Student: Stanciu Iulia-Cristina
Grupa: 421A

Hill-Valley

Scopul proiectului:

Scopul proiectului este clasificarea cu acuratețe cât mai mare a tendinței unui grafic format din 100 de puncte de a forma „un deal” (de a crește aproape constant pană la un anumit punct și apoi de a scădea în aceeași manieră) sau „o vale” (de a suferi o scădere a valorii pe axa y până la un moment, iar apoi, o creștere).

Baza de date:

Baza de date folosită este: “Hill-Valley Data Set”, disponibilă pe UCI datasets repository.
Link bază de date: UCI Machine Learning Repository: Hill-Valley Data Set

Fișiere componente ale bazei de date:

1. Fișier cu date despre problemă și seturile de antrenare și testare:
    - Hill-Valley.names
2. Fișiere conținând exemple (teoretice și grafice) ale problemei tratate:
    - Hill_Valley_sample_arff.text
    - Hill_Valley_visual_examples.jpg
3. Fișiere cu date de antrenare (atât atribute, cât și etichete):
    - Hill_Valley_without_noise_Training.data
    - Hill_Valley_with_noise_Training.data
4. Fișiere cu date de testare (atât atribute, cât și etichete):
    - Hill_Valley_without_noise_Testing.data
    - Hill_Valley_with_noise_Testing.data

Tipul fișierelor cu date: csv

Diferența între fișierele “with noise” și “without noise”:
Fișierele “with noise” includeau zgomot, adică abateri de la tendința generala de creștere sau scădere.

Tipul de problemă: Clasificare
Rezultatele posibile ale clasificării: Hill or Valley ca valori binare de 0 și 1.

Dimensiunile bazei de date:

- Numărul de instanțe: 606 în fiecare fișier (606*4 = 2424)
- Numărul de atribute: 100 + 1 (eticheta)

Caracteristici ale datelor din baza de date:

- Atributele sunt numere reale și reprezintă valorile coordonatei y ale unor puncte. Valorile coordonatei x sunt reprezentate de locul valorii pe fiecare linie a bazei de date (numărându-se de la 1 la 100).
- Etichetele sunt valori binare.
- Setul de date este împărțit în câte două fișiere date de testare și antrenare. Analiza datelor a fost făcută pe setul "fără zgomot", pe cel "cu zgomot", dar și pe setul complet de date (format din cele două prezentate anterior.
- Baza de date nu are valori lipsă.

Librării folosite în rezolvarea acestui proiect:

1. Pandas
2. Librăria Pandas a fost folosită procesarea bazelor de date și transformarea fișierelor de date .csv în Pandas DataFrame.
3. Numpy
   Librăria Numpy a fost folosită pentru separarea datelor din fișierele de antrenarea si testare în atribute și etichete. Cu ajutorul acestei biblioteci au fost alcătuite matrice de date cu atribute, respectiv etichete pentru seturile de date de antrenare și pentru cele de testare.
4. Scikit-learn
   Librăria Scikit-learn a fost folosită pentru aplicarea algoritmului MLP.
5. Scikit-learn Metrics
   Librăria Scikit-learn Metrics a fost folosită pentru a măsura performanța algoritmului folosind funcția de acuratețe.

Sistemul folosit și variația parametrilor:

Sistemul de clasificare folosit este acela de rețea neurala(MLP – Perceptron Multi-Strat).

Parametrii variați:

1. Numărul de straturi ascunse: 1 sau 2
2. Numărul de neuroni pe straturile ascunse: egal cu stratul anterior sau jumătate
3. Learning rate: 0.1 sau 0.01

Performanța sistemului:

Performanța algoritmului aplicat a fost măsurată cu ajutorul metricii: acuratețe (câte predicții identice cu etichetele inițiale a realizat sistemul).

Performanța maximă a sistemului a fost:

1. Pentru setul de date "fără zgomot": **51.32%**
2. Pentru setul de date "cu zgomot": **53.63%**
3. Pentru întreg setul de date: **50.50%**

Setul de date	Număr de straturi ascunse	Număr de neuroni pe straturile ascunse		Learning rate	Acuratețe
Setul de date “fără zgomot”	1	50		0.01	**0.5132**
		100			0.4868
		50		0.1	0.4868
		100			0.4868
	2	50	25	0.01	**0.5132**
		50	50		**0.5132**
		100	50		0.4868
		100	100		0.5281
		50	25	0.1	0.4868
		50	50		0.4868
		100	50		**0.5132**
		100	100		0.4868
Setul de date “cu zgomot”	1	50		0.01	**0.5363**
		100			0.5314
		50		0.1	0.4934
		100			0.4934
	2	50	25	0.01	0.4934
		50	50		0.5099
		100	50		0.4917
		100	100		0.5066
		50	25	0.1	0.5066
		50	50		0.5066
		100	50		0.4934
		100	100		0.4934
Setul de date complet (“cu zgomot” + “fără zgomot”)	1	50		0.01	**0.5050**
		100			**0.5050**
		50		0.1	**0.5050**
		100			**0.5050**
	2	50	25	0.01	0.4670
		50	50		**0.5050**
		100	50		**0.5050**
		100	100		**0.5050**
		50	25	0.1	**0.5050**
		50	50		**0.5050**
		100	50		**0.5050**
		100	100		**0.5050**