Metody Matematyczne w Transporcie 2018/19 Transport sem. II, studia magisterskie

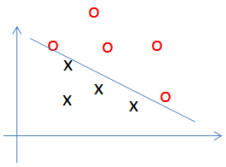
Laboratorium 7. Analiza wielowymiarowa

Analiza wielowymiarowa

Ranking

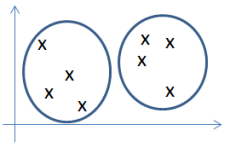
Uporządkowanie obiektów o *n* cechach od najlepszego do najgorszego

Klasyfikacja



Uczenie z nauczycielem

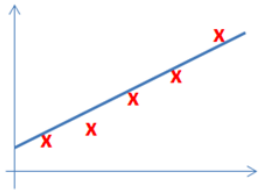
Grupowanie



Uczenie bez nauczyciela

Estymacja

Predykcja



$$h_\Theta(x) = \Theta_0 + \Theta_1 x$$

Wykrywanie reguł

$$p \rightarrow q$$

Klasyfikacja

Analiza dyskryminacyjna

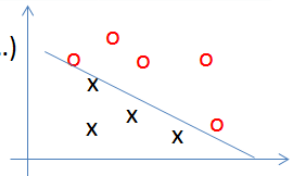
Przykład 3 – Badanie wiarygodności kredytowej klienta

Majątek	Dochód	Zadłużenie	Liczba członków rodziny	Liczba lat pracy u obecnego pracodawcy	Spłacił kredyt / nie spłacił kredytu
98	35	12	4	4	1 x
65	44	5	3	1	1 x
22	50	0	2	7	1 x
8	23	12	2	1	0 o
0	15	10	4	2	0 o
21	12	28	3	2	1 x
57	39	13	5	8	0 o
...					

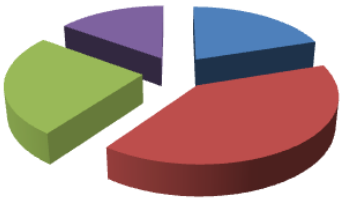
Nowy klient (majątek, dochód, zadłużenie, liczba członków rodziny, ...)

↓ ?

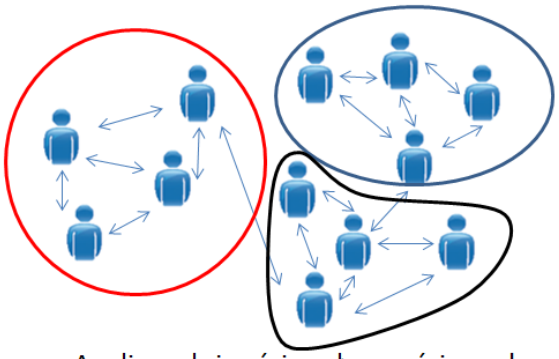
Spłaci kredyt / nie spłaci kredytu



Grupowanie - zastosowania

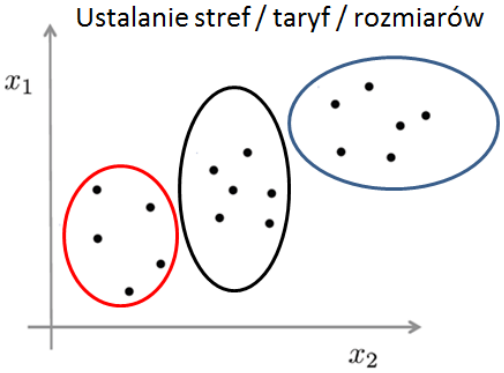


Segmentacja rynku

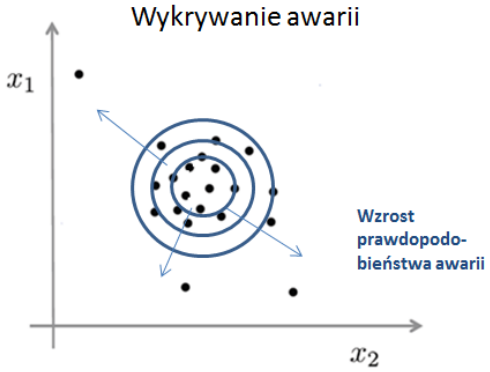


Analiza zależności społecznościowych

Ustalanie stref / taryf / rozmiarów



Wykrywanie awarii



Zadanie 1. Ranking

W oparciu o dane w załączonym pliku zbuduj ranking państw uwzględniający warunki życia.

