Michał Goworko
Jakub Jakóbczyk
Łukasz Lepak

Sieć neuronowa przewidująca klasę ryzyka zachorowania na raka szyjki macicy na podstawie czynników ryzyka

prowadzący: dr inż. Tomasz Trzciński

1. Opis problemu

Zadaniem projektowanej sieci neuronowej będzie określenie klasy ryzyka zachorowania na raka szyjki macicy przez konkretną pacjentkę. Zadany zbiór zawiera dane 858 przypadków. Dzielą się one da dane o 32 czynnikach ryzyka:

Nazwa czynnika	Opis
Age	Wiek
Number of sexual partners	Liczba partnerów seksualnych
First sexual intercourse	Wiek inicjacji seksualnej
Num of pregnancies	Liczba ciąż
Smokes	Czy pali
Smokes (years)	Ile lat pali
Smokes (packs/year)	Ile paczek rocznie wypala
Hormonal Contraceptives	Czy używa antykoncepcji hormonalnej
Hormonal Contraceptives (years)	Ile lat używa antykoncepcji hormonalnej
IUD	Czy stosuje IUD
IUD (years)	Od ilu lat stosuje IUD
STDs	Czy choruje na choroby przenoszone drogą płciową
STDs (number)	Na ile chorób przenoszonych drogą płciową choruje
STDs:condylomatosis	Czy choruje na te konkretne choroby przenoszone drogą płciową – dane o 12 chorobach
…	
STDs:HPV	
STDs: Number of diagnosis	Liczba diagnoz chorób przenoszonych drogą płciową
STDs: Time since first diagnosis	Czas od pierwszego zdiagnozowania
STDs: Time since last diagnosis	Czas od ostatniego zdiagnozowania
Dx:Cancer	Pacjentka miała wcześniej zdiagnozowany nowotwór
Dx:CIN	Pacjentka miała wcześniej zdiagnozowany CIN
Dx:HPV	Pacjentka miała wcześniej zdiagnozowane HPV
Dx	Jedna z 3 diagnoz powyżej jest pozytywna

Nowotwór jest chorobą tak bardzo złożoną i poważną, że diagnozuje się go wykonując niezależnie kilka Źle obliczone wynikiań. W tym konkretnym przypadku są to 4 testy:

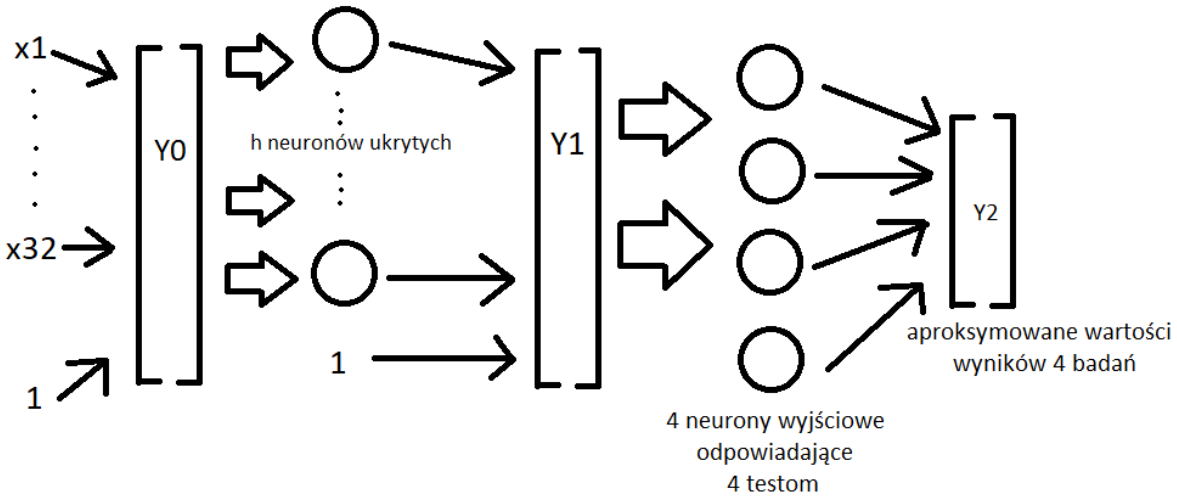
Nazwa testu	Opis
Hinselmann	Test Hinselmanna
Schiller	Test Shillera
Citology	Cytologia
Biopsy	Biopsja

Testy te dają wyniki negatywny (0) lub pozytywny (1). Klasa ryzyka zachorowania to liczba z zakresu 0-4 będąca ich sumą.

