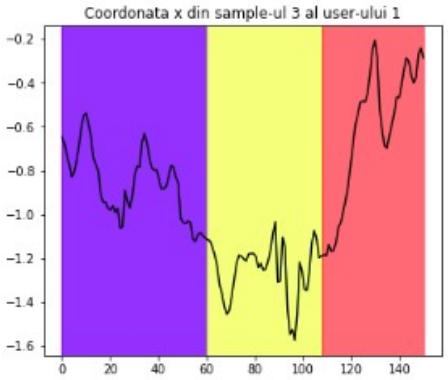
“Smartphone User Classification”
Documentatie pentru a doua solutie

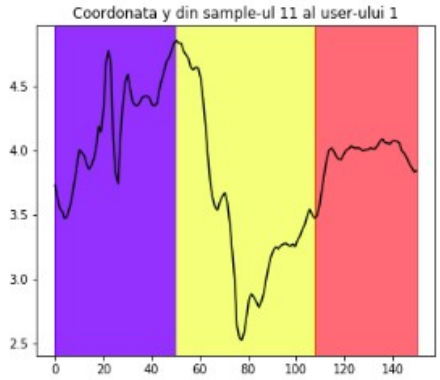
Student: Burcea Bogdan-Madalin
Grupa: 235

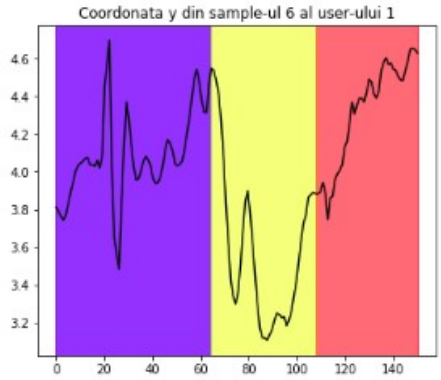
Modelul folosit pentru aceasta solutie a fost acela de Support-Vector Machine. Fiecare sample a trecut mai intai prin cateva faze de prelucrare, dupa care au fost asezate impreuna in doua tablouri bidimensionale corespunzatoare pentru aplicarea modelului SVM din libraria Scikit-Learn de Python.

Prelucrari aplicate:
- Aducerea la aceeasi forma:
Se observa ca intr-un sample nu sunt prezente intotdeauna 150 de observatii. Anumite prelucrari viitoare necesita ca fiecare sample sa aiba aceeasi forma, asa ca un prim pas a fost sa extind sau trunchez fiecare sample la 150 de observatii. Pentru a realiza acest lucru si a evita pierderea validitatii datelor initiale, am aplicat procedeul de interpolare liniara pentru fiecare coordonata / coloana a unui sample. Interpolarea s-a aplicat deci independent de 3 ori pentru fiecare sample.

- Extragerea de features din fiecare sample:
O plot-are a unora din datele de antrenare a scos in evidenta faptul ca forma graficului unei coordonate dintr-un sample poate sa se schimbe destul de mult. Astfel, graficul unei coordonate ajunge sa aiba zone care difera mult una de cealalta. In figurile urmatoare ferestrele colorate diferit difera de celalalte din cadrul aceeasi figuri.







Aceasta observatie a rezultat in incercarea de a extrage valori numerice ce reprezinta cat mai descriptiv un sample atat din tot ansamblul, cat si din cadrul fiecarei “ferestre”. Am ales sa fac fiecare fereastra de aceeasi dimensiune, iar prin experimentare am gasit ca numarul de 5 ferestre aduce rezultate mai bune. De asemenea, in final, aceste 5 ferestre sunt distincte, desi am incercat si o suprapunere a acestora. Suprapunerea incercata a adus rezultate mai slabe.

