Subject:
Human activities Recognition using smartphone data

Course: Machine Learning

Student: George Papadopoulos

Human activities Recognition using smartphone data - Introduction

Activities: *running*, *walking*, *cycling.*

End-to-end machine learning system with all phases.

Models: SVM, kNN classifiers.

Emphasis to data preparation, data exploration, feature extraction.

Use of smartphone accelerometer to collect data.

X, Y, Z axes

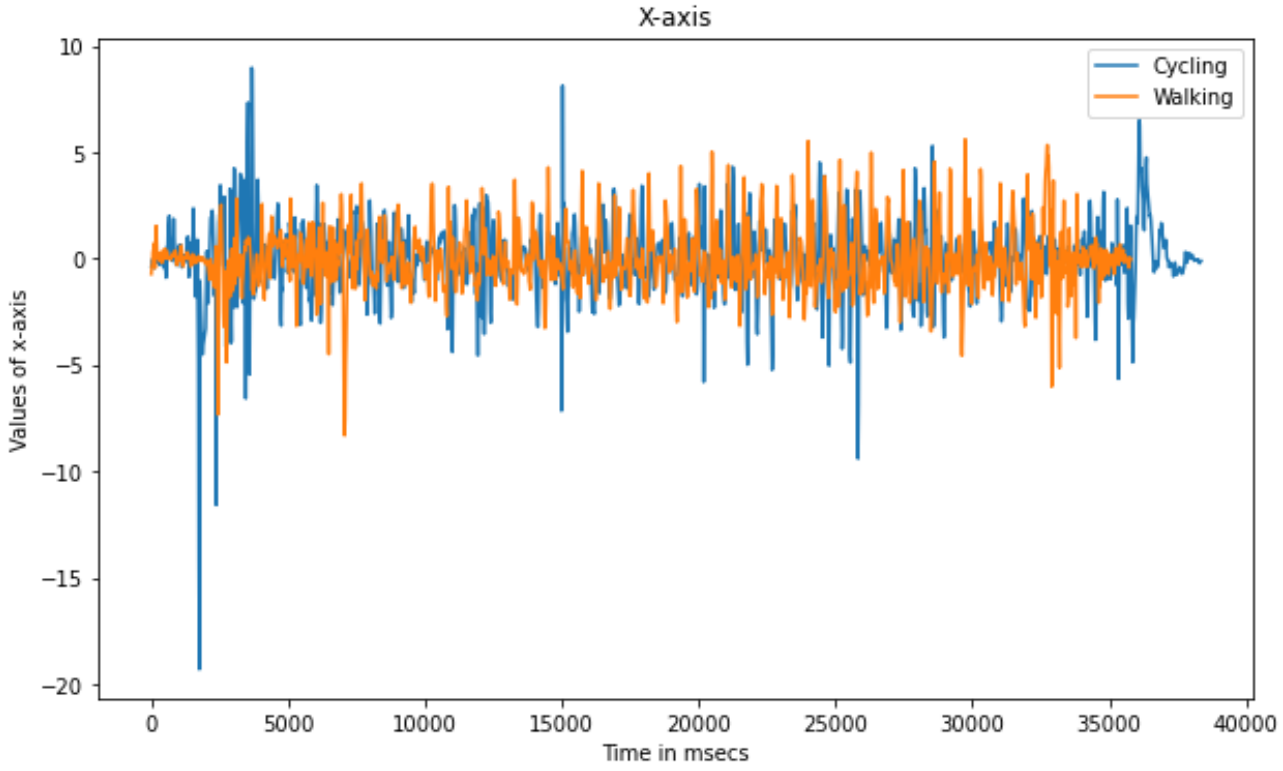
Get/Prepare Phase

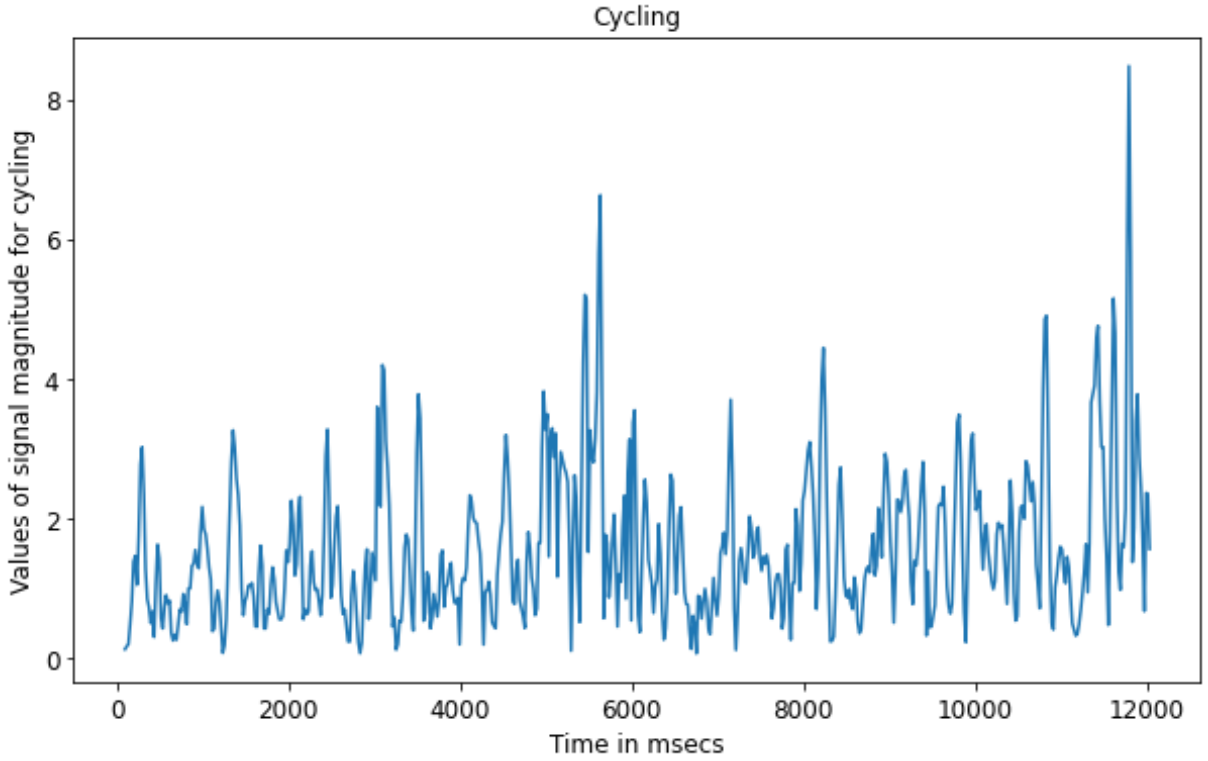
- 6 subjects-persons, 10 samples per activity, 180 samples totally.
