Uvod v strojno učenje

Jure Žabkar
jure.zabkar@fri.uni-lj.si

Predavanja

Torek, 17:00 Skype (oz. predavalnica P3, če/ko bo možno v živo)

April 14	April 21	April 28	
Maj 5	Maj 12	Maj 19	Maj 26

Consultations

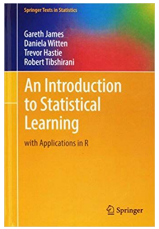
AI Lab @ FRI, R3.54

Please send an email including:

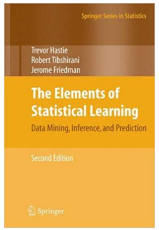
- the topic you would like to discuss
- your time constraints

to *jure.zabkar@fri.uni-lj.si*

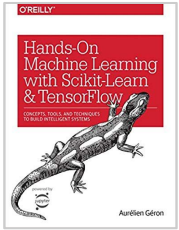
Literatura



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). ***An introduction to statistical learning*** (Vol. 6). NY: Springer.



Friedman, J., Hastie, T., & Tibshirani, R. (2009). ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction***. Springer Series in Statistics.



Geron, A. (2017). ***Hands-on machine learning with Scikit-Learn and TensorFlow***. O'Reilly.

Ocenjevanje

2 domači nalogi (2 x 15%)

Projekt (70%):

- Tema po izbiri
- Oddaja predloga na spletni učilnici
- Ustni zagovor

Podatkovne zbirke

UC Irvine Machine Learning Repository

Kaggle datasets

Amazon’s AWS datasets

http://dataportals.org/

http://opendatamonitor.eu/

http://quandl.com/

Strojno učenje

(angl. Machine Learning)

ParkinsonCheck™

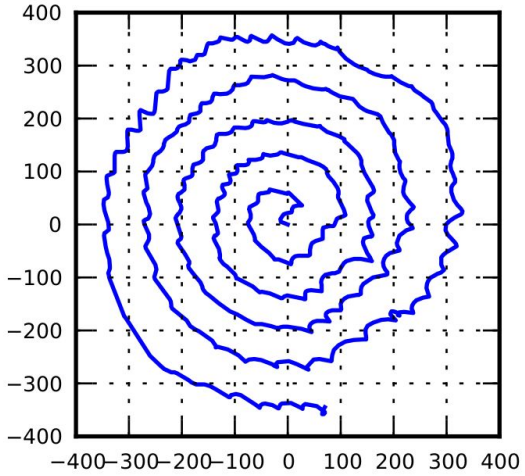
- A smartphone app for (early) detection of motoric signs of Parkinson's disease and some other tremors
- Freely available in Slovenia
- A built-in expert system enables users to use it in their home environment
- Fully standalone, no need to communicate with an outside server or sensor
- Based on spirography, but enhanced with other sensors, e.g. accelerometry

