2. PRAKTIKA: Eredu iragarlea eta bere kalitatearen estimazioa

Edukia

1 Materiala

- Weka aplikazioa
- Baliabide bibliografikoak:
  - Ebaluazio eskemak: [Witten et al., 2011, Sec. 5.0-5.4]
  - Informazio orokorra adibideekin: [Witten et al., 2011, Sec. 11.1, 11.2]
  - Kontsulta praktikoak: `https://waikato.github.io/weka-wiki/`
  - *Evaluation:* `http://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka`
- eGelatik eskuragarri:
  - Baliabide orokorrak: aplikazioaren eskuliburua
  - Praktikarako datu-sorta: `adult.train.arff` eta `adult.test.arff`

2 Helburuak

Aldez aurretiko konpetentziak: datu meatzaritzak burutu ahal dituen ataza ezberdinak bereizteko gai izan (sailkapen gainbegiratua, clustering ala sailkapen ez-gainbegiratua, asoziazioa) ataza bakoitzari buruzko adibideak ezagutu. Gainera, instantziak eta atributuak definitzeko gaitasuna behar da.

Datuetatik abiatuta, eredu iragarlearen inferentzia egitea Weka erabiliz. Eredu iragarlearen kalitatearen estimazioa egitea dauden ebaluazio eskema desberdinen bidez. Sailkatzaileen kalitatea neurtzeko ebaluazio neurriak interpretatzeko trebetasuna hartzea. Hurrengo konpetentziak landu:

- **Zeharkako konpetentziak:**
  - Lan autonomoa
  - Pentsamendu kritikoa
- **Konpetentzia espezifikoak:**
  - Sailkapen gainbegiratua definitzeko gai izatea
  - Eredu iragarlearen kalitatea estimatzeko ebaluazio eskemak bereiztea: train vs. test, hold-out, k-fold cross validation.
  - Ebaluazio neurri ezberdinak interpretatzeko gai izatea: accuracy, precision, recall, f-measure, ...

3 Gidoia

Praktika hau eredu iragarleak datuetatik sortu eta ereduen kalitatea estimatzen zentratzen da. Eredu iragarleari sailkatzaile gainbegiratu deritza. Zergatik deitzen zaio sailkapen "gainbegiratua"? [Witten et al., 2011, Chap 1] Arrazoia hau da: ereduak atributu konkretu bat iragartzeko erabiltzen da (atributu horri "klase" deritza) eta eredu iragarlea edo sailkatzaile gainbegiratua sortzeko erabiltzen diren datuetan klasea ezagutzea ezinbestekoa da. Alegia, ikasketa, gainbegiratutako (klasearen balioa daukaten) datuekin egiten da. Klase atributua zein den adierazi behar da (defektuz, Wekak azkena hartzen du).

3.1 Datuen Azterketa

Praktika honetako datu fitxategi nagusiak: `adult.train.arff` eta `adult.test.arff` dira, (eskuragarri *UCI Machine Learning Repository*[1] delakoan). Datu meatzaritzarekin hasteko, lortutako datu sorta analizatzea komeni da, bermatu atazarako datu adierazgarriak direla eta gogoan izan *missing, unique, different*, korrelazioak etab.

1 Ariketa. Datuen analisia
Arakatu atazarako eman diren fitxategiak eta hurrengo galderei erantzun:

[1] `https://archive.ics.uci.edu/ml/datasets/Adult`

1. *Zertan datza ataza? Zer motako ataza da (iragarpena, clustering, asoziazioa)?*
2. *Esku artean dugun datu sorta erabil daiteke sailkapen gainbegiratua aplikatzeko? Emandako instantziak sailkatuta daude?*
3. *Buruketako deskribapenaren arabera, klaseak zenbat balio har ditzake? Emandako datu sortan, zenbat balio erregistratu dira klaserako? zein da klaseko balioen distribuzioa entrenamendu multzoan? eta test multzoan?*
4. *Test multzoa deskribatzeko zehazki entrenamendu multzoan erabili diren atributuak erabili behar dira, hala da emandako multzoetan?*
5. *Zenbat instantzia daude entrenamendu multzoan? eta test multzoan?*

3.2 Sailkapen gainbegiratua

Sailkapen gainbegiratuan, **sailkatutako** datu multzo batetik abiatuta ezagutza (eredu iragarlea) lortzea ahalbidetzen da eta hori erabiltzea sailkatu gabeko datuak sailkatzeko.