2. Opis algorytmu

2.1 Opis sieci i obliczanie sieci

Przedstawione w poprzednim punkcie zagadnienie jest klasycznym przypadkiem problemu aproksymacji funkcji. Do jego rozwiązania użyta zostanie sieć neuronowa. Wybrany został perceptron dwuwarstwowy, ze względu na jego względną prostotę i dobre właściwości aproksymacyjne.



Warstwom neuronów odpowiadają macierze współczynników W1 dla warstwy ukrytej i W2 dla warstwy wyjściowej

Wstępne przetwarzanie danych wejściowych

Zadany zbiór danych w niektórych miejscach jest niekompletny. Czynniki, które są niewiadome są uzupełniane w następujący sposób.

I. Number of sexual partners, First sexual intercourse, Num of pregnancies wartości losowane są zgodnie z rozkładem normalnym $X \sim \mathrm{N}(\mu, \sigma)$, gdzie μ jest średnią arytmetyczną danego czynnika z wszystkich znanych danych a σ jest obliczana z wzoru

$$\text{wzór (w2.1)} \sigma = max(\frac{(max(czynnik) - \mu)}{3}, 0.1) \blacksquare$$

II. Zestawy czynników: Smokes, Hormonal Contraceptives, IUD rozpatrywane są grupami. Jako, że pierwszy czynnik w każdej grupie oznacza wartość prawda/fałsz jest on wyznaczany jako pierwszy, zgodnie z poniższym wzorem

$$\text{wzór (w2.2)} wartość = \begin{cases} 0, jeżeli\ X > \mu \\ 1, jeżeli\ X < \mu \end{cases} \blacksquare$$

gdzie X jest zgodna z rozkładem jednostajnym $X \sim \mathrm{U} < 0{,}1 >$ a μ oznacza wartość średnią. Następnie w zależności od wylosowanej wartości, jeżeli jest 0 w pozostałe miejsca w grupie również wpisywane są zera. Jeżeli wyznaczona jest jedynka kolejne wartości losowane są zgodnie z poniższym schematem.

schemat (s2.1) Wartość wylosowana jest zgodna z rozkładem normalnym $X \sim \mathrm{N}(\mu, \sigma)$, gdzie μ jest średnią arytmetyczną danego czynnika ze znanych danych, z pominięciem zer a σ jest obliczana zgodnie ze wzorem (w2.1).■

III. Czynniki STDs, STDs (number), STDs:condylomatosis, … , STDs:HPV rozpatrywane są w następujący sposób, pierwszy parametr (STDs) jest wyznaczany zgodnie ze wzorem (w2.2). Kolejny parametr (STDs (number)) wyznaczany jest zgodnie ze schematem (s2.1). Następnie konkretne choroby wyznaczane są iteracyjnie, w zależności od wyznaczonej ich liczby dla konkretnej osoby. Prawdopodobieństwo wylosowania konkretnej choroby jest wprost proporcjonalne do średniej dla całego zestawu danych. W przypadku gdy dana choroba nie występuje, średnia jest zawyżana do poziomu $X * 0{,}01$, gdzie $X \sim \mathrm{U} < 0{,}1 >$. Dodatkowo w przypadku gdy wylosowaną chorobą jest AIDS dodawany jest czynnik STDs:HIV.

IV. Czynniki nieznane, z grupy DX są zamieniane na 0

Algorytm obliczania wektora Y2 dla zadanego wektora wejściowego Y0:

1) $\mathrm{S1} = \mathrm{W1} * \mathrm{Y0}$
2) *Przekształcenie wszystkich wartości wektora S1 funkcją aktywacji*
   $\mathrm{Y1} = \mathrm{aktywuj(S1)}$
   Wybrana została funkcja aktywacji: $\psi(z) = \frac{e^z}{1+e^z}$
3) *Rozszerzenie wektora Y1 o element „1”*
4) $\mathrm{Y2} = \mathrm{W2} * \mathrm{Y1}$

2.2 Uczenie sieci

Uczenie sieci odbywa się w oparciu o metodę wstecznej propagacji gradientu. Przyjęta została standardowa dla takich przypadków funkcja kosztu: $q(y) = \frac{1}{2}\left||y - y^d|\right|^2$.

