Vilniaus Universitetas

Matematikos ir Informatikos fakultetas

Daugiamačių duomenų vizualizavimas

4 Užduotis

Daugiamatinės skalės
Multidimensional scaling (MDS)

Laimonas Beniušis
Informatika 1gr.

Turinys

1 Atributai

Duomenų aibė yra nėra pilna (255 iš 5172 įrašų). Ne visi atributai yra tinkami vizualizavimui, todėl palikti tik šie:

charSeq1	Raidinių žodžių kiekis, kurių ilgis 1 / žodžių kiekis
charSeq2	Raidinių žodžių kiekis, kurių ilgis 2 / žodžių kiekis
charSeq3	Raidinių žodžių kiekis, kurių ilgis 3 / žodžių kiekis
charSeq4	Raidinių žodžių kiekis, kurių ilgis 4 / žodžių kiekis
charSeq5	Raidinių žodžių kiekis, kurių ilgis 5 / žodžių kiekis
charSeq6	Raidinių žodžių kiekis, kurių ilgis 6 / žodžių kiekis
charSeq7	Raidinių žodžių kiekis, kurių ilgis 7 / žodžių kiekis
charSeq8	Raidinių žodžių kiekis, kurių ilgis 8 / žodžių kiekis
charSeq9	Raidinių žodžių kiekis, kurių ilgis 9 / žodžių kiekis
whiteSpaceCount	Tarpinių simbolių (*whitespace*) kiekis / simbolių kiekis
spam	{SPAM,NOTSPAM} E.laiško kategorija

Buvo panaudota *Pandas* Python techninė įranga.

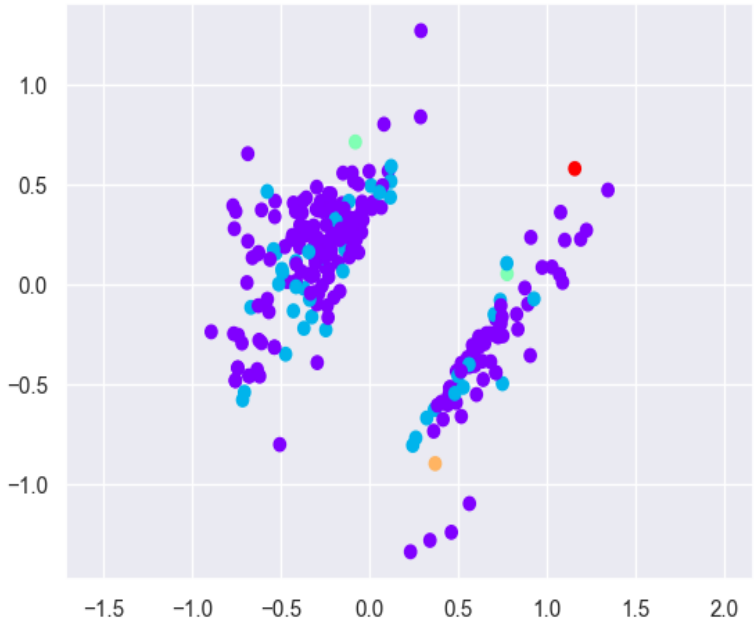
2 MDS vizualizavimas, interpretacija

Kategorijos (spam) požymis buvo pakeistas į skaitinį:

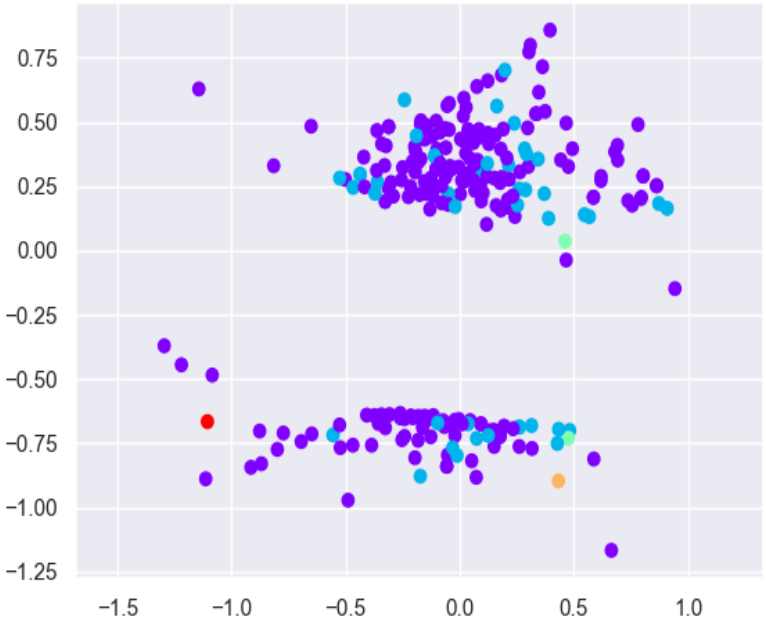
SPAM – 0

NOTSPAM – 1

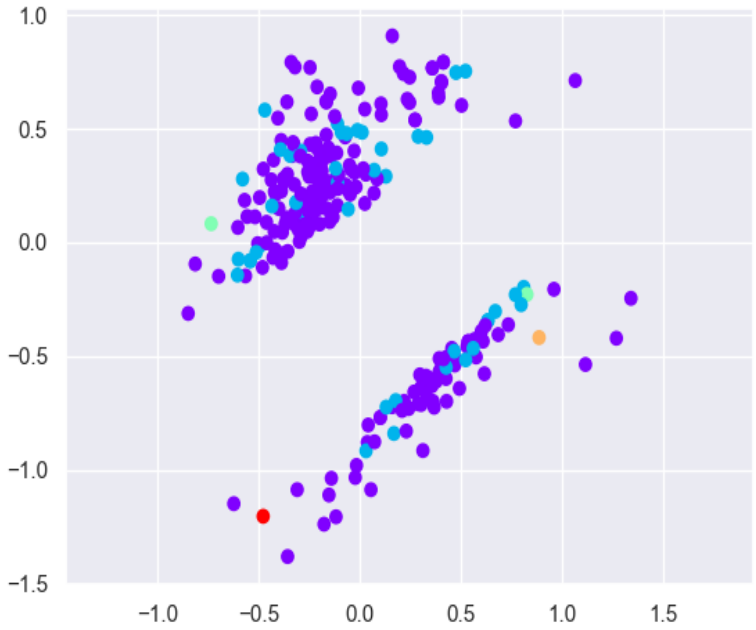
Buvo atliktas tų pačių duomenų daugiamatinių skalių metodo vizualizavimas su skirtingomis pradinėmis stadijomis (*random state).* Akivaizdžiai matome tarpą (duomenų taškų neturinčia sritį), kuris pasireiškia kiekviename iš paveikslėlių.. Taip pat, nuo pradinės stadijos mažai keičiasi duomenų išsidėstymas relatyviai (išlieka panaši forma), tačiau kinta orientacija. Pirmu atveju (žr 1. Pav.) vaizdas skiriasi ne tik orientacija, bet ir apsiverčia aplink ašį, kuri yra ortogonali minėtam tarpui.



1. Pav: MDS vizualizacija (atsitiktinė stadija 1)



2. Pav: MDS vizualizacija (atsitiktinė stadija 2)



3. Pav: MDS vizualizacija (atsitiktinė stadija 5)

3 Išvados

Galima pastebėti, kad nėra didelės duomenų dispersijos (dauguma įrašų yra panašūs). Tai galima pastebėti iš to, kad vyrauja dvi spalvos (mėlyna ir violetinė). Tik keli (4 iš 255) įrašai įgauna raudoną, žalią ar oranžinę kas gali parodyti, kad šie e. laiškai yra žymiai kitokio pobudžio, negu dauguma.

MDS metodas el. laiškų filtravimui nėra tinkamas, tačiau tinka įspūdžio apie laiškų turinio panašumą susidarymui.