3. Gainbegiratutako Sailkatzaileen Ebaluazioa

Espezialitatea: Konputazioa, hirugarren ikasmaila
Titulazioa: Informatika Ingeniaritzako Gradua
Konputazio Zientzia eta Adimen Artifiziala saila
Universidad del País Vasco - Euskal Herriko Unibertsitatea

Aurkibidea

1. Gainbegiratutako sailkapenerako notazioa

	X_1	...	X_i	...	X_n	C	C_M
1	x_1^1	...	x_i^1	...	x_n^1	c^1	c_M^1
...	...	...	...	...	...	...	...
j	x_1^j	...	x_i^j	...	x_n^j	c^j	c_M^j
...	...	...	...	...	...	...	...
N	x_1^N	...	x_i^N	...	x_n^N	c^N	c_M^N

Asmatutako kasu kopurua: $\sum_{j=1}^{N} \delta(c^j, c_M^j)$

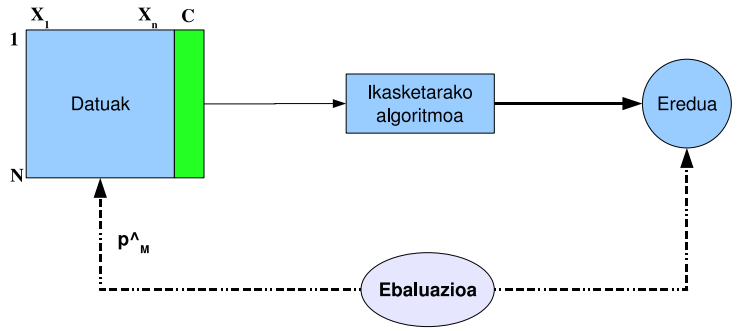
$$\delta(c^j, c_M^j) = \begin{cases} 1 & \text{baldin } c^j = c_M^j \\ 0 & \text{baldin } c^j \neq c_M^j \end{cases}$$

2. Erroreen matrizea

	$C = 1$	$C = 0$
$C_M = 1$	TP	FP
$C_M = 0$	FN	TN

- Asmatze-tasa: $\frac{TP+TN}{TP+FP+FN+TN}$ $\rightarrow$ Accuracy
- Errore-tasa: $\frac{FP+FN}{TP+FP+FN+TN}$
- True Positive Rate: $TPR = \frac{TP}{TP+FN}$ $\rightarrow$ Sensitivity
- False Negative Rate:$FNR = \frac{FN}{TP+FN}$
- False Positive Rate: $FPR = \frac{FP}{FP+TN}$
- True Negative Rate: $TNR = \frac{TN}{FP+TN}$ $\rightarrow$ Specificity

3.1 Estimazio-metodo ez-zintzoa

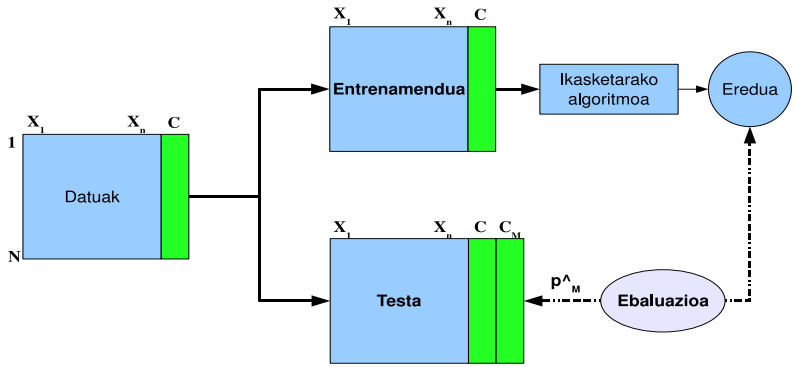


$$\hat{p}_M = \frac{1}{N}\sum_{j=1}^{N}\delta(c^j = c_M^j)$$

3.2 Holdout estimazio-metodoa: Entrenamenduan eta testean oinarritutakoa

- Holdout estimazio-metodoan, datuen multzoa bi azpimultzo disjuntutan banatzen da: **entrenamendurako azpimultzoa** eta **testerako azpimultzoa**. Adibidez, %60 entrenamendurako, %40 testerako
- Sailkapen-eredua induzitzeko datuen azpimultzo bat besterik ez da erabiltzen
- Kasu guztiak ez dira ebaluatuak izango
- Ebaluazioaren emaitza egindako entrenamendu-test banaketaren mendekoa izango da
- Datubasea txikia denean ez da holdout metodoa erabiltzea komeni

