Nick Videtti

nvidetti@syr.edu

Assignment 6

Naive Bayes

Naive Bayes and Decision Tree with Accuracy Comparisons

8/31/2022

| Deciphering Poorly Written Digits | |
| --- | --- |
| **Introduction** | Handwriting has always been an interesting topic. The education system teaches all students to write exactly the same way, but each student’s handwriting ends up being very unique. Many grade school teachers can tell which of their students handed in a written assignment by looking at handwriting alone. The quality of handwriting is based almost exclusively on legibility.  In recent years, cursive handwriting has been removed completely from the curriculum of many elementary schools, and it can be argued that the change was long overdue. Technology has brought society to a point where it can be argued that handwriting is not needed past the very early years of school. The computer has already had a much more substantial impact on the pen/pencil than the pen/pencil ever had on the quill pen. While technology continues to become more user-friendly and climate change continues to occur, going paperless will be as feasible and desirable as ever before, including in the classroom.  The downfall of this current period is that handwriting is still needed at times, but not many outside of academia practice their handwriting, which leads to poorer handwriting in adults than children. Furthermore, most are used to reading either their own handwriting or the perfect text found in books or on screens, which makes it more difficult for the given person to read even slightly sub-par handwriting.  While technology has caused the need for a solution, it also is the reason for a possible solution. Some examples of different numbers that were handwritten have been supplied, including some very heavy detail on how the numbers were written. Using these, can poor handwriting be deciphered? This example will explore digits 0-9, but the same process can be used for letters as well. |
| **Analysis** | Naive Bayes and Decision Tree models will be created for predicting what the written digits are most likely to have been as intended. The data provided are simply examples of different written digits, specifically information for each of the 784 pixels in a 28x28 pixel picture of the written digit. That is, for each row of data, there is a “label” column that indicates the actual intended digit, and 784 columns/attributes of corresponding pixel data.  There are two datasets provided, one for training and one for testing, but for each model, separate training and testing sets will be created and used. For both training sets, they will be copies of the main training set, with the “label” variable set as a factor variable. The label column in the main testing set will also be set to a factor variable, but the “label” column will be removed entirely from the actual testing datasets for each model. This way, the models can be tested on the model-specific testing datasets, but can also be compared to the main testing dataset to check for accuracy.  The differences between the datasets for Naive Bayes and for Decision Tree modeling comes down to the pixel variables. For Naive Bayes, those will be set to factors, but for Decision Tree, those will be kept as integers. The main motivation behind this was that the model simply did not run in a reasonable amount of time for the Decision Tree when the pixel variables were set to factors. The original plan was to have both the Naive Bayes and Decision Tree models have the pixel variables set as factors, but this turned out to be a good opportunity to see what method may or may not work better. While this is far from an apples-to-apples comparison since it is being used between different models, it may be better than no comparison at all. It is also possible that creating discrete bins representing ranges of values for each variable is better than either of the two methods used in these predictions, but that method will not be attempted.  To summarize, the first step is for a Naive Bayes model with factor attributes to be trained and tested on its respective training and testing datasets. Secondly, a Decision Tree model with integer attributes will be trained and tested on its respective training and testing datasets. Then, each model will be tested for accuracy by comparing to the actual values that still exist in the main testing dataset. Finally, the accuracies will be compared. |
| **results** | Given the datasets are so large, it is not necessarily useful to discuss the Naive Bayes model other than its accuracy. This will be discussed at the end of this section.  As for the Decision Tree, there is a very large one that is difficult to read. The screenshot below gives a preview of it and an idea of the structure, but for a full view, a PDF of this Decision Tree has been included with the submission of this paper.  The accuracy of each model was generally good, but far from perfect. While there are the considerations that were discussed in the “Analysis” section, the randomization seed was set in R using set.seed(1), which may also have an effect on the overall accuracy of iterated models. The results were as follows.    For this experiment, given the previously stated caveats/simplifications, the Naive Bayes model was more accurate than the Decision Tree model. This could be due to the chosen randomization seed, the difference in data types between the attributes of each dataset, or due to the modeling technique itself, but these results are unfortunately not conclusive that the Naive Bayes model is truly overall the more effective modeling technique for these predictions. |
| **conclusions** | After one attempt, the digits were predicted correctly for roughly 84% of the written digits. While this is not necessarily the best starting point, it certainly could be much worse. The key to improving this will be to try again until a better accuracy is found.  In the few places of adult life where writing digits is still required, it is paramount that problems like this are able to be solved with nearly flawless accuracy and nearly negligible time as the paperless world, and eventual paperless classroom discussed earlier, to become an efficient reality. |