http://www.parkinsoncheck.net/

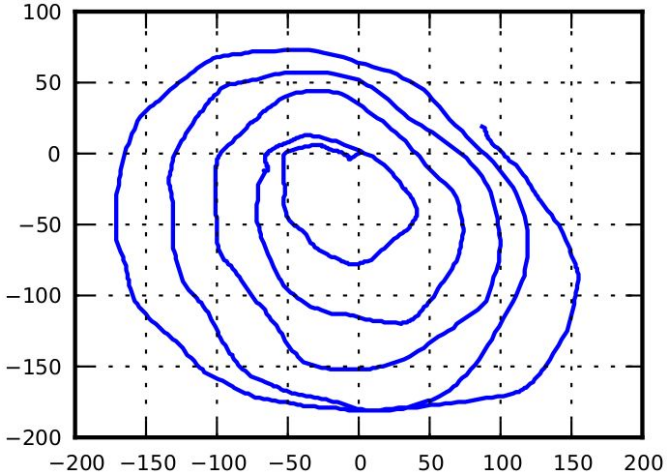
Early detection,

Patient monitoring

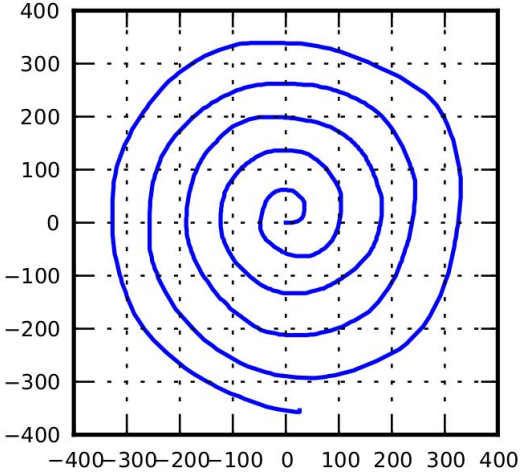
Essential tremor



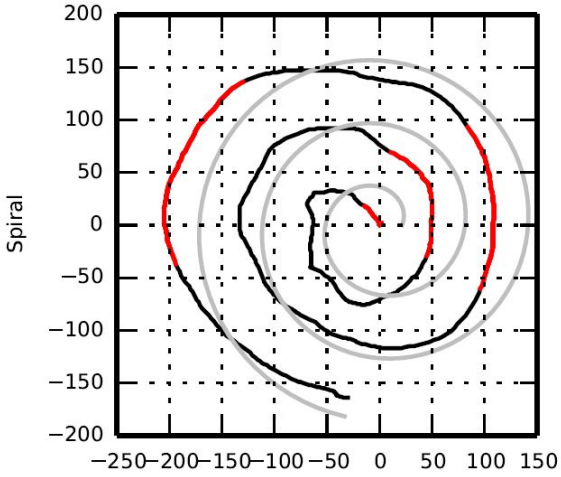
Parkinsonian tremor

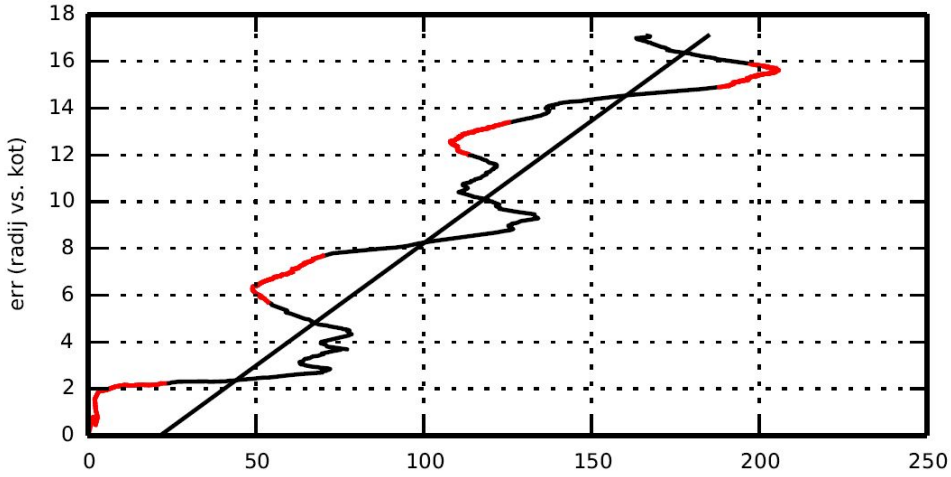


Healthy



control


Spiral
200
150
100
50
0
−50
−100
−150
−200
−250 −200 −150 −100 −50 0 50 100 150

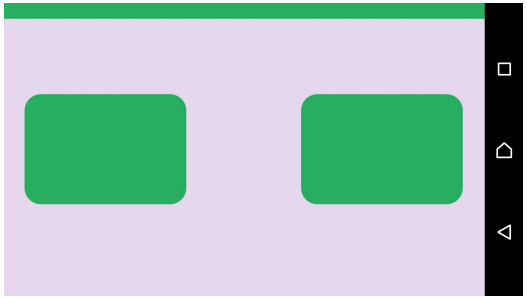

err (radij vs. kot)
18
16
14
12
10
8
6
4
2
0
0 50 100 150 200 250

The coefficients of the logistic regression with pre-discretization of attributes. Ten most influential attributes (by beta value) are given.

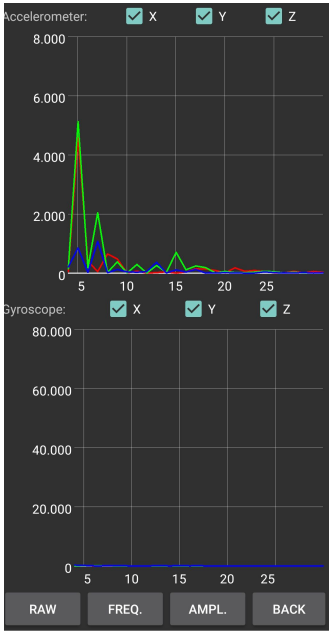
Attribute	Importance	General description
radSp.avgP.min	1.13	radial speed variability
tangSp.avgP.min	1.02	tangential speed variability
absSp.avgP.min	0.78	absolute speed variability
plrErrComCnt.avg	0.70	level of curvature/smoothness of the spiral
radSp.percNeg005.min	0.69	percentage of time the patient drew towards the centre
plrErrComCnt.max	0.67	level of curvature/smoothness of the spiral
plrErrFit.avg	0.65	general misfit from the ideal spiral (template)
tangSp.avgP.rng	0.65	tangential speed variability
tangSp.avgP.max	0.63	tangential speed variability
rot.avgP.min	0.62	number of times the spiral crosses itself

