**SPRAWOZDANIE**

**Zadanie 1**

**Wykorzystanie sztucznej sieci neuronowej**

**do wykrywania choroby niedokrwiennej serca (choroby wieńcowej, CHD)**

**Przedmiot**

**S**ztuczna Inteligencja

**Data wykonania**

01 grudnia 2019

**Autorzy**

Daniel Witkowski, nr. indeksu: 53831

Dominika Skrzyńska, nr indeksu: 40061

**1. Cel zadania**

Celem zadania jest wytrenowanie sztucznej sieci neuronowej tak, aby na podstawie określonych danych wykrywała wystąpienie choroby niedokrwiennej serca u pacjenta. Zadanie zostało wykonane w oparciu o retrospektywne dane grupy mężczyzn pochodzących z Prowincji Przylądkowej Zachodniej (Western Cape) w RPA, regionu charakteryzującego się wysokim ryzykiem chorób serca.[[1]](#footnote-1)

**2. Dane wejściowe**

Źródło danych: https://web.stanford.edu/~hastie/ElemStatLearn//datasets/SAheart.data?fbclid=IwAR29teRHELpnT4B4QNRGbgl72\_s0ChV\_ODqZ\_tR9NBs-\_ncx0TOrzqH50iI

Sieć była trenowana w oparciu o 10 czynników:

**sbp** (systolic blood pressure) – ciśnienie tętnicze, skórczowe

**tobacco** (cumulative tobacco (kg)) – tytoń skumulowany

**ldl** (low densiity lipoprotein cholesterol) - Cholesterol lipoproteinowy o niskiej gęstości

**adiposity** - otyłość

**famhist (**family history of heart disease (Present, Absent)) – występowanie chorób serca w rodzinnie (1 - obecna, 0 - nieobecna)

**typea (**type-A behavior) – osobowość typu A[[2]](#footnote-2)

**obesity** - otyłość

**alcohol** (current alcohol consumption) – poziom spożycia alkoholu

**age (**age at onset) – wiek na początku

**chd** (response, coronary heart disease) – choroba wieńcowa

Użyte dane zawierały przypadki osób ze stwierdzoną chorobą wieńcową, u których prowadzono terapię zmniejszającą wystąpienie chorób serca. W niektórych przypadkach pomiar był wykonany po leczeniu.

Metoda oceniania: hold-out, lambda=2/3 – gdzie 2/3 danych zbioru to dane trenowane, a pozostała część, 1/3 to dane testowe.

**3. Kod źródłowy**

Skrypt pobiera a następnie wczytuje bibliotekę AMORE.

**install.packages("AMORE")**

**library(AMORE)**

Następuje wczytanie do zmiennej ***Heart*** danych z pliku .csv wraz z nagłówkami. Plik znajduję się w R\_Lab1 w domyślnej lokalizacji ustawionej w RStudio.

**Heart = read.csv("~/R/R\_Lab1/Saheart.csv")**

**Ilosc.Danych=nrow(Heart)**

Metoda ***set.seed()*** ustala tzw. ziarno dla generatora liczb pseudolosowych, w celu uzyskania powtarzalności. Następuje podział danych wejściowych na dane trenujące ***idxTren*** – podłoga z 2/3 losowych indeksów i dane testowe ***idxTest*** – pozostałe 1/3 indeksów.

**set.seed(8) #Linux**

**set.seed(666) #Windows**

**idxTren<-sample(1:Ilosc.Danych,2\*Ilosc.Danych/3)**

**idxTest<-setdiff(1:Ilosc.Danych,idxTren)**

Funkcja ***target()*** przekształca dane do postaci liczbowej. Ilość danych stanowi wartości etykiet.

**target<-function(x)**

**{**

**n<-length(x)**

**wartosci<-levels(x)**

**l<-length(wartosci)**

**T<-matrix(0,nrow=n,ncol=l)**

**for(i in 1:l)**

**T[,i]<-(x==wartosci[i])**

**colnames(T)<-wartosci**

**return(T)**

**}**

Dane zostają zapisane do zmiennej ***wZadane***.

**wZadane<-target(Heart$chd)**

**wZadane**

**set.seed(3) #Linux**

**set.seed(777) #Windows**

Zostaje utworzona struktura sieci. W warstwie wejściowej podano 9 neuronów oraz 2 neurony w warstwie wyjściowej.

**siec<-newff(n.neurons=c(9,4,2),**

**learning.rate.global=0.02,**

**momentum.global=0.8,**

**hidden.layer="sigmoid",**

**output.layer="purelin",**

**method="ADAPTgdwm",**

**error.criterium="LMS")**

Następuje trenowanie sieci za pomocą danych wejściowych (Heart[idxTren, -10 – losowo pobrane wartości nie uwzględniające ostatniej 10 kolumny), oraz wartość wynikowa dla każdej próbki (wZadane[idxTren,]). Zadana ilość kroków wynosi 10000 (wyświetlona 1000).

