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

Bigarren praktika

Kudeaketaren eta Informazio Sistemen Informatikaren Ingeniaritzako Gradua

Erabakiak Hartzeko Euskarri Sistemak

Eredu iragarlea eta bere kalitatearen estimazioa

Urko Bidaurre

Irakaslea
Alicia Perez Ramirez

2026.eko urtarrilaren 27

Gaien aurkibidea

1 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, . . .

2 Gidoia

Praktika hau eredu iragarleak datuetatik sortu eta ereduen kalitatea estimatzen zentratzen da. Eredu iragarleari sailkatzaile gainbegiratu deritza. Zergatik deitzen zaio sailkapen "gainbegiratua"? Arrazoia hau da: eredua 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).

2.1 Datuen Azterketa

Praktika honetako datu fitxategi nagusiak: `adult.train.arff` eta `adult.test.arff` dira. 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. Zertan datza? Zer motako ataza da (iragarpena, clustering, asoziazioa)?

Ataza pertsona batek 50 mila € baino gehiago irabazten dituen ala ez iragartzean, haren datu demografikoetan oinarrituta. Iragarpen ataza bat da, izan ere, atributu espezifiko baten balioa iragartzean datza (klasea), beste atributuen balioetan oinarrituta. Sistemak erantzun zuzena ezagutzen den iraganeko adibideetatik ikasten du.

2. Esku artean dugun datu sorta erabil daiteke sailkapen gainbegiratua aplikatzeko? Emandako instantziak sailkatuta daude?

Bai erabil daitezke, izan ere, datu sorta sailkatuta dago (instantzia bakoitzeko bere atributuak eta dagokion klasea adirazten dira).

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?

Klaseak bi balio har ditzake: $\leq 50K$, $> 50K$. Entrenamendu multzoan, proportzioak %75'9 eta %24'1 dira. Test sortan, ordea, %76'4 eta %23'6.

4. Test multzoa deskribatzeko zehazki entrenamendu multzoan erabili diren atributuak erabili behar dira, hala da emandako multzoetan?

Bai, hala da. Test multzoa deskribatzeko, entrenamendu multzoan erabilitako atributu berberak erabiltzen dira.

5. *Zenbat instantzia daude entrenamendu multzoan? eta test multzoan?*

   Entrenamendu multzoan, 32.561 instantzia daude, eta test multzoan, 16.281.

2.2 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` (Zero Rules):** Oinarrizko algoritmoa da. Ez die atributuei kasurik egiten eta soilik klaseak har dezakeen balio probableena iragartzen du. Kasu honetan, < = 50K iragartzen du kasu guztietarako.
- **`OneR` (One Rule):** Atributu bakar baten balioetan oinarritzen da iragarpena egiteko, hain zuzen ere, errore baxuena duen atributuan.
- **`IBk` (Instance-Based *k*-Nearest Neighbors):** Ez du eredu "ikusgarririk"sortzen (erregelak), baizik eta datuak memorizatu. Instantzia berri baten klasea iragartzeko, k auzoko gertuenak bilatzen ditu eta hortik bozkatu egiten du.

2.3 Ebaluazio eskemak

Sailkatzailea ezezik, sailkatzailearen iragarpen gaitasunak ematea ezinbestekoa da. Sailkatzaile baten kalitatearen estimazioa egiteko hurrengo ebaluazio eskemak daude:

- **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
- **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 multzoarekin berarekin. Honek, estimatutako kalitatearen goi bornea emango luke, ez da kalitatearen estimazio errealista.
- **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 trainebaluazio 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*:

   Teknika honetan, datasetak tamaina bereko k partizio (folds) egiten ditu ausaz. Prozesua k aldiz errepikatzen da: iterazio bakoitzean, partizio bat erabiltzen da testerako, eta beste $k-1$ entrenamendurako. Azken emaitza k esperimentuen batuketa edo batazbestekoa da. Datu guztiak behin testatzeko erabiliko direla bermatzen du.

2. *Zer ezberdintasun dago k-fCV eta k aldiz errepikatutako hold-out artean?*

*k-fCV*n, instantzia bakoitza zehazki behin erabiltzen da test-a egiteko, hau da, partizioak osagarriak dira. Errepikatutako *hold-out*ean, ordea, ausazko partiketak egiten dira hainbat aldi independentez. Ondorioz, gerta liteke instantziaren bat beti train sortan edo test sortan banatzea.

3. Aztertu Wekako `Test options` *aukeren artean nola gauzatu aurreko eskema bakoitza.*

- **Train vs dev:** `Use training set` aukera hautatuz. Kontuz! Ebaluazio ez-zintzoa.
- **Hold-out:** `Percentage split` aukera hautatuz.
- **k-fold crossvalidation:** `Cross-validation` aukera hautatuz eta k-ren balioa zehaztuz.
- **Train vs test:** `Supplied test set` aukera hautatuz eta test fitxategia zehaztuz.

4. Aukeratu arestian aipatutako sailkatzaileetako bat eta `adult.test.arff` *erabili ebaluaziorako.*

- *Zabaldu testu editore batekin* `adult.test.arff` *fitxategia, sailkatuta daude instantziak? zergatik?*

Bai, klasifikatuta daude, izan ere, ebaluazioa egiteko (ereduak asmatu duen jakiteko), jakin behar dugu erantzun zuzena zein den (*Ground Truth*). Wekak erantzun hau ezkutatzen dio ereduari iragartzerako orduan, eta gero iragarpena alderatzen du erantzun zuzenarekin, ereduaren asmatze-tasa lortzeko.

2.4 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:
  - $TPRate = Recall = Sensitivity$: Iragarleak i esanaren eta i izanaren proportzioa.
  - $FPRate$: Iragarleak i esanaren eta j izanaren proportzioa.
  - $TNRate = Specificity$: Iragarleak j esanaren eta j izanaren proportzioa.
  - $FNRate$: Iragarleak j esanaren eta i izanaren proportzioa.
- Accuracy (Asmatze-tasa): Iragarleak i edo j esanaren eta i edo j izanaren proportzioa, hurrenez hurren.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Iragarleak i esanaren eta i izanaren proportzioa.

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

- Recall: Iragarleak i esanaren eta i izanaren proportzioa.
- F-measure: Precision eta Recall-en arteko harmoniarako batezbestekoa. ??????

$$F_{measure} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

5. Ariketa. *Klase bakoitzeko eta klaseka ponderatutako batazbestekoa*

TODO

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)*

2.5 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?