Digitizing UPDRS

Tapping test



Accelerometry



Finger tapping


QUIERO

H2020 project QUIERO (motivation)

- >30 million MRI scans per year in EURAMET countries,
- one of the most important tomographic tools adopted in clinical practice,
- MRI scan results interpreted by visual inspection,
- limited objectivity and comparability

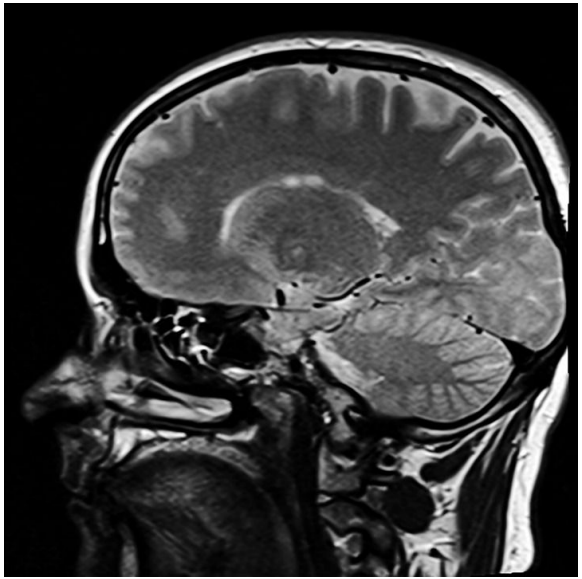
QUIERO (aim)

