Jeff Levesque

Professor Gates

IST 565

Homework #6

**Introduction**

Handwriting recognition provides various organizations the ability to convert text documents, into digitized collections. The conversion of this digitized format allows documents to be better archived, ease of sharing, and more complicated analysis. Amazon has taken advantage of handwriting recognition, through the kindle platform. More specifically, using the tool ‘Kindle Convert’, printed books can be converted to an equivalent ebook.

Though industry such as Amazon have taken advantage of this, numerous libraries have taken great strides converting massive quantities of printed works, into digital format. In 2009, the Library of Congress scanned it’s 25,000th printed document electronically. This movement ensures that the longevity of content, since the printed copy, may be brittle requiring more care.

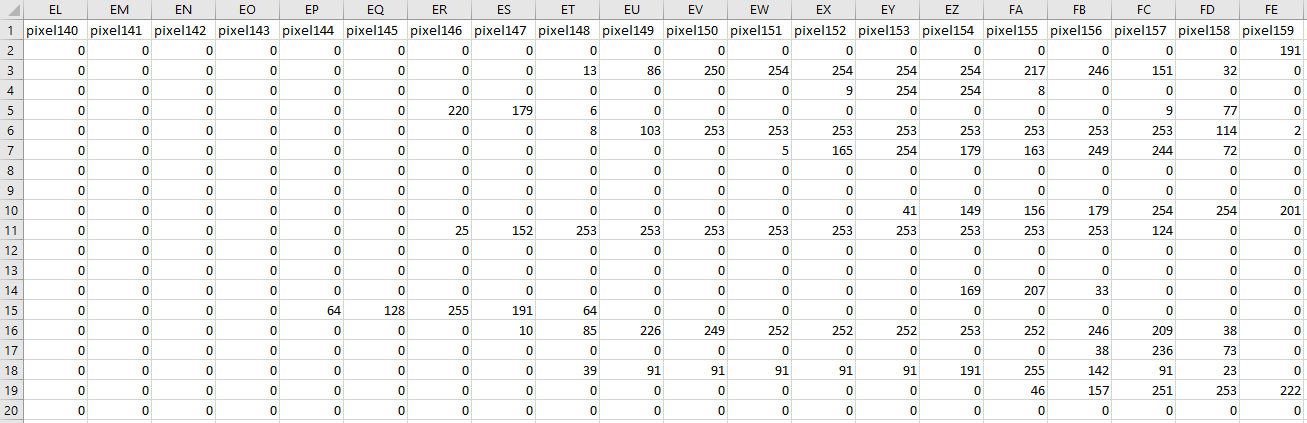
Other uses for handwriting recognition include various electronic devices. For example, many smartphones, as well as tablets have countless applications, allowing users to write directly on the device screen. The corresponding text, then undergoes conversion to the digitized character representation. Tools like these allow students, and professionals to easily take notes freeform.

Though, many different algorithms can be implemented for the classification task, this project will solely focus on decision trees, as well as naïve bayes. More specifically, these two algorithms will be used to classify printed numeric characters. Then, a comparison will be made between the accuracy of the algorithms, as well as the performance to generate a model, and predict.

**Analysis**

Data Preparation:

A csv dataset representing instances of numeric digits ranging from 0-9. More generally, each numeric digit was represented by a sequence of (roughly 780) pixels. Each pixel contains a numeric value associated with the pixel density. Therefore, higher values, are associated with darker pixels. The train dataset contained 42,000 records, where each row represented a numeric instance. Conversely, each column represented a pixel density. The test dataset was similarly structured, containing 28,000 numeric records.



Since pixel 0 through pixel 11, and pixel 780 through 783 of the train dataset contained a column sum of zero, they were removed from the dataframe.

**Results**

**Conclusions**