1. PRAKTIKA: Datuen deskribapen operatiboa

Edukia

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 espezifikoak:**
  - 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 aurreti honako lanak 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
- 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 eraucteko 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 datu-fitxategia: `heart-c.arff` (*UCI Machine Learning Repository*[1]).

1. Zein motatako informazioa (audio, irudiak, ...) dakar `.arff` fitxategiak? Zein da ARFF-ren esannahia? Zertarako erabiltzen dira mota honetako fitxategiak? [Witten et al., 2011, Sec. 2.1, 2.2, 11.1]
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?
   (b) Buruketako deskribapenaren arabera, zenbat balio har ditzake klaseak? Daukagun lagin multzoan, zenbat balio har ditzake klaseak?
   (c) `.arff` fitxategian '`%`' ikurrarekin hasten diren lerroak, fitxategiko parte eragile dira?
3. Definitu: "Instantzia" eta "Atributu" [Witten et al., 2011, Sec. 2.2, 2.3]
4. Zer motako atributuekin egiten du lan Wekak?
5. Wekan instantzia guztiek atributu kopuru bera dute?
6. Wekan zein da atributu baterako daturik ez dugula adierazteko ikurra?
7. Aztertzen ari garen atazarako:
   - Zenbat instantzia dago? (N=)
   - Instantziak karakterizatzeko zenbat atributu dago? (n=) Lehenengo 5 atributuetarako eta klaserako, galdera hauei erantzun:
     - Zein motakoa da atributua? (eg. nominala, zenbakizkoa, string, ...)
     - Atributu bakoitzerako aztertu zenbat instantziek ez duten baliorik atributu horretan (*missing values*). Zein portzentaian?
     - Zenbat balio desberdin erregistratu dira atributu bakoitzerako? (*distinct*)
     - Atributu bakoitzerako, badago behin baino erregistratu ez den baliorik? (*unique values*)
     - Histogramen gaineko zenbakiek zer adierazten dute?
     - Numerikoak diren atributuetarako zein da erregistratu den balio minimo, maximoa, batazbestekoa eta desbiderapena?

[1] `UCI-MLR:http://archive.ics.uci.edu/ml/`-n eskuragarri dago *Index of .arff Datasets* atalean: `http://repository.seasr.org/Datasets/UCI/arff/`

8. Atributuak klasearekiko histograma aztertu. [Witten et al., 2011, Sec. 11.2]

   - Intuitiboki, zeintzuk dira informazio gehien eskaintzen duten atributuak sailkapen problemari aurre egite aldera? Alegia, atributu gutxirekin iragarpenak egiteko gai izango ginen?
   - Badago korrelazioa aurkezten duten atributu-bikoteak? Korrelazionatutako atributuak erabiltzea erabilgarria izango da?

9. Atributuak bikoteka aurkeztu: `Visualize` (goian, eskuman):

   - Iragarri nahi den klasearen balioak ondoen diskriminatzen duten atributu bikoteak aukeratu.
   - Informazio gutxien eskaintzen dituzten 3 atributu ezabatu eta datu fitxategia gorde izen honekin: `heart_c_3attManuallyRemoved.arff`. Jarraian, hasierako datuak berreskuratu goiko botoia `Undo` sakatuz.

Erreferentziak

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

A Galdetegia

Erantzun laburrak eta zehatzak eman:

- Zertarako erabili datuak datu meatzaritzan?

- Deskribatu ataza hauetako bakoitza eta adibide bat eman azalpena argitzeko:
  - Iragarpena:
  - Clustering:
  - Asoziazioa:
- Zer erabiltzen da datuetako adibide bat deskribatzeko? Zer motako aldagaiak erabil daitezke datuak deskribatzeko?

- Iragarpen atazean, zer da klase aldagaia? Zer adierazten du aztertutako adibidean?