	Przeciętna długość życia w latach	Średni czas trwania edukacji w latach	Emisja CO2 ze spalania paliw w mln ton w 2009 r.	Studenci szkół wyższych na 10 tyś ludności w 2008/09	Użytkownicy Internetu na 1000 osób w 2009	Stopa bezrobocia długookresowego w	PKB na osobę	Wskaźnik zagrożenia ubóstwem	Wskaźnik zabójstw
Argentyna	75,7	9,3	166,6	546,964654	340	0	14603	30	5,5
Australia	81,9	12	394,9	508,765853	743	0,8	38692	10	0,51
Austria	80,4	9,8	63,4	299,865434	735	1	37056	16,6	0,51
Belgia	80,3	10,6	100,7	378,15115	762	3,5	34873	20,8	1,74
Białoruś	69,6	9,3	60,8	543,828687	274	0	12926	50	4,9
Brazylia	72,9	12,1	337,8	250,219016	392	0	4902	35	23
Bułgaria	73,7	9,9	42,2	310,336029	450	3	11139	41,6	1,91
Chile	78,8	9,7	64,93151	413,636968	413	0	13561	40	1,33
Chiny	78,8	9,7	6831,596	159,062028	289	0	13561	52,5	1,12
Chorwacja	76,7	9	19,76649	305,66959	506	5,1	16389	45	1,11
Estonia	73,7	12	14,66291	508,483861	725	3,8	17168	21,7	7,1
Finlandia	80,1	10,3	55,00907	590,164756	825	1,4	33872	16,9	2,3
Francja	81,6	10,4	354,3013	363,178206	716	3,3	34341	19,3	1,09
Grecja	79,7	10,5	90,21718	589,883989	445	3,9	27580	27,7	1,35
Hiszpania	81,3	10,4	283,3701	424,889875	626	4,3	29661	25,5	0,9
Irlandia	80,3	11,6	39,5	453	674	3,4	33078	38	1,35
Japonia	83,2	11,5	1092,859	320	780	1,4	34692	28	0,4
Kanada	81	11,5	520,7455	419	803	0,6	38668	10	2,4
Norwegia	81	12,6	37,3	467	921	0,5	58810	8	0,6
Polska	76	10	293,3	563	590	2,5	17803	27,8	1,29
Rumunia	73,2	10,6	78,4	437	366	2,2	12844	41,4	1,96
Stany Zjednoczone	79,6	12,5	5195	583,776443	780	1,5	25438	35	5
Szwajcaria	82,2	10,3	42,42341	286,450524	813	1,2	39849	12	0,66
Szwecja	81,3	11,6	41,70542	457,688005	908	1,1	36936	15	0,86
Turcja	72,2	6,5	256,3147	342,175944	364	3,5	13359	47	3,8
Ukraina	68,6	11,3	256,3894	615,108941	170	0	6535	67	7
Węgry	73,9	11,7	48,16015	434,646822	618	4,2	17472	29,9	1,39
Wielka Brytania	79,8	9,5	465,8019	392,761721	836	1,9	35087	23,1	1,17
Włochy	81,4	9,7	389,2823	346,435051	488	3,5	29619	24,5	0,87

Krok 1. Określ własną nominantę / nominanty (własne preferencje)

Krok 2. Określ stymulanty i destymulanty

Krok 3. Wykonaj normalizację

Krok 4. Określ własne wagi (subiektywna ocena znaczenia cechy)

Krok 5. Oblicz wartość syntetycznego wskaźnika

Krok 6. Podziel zbiór państw na 3 podzbiory: najlepsze, przeciętne, najgorsze pod względem obliczonej wartości wskaźnika.