To provide a full metrological characterization of Electric

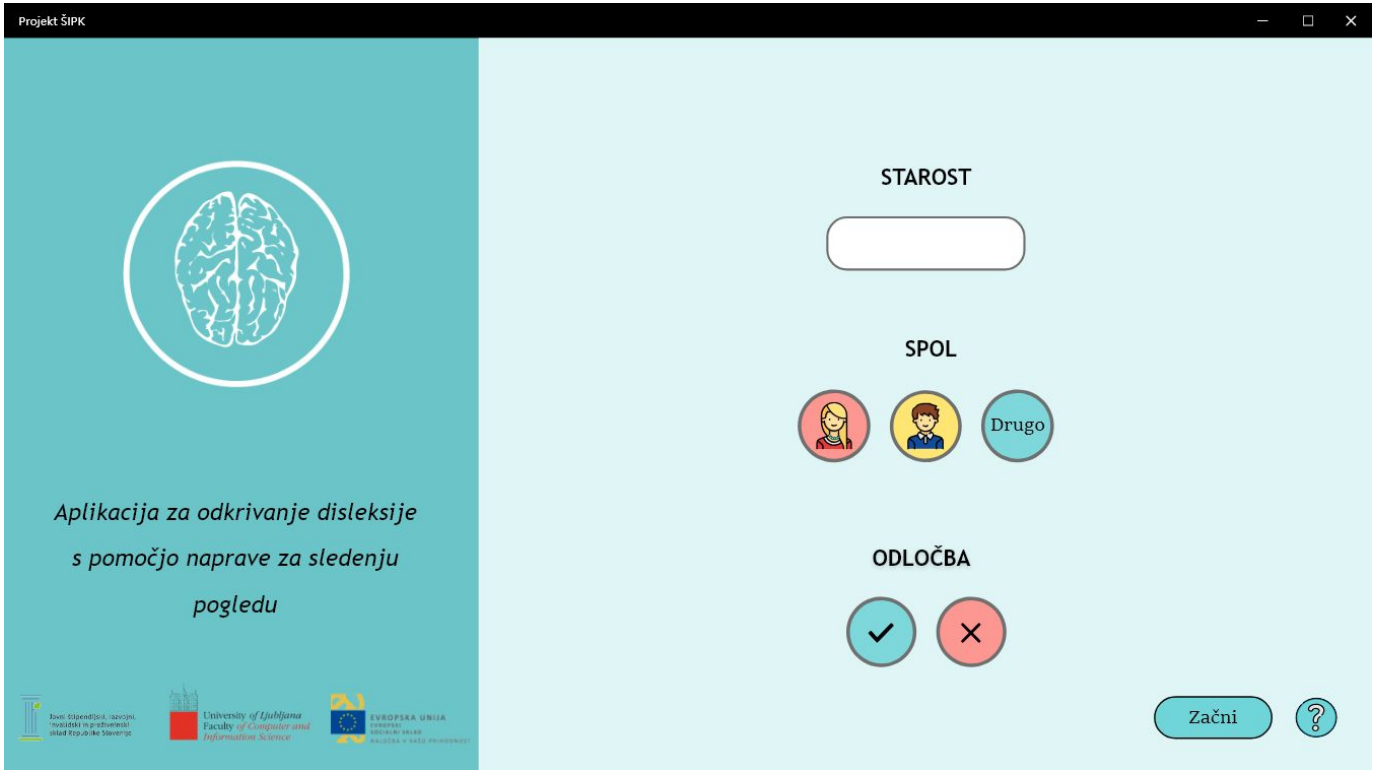
Properties Tomography (EPT) &

Magnetic Resonance Fingerprinting (MRF)

to underpin their use in clinical practice.



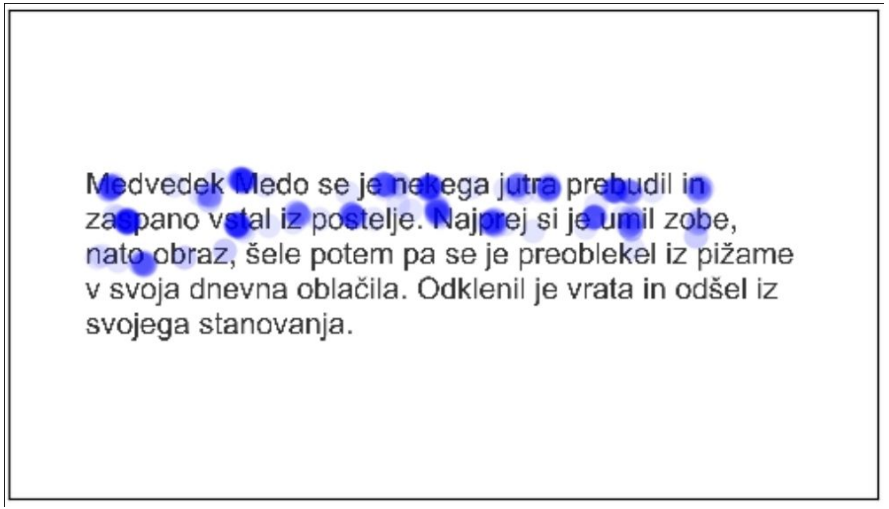
Dislex.AI, Screening for Dyslexia



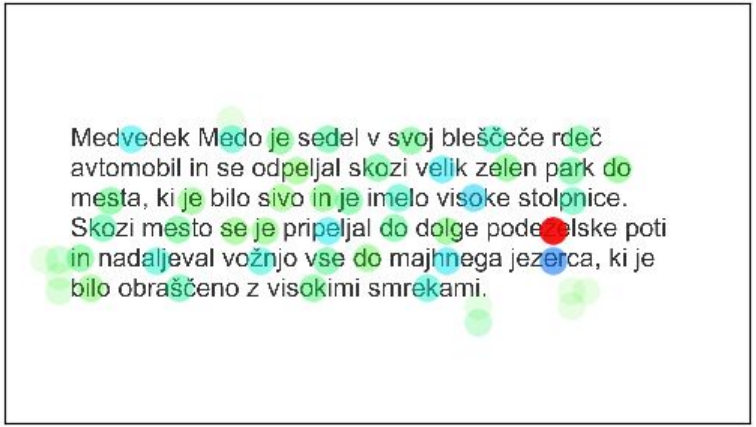
Screening for Dyslexia

AIM:

To develop a set of digital tests for assessing dyslexia in primary school children.