Potrzebne są wartości pochodnych funkcji kosztu po poszczególnych współczynnikach $\frac{dq(\bar{f}(x,\theta))}{d\theta'_{j,i}}$

Warstwa wyjściowa

Pochodne po wagach warstwy wyjściowej: $\frac{dq(\bar{f}(x,\theta))}{d\theta'_{k,j}} = (y_k^2 - y_k^d)y_j^1$. Co po przekształceniu na potrzebną do obliczeń postać macierzową daje:

$$P2 = (Y2 - YD)Y1^T$$

P2 to macierz pochodnych po współczynnikach warstwy wyjściowej.

Warstwa ukryta

$$\frac{dq}{d\theta'_{j,i}} = \frac{dq}{ds_j}\frac{\delta s_j}{\delta\theta'_{j,i}} = \frac{dq}{ds_j}y_i^0$$

$$\frac{dq}{ds_j} = \frac{dq}{dy_j^1}\frac{\delta y_j^1}{\delta s_j} = \frac{dq}{dy_j^1}\psi'(s_j)$$

$$\frac{dq}{dy_j^1} = \sum_k \frac{\delta q}{\delta y_k^2}\frac{\delta y_k^2}{\delta y_j^1} = \sum_k (y_k^2 - y_k^d)\theta''_{k,j}$$

Co w zwartej postaci macierzowej daje:

$$P1 = \left(W2^T(Y2 - YD)\right)\psi'(s_j)Y0^T$$

Z tym, że do obliczeń należy wziąć macierz W2 z obciętą ostatnią kolumną, która odpowiada współczynnikom wyrazu „1” i nie powinna być brana pod uwagę.

Algorytm uczenia

```
INFINITE LOOP:
    LOOP FOR ALL LEARNING EXAMPLES:
        CALCULATE NETWORK FOR CURRENT EXAMPLE
        CALCULATE DERIVATIVES
        ADD DERIVATIVES TO SUM

    DIVIDE DERIVATIVES SUMS BY NUMBER OF LEARNING EXAMPLE
    W1=W1 - beta * SUM_P1
    W2=W2 - beta * SUM_P2

    EVERY 20000 ITERATIONS:
        CALCULATE QUALITY INDEX
        IF QUALITY INDEX < PREDEFINED BREAK CONDITION:
            BREAK
```

Współczynnik uczenia „beta" ustalany jest ręcznie i jest stały w trakcie uczenia sieci. Na podstawie wielu testów udało się ustalić, że optymalne parametry sieci to 50 neuronów ukrytych, oraz współczynnik uczenia beta = 0.2.

Jakość sieci określana jest średnią funkcją kosztu: $q(y) = \frac{1}{2}\left||y - y^d|\right|^2$ po wszystkich przykładach uczących.

