Проект по Машинно обучение и вероятностно моделиране

Георги Дежов

ФМИ София

12 февруари 2020

Цел

- Да се създаде работещ модел за разпознаване на човешки дейности на база данните *WISDM Smartphone and Smartwatch Activity and Biometrics*.

Увод

- “WISDM Smartphone and Smartwatch Activity and Biometrics” включва данни, събрани от 51 субекта;
- Всеки е изпълнявал 18 дейност за по 3 минути всяка;
- Всеки е имал “умен” часовник поставен на доминиращата му ръка и смартфон в джоба си;
- Събрани са данни от сензорите на акселерометъра и на жироскопа от смартфона и от смарт часовника;
- Данните са записвани със скорост 20Hz,т.е. на всеки 50 ms;

Увод

- За смартфон е използван или Google Nexus 5 / 5X или Samsung Galaxy S5 с Android 6.0.
- За часовник е използван LG G Watch с инсталиран Android Wear 1.5.

Увод

Таблица 1: Обща характеристики на данните и процесът на събиране на данни

Брой субекта	51
Брой дейности	18
Минути за изпълнение на дейност	3
Скорост за запис	20Hz
Използван смартфон	Google Nexus 5/5x или Samsung Galaxy S5
Използван смарт часовник	LG G Watch
Брой измервания	15,630,426

Увод

Таблица 2: 18-те дейности представени в базата данни

Код	Дейност	Код	Дейност
ходене	A	ядене на паста	J
тичане	B	пиене от чаша	K
качване на стълби	C	ядене на сандвич	L
седене	D	ритане	M
стоене	E	хващане	O
писане на машина	F	дриблиране	P
миене на зъби	G	писане на ръка	Q
ядене на супа	H	ръкопляскане	R
ядене на чипс	I	сгъване	S

Методология

- На случаен принцип е избран един от 51-та човека, в конкретният случай се е паднал субект номер 5;
- Взити са данните записани от акселератора на телефона, който е бил в джоба му;
- Добавен е параметър, който да индикира разделението на дейностите на такива извършвани главно с ръце и такива, в които не се ползват ръце.

Методология

- В дейности активно изпозващи ръце попадат: писане на машина, миене на зъби, ядене на супа, ядене на чипс, ядене на паста, пиене от чаша, ядене на сандвич, хващане, дриблиране, писане на ръка, ръкопляскане и сгъване.
- В дейности не изпозващи ръце попадат: вървене, тичане, качване на стълби, седене, стоене и ритане на топка;

Методология

- Проверява се за липсващи данни и форматите на съответните величини;
- Нормализира се величините за линейно ускорение по X, Y и Z като се изважда средна стойност и се разделя на стандартното отклонение.