Dyslectic vs. non-dyslectic



Priporočilni sistemi

Prilagajanje uporabniku

Prepoznavanje obraza

Google

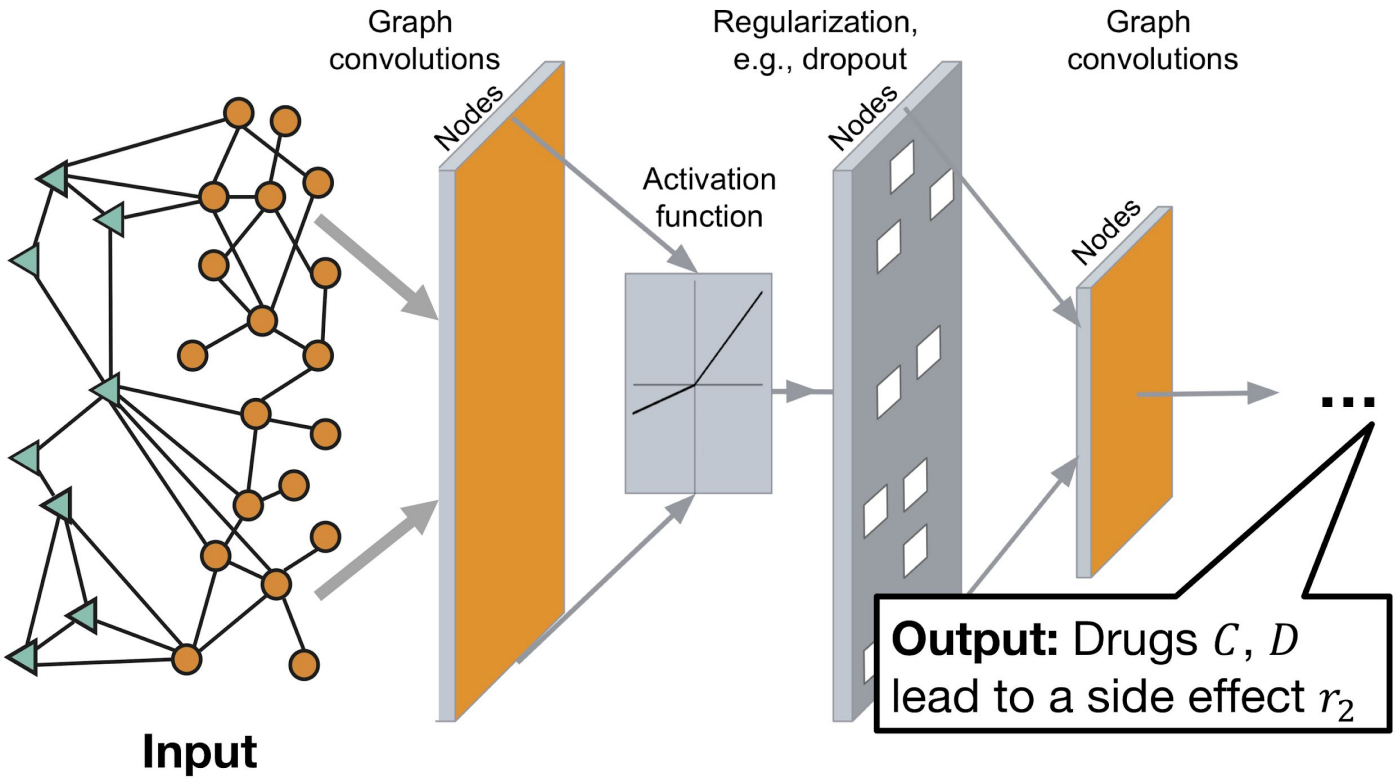
Prepoznavanje govora

amazon alexa

Robotika

Boston Dynamics

Odkrivanje zdravil



Igre, delnice

AlphaGo

IBM WATSON

Orodja

Orange