Ranking

- obiekt / zjawisko złożone, charakteryzujące się wieloma różnorodnymi cechami
- konieczność sporządzenia klasyfikacji obiektów złożonych od najlepszego do najgorszego

Przykład 1 – Ranking województw

Lp	Wojewó-dztwo	Lu-dność	PKB [mln zł]	Wynagr. Brutto [zł]	Liczba poja-zdów	Liczba stacji paliw	Drogi publi-czne	Drogi publiczne na 100 km^2
		[1]	[2]	[3]	[4]	[5]	[6]	[7]
1	Dolnośląskie		110448	3627.02	1655833	612	23468.0	117.7
2	Kujawsko-pomorskie		61721	3139.23	1253158	574	26480.8	147.3
3	Lubelskie		51082	3279.39	1322252	668	34012.7	135.4
4	Lubuskie		30358	3132.90	609231	338	13218.2	94.5
5	....							

Cechy [1-7]

Obiekt złożony

Cechy mają różne miana, ich wartości mogą znacząco różnić się

Przykład 2 – ranking marek samochodów

Lp.	Marka	Cena w tys zł	Moc silnika w KM	Pojemność bagażnika w dm^3	Przyspieszenie w s do 100 km/h	Zużycie paliwa na 100 km	Długość	Liczba poduszek powie-trznych
		X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	Ford Focus	56	100	350	11,5	8,5	4,4	4
2	Fiat Punto	38	60	200	13,5	7,0	3,8	1
3	Renault Clio	48	70	290	14,5	5,5	4,1	2
4	Opel Astra	36	60	440	14,0	9,5	5,0	0
5	Volkswagen Bora	46	75	350	13,0	6,5	4,5	2
6	Skoda Fabia	38	60	320	15,0	7,5	3,9	1
7	Seat Cordoba	50	75	500	13,0	7,0	4,2	2
8	Nissan Primera	56	90	470	11,0	8,5	5,4	2

Preferencje:
optymalna długość mieści się w przedziale <4 ; 4,6>

Normalizacja zmiennych - przypomnienie

Wyjściowy zbiór zmiennych						Y
x_1	x_2	x_3	X_4	x_5	x_6	
23	0.1	20000	2	120	24.3	
12	0.4	5000	3	234	12.1	
26	0.25	1000	20	13	10.9	

Podzbiór zmiennych diagnostycznych

Przekształcenie celem ujednolicenia

Normalizacja

$$x_{ij} = \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}}$$

Znormalizowany zbiór zmiennych						Y
x_1	x_2	X_3	x_4	x_5	x_6	
	0	1	0			
	1	0.2105	0.0505			
	0.5	0	1			

Etapy budowy rankingu 0-1

Założenia:

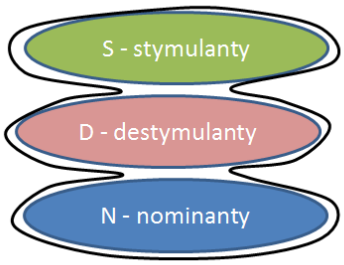
Dany jest zbiór O obiektów: $O = \{O_1, O_2, \dots, O_r\}$

Każdy obiekt jest opisany zbiorem zmiennych (cech) $X = \{X_1, X_2, \dots, X_s\}$

Etapy:

0. Wybór zmiennych diagnostycznych – zdefiniowanie zbioru X
1. Podział zbioru X na trzy podzbiory: S, D i N

$X = S \cup D \cup N$ spełniające warunek rozłączności $S \cap D = D \cap N = \emptyset$