Pentru fiecare fereastra (sau tot ansamblul) a unui sample am experimentat cu calcularea urmatoarelor feature-uri numerice si combinarea lor:
- Minimul / Maximul pentru fiecare coordonata sau doar pentru anumite coordonate;
- Media algebrica / patratica pentru fiecare coordonata sau doar pentru anumite coordonate;

- Toate diferentele dintre doua valori de coordonata adiacente.
- Minimul / Maximul dintre diferente pentru fiecare coordonata;
- Media algebrica / patratica dintre diferente pentru fiecare coordonata;
- Varianta diferentelor pentru fiecare coordonata;

- Minimul / Maximul pentru fiecare norma a unui vector de 3 coordonate;
- Media algebrica / patratica pentru fiecare norma a unui vector de 3 coordonate;
- Varianta pentru fiecare norma a unui vector de 3 coordonate;

- Numarul de “intoarceri” pe care le face graficul pentru fiecare coordonata;
- Corelatia pentru fiecare pereche de coordonate, fiecare tratata ca o variabila aleatoare.

Norma calculata pentru fiecare vector de 3 coordonate a fost aleasa ca norma L2, dar a fost incercata si norma L1, insa cu rezultate mai slabe.

Pentru tot ansamblul, s-a incercat in plus si calcularea urmatoarei valori:
- Varianta pe fiecare coordonata pentru o variabila aleatoare reprezentata de numarului de “intoarceri” ale graficului calculat pentru fiecare fereastra.

In final, prin experimentare, am ales urmatoarele feature-uri pentru fiecare window:
- Minimul / Maximul pentru fiecare coordonata;
- Media algebrica pentru fiecare coordonata;

- Minimul / Maximul dintre diferente pentru fiecare coordonata;
- Media algebrica dintre diferente pentru fiecare coordonata;

- Minimul / Maximul pentru fiecare norma a unui vector de 3 coordonate;;
- Media algebrica pentru fiecare norma a unui vector de 3 coordonate;;

- Numarul de “intoarceri” pe care le face graficul pentru fiecare coordonata;
- Corelatia pentru fiecare pereche de coordonate, fiecare tratata ca o variabila aleatoare.

Iar pentru tot ansamblul s-au ales urmatoarele feature-uri:
- Minimul / Maximul pentru fiecare coordonata;
- Media algebrica pentru fiecare coordonata;

- Minimul / Maximul dintre diferente pentru fiecare coordonata;
- Media algebrica dintre diferente pentru fiecare coordonata;

- Minimul / Maximul pentru fiecare norma a unui vector de 3 coordonate;;
- Media algebrica pentru fiecare norma a unui vector de 3 coordonate;;

- Numarul de “intoarceri” pe care le face graficul pentru fiecare coordonata;
- Corelatia pentru fiecare pereche de coordonate, fiecare tratata ca o variabila aleatoare.
- Varianta pentru fiecare norma a unui vector de 3 coordonate;
- Varianta pe fiecare coordonata pentru o variabila aleatoare reprezentata de numarului de “intoarceri” ale graficului calculat pentru fiecare fereastra.

Ca rezultat, fiecare sample s-a transformat din cele 150 de perechi tridimensionale de date de acceleratie (450 de valori numerice), intr-un vector linie de 166 de feature-uri numerice.
Pe vectorii linie rezultati s-a aplicat per feature o scalare de tip min max la intervalul (0, 1).
Alte tipuri de normalizari incercate: standardizare, normalizarea l1, normalizarea l2.

Functii de Kernel de SVM incercate: Linear, Radial basis function, Polynomial.
Pentru gasirea unui kernel bun, am aplicat procedeul de grid search pe parametrii unui kernel si apoi efectuarea unui 10-fold cross validation pentru gasirea unei indicator aproximativ de precizie a modelului. Pentru cross-validare, am compus fiecare dintre cele 10 bucket-uri din exact 45 de sample-uri pentru fiecare user din cei 20, in presupunerea ca si in datele de testare distributia este uniforma intre utilizatori.

Pentru aceasta solutie am ales in final Kernel = ‘linear’, C=10, acest model avand o acuratete medie din 10-fold cross validation de 96.7000%.

Dupa alegerea parametrilor de model, acesta a fost antrenat pe intreg setul de date de antrenare si aplicat apoi pe datele de testare pe care s-au realizat aceeasi extragere de feature-uri si scalarea acestora prin scaler-ul antrenat pe datele initiale.