Scikit-learn

orange

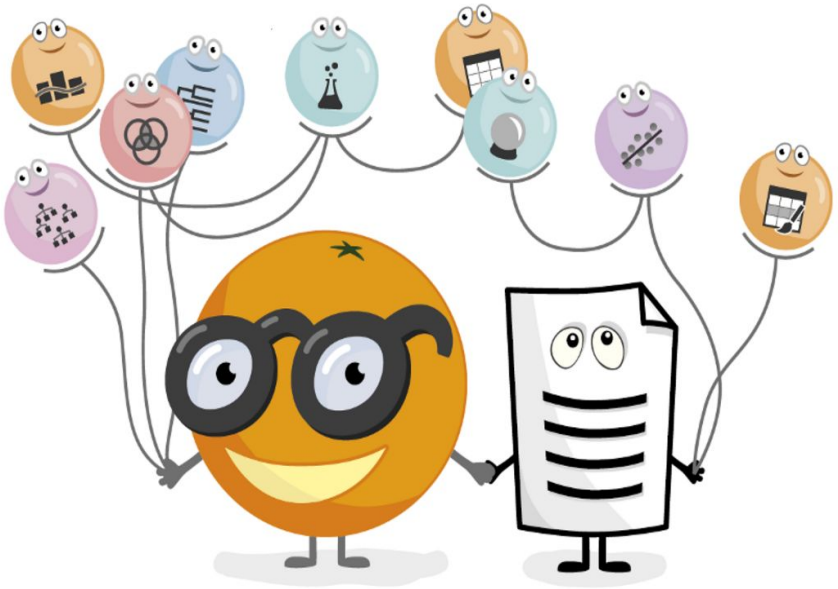
Screenshots Download Docs Blog

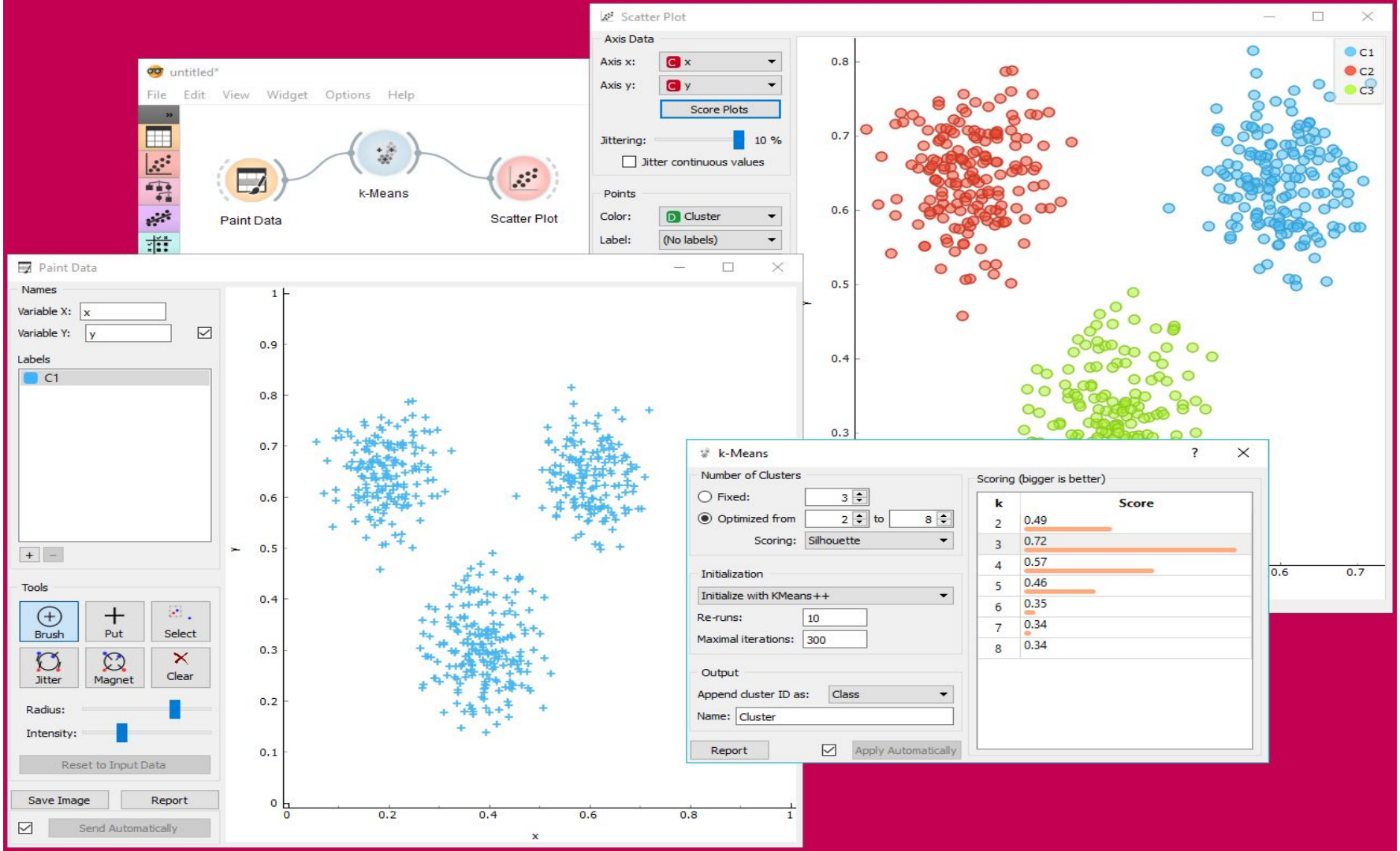
Data Mining
Fruitful and Fun

Open source machine learning and data visualization for novice and expert. Interactive data analysis workflows with a large toolbox.