  - Balanced dataset.
  - 4 subjects train-validation set, 2 subjects test set.
- Each sample 30+ secs. 0-10 preparation, 10-22 clear sample, 22-30+ stop.
- Use of the same smartphone to avoid differences of sensors.
- Sampling at 50 Hz. Measurement unit: m/s^2.
- Annotation after sampling.
- Upload to GitHub.
- Parser creation to read .txt samples files.

Prepare – Exploration phase

- Cutting of X, Y, Z signals and keep only 12 secs.

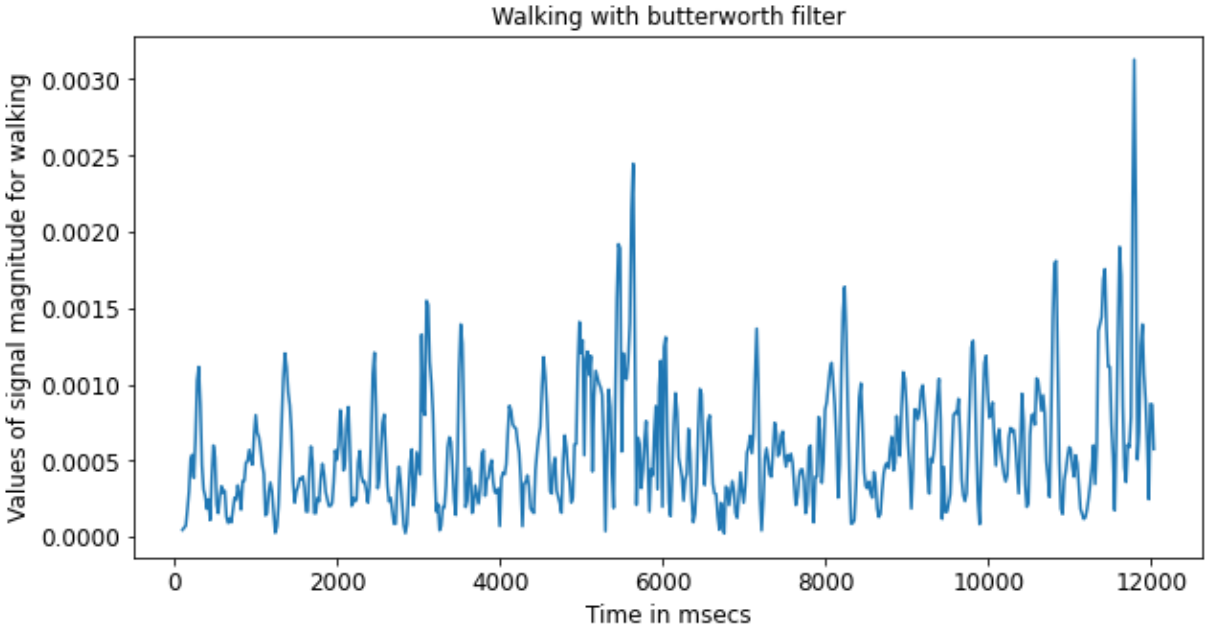
- Problem with orientation changes of smartphone.
- Computation of X-Y magnitude to solve it.