Wekako `Classify` atalean sartu. `Classifier` → `Choose`: bertan sailkatzaile algoritmo familiak agertzen dira. Hurrengo sailkatzaileak bilatu eta bilatu zertan oinarritzen diren iragarpenak egiteko. Informazioa bilatzeko: `More` botoian sakatu, *Wikispaces*en bilatu edo kontsulta-liburuan bilatu:

- `ZeroR`:
- `OneR`:
- `IBk`:

2 Ariketa. *Ereduak gorde*

1. Save *Zein eredu inferitu du algoritmo bakoitzak* `adult.train.arff` *datu sortatik? Eredu bitarra gorde.*
2. ⋆ *Kargatu (*Load*) eredua eta test multzoko iragarpenak egin (*visualize output predictions*).*

3.3 Ebaluazio eskemak

Sailkatzailea ezezik, sailkatzailearen iragarpen gaitasunak ematea ezinbestekoa da. Sailkatzaile baten kalitatearen estimazioa egiteko hurrengo ebaluazio eskemak daude: [Witten et al., 2011, Sec. 5.0-5.4]

- **Train vs dev**: gainbegiratutako bi multzo emanda, eredua multzo handiarekin entrenatu (Train multzoarekin) eta beste multzoaren gainean (development) ebaluatu iragarritako klasea klase errealarekin bat datorren edo ez.
  - **Ebaluazio teknika ez-zintzoa**: eredua ebaluatzeko entrenamendurako erabilitako multzoarekin berarekin. Honek, estimatutako kalitatearen goi bornea emango luke, ez da kalitatearen estimazio errealista.

- **Hold-out**: gainbegiratutako multzo bakar bat izanda, multzo hori ausaz desordenatu (*randomize*) eta bitan banatzen da adb. %66a Train gisa eta %33a Test bezala erabiltzeko Train vs. Test eskema erabiliz. Gomendagarria izaten da eskema hau n aldiz errepikatzea (adb. n=5) eta lortutako emaitza guztien batazbestekoa eta desbiderapen estandarra ematea.
- **K-fold crossvalidation**: ebaluazio gurutzatu anizkoitza (k-koitza).
  - **Leave-one-out**: *K-fold crossvalidation* eskemaren kasu berezia da non k-ren balioa multzoan dagoen instantzia kopurua den. Alegia, instantiza bezainbeste train-ebaluazio esperimentu egingo dira eta esperimentu bakoitzean erabiliko den test multzoak instantzia bakar bat baino ez du izango.

3 Ariketa. *Ebaluazio eskemak*

1. *Osatu k-fCV definizioa*
2. *Zer ezberdintasun dago k-fCV eta k aldiz errepikatutako hold-out artean?*
3. *Aztertu Wekako* `Test options` *aukeren artean nola gauzatu aurreko eskema bakoitza.*
4. *Aukeratu arestian aipatutako sailkatzaileetako bat eta* `adult.test.arff` *erabili ebaluaziorako.*
   - *Zabaldu testu editore batekin* `adult.test.arff` *fitxategia, sailkatuta daude instantziak? zergatik?*

3.4 Ebaluazio neurriak: nahasmen matrizea eta neurri eratorriak

4 Ariketa. *Meritu-figurak*
Definitu, formula matematikoen laguntzaz, hauetako bakoitza bi klasedun problemarako:

- Nahasmen-matrizea: m[i,j] (ala m[j,i] aplikazio batzuetan) iragarleak zenbat aldiz esan duen i klasea eta errealitatean j klasea zen. Zutabe eta errenkaden ordenari dagokionez hitzarmenik ez dagoenez, esplizituki adierazi behar da, izan ere, Wekak halaxe dakar: estimatutako "*classified as*" bezala denotatzen du.
- Nahasmen matrizean oinarrituta, definitu hurrengoak:
  - TPRate=*Recall*=*Sensitivity*: OSATU
  - FPRate: OSATU
  - TNRate=*Specificity*: OSATU
  - FNRate: OSATU
- Accuracy: OSATU

$$Accuracy = \frac{\text{OSATU}}{TP + FP + TN + FN} \tag{1}$$

- Precision: OSATU

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall:

- F-measure: OSATU

5 Ariketa. *Klase bakoitzeko eta klaseka ponderatutako batazbestekoa*

1. *Aztertu 3 klase edo gehiago duen datu sorta bat. Wekak emaitzak klase bakoitzeko ematen ditu, nola interpretatzen dira emaitza horiek?*
2. *Wekak batazbesteko ponderatuak ematen ditu, nola lortzen dira emaitza horiek?*
3. ⋆ Micro-average *eta* Macro-average *definitu meritu figurentzat (precision, recall, f-score)*

3.5 Kalitatea hobetzeko parametro erabakigarriak

`Classifier` atalean, sailkatzailea aukeratzean (`ZeroR` kasuan izan ezik), sailkatzailearen parametro sorta definitzen da.