Zbiór zmiennych diagnostycznych

Wzrost kojarzony jest ze wzrostem oceny obiektu, spadek ze spadkiem oceny

Wzrost kojarzony jest ze spadkiem oceny obiektu, spadek ze wzrostem oceny

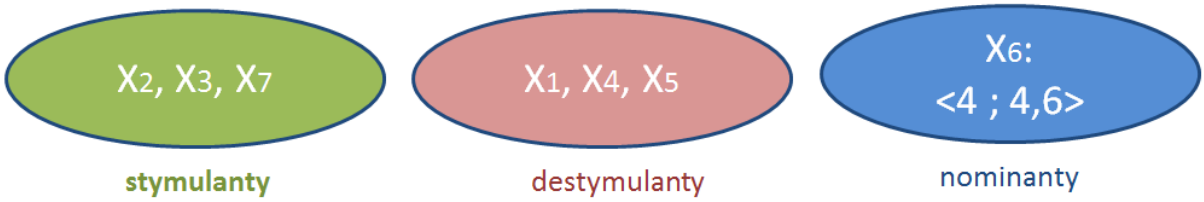
Posiada określony najkorzystniejszy przedział nominalny (wartość nominalną). Zarówno wartości większe jak i mniejsze kojarzone są ze spadkiem oceny obiektu.

Przykład 2 – ranking marek samochodów

Lp.	Marka	Cena w tys zł	Moc silnika w KM	Pojemność bagażnika w dm^3	Przyspieszenie w s do osiąg. 100 km/h	Zużycie paliwa na 100 km	Długość	Liczba poduszek powie-trznych
		X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	Ford Focus	56	100	350	11,5	8,5	4,4	4
2	Fiat Punto	38	60	200	13,5	7,0	3,8	1
3	Renault Clio	48	70	290	14,5	5,5	4,1	2
4	Opel Astra	36	60	440	14,0	9,5	5,0	0
5	Volkswagen Bora	46	75	350	13,0	6,5	4,5	2
6	Skoda Fabia	38	60	320	15,0	7,5	3,9	1
7	Seat Cordoba	50	75	500	13,0	7,0	4,2	2
8	Nissan Primera	56	90	470	11,0	8,5	5,4	2

Preferencje:

optymalna długość mieści się w przedziale <4 ; 4,6>



stymulanty | destymulanty | nominanty

Etapy budowy rankingu 2-4

2. Normalizacja stymulant $X_j \in S$

Normalizacja unitaryzacyjna, w której punktem odniesienia jest zakres zmiennej

<0,1> unitaryzacja zerowana

$$z_{ij} = \frac{x_{ij} - \min\limits_i x_{ij}}{\max\limits_i x_{ij} - \min\limits_i x_{ij}}$$

3. Normalizacja destymulant $X_j \in D$

$i = 1,2, \dots, r$

$j = 1,2, \dots, s$

$$z_{ij} = \frac{\max\limits_i x_{ij} - x_{ij}}{\max\limits_i x_{ij} - \min\limits_i x_{ij}}$$

4. Normalizacja nominant $X_j \in N$

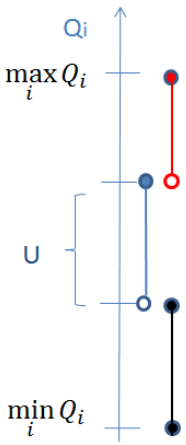
$$z_{ij} = \begin{cases} \dfrac{x_{ij} - \min\limits_i x_{ij}}{c_{1j} - \min\limits_i x_{ij}}, & gdy\ x_{ij} < c_{1j} \\ 1, & gdy\ c_{1j} \le x_{ij} \le c_{2j} \\ \dfrac{x_{ij} - \max\limits_i x_{ij}}{c_{2j} - \max\limits_i x_{ij}}, & gdy\ x_{ij} > c_{2j} \end{cases}$$

