FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

KATEDRA
KYBERNETIKY

Základy klasifikace neuronovými sítěmi

Západočeská Univerzita V Plzni
Katedra Kybernetiky
Projekt 4

Vladimíra Kimlová
2. ročník
20. března 2022

Obsah

Seznam obrázků

1 Zadání

Cílem projektu je naučit se pracovat s vybranými frameworky, které využívají principy neuronových sítí k řešení klasifikačních problémů. Získané znalosti umožní komplexní řešení obecné klasifikační úlohy od předzpracování dat po vyhodnocení výsledků a budou základem pro navazující bakalářskou práci.

Postup práce / podúlohy

- seznámení se s teorií a principy neuronových sítí
- poskládání dat pro klasifikační úlohu
- sestavení modelu neuronové sítě
- trénování sítě
- vyhodnocení výsledků klasifikace
- sepsání dokumentace (2-4 strany)

Další dovednosti získané v rámci práce na projektu

- verzovací systém Git (sdílení kódu přes repozitář na GitHubu)
- práce na vzdáleném stroji přes SSH připojení
- práce v programovacím jazyce Python ve virtuálním prostředí nástroje Anaconda
- psaní dokumentace v LaTeXu

Technologie

Python, Keras, Scikit-Learn, PyTorch, Git, SSH, Anaconda, LaTeX

Repo projektu

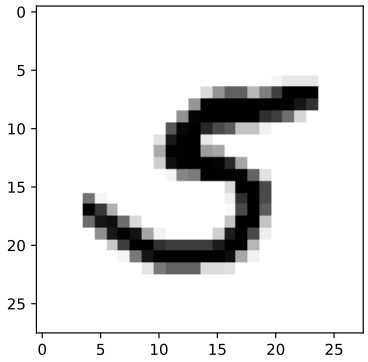
`https://github.com/kitt10/jivl`

2 Vypracování

Projekt byl vypracován v Pythonu – konkrétně v JupyterLabu. V několika prvních úlohách byl využíván vzorový kód Simple MNIST convnet a bylo pracováno i se samotným MNIST datasetem, tj. ručně psanými číslicemi. Ve druhé části projektu byl použit dataset Heart Disease UCI.

2.1 MNIST

V rámci tohoto úkolu byly analyzovány trénovací vzorky, tj. matice 28×28 s hodnotami 0 – 255 představující šedotónový obrázek ručně psaného čísla, a prvních deset "targets"z testovací části datasetu, tedy čísla od 0 do 9.



Obrázek 1: Trénovací vzorek – číslo 5

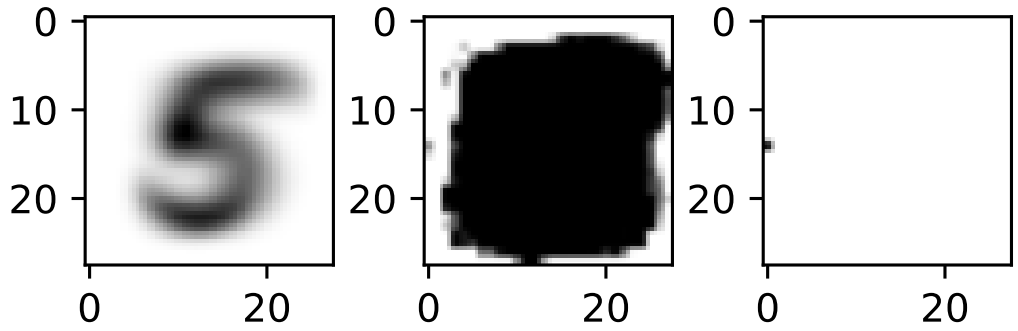
Dále bylo trénování neuronové sítě spuštěno pouze na prvních deseti tisících vzorcích. Tímto došlo k nepatrnému poklesu úspěšnosti klasifikace z původních 99 % (pro výchozích 60 000 trénovacích vzorků) na 98 %. Tento minimální úbytek úspěšnosti je způsoben tím, že třídy jsou v trénovací množině neseřazené, což má za následek téměř totožné zastoupení všech tříd v prvních deseti tisících vzorcích.

Následně byl klasifikační problém zúžen pouze na rozpoznávání čísel 8 a 6. Došlo tím k redukování počtu trénovacích dat na přibližně 12 000 a testovacích na přibližně 2 000, zastoupení obou tříd bylo rovnoměrné. Přesnost neuronové sítě po předložení takovéhoto klasifikačního problému byla téměř 100 %.

