eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

Lehenengo praktika

Kudeaketaren eta Informazio Sistemen Informatikaren Ingeniaritzako Gradua

Erabakiak Hartzeko Euskarri Sistemak

Datuen deskribapen operatiboa

Urko Bidaurre

Irakaslea
Alicia Perez Ramirez

2026.eko urtarrilaren 21

Gaien aurkibidea

1 Materiala

- Weka aplikazioa
- Baliabide bibliografikoak:
  - Informazio orokorra adibideekin: [Witten et al., 2011, Chap. 2]
  - Kontsulta praktikoak: `https://waikato.github.io/weka-wiki/`
- eGelatik eskuragarri:
  - Baliabide orokorrak: aplikazioaren eskuliburua
  - Praktikarako datu-sorta: heart-c.arff

2 Helburuak

Praktika honen helburuak datu meatzaritzarako ikuspegi orokorra ematea da Weka aplikazioaren bitartez. Honetarako datu meatzaritzan informazioa erauzteko hiru teknika nagusiak aipatuko dira: **iragarpena**, **clustering** eta **asoziazioa**. Wekarako sarrera gisa ARFF fitxategien kudeaketan sakonduko dugu iragarpen ataza baten bitartez.

Hurrengo konpetentziak landu:

- **Zeharkako konpetentziak**:
  - Lan autonomoa
  - Pentsamendu kritikoa
- **Konpetentzia orokorrak:**
  - Ikasketa automatikoaren funtsa deskribatzeko gai izatea
  - Datuen deskribapen operatiboa emateko gai izatea
  - Wekarako sarrera: atal ezberdinak bereizteko gai izatea

3 Gidoia

3.1 Aldez aurretiko lana

Praktika hau egiten hasi aurretik honako lanal eskatzen dira:

1. Gai hauei buruzko informazioa irakurri
   - Machine learning: datuetatik ezaguerara . Ikasketa automatikoaren funtsa, datuetatik erabiliz ezaguera edo informazioa erauztea da. Datuek, lortu nahi den ezagueraren adierazgarri izan behar dute. Lagin-espazioko adibide esanguratsuak. Irakurri: [Witten et al., 2011, Chap. 1]
   - Weka-ko datuen formatua: ARFF. Atributuak erabiltzen dira datuen deskribapen operatiboa emateko. Izan ere, atributuen bitartez deskribatutako datuei buruketako instantzia (edo adibide) deritze. Alegia, instantziak karakterizatzeko atributuak erabiltzen dira. Irakurri: [Witten et al., 2011, Chap. 2]
2. Weka deskargatu eta instalatu: `http://www.cs.waikato.ac.nz/ml/weka/`

3.2 Datu meatzaritzako paradigmak

Datu meatzaritzak mota honetako atazak ebazteko balio du:

- Gene batzuen presentziaren arabera, etorkizunean gaixotasun bat izateko probabilitatea eman.
- Biometria: begiko irisaren ezaugarri batzuen arabera, pertsona identifikatu
- Bezero baten erosketen arabera, beste produktu batzuk gomendatui
- Espezie bateko ezaugarrien arabera, bariedadeak bereiztu, alegia, taxonomiak deskubritu
- Aseguru etxeetan antzeko jokaerak dituzten bezeroei antzeko produktuak eskaini
- Iraganean entzundako musikaren arabera, musika gomendatu

Datuetatik informazioa erauzteko hiru paradigma nagusi bereizgten dira: iragarpena (edo sailkapen gainbegiratua), clustering (sailkapen ez-gainbegiratua) eta asoziazioa. Aurreko atazak hauetako batean sartzen dira. Hiru paradigmak deskribatu eta bakoitzerako adibideak eman, horretarako, iturri hau erabilgarria da: [Witten et al., 2011, Sec. 2.1 y Sec. 1.3]

3.3 Datuen deskribapen operatiboa

Praktika honetarako erabiliko dugun fitxategia `heart-c.arff` da.

1. **Zein motatako informazioa (audio, irudiak, ...) dakar .arff fitxategiak? Zein da ARFFren esannahia? Zertarako erabiltzen dira mota honetako fitxategiak?** [Witten et al., 2011, Sec. 2.1, 2.2, 11.1]
   - Informazioa: ARFF artxiboak ez du ez audiorik ez irudirik. Pazienteen ezaugarri klinikoak deskribatzen dituzten datu egituratu alfanumerikoak ditu.
   - ARFFren esanahia: *Attribute-Relation File Format*-en siglak. Wekaren jatorrizko formatua da, atributu-multzo bat partekatzen duten instantzien zerrenda deskribatzeko.
   - Erabilera: Ikasketa automatikoko algoritmoek (Machine Learning) patroiak aurkitzeko prozesatuko dituzten datuak gordetzeko erabiltzen dira.
2. **Editatu `.arff` fitxategia testu editore batekin. Burukoan agertzen den atazako deskribapena aztertu eta ondorengo galderei erantzun:**
   (a) Zertan datza ataza? Iragarpen (*prediction*), taldekatze (*clustering*) ala elkarketa (*association*) buruketa da?
   Diagnostikoa adierazten duen `num` (bihotzeko gaixotasuna) izeneko azken atributua dagoenez, eta besteetan oinarrituta iragarri nahi dugunez, Iragarpen ataza bat dela esan dezakegu, eta zehazkiago, saikapen ataza bat (irteera etiketa bat delako, ez zenbaki jarraitu bat).
   (b) Buruketako deskribapenaren arabera, zenbat balio har ditzake klaseak? Daukagun lagin multzoan, zenbat balio har ditzake klaseak?
   `num` klaseak 5 balio posiblerekin definituta dago: '< 50', '$> 50_1$', '$> 50_2$', '$> 50_3$', '$> 50_4$'.
   (c) `.arff` fitxategian `%` ikurrarekin hasten diren lerroak, fitxategiko parte eragile dira?
   Ez, ez dira fitxategiko parte eragile. Datuak kargatzerakoan *Weka*-k baztertzen dituen komentarioak dira.
3. **Definitu: "Instantzia" eta "Atributu"** [Witten et al., 2011, Sec. 2.2, 2.3]
   - **Instantzia:** Datu-errenkada bakoitza da (`@data`-ren ondoren agertzen diren errenkada bakoitza). Adibide zehatz bat adierazten du; kasu honetan, paziente espezifiko bat, bere datu klinikoekin.
   - **Atributu:** Definitutako zutabe edo propietate bakoitza da (`@attribute` erabiliz). Pazienteak deskribatzen dituzten ezaugarriak irudikatzen dituzte, hala nola adina, kolesterola eta abar.

Erreferentziak

[Witten et al., 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition.

- *https://www.lystloc.com/blog/what-is-a-travelling-salesman-problem-tsp/*
- *SYMPLEX: SIMPLEX SOLVER*