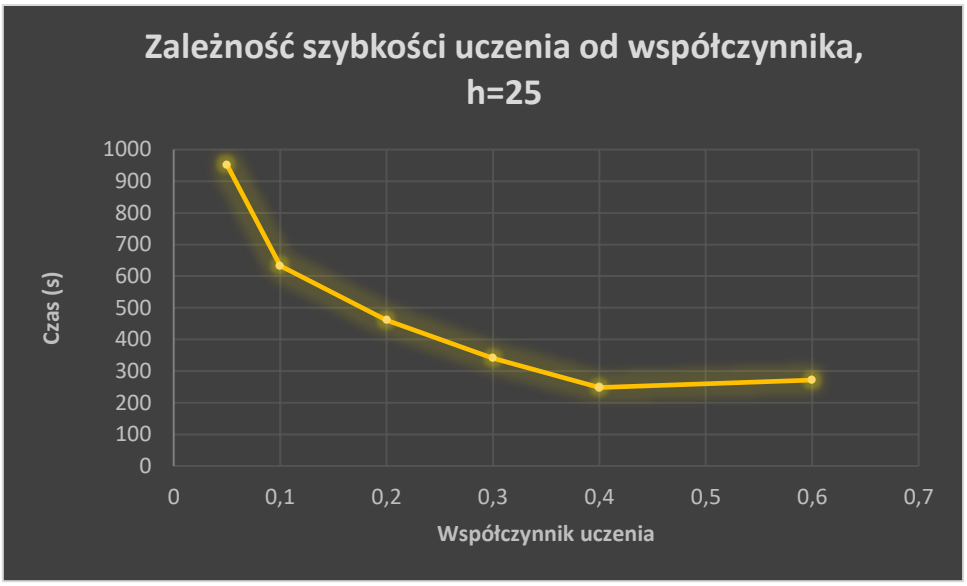
3. Wybór parametrów sieci

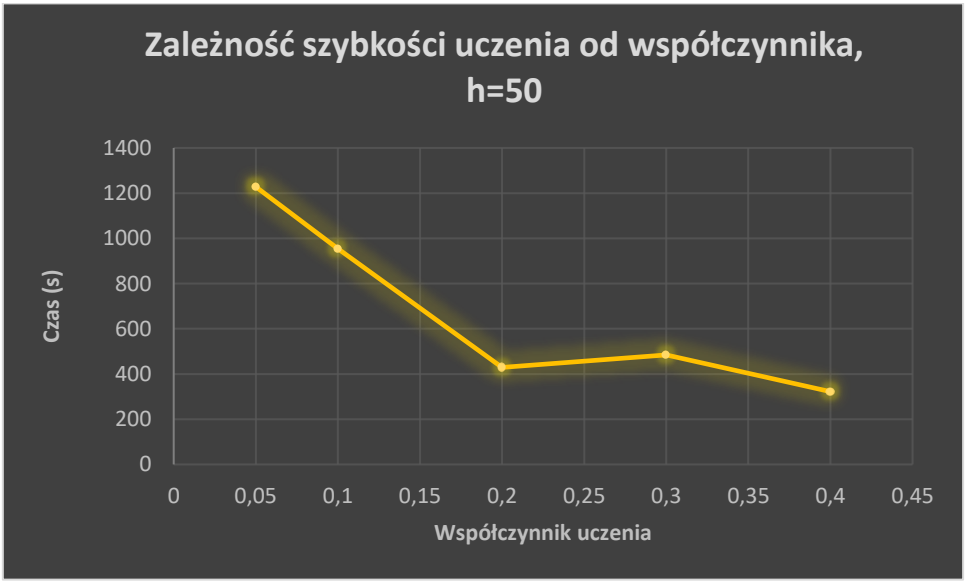
Przed rozpoczęciem procesu uczenia sieci wykonany został szereg testów, mających na celu ustalić wartości parametrów, przy których sieć działa optymalnie. Zbiorem uczącym w czasie testów był zbiór 100 przykładów rozpoczynający się w 3/5 zbioru danych.

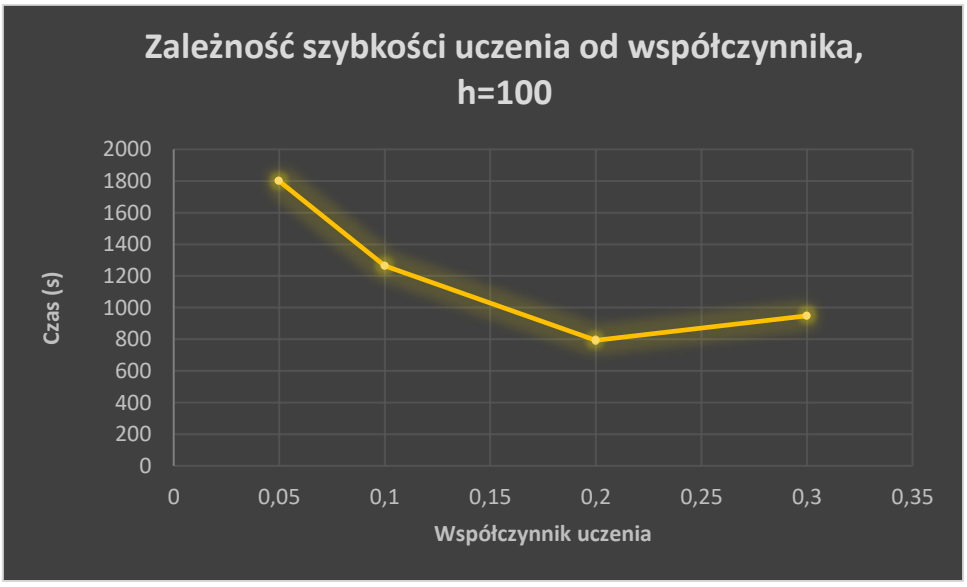
		Współczynnik uczenia					
		0,05	0,1	0,2	0,3	0,4	0,6
Wielkość warstwy ukrytej (h)	10	-> inf	511 500 1397 s	189 000 497 s	-> inf	198 500 503 s	-> inf
	25	321 000 952 s	210 000 633 s	161 500 461 s	106 000 340 s	84 500 248 s	93 000 272 s
	50	365 500 1228 s	276 000 954 s	128 500 429 s	132 500 485 s	96 500 322 s	-> inf
	100	362 000 1799 s	276 500 1263 s	171 500 793 s	206 000 947 s	-> inf	-> inf

W przypadku niskich wartości współczynnika uczenia (do 0,1) nie są widoczne oscylacje, natomiast sieć uczy się wolno. Wartość 0,2 jest kompromisem pomiędzy oscylacjami, a szybkością uczenia. Przy wartościach powyżej 0,2 sieć uczy się szybko, ale wpada w oscylacje, co często kończy się niepowodzeniem procesu uczenia. Warto rozpatrywać liczby neuronów ukrytych 25 i większe, dla mniejszej ich liczby sieć działa nieprzewidywalnie i wolno.

