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.

4. **Zer motako atributuekin egiten du lan Wekak?**

   Wekak nagusiki honako atributuekin egiten du lan:

   - **Numerikoak:** Zenbakizko balioak (errealak edo osoak).
   - **Nominalak:** Etiketaz osatutako aurrez definitutako zerrenda (adbz. `arra`, `emea`).
   - Baita ere existitzen dira *String*, *Date* eta *Relational*, nahiz eta fitxategi honetan numerikoak eta nominalak soilik egon.

5. **Wekan instantzia guztiek atributu kopuru bera dute?**

   Bai. ARFF fitxategi estandar batean (trinkoa), instantzia guztiek balio kopuru bera izan behar dute, eta goiburuan definitutako atributuen ordenari dagokio zehazki. Datu bat falta bada, esplizituki markatu behar da, ez da zutabea ahazten.

6. **Wekan zein da atributu baterako daturik ez dugula adierazteko ikurra?**

   Atributu baterako daturik ez dugula adierazteko galdera ikurra (*?*) erabiltzen da.

7. **Aztertzen ari garen atazarako:**

   - **Zenbat instantzia dago?** (N= 303)
   - **Instantziak karakterizatzeko zenbat atributu dago?** (n= 14) **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?

Atributua	**Mota**	**Missing**	**Distinct**	**Iruzkinak**
age	Numeric	0 (0%)	41	Tartea: 29-77 años
sex	Nominal	0 (0%)	2	{female, male}
cp	Nominal	0 (0%)	4	{typ_angina, asympt, non_anginal...}
trestbps	Numeric	0 (0%)	49	Presio arteriala atseden egoeran
chol	Numeric	0 (0%)	152	Kolesterol serikoa (mg/dl)
num (Klasea)	Nominal	0 (0%)	5	{<50, >50_1, >50_2, >50_3, >50_4}

1 Taula: Lehenengo 5 atributuen deskribapen operatiboa eta `heart-c` dataset klasea.

   - Histogramak: Barren gaineko zenbakiek adierazten dute zenbat instantzia (paziente) erortzen diren maila edo kategoria horretan.
   - Numerikoak diren atributuentzat (*age*, *trestbps*, *chol*, ...):

Atributua	**Minimoa**	**Maximoa**	**Batazbestekoa**	**Desbiderapena estandarra**
age	29	77	54.366	9.082
trestbps	94	200	131.624	17.538
chol	126	564	246.264	51.831

2 Taula: Estatistika deskriptiboak lehenengo 3 atributu numerikoetarako.

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?**

     Intuitiboki, atributu bakoitzaren histogramak aztertuz, koloreak gehien desberdintzen edo bereizten dituzten atributuak izango dira sailkapen problemarekiko adierazgarriagoak. Hau da, atributu konkretu batek barra bat baldin badu, zeinetan ia guztia urdina den, eta beste bat zeinetan ia guztia gorria den, honek informazio asko ematen du, izan ere, atributu honek hartzen duen balioak azken sailkapenean eragin handia duela adierazten du. Ordea, atributu baten barra guztiak nahiko orekatuak baldin badaude (koloreari dagokionez), honek informazio gutxi ematen duela ondoriozta dezakegu, izan ere, atributu honen balioak aldatzeak ez du azken sailkapenean eragin handirik izango.

   - **Badago korrelazioa aurkezten duten atributu-bikoteak? Korrelazionatutako atributuak erabiltzea erabilgarria izango da?**

     Bi atributuk grafiko oso antzekoak badituzte edo begi-bistako erlazio lineal bat badute (adibidez, "adina urteetan"eta "adina hilabeteetan"), korrelazioan egongo lirateke eta bat soberan egongo litzateke.

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

   - Iragarri nahi den klasearen balioak ondoen diskriminatzen duten atributu bikoteak aukeratu.
     [TODO]
   - 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.
     [TODO]

A Galdetegia

Zertarako erabili datuak datu meatzaritzan?

Datuak dira funtsezko lehengaia. Ezkutuko ereduak eta agerikoak ez diren harremanak aurkitzeko erabiltzen dira, informazio gordin hori ezagutza erabilgarri bihurtzeko.

- Helburua datu horiek (adibide historikoak) prozesatzea da, etorkizuneko erabakiak hartzeko balio duten ereduak orokortu eta sortu ahal izateko.
- Datu adierazgarririk gabe (adibide esanguratsuak), ezin da ikaskuntza automatikoa egin.

Deskribatu ataza hauetako bakoitza eta adibide bat eman azalpena argitzeko:

- **Iragarpena** (*Prediction / Supervised Classification*): Atributu espezifiko baten balioa iragartzean datza ("klase"deitua), beste atributuen balioetan oinarrituta. Sistemak erantzun zuzena ezagutzen den iraganeko adibideetatik ikasten du.
- **Clustering** (*Unsupervised Classification*): Datuen multzo bat talde edo kluster desberdinetan banatzean datza, non talde barruko instantzien arteko antzekotasuna handia den eta talde desberdinen artekoa txikia. Ez dago aurrez definitutako klase etiketarik.
- **Asoziazioa** (*Association*): Elementuen arteko erlazio edo eredu bateratuak bilatzen dira. Erregela gisa adierazi ohi dira → "X pasatzen bada, orduan Y pasatu ohi da".

Zer erabiltzen da datuetako adibide bat deskribatzeko? Zer motako aldagaiak erabil daitezke datuak deskribatzeko?

Atributuak (edo ezaugarriak) erabiltzen dira. Instantzia (adibide) bakoitza atributu multzo finko batek karakterizatzen du. Wekak eta datu-meatzaritzak batez ere bi mota erabiltzen dituzte:

1. **Zenbakizkoak (*Numeric*):** Balio jarraituak edo neurgarriak (adib. age, chol).
2. **Nominalak (*Nominal*):** Aurrez definitutako zerrenda bateko kategoriak edo etiketak (adib. sex female, male).

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

Klasea iragarri nahi dugun atributu helburua (*target*-a) da. Ereduak ebatzi behar duen ezezaguna da, gainerako atributuen informazioan oinarrituta.

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*