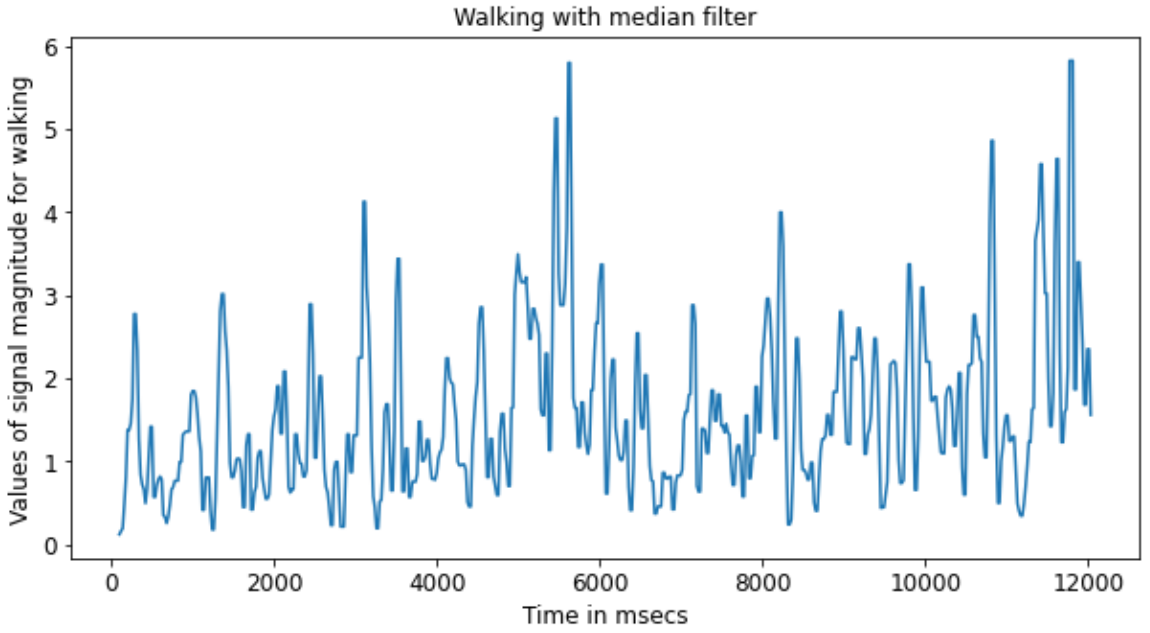




Preparation phase - Denoising

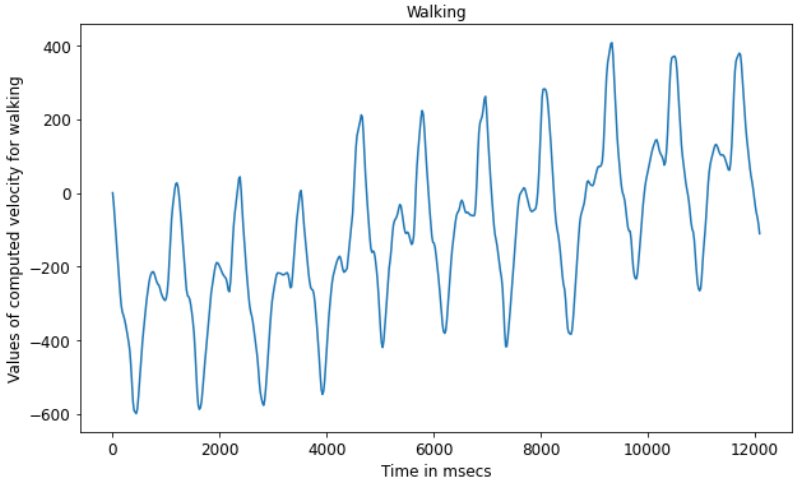
- Experimentation with Butterworth low pass filter and Median filter.
- Choose Median filter which smooths the signal.