Download Orange

The old version, Orange 2.7, is still available.




untitled*
File Edit View Widget Options Help
Paint Data
k-Means
Scatter Plot
Scatter Plot
Axis Data
Axis x: x
Axis y: y
Score Plots
Jittering: 10 %
Jitter continuous values
Points
Color: Cluster
Label: (No labels)
C1
C2
C3
Paint Data
Names
Variable X: x
Variable Y: y
Labels
C1
Tools
Brush
Put
Select
Jitter
Magnet
Clear
Radius:
Intensity:
Reset to Input Data
Save Image
Report
Send Automatically
k-Means
Number of Clusters
Fixed: 3
Optimized from 2 to 8
Scoring: Silhouette
Initialization
Initialize with KMeans++
Re-runs: 10
Maximal iterations: 300
Output
Append cluster ID as: Class
Name: Cluster
Report
Apply Automatically
Scoring (bigger is better)
k Score
2 0.49
3 0.72
4 0.57
5 0.46
6 0.35
7 0.34
8 0.34

Kaj je strojno učenje?

Klasično programiranje

Klasično programiranje

Strojno učenje

Klasično programiranje

Strojno učenje

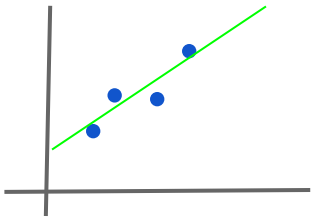
"Field of study that gives computers the ability to learn without being explicitly programmed"

- Arthur Samuel, 1959

Strojno učenje

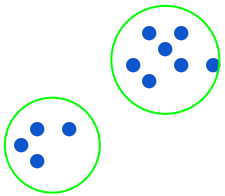
Nadzorovano

Regresija, Klasifikacija

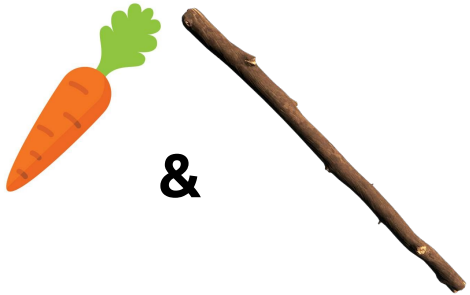


Nenadzorovano

Gručenje, povezovalna pravila



Spodbujevalno učenje



Regresija

Klasifikacija

Predstavitev značilk

Značilke nikoli popolnoma ne opišejo domene.

“All models are wrong, but some are useful.” - George Box

Predstavitev značilk

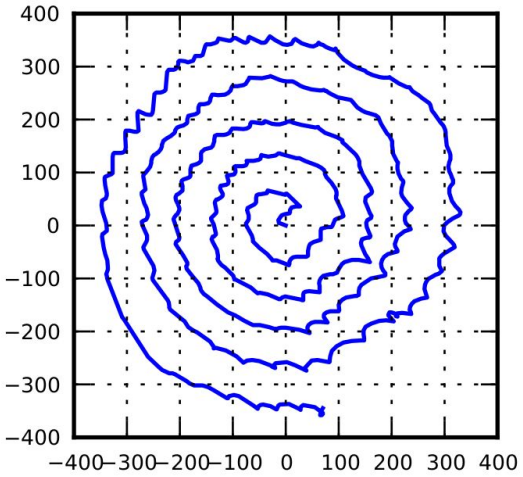
- Vektorji značilk predstavljajo učne primere.
- Na podlagi katerih značilk bi lahko napovedovali uspešnost študentov pri "Uvodu v strojno učenje" na začetku semestra?
- Npr. "povp. ocena že opravljenih izpitov", "znanje programiranja", ... se zdita smiselni značilki?

Predstavitev značilk

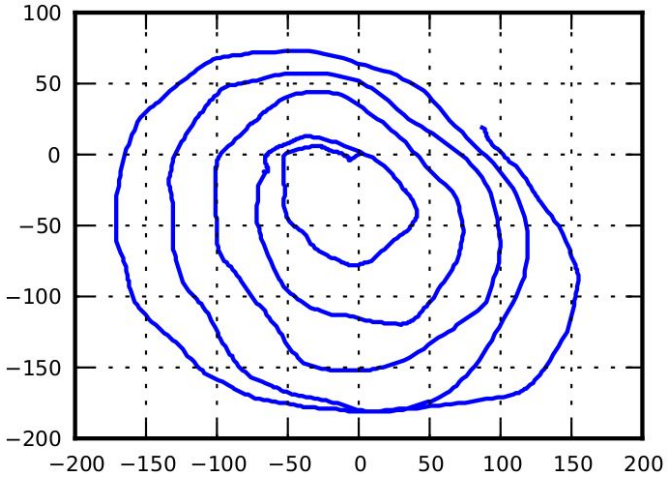
- Kaj pa "mesec rojstva" ali "barva oči"?
- Pretirano prileganje podatkom ali ...
- ... napačna interpretacija
- "some features matter, others don't":

  zakaj ne vržemo kar vseh noter in vidimo, kaj pade ven?

