II užduotis (Vieno neurono mokymas sprendžiant klasifikavimo uždavinį)

Užduoties tikslas – apmokyti vieną neuroną spręsti dviejų klasių uždavinį ir atlikti tyrimą su dviem duomenų aibėm.

Šiai užduočiai atlikti reikės naudoti dvi duomenų aibes (Irisų ir Krūties vėžio):

- **Irisų duomenų aibę** galima parsiųsti iš https://archive.ics.uci.edu/dataset/53/iris. Šiuose duomenyse yra trys klasės Setosa, Versicolor ir Virginica. Analizei reikia imti tik dvi: Versicolor ir Virginica.
- **Krūties vėžio duomenų aibę** galima parsiųsti iš https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original. Šiuose duomenyse yra dvi klasės: 2 – nepiktybinis navikas , 4 – piktybinis navikas.

SVARBU: Pagal studento numerį reikia **pasirinkti vieną iš variantų** (studento numerio paskutinio skaitmens dalybos iš trijų **liekana**).

Variantai:
0) Neurono mokymui naudoti **paketinį** gradientinį nusileidimą ir **sigmoidinį** neuroną.
1) Neurono mokymui naudoti **stochastinį** gradientinį nusileidimą ir **sigmoidinį** neuroną.
2) Neurono mokymui naudoti **stochastinį** gradientinį nusileidimą ir **ADALINE** mokymo taisyklę.

Pastaba Nr. 1. Sigmoidinio neurono atveju 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).

Pastaba Nr. 2. Taikant ADALINE mokymo taisyklę, naudojama tiesinė aktyvacijos funkcija. Neurono išėjimams (*output*, *predicted class*) apskaičiuoti naudojama slenkstinė funkcija.

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. Kadangi irisų duomenų yra nedaug (po 50 kiekvienos klasės įrašų), reikia padidinti kiekvienos klasės įrašų (eilučių) kiekį taip, kad duomenų pasiskirstymas klasėse išliktų toks pat. Po padidinimo kiekvienos klasės įrašų skaičius turi būti nemažesnis nei 200. Aprašykite, kokiu būdu tai atlikote. Atvaizduokite duomenų du požymius (stulpelius) XY koordinačių sistemoje, įsitikinkite, kad duomenų pasiskirstymas klasėse išliko daug nepakitęs.

3. Sukurti programą (parašyti kodą), kuri įgyvendintų vieno neurono mokymo ir testavimo procesą, sprendžiant klasifikavimo uždavinį. Galima naudoti bet kurią programavimo kalbą. Turi būti įgyvendintas neurono mokymas ir testavimas, todėl turimi duomenys turi būti padalinami į mokymo ir testavimo aibes. Įprastai santykis tarp mokymo ir testavimo aibių yra 80:20 arba 70:30.

   Programoje turi būti įgyvendinta:
   - Duomenų nuskaitymas iš failo.
   - Galimybė keisti tokius hiperparametrus, kaip mokymo greitis (*learning rate*), epochų skaičius.

   Programos rezultatas turi būti:
   - Gauti svoriai (analizuojant irisų duomenis, būtų penki svoriai w_0, w_1, w_2, w_3, w_4, čia w_0 – poslinkis (*bias*); analizuojant krūties vėžio duomenis, būtų 10 svorių).
   - Gautos paklaidos po kiekvienos epochos mokymo duomenims (žr. https://emokymai.vu.lt/mod/resource/view.php?id=12252, 46 skaidrė).

- Gauta paklaida testavimo duomenims (žr. https://emokymai.vu.lt/mod/resource/view.php?id=12252, 48 skaidrė).
- Gautas klasifikavimo tikslumas po kiekvienos epochos mokymo 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).

4. Atlikti tyrimus ir vienai, ir kitai duomenų aibei (Irisų ir Krūties vėžio). Kiekvieno tyrimo rezultatus (klasifikavimo tikslumas ir paklaida) pateikti lentelėse arba grafikuose su atitinkamais komentarais. Tie rezultatai, kuriuos lengviau ir greičiau galima suprasti iš grafikų, jie turi būti pateikti grafikuose, pvz., rezultatų (paklaidos ir klasifikavimo tikslumo) priklausomybė nuo tiriamų hiperparametrų.

   Tyrimų metu būtina nustatyti ir aprašyti:
   - Kaip paklaidos reikšmės priklauso nuo epochų skaičiaus? Pateikti grafiką, kurio x ašyje yra atidėdamos epochos, o y ašyje paklaidų reikšmės mokymo duomenims.
   - Kaip klasifikavimo tikslumas priklauso nuo epochų skaičiaus? Pateikti grafiką, kurio x ašyje yra atidėdamos epochos, o y ašyje klasifikavimo tikslumas mokymo duomenims.
   - Kaip rezultatai priklauso nuo skirtingų mokymosi greičio reikšmių? 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.

Užduoties ataskaitoje:
- Aprašyti, kokie duomenys buvo naudojami, kiek yra duomenų įrašų (eilučių), požymių (stulpelių), kaip buvo padidintas Irisų duomenų įrašų skaičius, kaip duomenys buvo padalinti į mokymo ir testavimo aibes. Pateikti kitą, jūsų manymu, svarbią informaciją.
- Nurodyti jūsų studento numerį. Aprašyti, kurį variantą jūs turėjote atlikti pagal jūsų numerį.
- Paaiškinti, kaip buvo parinktos pradinių svorių reikšmės.
- Pateikti programos tekstą su išsamiais komentarais.
- Paaiškinti kas yra paketinis gradientinis nusileidimas (jei reikėjo atlikti 0-inį variantą) arba kas yra stochastinis gradientinis nusileidimas (jei reikėjo atlikti 1-ą arba 2-ą variantą). Paaiškinti, kas jūsų variante yra epocha.
- Paaiškinti, kas yra sigmoidinis neuronas (jei reikėjo atlikti 0-į arba 1-ą variantą) arba kas yra ADALINE taisyklė (jei reikėjo atlikti 2-ą variantą).
- Detaliai aprašyti atlikto tyrimo rezultatus (žr. 3 punktą). Aprašykite, kaip rezultatai skiriasi skirtingoms duomenų aibėms.
- Apmokius neuroną ir nustačius atvejį, kada gaunamas didžiausias klasifikavimo tikslumas ir mažiausia paklaida mokymo duomenims, pateikti:
  - gautus svorius,
  - epochų skaičių,
  - paklaidą paskutinėje epochoje mokymo duomenims,
  - klasifikavimo tikslumą paskutinėje epochoje mokymo 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 aiškios 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.