances in a geometric space. I project the data to a lower dimension so that data points that are close to each other, in terms of Euclidean distance, in the higher dimension are also close in the lower dimension.

- Isometric Feature Mapping (Isomap)

It projects the data to a lower dimension while preserving geodesic distance, rather than Euclidean distance as in MDS. Geodesic distance is the shortest distance between two points on a curve.

- Locally Linear Embedding (LLE)

Recovers the global nonlinear structure of linear fits. Each local patch of the collector can be written as a linear, weighted sum of its neighbors with sufficient data.

- Hessian Maps (HLLE)

I project the data to a lower dimension while preserving the local neighborhood like LLE, but uses the Hessian operator to improve this result and hence the name.

- Spectral embedding (Laplacian maps)

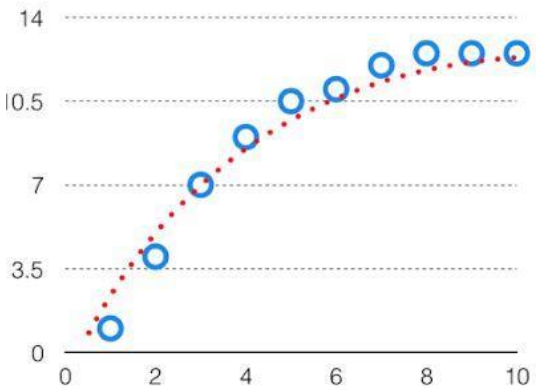
It uses spectral techniques to perform dimensionality reduction by mapping nearby inputs to nearby outputs. It preserves locality rather than local linearity.

- t-Distributed Stochastic Neighbor Embedding (t-SNE)

It calculates the probability that pairs of data points in high-dimensional space are related and then chooses a low-dimensional embedding that produces a similar distribution.

- Explain the bias-variance trade-off.

If the algorithm is too simple, i.e., if its hypothesis is fitted by a linear equation, it may fall into a high bias and low variance condition, which means that it is error prone. On the other hand, if the algorithm is too complex, using a hypothesis with a high degree equation, it may fall into a high variance and low bias condition. In the latter condition, the algorithm may not perform well with new inputs. There is an intermediate balance between these two conditions, known as a "trade-off" or "balance between bias and variance". This balance in complexity is the reason why there is a trade-off between bias and variance. An algorithm cannot be simultaneously more complex and less complex.



Graph with a perfect tradeoff.

References

- Follow, D. (2017, noviembre 23). *ML*. GeeksforGeeks. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/
- Gonzalez, L. (2020, abril 30). *Reducción de la Dimensionalidad - 🤖 Aprende IA*. 🤖 Aprende IA. https://aprendeia.com/reduccion-de-la-dimensionalidad-machine-learning/
- Follow, P. (2020, febrero 3). Bias-variance trade off - machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/ml-bias-variance-trade-off/