Методология

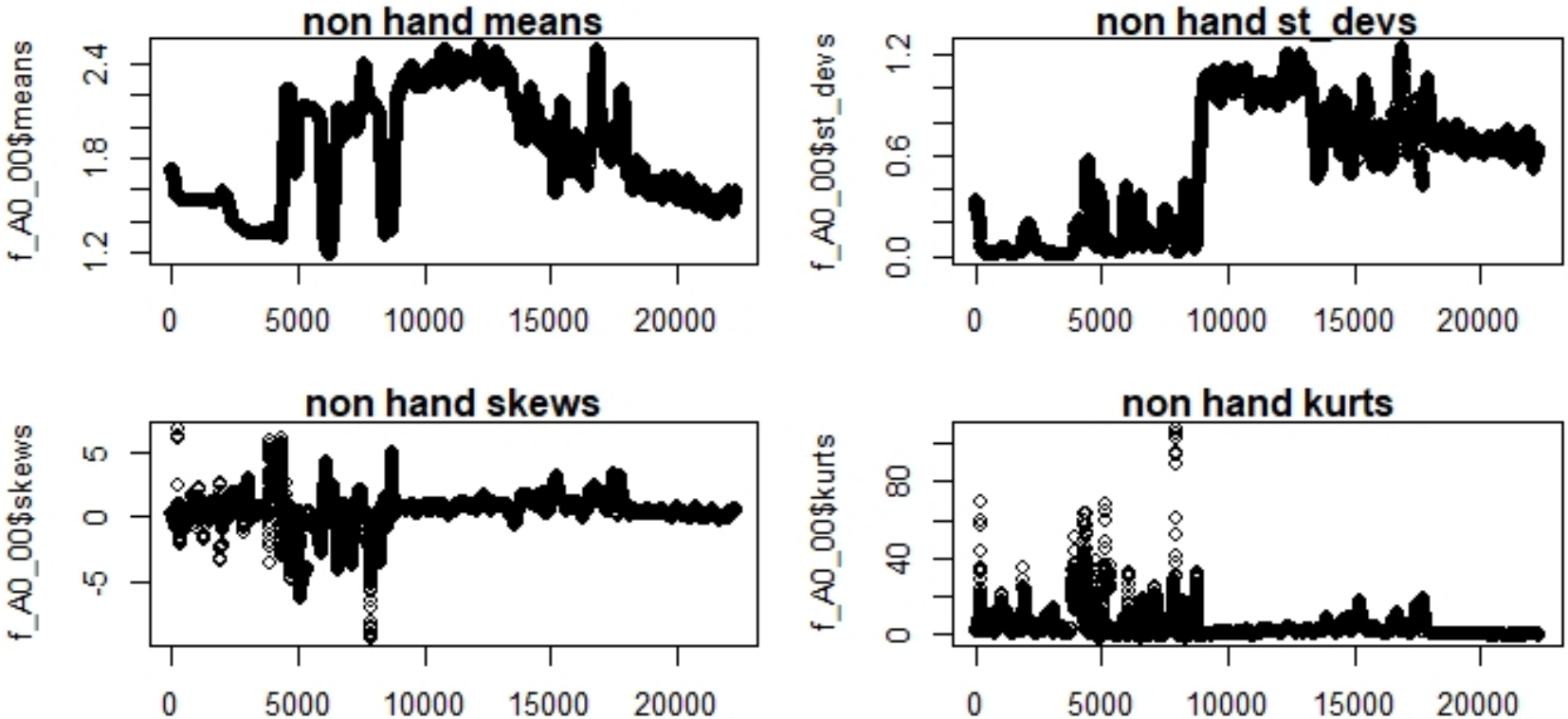
- Създаваме се един нов параметър, който е корен квадраен от сумата на квадратите на X, Y и Z наречен “dist”.
- Използва се метода на плъзгащия се прозорец с 200 припокриващи се точки;
- Използват се главните моменти: средно, стандартно отклонение, асиметрия и ексцес, за да се характеризира рапределението на параметъра “dist” по редове;
- Създават се две подбази за трениране на модела и за тестване;
- Създават се модели на логистична регресия, които се сравняват и се изпозват за прогнози на нови данни.

Резултати

- Проверка за качаството на данните не показва липсващи данни.
- Променливата Z се представя в числов вид;
- Индикаторната променлива се представя като фактор.

