Support Vector Machines
Homework 05

Due March 15, 2016

Robert Brown

For the midterm assignment, I plan on doing multi-class classification of the chars74k dataset – a character set collected from google street view which exhibits tremendous irregularity. These irregularities include different fonts, colors, image sizes, and off-axis characters. For learning to be feasible on this dataset, it is neccesary to somehow normalize all of these variable parameters. To do this, I converted the images to a binary color scheme using Otsu Thresholding (https://en.wikipedia.org/wiki/Otsu's_method). From here, it is neccesary to determine the character and background colors (black or white) and normalize them across all images. This was done by finding the most frequent value (black or white) along the perimeter, and assuming that this is the background color. This can be easly inverted, forcing all images backgrounds black and characters white. The last step of normalization is scaling the images to all be the same pixel count (20x20), thus building a uniform feature vector (n=400).

Figure 1: Training image #1122

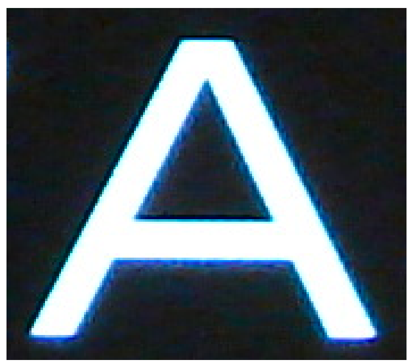

Figure 2: Training image #1653

Figure 3: Training image #874

As this project is meant to be a binary classification problem, note that the largest single-character subset of the dataset could be taken and be used to train a binary classification support vector machine. For an overview of the character distribution of the dataset, see the histogram below.

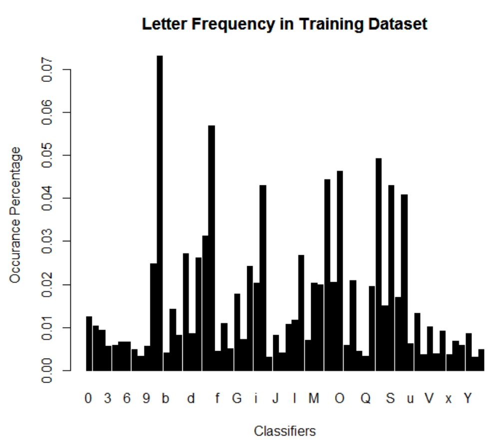

Figure 4: Normalized image #1122

Figure 5: Normalized image $1653

Figure 6: Normalized image #874

Above, the normalized images can be seen. Below, the matrix representations of these images can be seen (in the same order as above). The feature vector is created by flattening these matricies into one-dimensional structures. Lastly, note that for this type of high dimensional binary dataset, plots and summary statistics of the features are not particularly useful as there are 400 features that can take only 0 or 1.

$$
\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&0&1&1&1&1&0&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&1&0&1&1&1&1&1&0&0&0&0\\
0&0&0&0&0&1&1&1&1&0&0&0&1&1&1&1&1&0&0&0\\
0&0&0&0&1&1&1&1&1&0&0&0&1&1&1&1&1&0&0&0\\
0&0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0&0\\
0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0\\
0&0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0\\
0&0&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&0&0\\
0&0&1&1&1&1&1&0&0&0&0&0&0&0&1&1&1&1&0&0\\
0&1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix} \tag{1}
$$

$$\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0\\
0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0\\
0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0\\
0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0\\
0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0\\
0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0\\
0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0\\
0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0\\
0&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0\\
0&0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0&0\\
0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0\\
0&0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0\\
0&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix} \tag{2}$$

$$\begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0\\
0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0\\
0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0\\
\boxed{1}&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0\\
\boxed{1}&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0\\
\boxed{1}&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0\\
\boxed{1}&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0\\
\boxed{1}&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0\\
\boxed{1}&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0\\
\boxed{1}&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0\\
0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0\\
0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0\\
0&0&\boxed{1}&\boxed{1}&\boxed{1}&\boxed{1}&0&0&0&0&0&0&0&0&0&0&\boxed{1}&\boxed{1}&\boxed{1}&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix} \tag{3}$$