Etapy budowy rankingu 5-6

5. Obliczenie wartości zmiennej agregatowej (syntetycznej) dla każdego obiektu

$$Q_i = \sum_{j=1}^{s} z_{ij} \qquad i = 1,2, \dots, r$$

6. Podział zbioru obiektów na 3 grupy: najlepszych, przeciętnych i najgorszych



$$U = \frac{\max\limits_i Q_i - \min\limits_i Q_i}{3}$$

$Q_i \epsilon \left(\max\limits_i Q_i - U\ , \max\limits_i Q_i \right\rangle$ Obiekty najlepsze

$Q_i \epsilon \left(\max\limits_i Q_i - 2U\ , \max\limits_i Q_i - U \right\rangle$ Obiekty przeciętne

$Q_i \epsilon \left\langle \min\limits_i Q_i,\ \max\limits_i Q_i - 2U \right\rangle$ Obiekty najgorsze

Przykład 2

Lp.	Marka	Cena w tys zł	Moc silnika w KM	Pojemność bagażnika w dm^3	Przyspieszenie w s do 100 km/h	Zużycie paliwa na 100 km	Długość	Liczba poduszek powie-trznych	Qi
		Z1	Z2	Z3	Z4	Z5	Z6	Z7	
1	Ford Focus	0,000	1,000	0,500	0,875	0,250	1,000	1,000	**4,625**
7	Seat Cordoba	0,300	0,375	1,000	0,500	0,625	1,000	0,500	**4,300**
5	Volkswagen Bora	0,500	0,375	0,500	0,500	0,750	1,000	0,500	**4,125**
8	Nissan Primera	0,000	1,000	0,900	1,000	0,250	0,000	0,500	**3,650**
3	Renault Clio	0,400	0,250	0,300	0,125	1,000	1,000	0,500	**3,575**
4	Opel Astra	1,000	0,000	0,800	0,250	0,000	0,500	0,000	**2,550**
6	Skoda Fabia	0,900	0,000	0,400	0,000	0,500	0,500	0,250	**2,550**
2	Fiat Punto	0,900	0,000	0,000	0,375	0,625	0,000	0,250	**2,150**

$$U = \frac{4{,}625 - 2{,}150}{3} = 0{,}825$$

Zadanie 2. Estymacja parametrów modelu

Należy zbudować model do oceny warunków ruchu tzn. prognozujący średnią prędkość i średnią prędkość pojazdów ciężarowych, przy danej gęstości na podstawie: przewidywanego udziału pc, dnia tygodnia, pory dnia i pory roku.

Dane pomiarowe S6 Gdańsk Osowa. Średnia prędkość i liczba pojazdów.

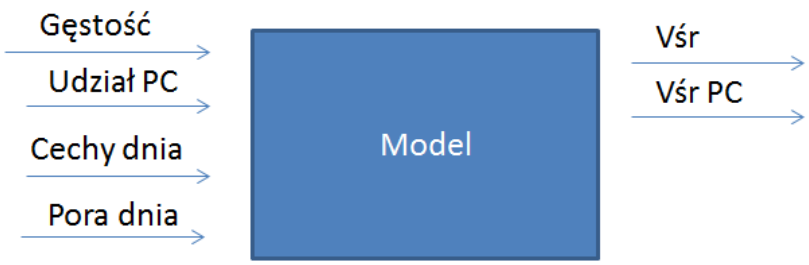
	A	B	C	D	E	F	G
1	interval	Udzial	Gestosc	avg_speed	avg_speed_pc	volume	volume_pc
2	2014-03-01 00:00	0,02	1,66	103,442718	103,80238	172	168
3	2014-03-01 01:00	0,21	1,57	96,56753247	102,609005	152	120
4	2014-03-01 01:15	0,03	1,34	92,65929178	93,16856989	124	120
5	2014-03-01 02:45	0,10	2,41	97,83729823	100,216316	236	212
6	2014-03-01 04:00	0,02	4,00	102,113448	102,4531379	408	400
7	2014-03-01 04:15	0,07	3,98	102,5476042	104,7450615	408	380