In urma executarii unei 3-fold cross-validation - unde fiecare bucket contine un numar egal de samples de la fiecare utilizator - se obtin urmatoarele matrice de confuzie.
- Pentru primul bucket ca date de validare:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	145	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	140	0	6	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1
3	0	0	141	0	2	0	0	0	0	0	0	0	6	0	0	0	0	0	1	0
4	0	9	0	134	0	0	0	0	0	0	1	0	0	0	5	0	0	0	0	1
5	0	0	1	0	137	0	0	0	1	0	0	0	11	0	0	0	0	0	0	0
6	0	0	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	4	1	0	0	0	1	143	0	0	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	0	0	0	149	0	0	0	0	0	1	0	0	0	0	0	0
9	0	0	3	0	2	0	0	0	140	0	0	0	1	0	0	0	0	0	4	0
10	0	1	0	0	0	1	0	0	0	147	0	0	0	0	0	0	0	0	0	1
11	0	0	0	0	0	0	0	0	0	0	149	0	0	0	0	1	0	0	0	0
12	0	0	0	0	1	0	0	0	0	0	0	149	0	0	0	0	0	0	0	0
13	0	0	1	0	6	0	0	0	5	0	0	0	134	0	0	0	0	0	4	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0	0
15	0	6	0	3	0	1	0	0	0	2	0	0	0	0	138	0	0	0	0	0
16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	149	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	146	0	2	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0
19	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	6	0	140	0
20	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	146

```
==============================
accuracy of validation =  0.959
```

- Pentru cel de al doilea bucket ca date de validare:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	142	1	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	134	0	4	0	1	0	0	0	0	0	0	0	0	10	0	0	0	0	1
3	0	0	144	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0
4	0	3	0	142	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	1
5	0	0	0	0	141	0	0	0	1	0	0	0	8	0	0	0	0	0	0	0
6	1	0	0	0	0	146	0	0	0	1	0	0	0	0	2	0	0	0	0	0
7	4	0	0	0	0	0	145	0	0	0	0	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	147	0	0	0	0	0	3	0	0	0	0	0	0
9	0	0	4	0	3	0	0	0	141	0	0	0	0	0	0	0	0	0	2	0
10	0	1	0	0	0	0	0	0	0	148	0	0	0	0	1	0	0	0	0	0
11	0	0	0	2	0	0	0	0	0	0	147	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	1	0	0	149	0	0	0	0	0	0	0	0
13	0	0	3	0	13	0	0	0	3	0	0	0	129	0	0	0	0	0	2	0
14	0	0	0	0	0	0	0	1	0	0	0	0	0	149	0	0	0	0	0	0
15	0	7	0	0	0	2	0	0	0	1	0	0	0	0	139	0	0	0	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	149	0	0	0
18	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	146	0	3
19	0	0	0	0	0	0	0	0	1	0	0	0	7	0	0	0	3	0	139	0
20	0	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	1	143

```
==============================
accuracy of validation =  0.9566666666666667
```

- Pentru cel de al treilea bucket ca date de validare:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	142	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	142	0	2	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0
3	0	0	145	0	1	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0
4	0	7	0	140	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0
5	0	0	3	0	140	0	0	0	3	0	0	0	4	0	0	0	0	0	0	0
6	0	1	0	2	0	145	0	0	0	0	0	0	0	0	2	0	0	0	0	0
7	5	0	0	0	0	2	143	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	148	0	0	0	0	0	2	0	0	0	0	0	0
9	0	0	2	0	3	0	0	0	138	0	0	0	5	0	0	0	0	0	2	0
10	0	2	0	0	0	3	1	0	0	143	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	149	0	0	0	0	1	0	0	0
13	0	0	8	0	11	0	0	0	2	0	0	0	128	0	0	0	0	0	1	0
14	0	0	0	0	0	0	0	6	0	0	0	0	0	144	0	0	0	0	0	0
15	0	8	0	3	0	2	0	0	0	3	0	0	0	0	133	0	0	0	0	1
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	140	0	5	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	0	1
19	0	0	0	0	1	0	0	0	1	0	0	0	2	0	0	0	1	0	145	0
20	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	147

```
==============================
accuracy of validation =  0.9536666666666667
```

Acuratetea medie care se obtine pentru acest 3-fold cross-validation este de 95.6444%.