**wynik<-train(siec,**

**Heart[idxTren,-10],**

**wZadane[idxTren,],**

**error.criterium="LMS",**

**report=TRUE,**

**show.step=10,**

**n.shows=1000)**

Skrypt generuje wykres liniowy przy pomocy funkcji ***plot().***

**plot(wynik$Merror,**

**type="o",**

**xlab="Iteracja (x100)",**

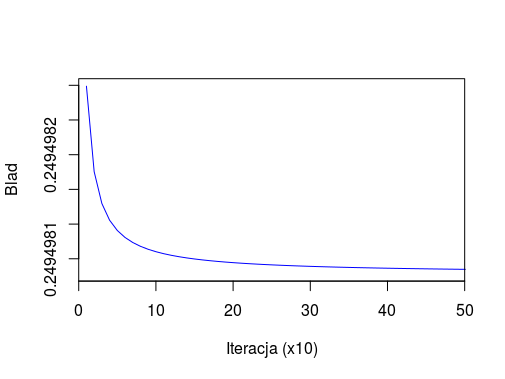
**ylab="Blad",**

**col="blue",**

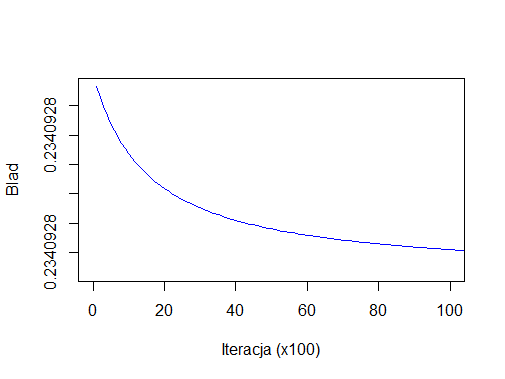
**xlim = c(0,100)**

**)**

Wykres krzywej błędu (Linux):



Wykres krzywej błędu (Windows):



Skrypt twozy zmienną ***y,***  która zawiera wyniki klasyfikacji w postaci wartości liczbowych jednego wyjścia dla każdego przypadku testowego. Wytrenowana sieć jest następnie wykorzystana do danych testowych:

**y<-sim(wynik$net,Heart[idxTest, -10])**

**y**

Następuje definicja funkcji oceny klasyfikacji. Argumenty funkcji: ***zad*** – wartości wzorcowe (zadane), ***wy*** – otrzymane wyjście sieci. Wynik funkcji zapisany w zmiennej ***wynik***.

**test.klasyf<-function(zad,wy)**

**{**

**zadane<-max.col(zad)**

**rozpoznane<-max.col(wy)**

**print(table(zadane,rozpoznane))**

**}**

**wynik<-test.klasyf(wZadane[idxTest,],y)**

Skrypt wylicza dokładność klasyfikacji.

**cat("Dokladnosc klasyfikacji:",**

**sum(diag(wynik))/sum(wynik)\*100, "%\n")**

**4. Podsumowanie i wnioski**

Dokładność klasyfikacji 67.53247 % na systemie Linux

Dokładność klasyfikacji 69.48052 % na systemie Windows

Zauważyliśmy różnice w dokładności wyników i krzywej błędu pomiędzy systemami operacyjnymi Linux oraz MS Widnows 10 dlatego użyliśmy różnych seedów dla obu systemów. Nie zauważyliśmy żadnych istotnych różnic podczas zmiany ilości neuronów w warstwie ukrytej. Największe różnice dokładności uzyskaliśmy dzięki dobraniu odpowiednich seedów.

**5. Bibliografia:**

1. Dokumentacja biblioteki AMORE: <https://www.rdocumentation.org/packages/AMORE/versions/0.2-15>
2. Dokumentacja języka R: <https://www.rdocumentation.org/>

1. „A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.” Źródło: https://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.info.txt?fbclid=IwAR3stqS9EnxjoQbwPvB1rCBOFMZNVrp3BKGcUYQ9-ksqayR-vOXOfJrBXbM [↑](#footnote-ref-1)
2. „Osobowość typu A, WZA (wzór zachowania A), zachowanie typu A to typ osobowości, charakteryzującej się wysokim poziomem stresu, wywołanym presją czasu, tendencją do zachowań rywalizacyjnych, wysokim poziomem ambicji, agresywnością i wrogością wobec innych. Osoba z WZA postrzega otoczenie jako zagrażające i żyje w nieustannej reakcji alarmowej (fazie stresu). W kategoriach psychologicznych konstruktów teoretycznych WZA charakteryzuje neurotyczna ekstrawersja (EPQ-R) i niski poziom ugodowości (NEO-FFI).”

   żródło: https://pl.wikipedia.org/wiki/Osobowo%C5%9B%C4%87\_typu\_A [↑](#footnote-ref-2)