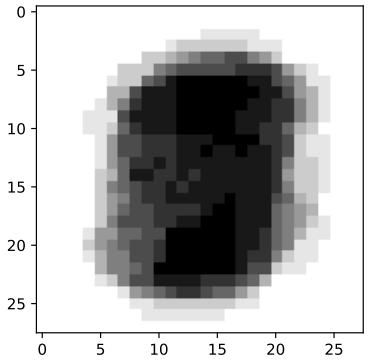
Jednotlivé vzorky byly dále analyzovány pro každou třídu separátně, tedy pro jednotlivé číslice 0 až 9. Byly zjištěny průměrné hodnoty, maximální hodnoty pro každý pixel přes všechny vzorky pro jednotlivé třídy a unikátní pixely, tj. pixely, které jsou nenulové pro vzorky jen jediné třídy (viz obrázek 2).

Nakonec byl pro každý pixel v obrázku 28×28 znázorněn počet aktivních tříd, aktivnost třídy byla dána deseti, stem, tisícem a pěti tisíci nenulovými hodnotami pro daný pixel přes všechny třídy (viz obrázek 3).

Na základě takovéhoto rozboru lze pak určit důležitost jednotlivých pixelů, a navíc i jejich klíčovou roli v rámci jediné třídy.



Obrázek 2: Průměrné hodnoty, maximální hodnoty a unikátní pixely pro třídu 5



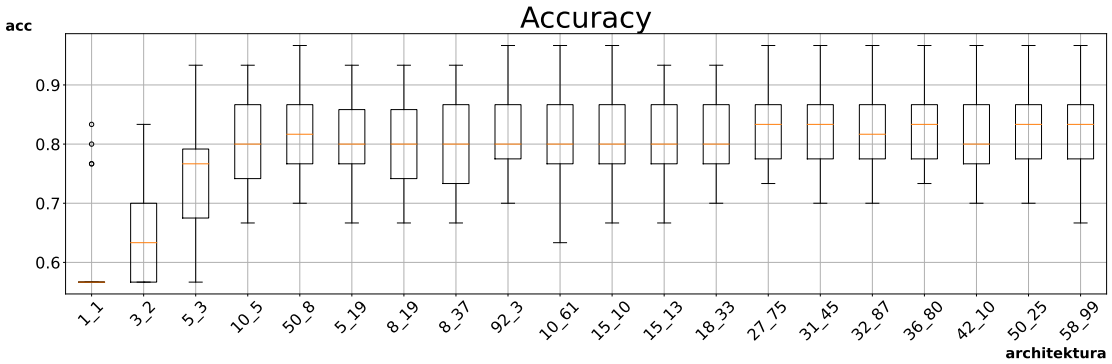
Obrázek 3: Počet aktivních tříd pro 1 000 hodnot

2.2 Heart

Tato část projektu byla zaměřena na binární klasifikaci, konkrétně na určení, zda se u daného pacienta jedná o chorobu srdce či nikoliv. Dataset obsahoval 303 pacientů s třinácti diagnostickými údaji (věk, pohlaví, krevní tlak, ...). Tato data byla nejprve normalizována, tj. převedena na hodnoty od 0 do 1. Dále byla rozdělena na data trénovací, testovací a validační v poměru 8 : 1 : 1. Architektura neuronové sítě byla zvolena jako Multilayer perceptron s aktivační funkcí sigmoid. Parametry neuronové sítě byly postupně měněny za účelem jejího natrénování na co nejlepší výsledky.

Nejprve byly měněny pouze počty neuronů ve skrytých vrstvách od 1 do 1000, tyto změny neměly diametrální vliv na accuracy a loss (kromě úplně nejmenších sítí - viz obrázek 4).

Posléze bylo vybráno 10 kombinací poměrně malých neuronových sítí (počty neuronů ve skrytých vrstvách se pohybovaly do 50), u nichž byly měněny i další parametry jako learning rate, batch size, počet epoch a optimizery. Celkový počet těchto testovaných neuronových



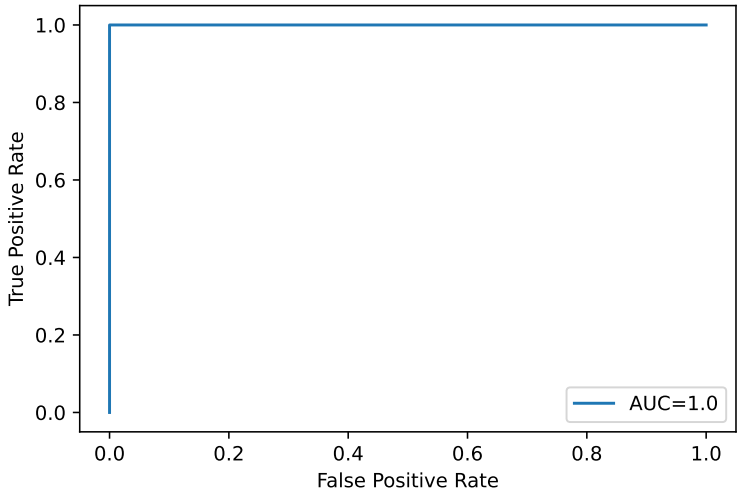
Obrázek 4: Hodnoty accuracy pro různě velké sítě