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	interval	Udzial	Gestosc	avg_speed	avg_speed_pc	volume	volume_pc	Gęstość_norm	AVG_speed	AVG_speed	CzyNoc	CzyLiczyt	CzyZima	CzyWakacj	CzyWeekend
2	2014-03-01 00:00	0,02	1,66	103,442718	103,80238	172	168	0,01108504	0,68961812	0,692015867	1	0	0	0	1
3	2014-03-01 01:00	0,21	1,57	96,56753247	102,609005	152	120	0,01049352	0,64378355	0,684060033	1	0	0	0	1
4	2014-03-01 01:15	0,03	1,34	92,65929178	93,16856989	124	120	0,008921573	0,617728612	0,621123799	1	0	0	0	1
5	2014-03-01 02:45	0,10	2,41	97,83729823	100,216316	236	212	0,01608112	0,652248655	0,668108773	1	0	0	0	1
6	2014-03-01 04:00	0,02	4,00	102,113448	102,4531379	408	400	0,02663704	0,68075632	0,68302092	1	0	0	0	1
7	2014-03-01 04:15	0,07	3,98	102,5476042	104,7450615	408	380	0,026524267	0,683650695	0,69830041	1	0	0	0	1
8	2014-03-01 04:30	0,10	3,67	98,16229281	100,5056302	360	324	0,024449307	0,654415285	0,670037535	1	0	0	0	1
9	2014-03-01 06:00	0,11	7,25	94,28510935	97,50033516	684	608	0,048363947	0,628567396	0,650002234	0	0	0	0	1
10	2014-03-01 07:30	0,05	15,66	97,08658488	98,41238946	1520	1440	0,104374187	0,647243899	0,656082596	0	1	0	0	1
11	2014-03-01 08:00	0,06	11,33	102,3482211	104,4272982	1160	1096	0,07555904	0,682321474	0,696181988	0	1	0	0	1

Na podstawie średniej prędkości i natężenia została obliczona gęstość. Na podstawie liczby pojazdów i liczby pojazdów ciężarowych został obliczony udział PC.

Na podstawie daty i godziny można wyznaczyć porę roku, dzień tygodnia, porę dnia itd.

Schemat:



Jak zbudować model? np. w oparciu o sieci neuronowe.

