Parker Williamson

1/31/2018

Springboard Data Science Career Track

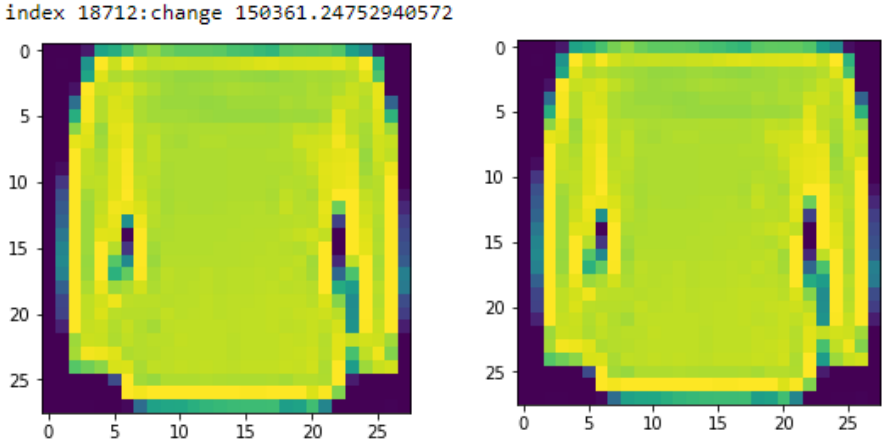
Capstone Project 1 – Data Wrangling

The basic data of my project is the fashion.mnist dataset from Kaggle. As an image dataset there is a limited about of data wrangling that can be done to improve the results. The main techniques for image improvement are Normalization and standardization. Normalization spreads the calibration of an image over the range of 0-255, so that the full range of values possible in the image is used. Standardization adjusts the images to fit the statistical model of a bell curve. That helps for these images because many of the images are heavily weighed to the low side of the spectrum. Standardization pulls those low points up so they are more visible.

The effect of each technique is different depending on which classification model you are using. For example C-Support Vector Classification (SVC) and Logistic Regression classification models expect normalized images to run the most efficiently.

To get an Idea of how the different preprocessing steps would affect the results I also plotted histograms for each image with a bin size of 10. That showed that most of the images had points on their whole range, but that the low end of the range was more heavily weighted. There were also some images that didn't fill out the full 0-255 intensity range of the pixels. Because the range wasn't full for all the images, normalizing them would improve the results by giving them the same initial intensity calibration. Standardization should help as well because the images are heavily weighted to the 0 end of the range and it should help to distribute more evenly the low end. Using both techniques at once would mean that one overwrote the other, so comparing each to see which one is the most effective on a representative training set is the best way to see which should be used.

The photo with the maximal absolute difference caused by normalization is image 18712. It seems to have a large area in the center brighten a small amount. That seems to make it slightly closer to a uniform shape.



I tested Multi-layer Perceptron (MLP), SVC and Logistic regression to compare the rough results of the classification methods. MLP was the most efficient with Normalization as the image preprocessing. The size of hidden layers on the MLP

MLP (size of hidden layers 784-100-100):

- With no preprocessing – 87.65% accuracy
- Normalized – **90.37% accuracy**
- Standardized – 90.09% accuracy

SVC:

- With no preprocessing – 70.67% accuracy
- Normalized – 85.57% accuracy
- Standardized – 81.87%

Logistic Regression:

- With no preprocessing – supposed to be normalized
- Normalized – 84.45% accuracy
- Standardized – 85.19% accuracy

From my tests I learned that Normalization and standardization are roughly equivalent, and both improve the results. Given that Normalization is the most effective I will plan to use that. Surprisingly the standardization of the logistic regression model is more effective despite sklearn stating that normalization it important for it (http://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html). As a bonus it looks like MLP is the most effective of the classifiers and I will need to test that against CNNs for the final project. An example of all 10 classes of images are on the following pages, clearly shirt, t-shirt, and pullover categories are a lot more similar than some of the other classes.