Data science project semester B

Submitter: Eitan Kats

Submission Date: 29.08.2021

Table of content

Classification improvement from semester A

In this notebook I have used the following methods to improve the performance:

1. Voting (soft,hard)
2. XGBoost
3. AdaBoostClassifier
4. BaggingClassifier
5. StackingClassifier

after using ensemble I have managed to reach 93.75% accuracy compared to 92.81% in semester A
Which is a 0.94% improvement. The classifier that reaches this score was hard voting Classifier with the following models:
Knn, randomForest, adaboost,xgboost

Fashion mnist:

In this notebook we have had to tackle the fashion mnist and classify clothes without the use of neural networks and SVM.
The Fmnist data is balanced that is why I have used the accuracy metric in the notebook.

The tools that were used:

1. Voting
2. XGBoost
3. KNN
4. RandomForestClassifier
5. PCA
6. Stacking

Pre-processing: A pipeline that consisted of standard scaler and PCA to 90% of the variance.
The best score in this notebook is: 89.37%
Best model: tuned XGBoost

Cats Vs Dogs:

In this notebook we have had to tackle the dogs vs cats dataset and classify pictures of dogs and cats without neural networks and SVM.
The Cats Vs Dogs data is balanced that is why I have used the accuracy metric in the notebook.

The tools that were used:

1. KNN
2. XGBoost
3. RandomForest
4. Stacking
5. K Means (pre-processing)
6. PCA
7. openCV
8. Bagging

Pre-processing: A pipeline that consisted of standard scaler and PCA to 90% of the variance.
I have seen that the images in black and white have performed worse when fitting them into the basic models, that is why I have decided to stop looking for ways to optimize the models.

The best score in this notebook is: 68.72%
Best model: Bagging classifier with XGBoost as the estimator, the XGBoost was already tuned to the best hyperparameters using gridSearch.

Hand Synchronization:

In this notebook we have had to use data from an academic study done on hand movements.
The data conducted of 3 states:

1. Spontaneous - 2 people moving their hands spontaneously in front of each other
2. Sync - 2 people trying to sync their hand movement
3. Alone - a person that moves his hand randomly

We have had to determine whether we can use machine learning to see if we can understand whether the hands are moving spontaneously, in sync, or just alone.

Pre processing: first of all we were given instruction on how to perform the pre processing.

1. Clear the data (nulls and bad hand data (2 hands in the alone state for instance))
2. Merge the alone data of each person with the rightHand dataframe which was recorded separately
3. Merge all the data together

The tools that were used:

1. KNN
2. XGBoost
3. RandomForest

4. Stacking
5. PCA
6. Voting(hard/soft)

The best score: 87.33%
The best model: soft voting with 3 models : knn,tuned randomforest, tuned xgb