

## **II užduotis (Vieno neurono mokymas sprendžiant klasifikavimo uždavinį)**

**Užduoties tikslas** – apmokyti vieną neuroną spręsti dviejų klasių uždavinį ir atlikti tyrimą.

Šioje užduotyje bus naudojama **krūties vėžio duomenų aibė**, kurią galima parsisiųsti iš <https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original> (iš \*.zip failo rinktis *breast-cancer-wisconsin.data*). Šiuose duomenyse yra dvi klasės: 2 – nepiktybinis navikas, 4 – piktybinis navikas.

Užduotyje reiks mokyti sigmoidinį neuroną. Sigmoidiniame neurone naudojama sigmoidinė aktyvacijos funkcija. Neurono išėjimams (*output, predicted class*) apskaičiuoti taip pat naudojama sigmoidinė funkcija. Neurono išėjimo reikšmės bus intervale (0; 1), tad norint nustatyti spėjamą klasę (*predicted class*), šias reikšmes reiks suapvalinti iki artimiausio sveiko skaičiaus (0 arba 1).

Sigmoidinio neurono mokymui reiks sukurti du algoritmus: (1) naudoti paketinį gradientinį nusileidimą, (2) naudoti stochastinį gradientinį nusileidimą

(žr. <https://emokymai.vu.lt/mod/resource/view.php?id=171951>, 43 ir 45 skaidrės).

### **Užduoties punktai:**

1. Parsisiųsti ir paruošti duomenis. Klasių žymes (*label*) pakeiskite 0 arba 1. Ištrinkite nereikalingus atributus (stulpelius), pvz., ID. Krūties vėžio duomenyse yra trūkstamų reikšmių, tose vietose yra klaustukai (?), todėl tokius įrašus (eilutes) reikia ištrinti. Duomenų eilutes reikia permaišyti atsitiktine tvarka, kad klasių įrašai būtų atsitiktinai išdėstyti, t. y., nebūtų taip, kad pradžioje tik vienos klasės duomenys, paskui – kitos.
2. Sukurti programą (parašyti kodą), kuri įgyvendintų vieno sigmoidinio neurono mokymo, validavimo ir testavimo procesą, sprendžiant klasifikavimo uždavinį. Reikia įgyvendinti du atvejus: (1) kai neuronas mokomas naudojant paketinį gradientinį nusileidimą, (2) kai neuronas mokomas naudojant stochastinį gradientinį nusileidimą. Turi būti įgyvendintas neurono mokymas, validavimas ir testavimas, todėl turimi duomenys turi būti padalinami į tris aibes. Galima taikyti santykį 80:10:10.

Programos rezultatas turi būti:

- Gauti neurono svoriai ir poslinkis (*bias*). Kadangi krūties vėžio duomenyse yra 9 požymiai, tai rezultate turi būti 9 svoriai ir vienas poslinkis.
- Gautos paklaidos po kiekvienos epochos mokymo duomenims (žr. <https://emokymai.vu.lt/mod/resource/view.php?id=171951>, 46 skaidrė).
- Gautos paklaidos po kiekvienos epochos validavimo duomenims.
- Gauta paklaida testavimo duomenims (žr. <https://emokymai.vu.lt/mod/resource/view.php?id=171951>, 48 skaidrė).
- Gautas klasifikavimo tikslumas po kiekvienos epochos mokymo ir validavimo duomenims.
- Gautas klasifikavimo tikslumas testavimo duomenims.

Klasifikavimo tikslumas (*accuracy*) – tai santykis tarp teisingai klasifikuotų ir visų duomenų. Norint jį gauti, reikia kiekvienam duomenų įrašui paskaičiuoti klasę pagal gautus neuronų svorius. Naudojant sigmoidinę funkciją, neurono išėjimo reikšmės yra intervale (0; 1), tad skaičiuojant klasifikavimo tikslumą, šias reikšmes reiktų suapvalinti iki artimiausio sveiko skaičiaus (0 arba 1).

- Apskaičiuojamas neurono mokymosi laikas.