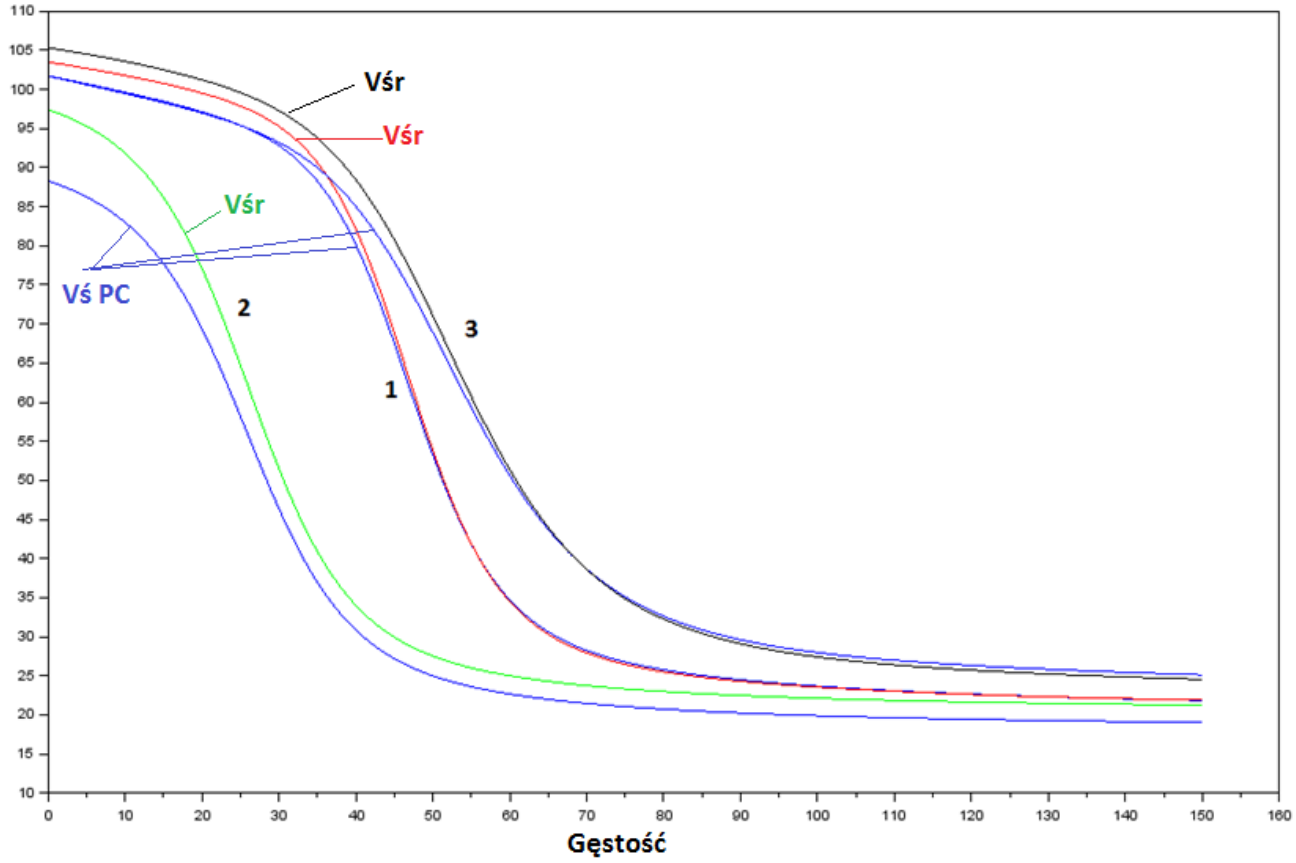
Budowa modelu w skrócie:

1. Kodowanie / reprezentacja danych wejściowych

   Obejmuje normalizację (mapowanie wartości wejściowych na przedział <0,1>). W przypadku naszych danych kodowanie może wyglądać następująco:

   Przyjmujemy gęstość min = 0 i gęstość max = 150, normalizujemy gęstość wg wzoru gęstość_norm = (gęstość – gęstość_min)/(gęstość_max – gęstość_min). Przyjmujemy prędkość min=0 i prędkość max = 150 (przykładowo), normalizujemy prędkość.

2. Dobór architektury sieci
3. Uczenie sieci (zbiór uczący ok. 70% danych)
4. Ocena wyników (zbiór testowy ok. 30 % danych)
5. Odpowiedzią modelu będą 2 krzywe (Vśr i Vśr PC). Poniżej przykład 3 różnych odpowiedzi modelu.



3 odpowiedzi uzyskano podając na wejściu modelu 3 różne zestawy parametrów wejściowych (Cechy dnia, Pora dnia, Udział PC). Odpowiedź modelu powinna pozwolić na ocenę warunków ruchu.