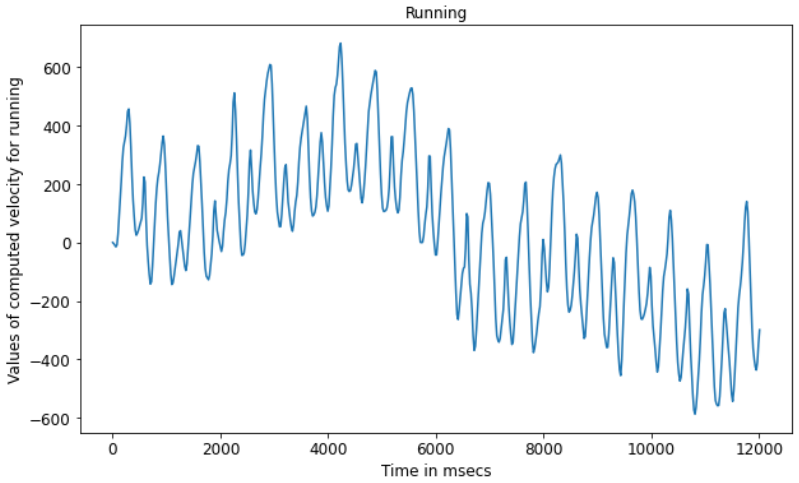




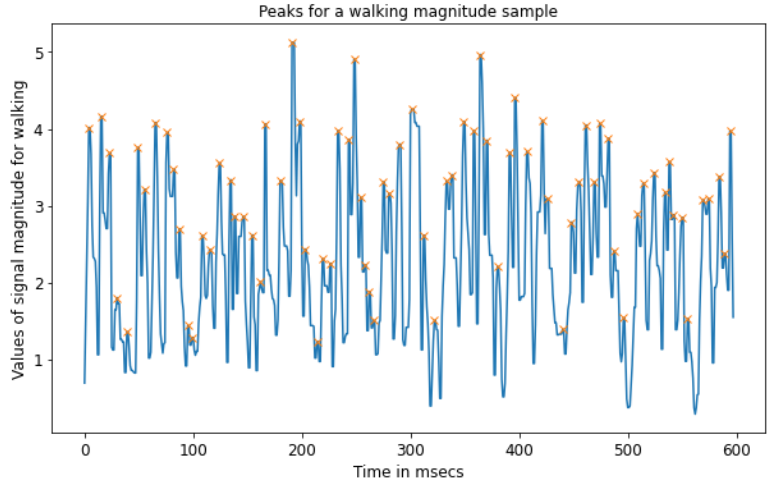
Exploration phase – Discover patterns

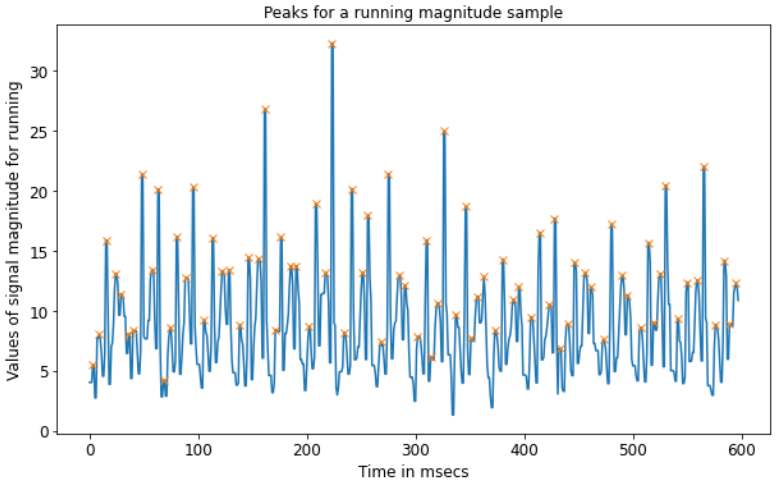
- *Velocity* considered significant to find patterns.
- Unfortunately, depends on initial velocity.
- But *dominant frequency* can be used.





