

## KONPUTAZIO-ZIENTZIEN METODO MATEMATIKOAK

### 2011.ko ekaina. 2. deialdia

(Azterketa honetan gehienez 5 puntu lortuko dira. Praktikan ere gehienez 5. Biak gainditu behar dira irakasgaia bere osotasunean gainditzeko.)

1. (0.75 puntu) Informazio teoria.
  - (a) Entropia,  $H(C)$ : Bi balio posible har ditzakeen  $C$  aldagai diskretu bat izanik, bere entropiaren kalkuluak ze informazio ematen digu? Nola kalkulatu da? 2. ariketako datubasean oinarrituz, kalkula ezazu  $C$  klase-aldagaiaren entropia.
  - (b) Elkarrekiko-informazioa,  $I(X, C)$ : Entropiaren kontzeptutik abiatuz, nola definitzen da elkarrekiko informazioa? Eman ezazu definizio formala eta interpretazio intuitiboa.
2. (2 puntu) Hiri bateko eraikuntza berezien datubase bat daukagu. Bertan 12 eraikuntzari buruzko ezaugarriak ditugu: logela kopurua, komun kopurua, mota eta zonaldea. Eraikuntzak bi multzotan sailkatuak izan dira: A klasekoak eta B klasekoak.

| Eraikuntza | Logelak (L) | Komunak (K) | Mota (M)    | Zonaldea (Z) | C |
|------------|-------------|-------------|-------------|--------------|---|
| 1          | 1           | 2           | etxebizitza | erdigunea    | A |
| 2          | 1           | 1           | txaleta     | kanpoaldea   | B |
| 3          | 1           | 1           | etxebizitza | erdigunea    | B |
| 4          | 2           | 2           | etxebizitza | erdigunea    | A |
| 5          | 2           | 2           | txaleta     | kanpoaldea   | B |
| 6          | 2           | 1           | txaleta     | kanpoaldea   | B |
| 7          | 3           | 1           | etxebizitza | erdigunea    | B |
| 8          | 3           | 2           | txaleta     | erdigunea    | A |
| 9          | 2           | 1           | txaleta     | kanpoaldea   | B |
| 10         | 1           | 2           | etxebizitza | erdigunea    | A |
| 11         | 3           | 1           | etxebizitza | erdigunea    | B |
| 12         | 3           | 2           | etxebizitza | kanpoaldea   | B |

- 2.1 Emandako datubasetik abiatuta, aldagai-aukeraketa egin behar duzu “Filter” izeneko teknikari jarraituz. Horretarako, aldagai iragarleek  $C$  klase-aldagaiarekiko duten elkarrekiko informazioak kalkula itzazu:  $I(L, C)$ ,  $I(K, C)$ ,  $I(M, C)$ ,  $I(Z, C)$ . Informazio horren arabera, zein aldagai-aukeraketa egingo zenuke? Aldagai iragarle guztiak esanguratsuak direla esango al zenuke?
- 2.2 Aurreko ataleko kalkuluetan oinarrituz, aldagai-aukeraketan aukera itzazu bi aldagai iragarle esanguratsuenak. Ondoren, ID3 algoritmoa erabiliz, sailkapen-zuhaitz bat eraiki ezazu.
- 2.3 Aurreko atalean eraikitako sailkapen-zuhaitza erabiliz, datubaseko 12 kasuak testeatuko bagenitu, zein asmatze-tasa lortuko genuke? Eraiki ezazu erroreen matrizea, egin ezazu TPR eta FPR koefizienteen estimazioak eta egin ezazu sailkapen-zuhaitzaren ROC kurba bidezko adierazpena. Zein ondorio aterako zenuke ROC kurba aztertuz?

3. (1.75 puntu) Sailkatzaile Bayestarrak.

3.1 Bigarren ariketako eraikuntzen datubasetik abiatuz, Naive Bayes eredua eraiki ezazu. Ondoren, erabil ezazu eredua honako ezaugarriak dituen eraikuntzaren (kasu berriaren)  $C$  klase-aldagaiaren iragarpena egiteko:  $L = 2$ ,  $K = 2$ ,  $M$ =txaleta,  $Z$ =erdigunea. Probabilitateen estimazioak kalkulatzeko, ez ezazu Laplaceren zuzenketa (rule of succession) erabili.

3.2  $C$  klase-aldagaiaren baldintzapean bi aldagai iragarleren artean dagoen elkarrekiko informazio kantitatea ere ezaguna da:  $I(L, K|C) = 0.0817$ ,  $I(L, M|C) = 0.7704$ ,  $I(L, Z|C) = 0.2400$ ,  $I(K, M|C) = 0$ ,  $I(K, Z|C) = 0.1363$ ,  $I(M, Z|C) = 0.3659$ . Emandako datubasetik abiatuz, eraiki ezazu TAN eredua. Ondoren, erabil ezazu eredua honako ezaugarriak dituen eraikuntzaren (kasu berriaren)  $C$  klase-aldagaiaren iragarpena egiteko:  $L = 2$ ,  $K = 2$ ,  $M$ =txaleta,  $Z$ =erdigunea. Probabilitateen estimazioak kalkulatzeko, Laplaceren zuzenketa (rule of succession) erabil ezazu.

3.3 Teoria. Azal ezazu, modu laburrean, zertan datzan Naive Bayes sailkatzailea. Azal ezazu baita, zertan datzan TAN eredua. Zein da bi sailkatzaile Bayestar horien arteko diferentziarik nagusia?

4. (0.5 puntu) Taulan dauden datuekin Forgy-ren algoritmoa erabili nahi dugu  $k = 3$  klaseak aurkitzeko. Hasierako partiketa zoriz eginez, azken zutabea (klasea) dagoen partiketa atera da. Jarrai itzazu algoritmoaren pausuak amaitu arte.

| Kasua          | $X$ | $Y$ | Klasea |
|----------------|-----|-----|--------|
| $\mathbf{x}_1$ | 1   | 1   | 1      |
| $\mathbf{x}_2$ | 1   | 6   | 1      |
| $\mathbf{x}_3$ | 4   | 3   | 2      |
| $\mathbf{x}_4$ | 5   | 5   | 2      |
| $\mathbf{x}_5$ | 6   | 1   | 3      |
| $\mathbf{x}_6$ | 6   | 7   | 2      |