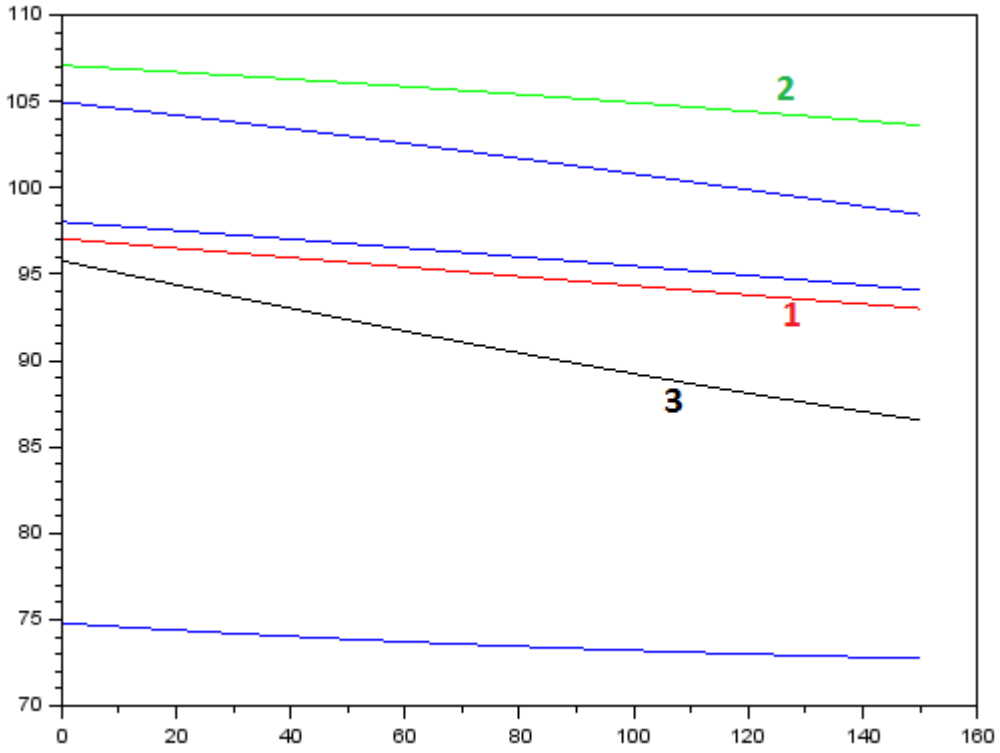
Opis przygotowania modelu w SciLabie

1. Plik dane_15.xlsx zawiera dane pomiarowe
2. Plik dane_15a.xlsx zawiera dane po normalizacji oraz z cechami dnia/godziny. Ponieważ gęstość min=0 i V_min = 0 to wzory normalizacyjne mają prostą postać wartość / wartość_max

3. Na podstawie pliku dane_15a.xlsx powstał plik dane_15c.txt. Zawiera dane przygotowane do wczytania w SciLabie. Posiada 9 kolumn (gestosc,udział,czyNoc,czySzczyt,czyZima,czyWakacje,czyWeekend,avg_speed,avg_speed_pc) i 102253 wiersze.
4. Załączam przykładowy plik w SciLabie sieci_gestosc, w którym następuje wczytanie 700 wierszy do zbioru uczącego i 300 wierszy do zbioru testowego, wybrana jest architektura sieci, przeprowadzony jest proces uczenia, wyznaczona jest odpowiedz sieci na zbiór uczący i testowy z policzeniem błędu. W przykładzie podałam wczytanie 1000 wierszy. Docelowo trzeba korzystać z całego zbioru. Podział wierszy na zbiór testowy i uczący powinien być wykonany losowo.
5. Plik odpytywanie (gęstości nie podajemy, zostanie wstawiona automatycznie)

Udział	czyNoc	czySzczyt	czyZima	czyWakacje	czyWeekend	Opis
0.3	0	1	1	0	0	Pytamy jaka będzie zależność Vśr od gęstości w trudnych warunkach (duży dział PC, szczyt, zima, dzień roboczy)
0.05	1	0	0	0	1	Niski udział PC, noc, weekend
0.1	0	0	0	1	0	Średni udział PC, wakacje, dzień

6. Wynik



Wynik odbiega od prezentowanego wyżej. Możliwe przyczyny: zbyt mała liczba wierszy, nieodpowiedni dobór cech opisujących dzień, zbyt niska złożoność architektury sieci itd. W 1000 wierszy mogły nie znaleźć się wiersze z wakacjami, szczytem itd.

W ramach zadania należy przeprowadzić eksperymenty z różnymi architekturami sieci oraz różnym zestawem zmiennych wejściowych. W ocenie wyników modelu należy posłużyć się miarą błędu.