Projekt NAI 2 Luty 2019

Mateusz Gruszka

Paweł Mazerant

Raport

Projekt: Rozpoznawanie chorób serca

Celem projektu jest stworzenie klasyfikatora binarnego do identyfikacji czy pacjent jest zdrowy czy chory na chorobę serca. Do rozpoznania wykorzystujemy dane zawierające objawy chorobowe i profile pacjentów.

Opis danych

Dane pochodzą ogólno dostepnego repozytorium[1] UC Irvine Machine Learning. Posiadamy 303 rekordy stworzone przez V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Każdy pacjent opisywany jest przez 76 atrybutów, 14 z nich jest istotnych, są to pokolei:

age - wiek (podany jako liczba).
sex - płeć (1 - mężczyzna, 0 - kobieta).
cp - dolegliwość w klatce (np. ból dławicowy typowy, ból dławicowy nietypowy, niedusznicowy, bezobjawowy. Odpowiednio otykietowane liczbami 1,2,3 i 4).
trestbps - ciśnienie krwi w stanie spoczynku (liczba wyrażona w jednostce mm/Hg).
chol - poziom cholesterolu w surowicy(liczba wyrażona w jednostce mg/dl).
fbs - cukier we krwi na czczo(Jeśli powyżej 120 mg/dl, etykieta 1 w przeciwnym wypadku 0).
restecg - wynik elektrokardiografu w stanie spoczynku (wynik 0 jeśli normalny, 1 posiadanie niepoprawnej fali ST-T, 2 - prawdopodobieństwo hipertorfii).
thalach - osiągnięte maksymalne tętno(wyrażone w jednostce uderzenia na minute).
exang - ból dławicowy wywołany ćwiczeniami (tak - 1, nie - 0).
oldpeak - zaobserowanie krzywej ST podczas ćwiczeń przypominającą spoczynek.
slope - nachylenie krzywej ST(1 - nachylenie w górę, 2 - płaska, 3 - nachylenie w dół).
ca - ilość głownych naczynek (liczba między 1 a 3).
thal - 3 - normalna wada, 6 - stała wada, 7 - odwracalna wada.
num - przewidywana diagnoza (powyżej 50% chory - 1,2,3,4, ponizej 50% zdrowy - 0).

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0

Przykładowe 5 rekorów.

Na potrzeby zadania dane podzielimy na zbiór treningowy i testowy (80% i 20%). Dodatkowo pominiemy rzędy danych w których dane są niekompletne. Końcowo mamy 237 rzedów do wykorzystania na trening modelu i 60 do przetestowania jakości rozwiązania.

Opis rozwiązania

Do rozwiązania wykorzystamy w pełni połączoną (ang. fully-connected) sieć neuronową.
W ramach eksperymentu przetestujemy i porównamy 10 rozwiązań na zbiorze testowym.

Test 1: liczba warstw: 3 (10 neuronów, 10 neuronów, 1 neuron), optymalizacja: sgd, współczynnik uczenia: 0.01).

Test 2: liczba warstw: 5 (14 neuronów, 12 neuronów, 10 neurów, 8 neuronów, 1 neuron), optymalizacja: sgd, współczynnik uczenia: 0.02.

Test 3: liczba warstw: 3 (13 neuronów, 5 neuronów, 1 neuron), optymalizacja: sgd, współczynnik uczenia: 0.7.

Test 4: liczba warstw: 6 (8 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 1 neuron), optymalizacja: sgd, współczynnik uczenia: 0.001.

Test 5: liczba warstw: 6 (13 neuronów, 10 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 1 neuron), optymalizacja: sgd, współczynnik uczenia: 0.001.

Test 6: liczba warstw: 14 (18 neuronów, 16 neuronów, 14 neuronów, 12 neuronów, 10 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 6 neuronów, 6 neuronów, 6 neuronów, 6 neuronów, 1 neuron), optymalizacja: sgd, współczynnik uczenia: 0.001.

Test 7: liczba warstw: 6 (13 neuronów, 10 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 1 neuron), optymalizacja: adam, współczynnik uczenia: 0.001.

Test 8: liczba warstw: 8 (13 neuronów, 10 neuronów, 8 neuronów, 8 neuronów, 8 neuronów, 10 neuronów, 10 neuronów, 1 neuron), optymalizacja: adam, współczynnik uczenia: 0.0001.

Test 9: liczba warstw: 4 (13 neuronów, 8 neuronów, 8 neuronów, 1 neuron), optymalizacja: adam, współczynnik uczenia: 0.001.

Test 10: liczba warstw: 6 (13 neuronów, 8 neuronów, 8 neuronów, 5 neuronów, 5 neuronów, 1 neuron), optymalizacja: adam, współczynnik uczenia: 0.002.

	Test 1	**Test 2**	**Test 3**	**Test 4**	**Test 5**	**Test 6**	**Test 7**	**Test 8**	**Test 9**	**Test10**
Trafność	53,3%	46,6%	53,3%	68%	71,6%	53,3%	78,3%	70%	80%	83%

Najlepsza konfiguracja z 10 eksperymentów:

Nazwa: Model Test 10

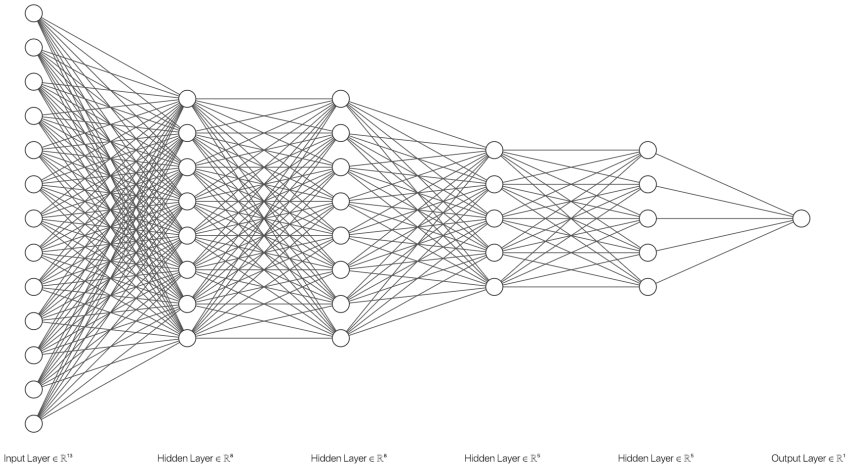
Warstw: 6

Neuronów: (13,8,8,5,5,1)

Optymalizacja: Adam[2]

Współczynnik uczenia: 0,002

Trafność: 83%



Referencje i źródła:

1. *https://archive.ics.uci.edu/ml/datasets/heart+Disease*
2. *https://arxiv.org/pdf/1412.6980v8.pdf*
3. *https://keras.io/getting-started/sequential-model-guide/*