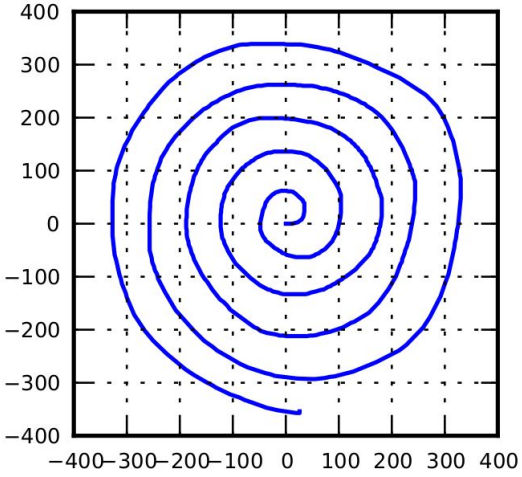
Esencialni tremor

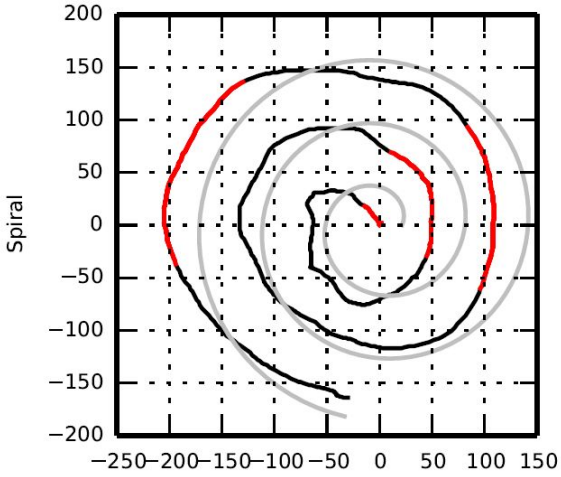


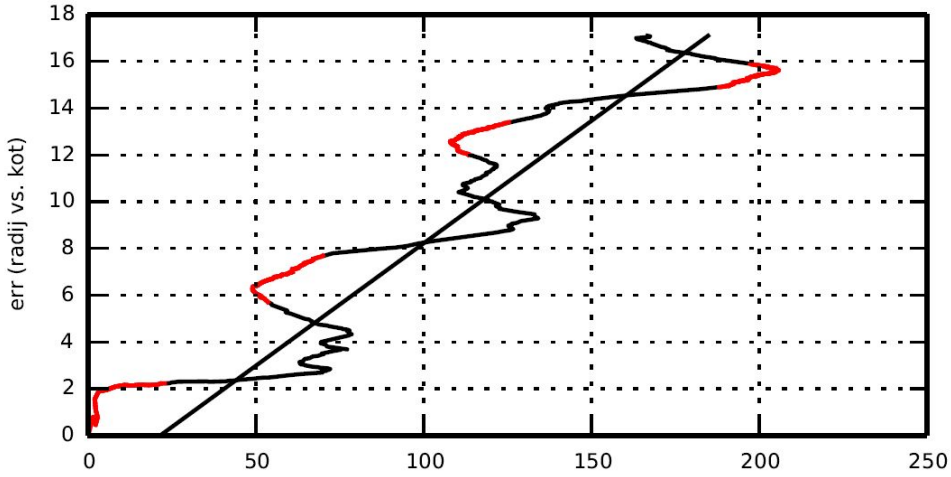
Parkinsonski tremor



Zdrava kontrola




Spiral
200
150
100
50
0
−50
−100
−150
−200
−250 −200 −150 −100 −50 0 50 100 150


err (radij vs. kot)
18
16
14
12
10
8
6
4
2
0
0 50 100 150 200 250

Attribute	Importance	General description
radSp.avgP.min	1.13	radial speed variability
tangSp.avgP.min	1.02	tangential speed variability
absSp.avgP.min	0.78	absolute speed variability
plrErrComCnt.avg	0.70	level of curvature/smoothness of the spiral
radSp.percNeg005.min	0.69	percentage of time the patient drew towards the centre
plrErrComCnt.max	0.67	level of curvature/smoothness of the spiral
plrErrFit.avg	0.65	general misfit from the ideal spiral (template)
tangSp.avgP.rng	0.65	tangential speed variability
tangSp.avgP.max	0.63	tangential speed variability
rot.avgP.min	0.62	number of times the spiral crosses itself

Problemi

- kako preprečiti pretirano prilagajanje?
- kako ocenjevati model?
- kako izbrati najboljše značilke?