sítí dosáhl 3 600, pro každou z nich bylo ještě provedeno 5 realizací, aby statistika byla více vypovídající. Nejlepších výsledků dosáhla síť s následujícími parametry:

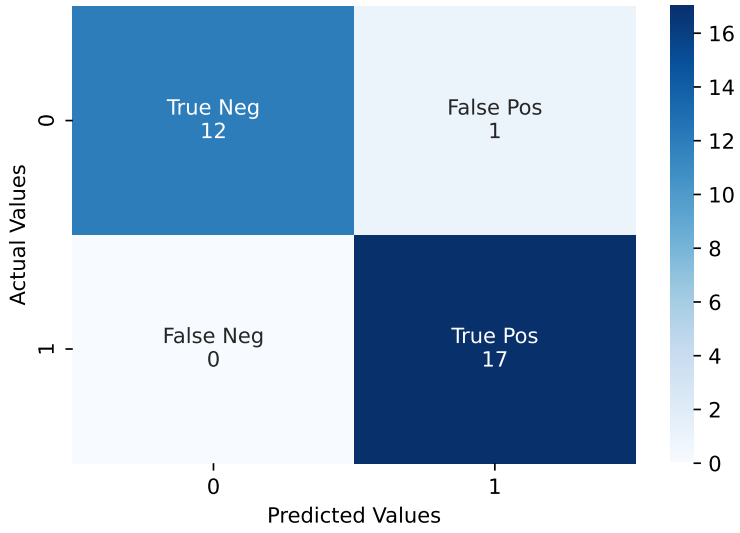
počty neuronů	[5, 3]
learning rate	0.1
batch size	128
epochy	200
optimizer	Adamax

V nejlepším případě (pro vhodný random seed) dosáhla tato síť accuracy 97 % a loss 14 %.

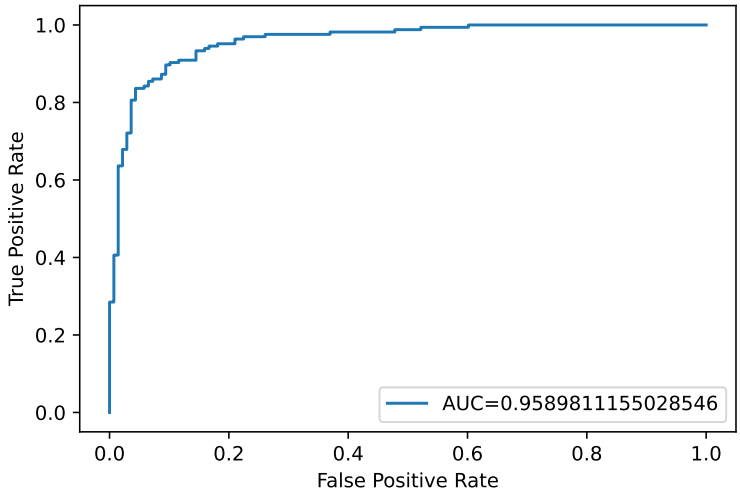
Na závěr byla provedena evaluace této sítě, a to na testovacích datech, a pak i na celém datasetu. Pro testovací data dosáhla síť učebnicových výsledků – ROC křivka měla ideální tvar (viz obrázek 5) a obsah pod touto křivkou byl roven jedné. Byla také napočtena matice záměn (viz obrázek 6) a další vyhodnocovací kritéria jako recall, percision, F1 score, specificity, atd. Výsledky vyhodnocení této sítě pro celkový dataset byly již realističtější, ale pořád příznivé (viz obrázek 7).



Obrázek 5: ROC křivka pro testovací data



Obrázek 6: Matice záměn pro testovací data



Obrázek 7: ROC křivka pro celý dataset

3 Závěr

V rámci projektu byla vyzkoušena práce s různými datasety, dále předzpracování dat, tvorba modelu neuronové sítě a také vyhodnocení jejích výsledků. Nejlepší získaný model dosahoval vysoké přesnosti. Testovacích dat je ale málo a výsledná neuronová síť tak nemá širší uplatnění.