6 Ariketa. *Parametro karakteristikoak eta beste faktore erabakigarri:*

1. Non topatu ahal da algoritmo bakoitzaren parametro karakteristikoei buruzko informazio gehiago?
2. Zertarako dira parametro horiek sailkatzaile bakoitzean?
3. Sailkatzailearen parametroak aldatuz, aldatzen dira lortutako emaitzak?
4. Aztertu `Classifier output` atalean agertzen den informazioa. Bertan, hasieran sailkatzeilearen parametro batzuk zehazten dira.
   (a) Zer adierazten dute parametro hauek?
   (b) Bilatu parametroak sailkatzeileetan, aldatu eta egiaztatu informazio hau aldatu dela atal horretan.
5. Eredu iragarle baten kalitatea eredu hori lortzeko erabili den algoritmoak determinatzen du neurri handi batean, baina algoritmoak ez ezik, algoritmo horretarako ezarritako parametroak eta ikasteko eskuragarri dagoen datu sorta ere erabakigarriak izaten dira.
   (a) Instantzien %30 kenduta, zenbat deterioratzen dira emaitzak? (aztertu `remove` filtroak)
   (b) Atributu gutxiago erabilita, emaitzak deterioratzen dira orokorrean? kasu guztietan?

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.

4 Galdetegia

Aurreko ariketak eginda, trebatzeko, erantzun laburki hurrengo galderei erantzunak justikatuz:

1. Egia ala gezurra? *Clustering* atazean (alegia, sailkapen ez-gainbegiratuan) egiteko behar diren datuak .arff formatuan klase atributua izango dute, sailkapen gainbegiratuan bezala.
2. Zein ezberdintasun dago *Hold-out* eta *10-fold crossvalidation* artean?
3. Zer adierazten du nahasmen matrizeak? Zer agertzen da elementu bakoitzean? Klaseak bi balio har ditzakeen ataza baterako azaldu.
4. Azaldu, formularen bidez, ebaluazio neurri hau: *recall*.
5. Aipatu ezagutzen duzun korrelazio neurriren bat eta horren kalkulurako erabilitako formula eman.
6. Demagun datu multzoan korrelazionatutako bi atributu ditugula. Intuizioaren arabera hurrengoen artean zeintzuk dira hartuko zenituzkeen neurriak eredu iragarlea lortzeko?
   (a) Atributuetako bat kendu, biek baitakarte informazio bera.
   (b) Atributu biak mantentzea komeni da, zenbat eta informazio gehiago izan, hainbat emaitza hobeak lortuko ditu sailkatzaileak.
   (c) Atributu biak ezabatzea komeni da, ez baitute iragarpen atazarako informaziorik eskaintzen.
7. Eredu iragarlea gainbegiratutako datu guztiekin entrenatzen da: Gorde eredu iragarlea eta berari dagokion kalitatearen estimazioa. Gogoan izan, kalitatea estimatzeko eredu lagungarri bat erabili dela, hori entrenatzeko, ordea, ez dira datu guztiak erabili, zentzuzkoa da metodologia hau?
8. Egia ala gezurra? Ebaluazio ez-zintzoak ez digu informazio erabilgarririk eskaintzen.
9. Zeintzuk dira kalitate oneko eredu iragarlea lortzeko faktore erabakigarriak:
   (a) Erabilitako algoritmoa bera?
   (b) Algoritmoaren parametroak?
   (c) Instantzia kopurua? eta instantzien kalitatea? Erantzuna arrazoitu hurrengo faktoreak aipatuz *missing values, unique values, different values* etab.
   (d) Atributuak? Erantzuna arrazoitu hurrengo faktoreak aipatuz: atributuen eta klasearen arteko korrelazioa aipatu baita atributuen arteko korrelazioa.