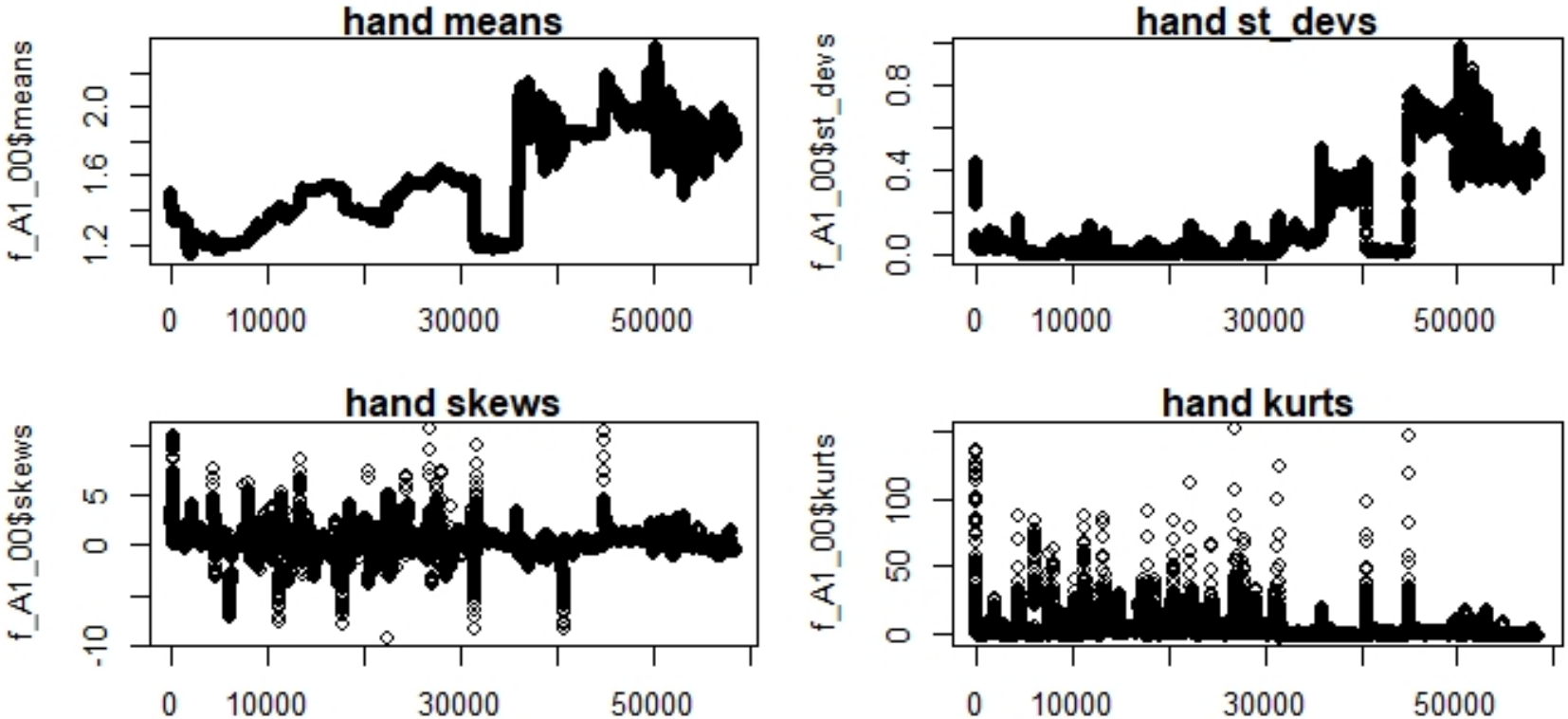
Резултати

Визуализация на средно, стандартно отклонение, асиметрия и ексцес за дейности не извършвани активно с ръце



Резултати

Визуализация на средно, стандартно отклонение, асиметрия и ексцес за дейности извършвани активно с ръце



Резултати

Логистична регресия с предиктор средно
Call:
$glm(formula = hand \sim means, family = "binomial\ data = df\_train)$
Deviance Residuals:
Min 1Q Median 3Q Max
-2.2432 -0.8353 0.5247 0.8471 1.5583
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.80702 0.05929 97.95 <2e-16 ***
means -2.83825 0.03331 -85.21 <2e-16 ***
—
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
AIC: 58105
Number of Fisher Scoring iterations: 4

Резултати

Логистична регресия с предиктор средно и стандартно отклонение

Call:

$glm(formula = hand \sim means + st\_devs, family =$

$"binomial\ data = df\_train)$

Deviance Residuals:

Min 1Q Median 3Q Max

-2.1557 -0.5930 0.4775 0.5328 1.8166

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.98201 0.07356 40.54 <2e-16 ***

means -0.55445 0.04829 -11.48 <2e-16 ***

st_devs -3.24003 0.04691 -69.07 <2e-16 ***

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

AIC: 52674

Number of Fisher Scoring iterations: 4

Резултати

Логистична регресия с предиктор средно, стандартно отклонение и асиметрия

Call:

$glm(formula = hand \sim means + st\_devs + skews, family =$ "$binomial\ data = df\_train$)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.3217 -0.5927 0.4710 0.5363 1.8027

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.78808 0.07593 36.719 <2e-16 ***

means -0.42171 0.05004 -8.428 <2e-16 ***

st_devs -3.45559 0.05198 -66.480 <2e-16 ***

skews 0.12124 0.01206 10.053 <2e-16 ***

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

AIC: 52574

Резултати

Логистична регресия с предиктор средно, стандартно отклонение асиметрия и ексцес

Call:

$glm(formula = hand \sim means + st\_devs + skews + kurts, family =$ "$binomial\ data = df\_train$)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2322 -0.5857 0.4570 0.6217 2.4507

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.884407 0.076560 37.675 < 2e-16 ***

means -0.391812 0.050393 -7.775 7.53e-15 ***

st_devs -3.616331 0.052652 -68.683 < 2e-16 ***

skews 0.144376 0.011064 13.049 < 2e-16 ***

kurts -0.034541 0.001669 -20.692 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

AIC: 52157

Резултати

Таблица 3: Реални към предсказани стойности от данните за трениране

	Реални	
Предсказани	0	1
0	7721	3926
1	7923	37047

Резултати

Таблица 4: Реални към предсказани стойности от данните за тестване

	Реални	
Предсказани	0	1
0	3358	1649
1	3347	15912

Заключение

- Четирите модела са статистически значими;
- Стойността на “AIC” намалява;
- Ще избререм модела с най-малко AIC, четвърти в случая;
- Моделът познава средно в 79% от случаите при нови данни;
- Моделът познава 90% от случаите когато се използват ръце;
- Моделът познава 50% от случаите когато не се използват ръце.