Do uczenia sieci wybrane zostały parametry: współczynnik uczenia 0,2 oraz h=50. Zapewniają one optymalną szybkość przy ograniczonych oscylacjach.


Zależność szybkości uczenia od współczynnika, h=25
Czas (s)
1000
900
800
700
600
500
400
300
200
100
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
Współczynnik uczenia


Zależność szybkości uczenia od współczynnika, h=50
Czas (s)
1400
1200
1000
800
600
400
200
0
0
0,05
0,1
0,15
0,2
0,25
0,3
0,35
0,4
0,45
Współczynnik uczenia


Zależność szybkości uczenia od współczynnika, h=100
Czas (s)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
0,05
0,1
0,15
0,2
0,25
0,3
0,35
Współczynnik uczenia

4. Walidacja k-fold

Wyniki uzyskane dla współczynnika uczenia beta równego 0.2 oraz dla 50 neuronów ukrytych w warstwie ukrytej poddano walidacji k-fold. Wybrano k = 5. Posiadane dane podzielono na 5 równych podzbiorów, z których cztery stanowiły zbiór uczący, a jeden stanowił zbiór testujący. Każdy z pięciu zbiorów był chociaż raz zbiorem testującym dla pozostałych czterech zbiorów stanowiących zbiór testowy. Średni czas nauki przez jeden zbiór uczący w naszym przypadku wyniósł ok. 25 godzin. Po przeprowadzeniu walidacji k-fold otrzymano następujące rezultaty:

n = 0
Dobrze obliczone wyniki: 168
Źle obliczone wyniki: 4
Średni błąd obliczenia: 0.357640848819
Wartość funkcji kosztu: 0.0278923521

n = 1
Dobrze obliczone wyniki: 107
Źle obliczone wyniki: 65
Średni błąd obliczenia: 1.62645283891
Wartość funkcji kosztu: 0.0354388382

n = 2
Dobrze obliczone wyniki: 165
Źle obliczone wyniki: 7
Średni błąd obliczenia: 0.505780538859
Wartość funkcji kosztu: 0.0251324243

n = 3
Dobrze obliczone wyniki: 159
Źle obliczone wyniki: 13
Średni błąd obliczenia: 0.5
Wartość funkcji kosztu: 0.0270724182

n = 4
Dobrze obliczone wyniki: 167
Źle obliczone wyniki: 5
Średni błąd obliczenia: 0.304997140665
Wartość funkcji kosztu: 0.03104285571

Opis oznaczeń:
n – oznaczenie zbioru, od którego zaczynały się dane uczące – były one w przedziale (n – (n + 3)) mod 5, zbiór testowy był zbiorem w kolejności (n + 4) mod 5.

Dobrze obliczone wyniki – ilość wyników, dla których sieć dała dobry wynik.

Źle obliczone wyniki – ilość wyników, dla których sieć dała zły wynik.

Średni błąd obliczenia – uśredniony błąd obliczeń sieci wyliczony, ze wzoru:

$$\sqrt{\frac{\sum(obliczonyWynik - rzeczywisty\ wynik)^2}{N}}$$

gdzie N oznacza ilość przykładów testujących.

Wartość funkcji kosztu – ostatnia zapisana wartość funkcji kosztu obliczanej wyżej podanym wzorem po wyuczeniu sieci danym zbiorem uczącym.

Z otrzymanych wyników widać, że sieć generalnie dobrze aproksymuje ilość badań na raka szyjki macicy, które wykażą wynik pozytywny, oprócz zbioru n = 1, dla którego ok. 37% aproksymacji było błędnych. Dla reszty zbiorów ilość błędnych aproksymacji nie przekraczała 10% liczności zbioru testującego. Duży błąd w zbiorze n = 1 mógł być spowodowany niewystarczającą różnorodnością danych uczących bądź zbyt dużymi różnicami między danymi uczącymi a testującymi, przez co sieć nie mogła nauczyć się dobrej aproksymacji dla każdego przypadku.