Micah Harlan
Stephen Davies
CPSC-419

Homework #5

Logistic Regression: Predicting a Country.

The task of my classifier was to predict the country a Tsunami occurred in. My homework 3 classifier was terrible. The classifier has a cross validation score of 0.120. This classifier was doomed from the start, there are 98 countries to choose from. None of the features in my classifier are essentially good. The ones I used are, deaths, houses destroyed, and the height of each wave. Each variable performs equally as well when running them alone.

When evaluating the confusion matrix it was too big to read. I calculated the f1 score which was .002. Which told me my classifier is not better than a random guess. The recall score was 0.10 and the precision was even worse at .003. Both recall and precision need to be improved.

I started off my PCA by throwing random features from my data set into it To compare improvement I also ran a 2 fold cross validation (any higher number for k takes 20+ minutes run). The score for my linear model with PCA was .1809 and the cross validation score was .178. Just adding these features in have increased the score. The next thing i did was removing some of these features I added to see which ones are useful, or don't help at all. Below is how each feature effected the classifier when it was absent from the list of features. The whole feature list that was added is, normalized total deaths, normalized total destroyed houses, normalized wave height, city names, normalized year, normalized validity of the tsunami.

- Removing normalized total deaths PCA score: 0.1814%.
- Removing normalized total destroyed houses PCA score: 0.1809%.
- Removing normalized wave height PCA score: 0.1785%.
- Removing city names PCA score: 0.1751%.
- Reemoving normalized PCA normalized year: .1820%.
- Removing normalized PCA normalized validity: 0.1536%.

From these results I removed the features that increased or had no impact on the score when absent. These features were: total houses destroyed and year, the score is now 0.1815%, which as good only removing the normalized year. I decided to put normalized houses back into my feature list bringing it back up to 0.1820%. In the previous runs The number of principle components was 5 (except when the I removed to features I dropped it to 4). The next thing I did to improve the score was running a gridsearch to find the best number of principle components.

This gridsearch took a while but it returned the value 4 for the best number of principle components, leaving it at the same score as before.

Final Features Chosen

- Normalized deaths
- Normalized height of waves
- City Names
- Normalized validity
- Normalized number of houses destroyed

Final Score: 0.1820

Starting off my logistic regressors score was 0.12% accurate. I was able to bring the score up to 0.1820 by removing and adding features. There was a 0.06% increase in the accuracy, which is not a lot, but with a classifier with a very low score anything will help it. What I learned from doing this PCA is how beneficial it can be to use it. It brought the score up on my logistic regressor 0.06% with out really doing anything to it.

Support Vector Machine

For my support vector machine I changed the target variable to make a Tsunami in the USA detector to make my target binary, instead of it consisting of 98 target choices. The first run of my program was an SVM with a PCA inside of it using the LinearSVC from sklearn gave a cross validation score of 0.871. This was really hard to improve from just messing around with the features. I went through and removed each feature but, kept the remaining features, each feature removed gave me a score of 0.871. The list of features is normalized deaths, normalized height of waves, city names, normalized validity, and normalized number of houses destroyed. I next decided to add features, I chose to add latitude and longitude. When I added latitude and longitude and it went up to 92% accurate I decided to not keep it. Even though it made my classifier really good, I would like to challenge myself. Predicting location based off of location is not too impressive and to me it feels like the easy way out.

The next thing I did was finding which SVM kernel is the best one to use, which returned polynomial for the kernel, this returned the same score as before. In that same gridsearch I also tried to find the best C value from a list containing 0.1, 1, 10, 100, and 1,000. This returned 0.1. The next gridsearch I did returned 1 for the best value of PCA components.

Final Features Chosen

- Nomalized deaths
- Normalized height
- Normalized Validity
- Normalized Destoryed Houses

Final Score: 0.871

My SVM classifier had little improvement after doing various grid searches and messing around with the features. Both my linear model and SVM perform the same. 0.87% accuracy is not too bad of a classifier. From both my SVM and logistic regression I learned that PCA is a very useful tool to improve scores.