3. Atlikti tyrimus. Kiekvieno tyrimo rezultatus (klasifikavimo tikslumas ir paklaida) pateikti grafikuose arba lentelėse su atitinkamais komentarais. Tie rezultatai, kuriuos lengviau ir greičiau galima suprasti iš grafikų, jie turi būti pateikti grafikuose. Pavyzdžiui, paklaidos ir klasifikavimo tikslumo priklausomybę nuo epochų reiktų vaizduoti linijiniu grafiku. Norint pamatyti skirtumus ar panašumus tarp rezultatų, esant įvairioms tam tikro hiperparametro reikšmėms, rezultatus vertėtų pateikti stulpelinėse diagramose. Tinkamas rezultatų vaizdavimas gali įtakoti užduoties įvertinimą.

Tyrimų metu būtina nustatyti ir aprašyti:

- Kaip paklaidos reikšmės priklauso nuo epochų skaičiaus mokymo ir validavimo duomenims? Pateikti grafiką, kurio x ašyje yra atidėdamos epochos, o y ašyje paklaidų reikšmės mokymo ar validavimo duomenims.
- Kaip klasifikavimo tikslumas priklauso nuo epochų skaičiaus mokymo ir validavimo duomenims? Pateikti grafiką, kurio x ašyje yra atidėdamos epochos, o y ašyje klasifikavimo tikslumas mokymo ar validavimo duomenims.
- Kaip rezultatai priklauso nuo skirtingų mokymosi greičio reikšmių? Tinkamą vaizdavimo būdą (lentelės, grafikai) galima pasirinkti patiems. Mokymosi greičio reikšmė turėtų būti intervale (0, 1), eksperimentų metu rekomenduojama panagrinėti atvejus su mažiausiai trimis skirtingomis mokymosi greičio reikšmėmis.
- Kaip rezultatus (paklaidą ir klasifikavimo tikslumą) įtakoja taikomas gradientinis nusileidimas (paketinis ir stochastinis).
- Kaip taikomas gradientinis nusileidimas (paketinis ar stochastinis) įtakoja mokymo laiką esant vienodam epochų skaičiui.

#### **Užduoties ataskaitoje:**

- Aprašyti, kokie duomenys buvo naudojami, kiek yra duomenų įrašų (eilučių), požymių (stulpelių).
- Aprašyti, kaip duomenys buvo padalinti į mokymo, validavimo ir testavimo aibes. Paaiškinti, kokia kiekvienos aibės paskirtis.
- Paaiškinti, kaip buvo parinktos pradinių svorių reikšmės.
- Pateikti programos tekstą su išsamiais komentarais.
- Paaiškinti, kas yra paketinis gradientinis nusileidimas ir stochastinis gradientinis nusileidimas. Paaiškinti, kas kiekvienu atveju yra epocha.
- Detaliai aprašyti atlikto tyrimo rezultatus (žr. 3 punktą). Aprašykite, kaip rezultatai skiriasi naudojant vieną ar kitą gradientinį nusileidimą.
- Apmokius neuroną ir nustačius atvejį, kada gaunamas didžiausias klasifikavimo tikslumas ir mažiausia paklaida validavimo duomenims, pateikti:
  - gautus svorius,
  - epochų skaičių,
  - paklaidas paskutinėje epochoje mokymo ir validavimo duomenims,
  - klasifikavimo tikslumą paskutinėje epochoje mokymo ir validavimo duomenimis,
  - paklaidą testavimo duomenims,
  - klasifikavimo tikslumą testavimo duomenimis,
  - kiekvienam testavimo duomenų įrašui nurodyti, kokias klases nustatė neuronas ir kokia turėjo būti.

**Pastaba:** gali būti, kad nepavyks rasti varianto, kad vienu metu klasifikavimo tikslumas būtų didžiausias, o paklaida – mažiausia. Tuomet savo nuožiūra reikia išrinkti variantą, kad klasifikavimo tikslumas būtų kiek galima didžiausias, o paklaida – kiek galima mažiausia.

- Pateikti atliktų tyrimų išvadas. Išvados turi būti aiškos ir lakoniškos, tiksliai atspindinčios gautus rezultatus. Išvadų apimtis – 4–10 sakinių.

**P.S.** Ataskaitoje turi būti aprašytas kiekvienas atliekamas veiksmas, pateikti žymėjimų aprašymai ir kita, jūsų manymu, svarbi informacija. Privalote nurodyti, kurios kodo ar ataskaitos dalys yra sugeneruotos su *chatGPT*, *DeepSeek* ar kitais analogiškais įrankiais.