Data Ethics Midterm Project

Sam Pell

2024-03-18

A Moral Evaluation of 'Classification of Cucumber Leaves Based on Nitrogen Content Using the Hyperspectral Imaging Technique and Majority Voting' Sam Pell

Introduction

In the paper 'Classification of Cucumber Leaves Based on Nitrogen Content Using the Hyperspectral Imaging Technique and Majority Voting,' K Nearest Neighbors is used as one of five major classification techniques. In the study, eighteen cucumber plants were grown in unique pots; nine of these plants were then dosed with Nitrogen in their soils, while the others were left as controls. From these plants, hyperspectral images were taken over the course of four days, labeled as D0 (first day), D1, D2, and D3 (final day); the idea was to use classification to identify how many days it had been since the Nitrogen was added, based on the hyperspectral images. These hyperspectral image datapoints were then fed into the k-nearest neighbor classification algorithm, although much of the information about the researchers' methods from here is opaque. To be specific, there was no training-testing split specified. There was no mention of cross validation. There was not even specified value of k. There was a distance metric specified, Euclidean, but no reasoning given for its usage. This paper also provides a rushed explanation of the method, though it clearly fails to tackle many of the significant ethical issues with misuse of the algorithm. The ethical concerns are as such: 1) the disillusionment about the process and ethics of K Nearest Neighbor classification, and 2) the opacity regarding their own methods and lack of substantial documentation.

Analysis of Methods

K Nearest Neighbors is a commonly used technique in statistics and data analysis, primarily to classify data points into pre-determined groups. The technique sorts new data points using a training dataset, where each point's class label is already known. The algorithm is then applied to the new data points, which come from a separate testing set. With each new point, the algorithm uses a specified distance metric to find the k nearest points to the new one. Of these k nearest points, referred to as the k nearest neighbors, the algorithm finds the most popular class label (the mode), and assigns the new point to that class.

This technique can result in a number of ethical issues. For one, if the training dataset is not properly representative of the testing dataset, some classifications can be skewed, or even missing. For instance, if a dataset contains the variable "color," which can be split up into "red," "green," "blue," or "pink," but the training dataset does not contain any objects with the color "pink," then it is not possible for the algorithm to predict "pink" values, even if they are present in the testing dataset. Many researchers use a 70-30 or 80-20 training-testing split, that is, 70% or 80% of their data is allocated to training the algorithm. In some

cases, a technique known as cross validation is used to help prevent training datasets from differing too harshly from the testing data; this method takes all of the training and testing data, combines it into one large dataset, and then splits it back into training and testing sets randomly. This process is then repeated multiple times, essentially creating multiple unique training-testing splits, each of which can then be used for classification. This additional use of cross validation is used to mitigate the effects of misrepresentative training-testing splits.

Another major issue with the method is choosing the proper value of k. While k can be any positive integer, it is better in many cases to choose k such that it is odd. This prevents sticky situations, such as when k = 2; points lying on the border between classes may end up tied, with each neighbor representing a unique class. In cases like this, many computer programs default to random selection of class. This randomness implies that the method is, in some cases, no better than a coin flip or roll of the die. There are still significant concerns with other values of k. For one, it is necessary to choose a value of k that is not divisible by the number of unique class labels in the dataset, so as to prevent the aforementioned border issue. Choosing k = 1 can also be problematic, where many new points are simply matching their nearest neighbor. This can lead to overfitting, where the training dataset may not be able to fully represent either the testing data or the overall population that the algorithm could be applied to. Thus, k = 3 is likely the best minimum value of k for preventing major issues in k nearest neighbors. Similar issues exist for the upper bounds of k. As k grows, it represents more and more of the training dataset. This can lead to the inclusion of points that are relatively distant in the classification process, and can lead to greater misclassification rates. These large values of k would also trend toward the overall mode of the training dataset, since they are less and less able to classify minority groups as k grows.

To continue, the distance metric used can be contentious. While many people are used to Euclidean distance (where a^2 = b^2 + c^2), also known as the L2 norm, this is not always the best method. Euclidean distance is often the best method in cases of relatively lower dimensionality, that is, when there are fewer predictive variables involved. However, when dimensionality increases, the so-called Manhattan distance (a = b + c), or L1 norm, may be better. Similarly, in cases where high dimensionality is a concern, the L-infinity norm can be used (a = max(b, c)), which seeks the closest point by distance in any dimension. There are mounds of unique distance metrics, but these three often cover the vast majority of use cases.

Analysis of Normative Consideration

It should already be evident that there is significant moral concern in the aforementioned paper. However, it is important to support these claims of immorality within the framework of a greater body of ethics. Deontology, a form of ethics developed by Immanuel Kant, focuses on the intentions behind actions, and bases itself upon two primary rules: 1) every action taken must still be effective if universalized, and 2) do not treat people simply as a means to an end, but also as ends unto themselves. Using deontology, this paper can be fairly assessed for its merits based upon the intent of the authors.

First, the authors give a halfhearted attempt to explain the process of K Nearest Neighbors. Not mentioning how to select k, how to partition the dataset, and how to select a proper distance metric means that a lot of the most essential bases are not covered by their explanation of the method. Here, it is clear that the authors did not give an honest effort to help their readers. If this workmanship were universalized, many of the statistical processes used today would fail to provide genuine insight. Furthermore, this act of caring about reader ability serves to use readers of the study as a means to an end, helping the authors generate references without substantively aiding their readers. An example of a study that avoids these ethical issues comes from 'K-nearest-neighbor algorithm to predict the survival time and classification of various stages of oral cancer: a machine learning approach,' wherein the authors give detailed explanations about facets of the K Nearest Neighbor method; this includes cross validation, proper selection processes for finding k, and dataset splitting ethics. This study, being open about the reasoning behind each parameter's value in K Nearest Neighbor classification, is a much better standard and puts the reader first.

The other moral issue with this paper comes via its opacity. The authors do not share much about their methods, turning a typically open-soure technique into a black box algorithm. If universalized to all studies, black box algorithms would cause tremendous issues for reproducibility. Furthermore, the utilization of black box algorithms could be argued to use readers as a means to an end, only providing them with some unknown framework to an algorithm they cannot use, rather than enabling readers to access new information and generate their own data-driven results.

Conclusion

In the process of performing K Nearest Neighbor classification, it is evident that statisticians have some onus to provide quality information about their methods. While this paper regards these moral issues in regards to deontology, it goes without saying that there are other frameworks which would find the same issues: that the paper examined does not provide accurate and forthcoming information about the techniques used. The shoddy explanation of K Nearest Neighbor brings into question the authors' ability to correctly perform the method, and the shrouding of specifics (like the value of k used or the disregard for distance formulas) indicates that the authors are not being forthcoming with their methodology and results.