- *Peaks* of signals seems to have different values and *widths* between them.



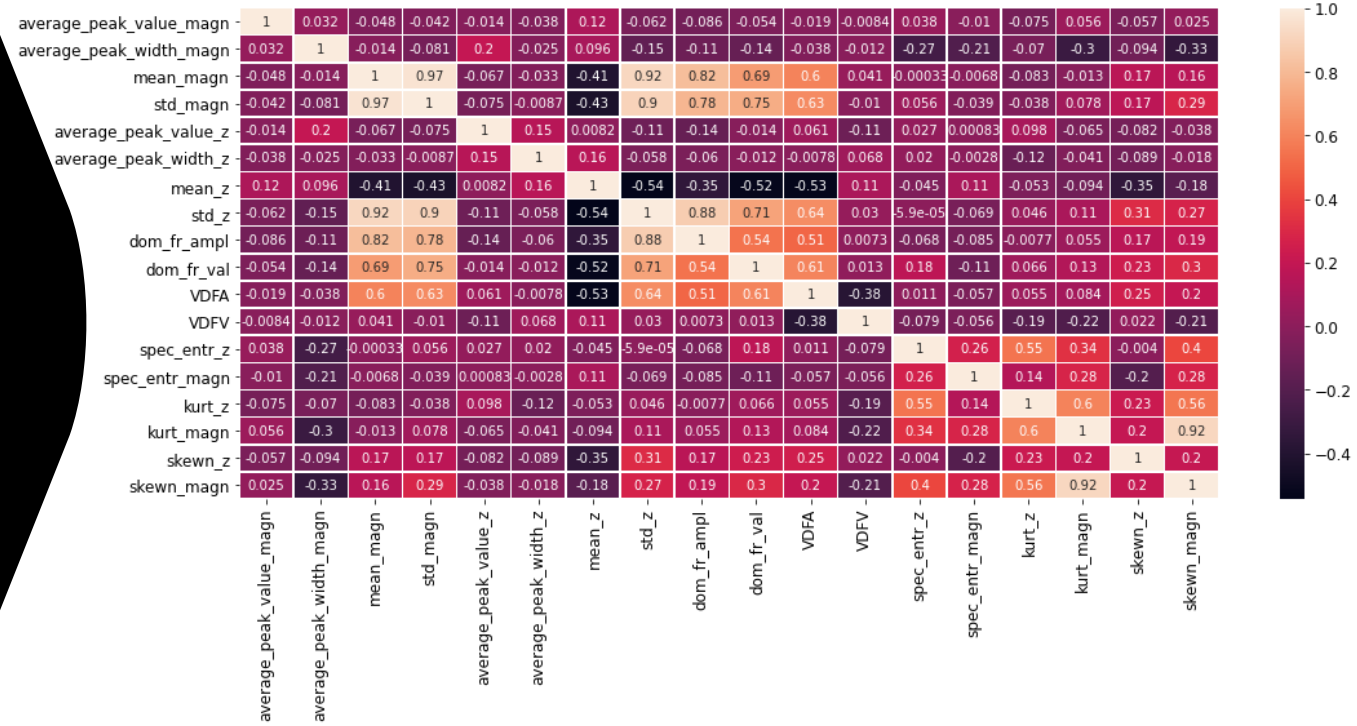


Exploration phase – Feature examination

- Possible features:
  - Statistics of Z-axis signal and X-Y magnitude: *mean*, *standard deviation*, *kurtosis*, *skewness.*
  - *Average peak values, average peak widths.*
  - *Spectral Entropy.*
  - *Dominant Frequency value* and *amplitude* for Z-axis, X-Y magnitude and velocity using Fourier Transformation.
- Examine correlations using *correlation matrix.*
  - Highly correlated:

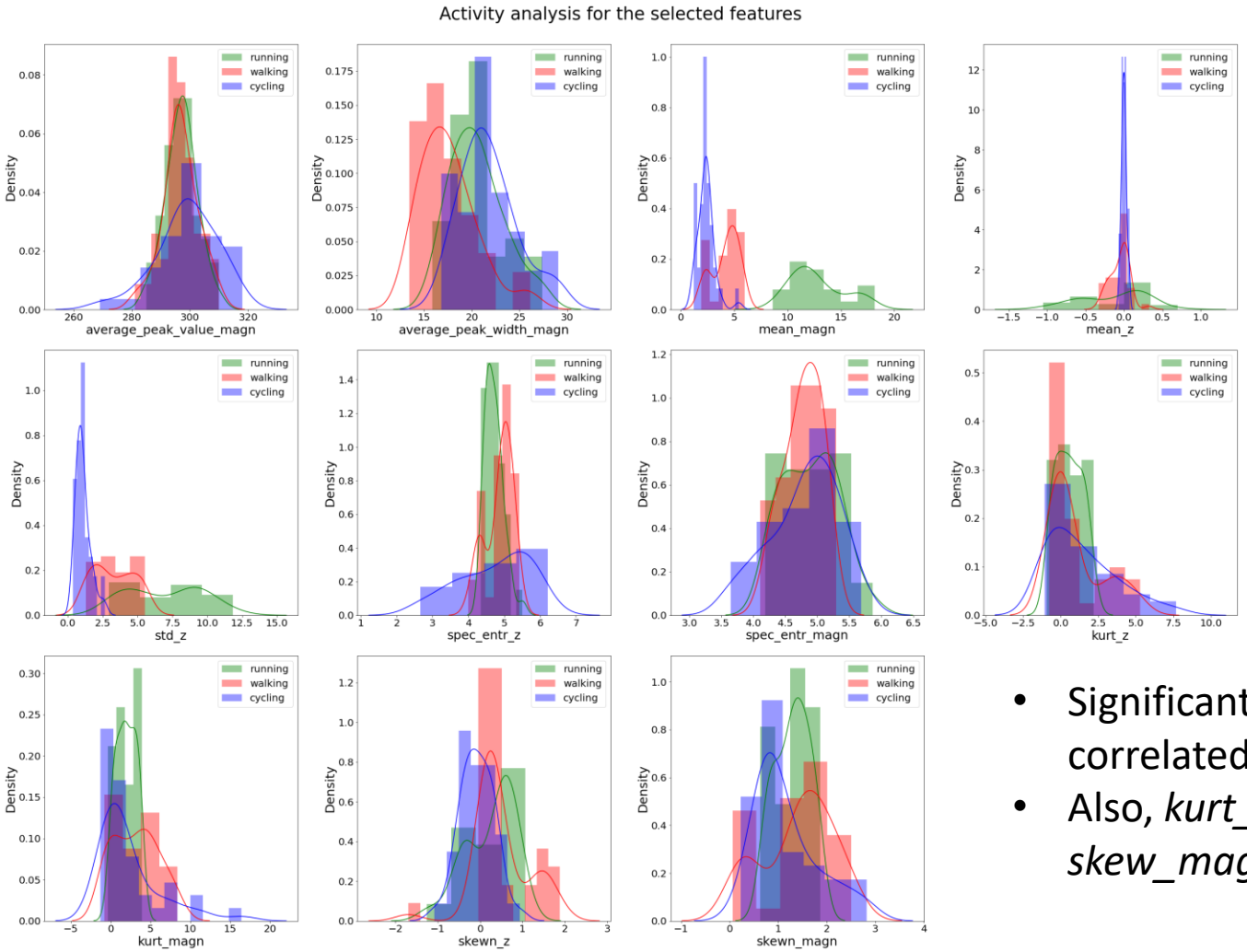
  mean_magn – std_mean – std_z

  kurt_magn – skew_magn



Feature selection

- Selection using distribution plots, backward Elimination, SelectKBest

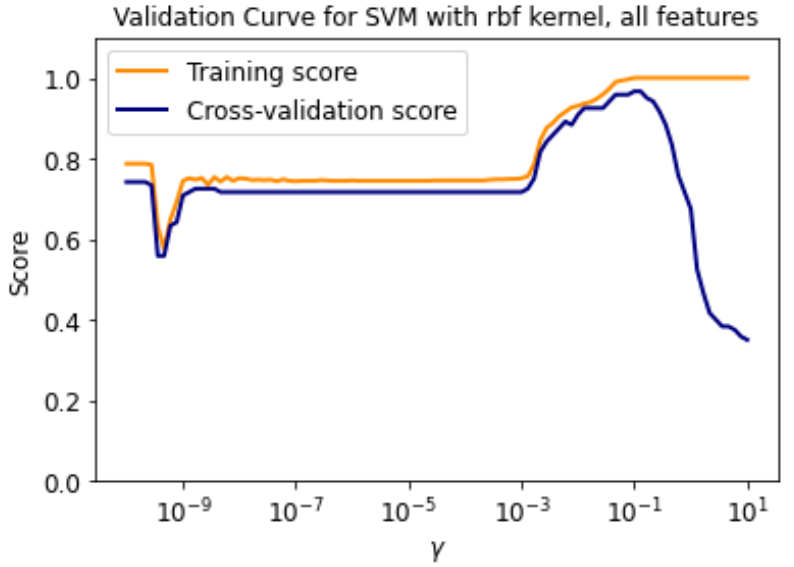


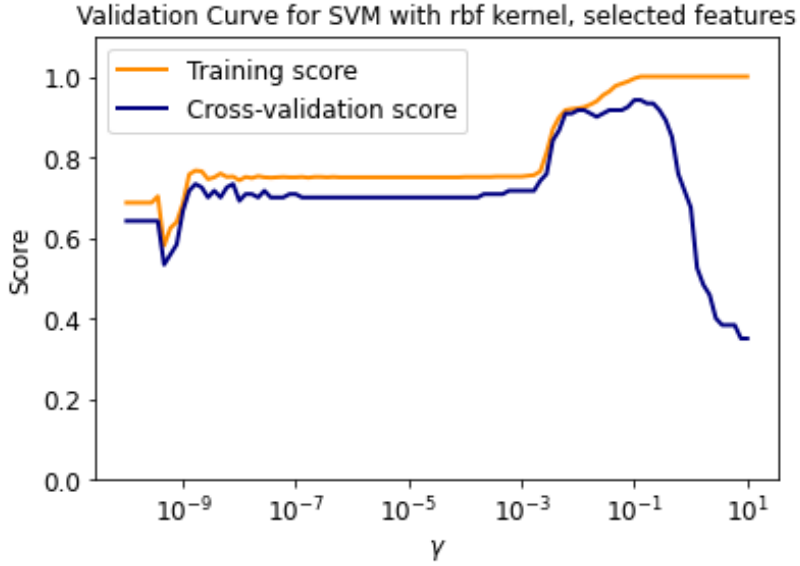
	Feature_Name	Score
2	mean_magn	446.725187
3	std_magn	179.601351
7	std_z	112.165381
8	dom_fr_ampl	70.188456
9	dom_fr_val	24.900838
1	average_peak_width_magn	22.306325
10	VDFA	17.493667
16	skewn_z	10.604691
6	mean_z	2.647377
14	kurt_z	2.422728
15	kurt_magn	1.835124
11	VDFV	1.784572
4	average_peak_value_z	1.590847
17	skewn_magn	1.431152
0	average_peak_value_magn	0.917507
13	spec_entr_magn	0.806367

- Significant *mean_magn, std_magn* and *std_z,* but strong correlated. *mean_magn* more important, exclude other two.
