Case Study 2

Joshua Levy

Objective

- Learn if more data is useful for a classification task.
- Learn about classification algorithms.
- Learn the pros and cons of data normalization.
- Given the UJIIndoorLoc Data Set from the UCI machine learning depository. Classify indoor location based on Wireless Access Points (WAPs) and the recorded Received Signal Strength Intensity (RSSI).

Data Features and Targets

- The bulk of our data comes in the form of 520 WAPs columns and the corresponding RSSI.
  - Signal strength of 100 means no WAP was detected on the device
- Longitude
- Latitude
- Floor
- Building ID
- SpaceID - represents where capture taken (office/classroom/corridor)
- Relative Position - 1 indicates capture taken inside, 2 outside in front of the door.
- UserID
- PhoneID
- Timestamp (in unix time) of when the capture was taken

Data Features and Targets

- The goal of this classification task is to use the WAP data to determine location. A new column was created called Building Code to act as the dependent variable.
    - Building Code is the result of concatenating the columns BuildingID and FloorID into a single set of digits (the numbers themselves were NOT added together).
- All 520 WAP columns were used in the experiments
- Ultimately, no other columns were used in the experiments as any inclusion from them resulted in a drop in accuracy.

Model Selection

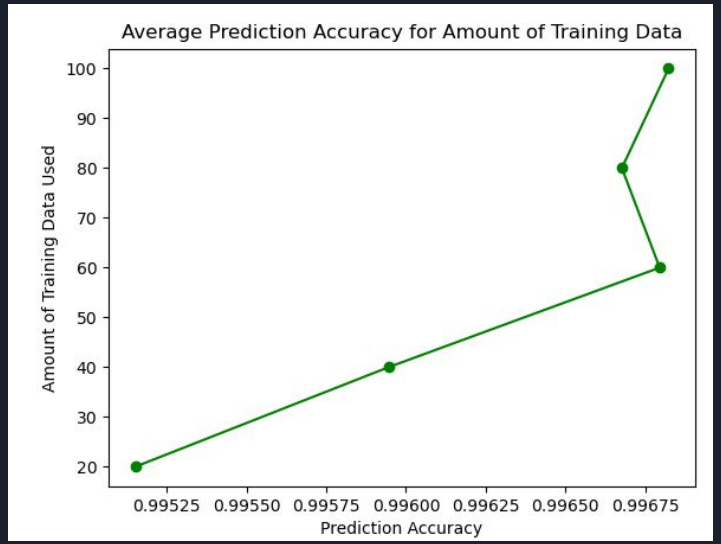
- The top three models I tried were Linear SVC, Logistic Regression, and Random Forest.
- I tried using both normalized and unnormalized data in each of the models.
  - Data was normalized using the max absolute scaler which normalizes data between -1 and 1.
- Linear SVC performed the worst out of the three models with an F-1 score of about 0.69 on the normalized data and 0.66 on the unnormalized data
- Logistic Regression won second place with an F-1 score of about 0.78 on the normalized data vs 0.73 on the unnormalized data.
- Random Forest performed the best with an F-1 score of about 0.8 on both the normalized and unnormalized sets.
  - The accuracy score was also calculated and the normalized performed better by the 6th decimal place.
  - Ultimately, I choose to use normalized sets throughout the rest of the experiments since the results were comparable and it came down to personal preference.

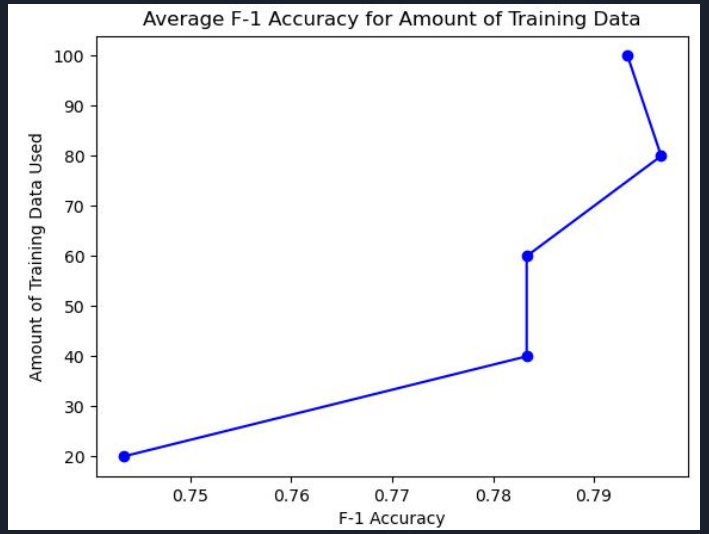
Findings

- Occasionally a Logistic Regression trial would result in an F-1 score over the 0.8 threshold but Random Forest was more consistent in reaching that threshold.
- When I had to average the scores of the three experiments the average F-1 score for Random Forest was below the 0.8 threshold in the experiments where it was accurate enough to reach that threshold (80% - 100% training data used).
- Most classifiers performed poorly in classifying location (KNN and SVC were also tried out but not included).
- All this said there was probably more I could have done to get better accuracy for my models.
- More data was not useful for this classification task.
- Data normalization did improve accuracy and F-1 scores in the algorithms but, not enough to be useful to our study.

Findings

Note: Prediction accuracy = number of correct predictions / all predictions.





Interestingly at 80% of training data used there was a drop in prediction accuracy but the F-1 Score actually went up. The reverse occurred at 100% training data used.