- Also, *kurt_magn* and *skew_magn* correlated, exclude *skew_magn.*

Model phase – SVM train

- Validation curves for two dataset with selected features and all features, for different *gamma* parameters.
  - Examine where model begins overfitting and which dataset is best.
  - Use of *Cross Validation*, *pipeline* and *StandardScaler.*
- Use of *GridSearchCV* to determine the best values of parameters *gamma* and *C* for *rbf kernel.*
  - The same for *polynomial kernel* with extra parameter *degree.*
- Score: *mean accuracy* of subclasses.
- Use of *cross_validation_score* to compare two kernels.
- Rbf: 94%, Polynomial: *87%.*

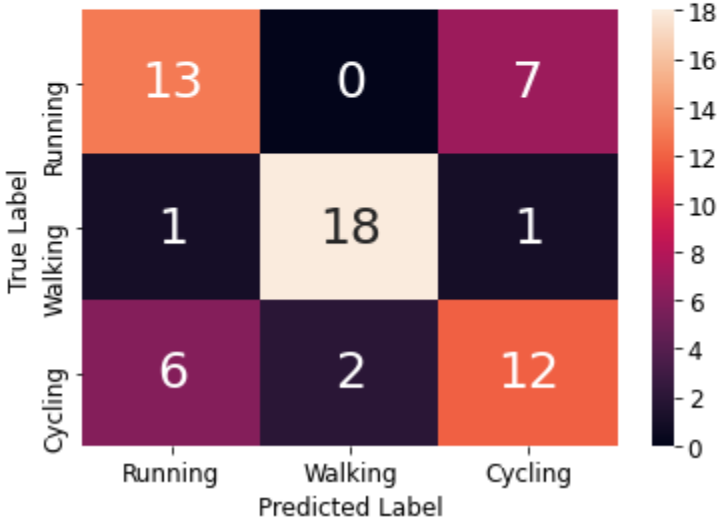




Model phase – SVM test

- Use of test set for first time to test SVM performance.
- Creation of classification report and confusion matrix.

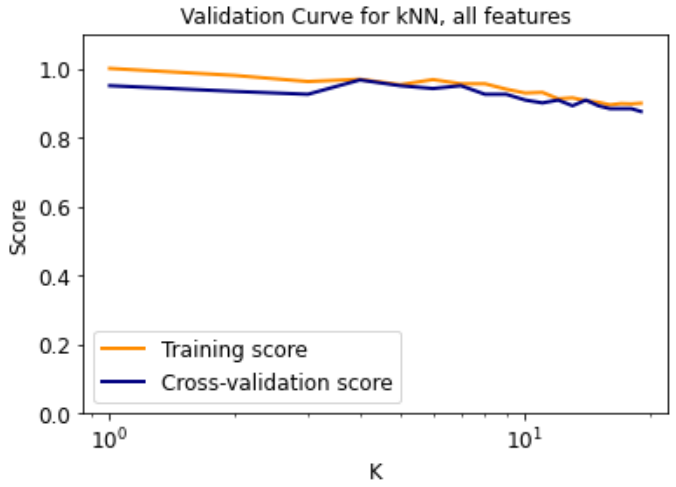
	precision	recall	f1-score	support
Running	0.650	0.650	0.650	20
Walking	0.900	0.900	0.900	20
Cycling	0.600	0.600	0.600	20
accuracy			0.717	60
macro avg	0.717	0.717	0.717	60
weighted avg	0.717	0.717	0.717	60

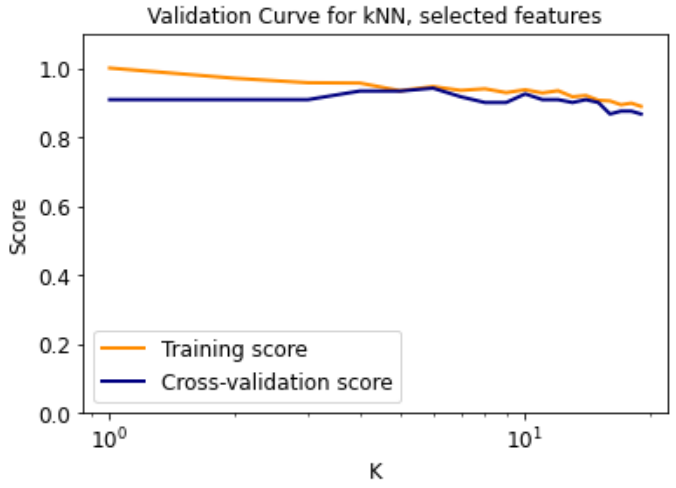


- Scale test set separately.
- Unexpected results, walking class has the least misclassified samples and so the highest scores.
- Cycling has the lowest scores.
- Classifier is confused between running and cycling.
- Interesting, symmetrical precision-recall.
- Overall f1 71%. Great difference between train-validation and test scores. More samples for training might be needed.

Model phase – kNN train and test

- Slight discrepancies between two datasets for different values of parameter *K-nearest neighbors*.
- Determine best value with *GridSearchCV* as before.





	precision	recall	f1-score	support
Running	0.750	0.600	0.667	20
Walking	0.792	0.950	0.864	20
Cycling	0.550	0.550	0.550	20
accuracy			0.700	60
macro avg	0.697	0.700	0.693	60
weighted avg	0.697	0.700	0.693	60

- Same rank as SVM, but different scores.
- Cycling almost equal scores to random classification.
- Walking high recall due to only 1 FN, lower precision due to 5 FP.
- Running, walking overlapping cycling
- Overall f1 70%, almost equal to SVM.

Models' comparison – Communicate results

- ROC curves: Each class against the others.
- From ROC curves can be concluded SVM has greater performance for all classes.
- Highest SVM ROC-AUC score : 98% for running. Meaning from 100 samples predicted as *running*, 2 will not belong to *running* class.
- Highest kNN: 95% for *running*.
- Concluding: Using smartphone in trouser pocket, accurate enough to monitor activities especially *running* and *cycling*. Building of a simple app to use by athletes or someone who cares about daily activity performances and times.
- Future work: Improvement of *walking* class predictions using more training data, longer duration. Or/And time-window based model.