Holdout estimazio-metodoa



$$\hat{p}_M = \frac{1}{h}\sum_{j=1}^{h}\delta(c^j = c_M^j)$$

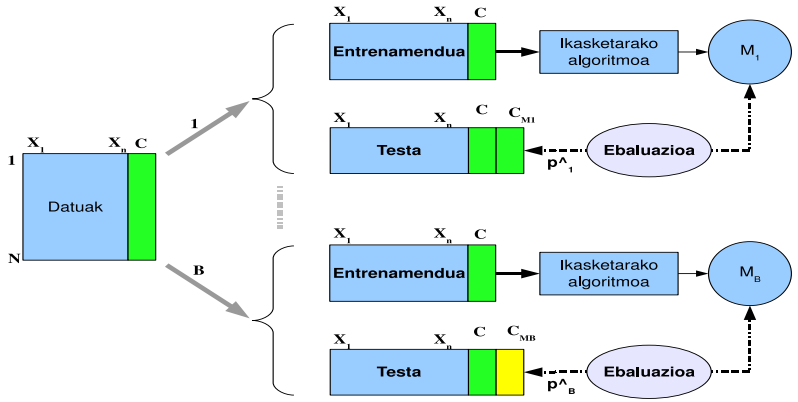
h izanik testerako azpimultzoaren tamaina

Holdout estimazio-metodoa hainbat aldiz errepikatua

- Holdout estimazio-metodoa B aldiz errepikatua izango da
- $i$garren entrenamendu-test banaketarekin Holdout egitean, $\hat{p}_i$ asmatze-tasa lortuko da, $i = 1, \ldots, B$
- B aldiz Holdout aplikatu eta gero, asmatze-tasa horien guztien batazbestekoak emango digu estimazio-metodo honen azken asmatze-tasa

$$\hat{p}_M = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i$$

Holdout hainbat aldiz errepikatua



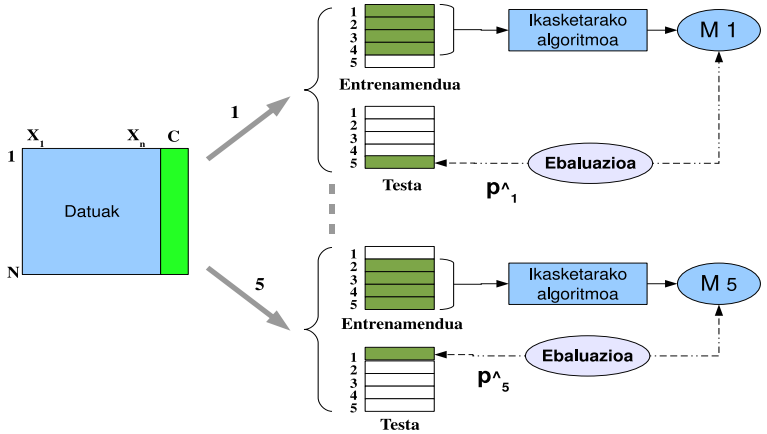
3.3 k geruzako balidazio gurutzatua (k-fold cross validation)

- D datubasea beren artean disjuntu eta tamainaz antzeko diren $D_1, D_2, ..., D_k$ geruzetan (fold) banatzen da
- Geruzetako kasuak zoriz aukeratzen dira
- Sailkatzailea k aldiz entrenatua eta testeatua izango da. Aldiero, $D \setminus D_i$ kasu erabiliko dira entrenamendurako eta D_i testerako ($i \in \{1, 2, ..., k\}$)

$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^{k} \hat{p}_i$$

- Datubaseko kasu guztiak behin testeatuak izango dira
- Entrenamendurako erabili diren k azpimultzoak beren artean nahiko antzekoak dira
- $k = N$ denean, k geruzako balidazio gurutzatuari “leave one out” esaten zaio

k geruzako balidazio gurutzatua. Adibidea, $k = 5$

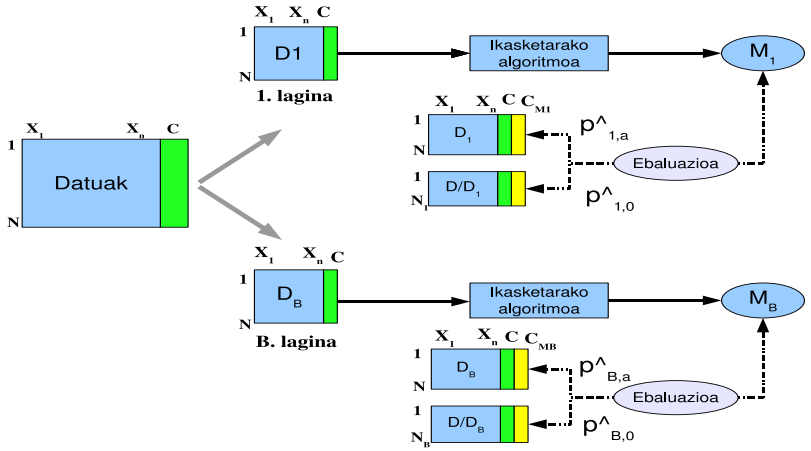


3.4 0,632 bootstrap bidezko estimazio-metodoa

- N kasu dituen datubase batean, entrenamendurako D_i **bootstrap lagina** horrela aukeratzen da: N kasu aukeratu, kasua birjarriz (berriz aukeratua izan daiteke)
- Testerako: entrenamendurako aukeratuak izan ez direnak, $D - D_i$
- Bootstrap laginerako aukeratua izateko probabilitatea:$1/N$
- Bootstrap laginerako aukeratua EZ izateko probabilitatea: $1 - (1/N)$
- N aldiz errepikatu, bootstrap lagin osoa osatzeko. Hortaz, guztira kasu bat aukeratua EZ izateko probabilitatea: $(1 - 1/N)^N$
- Datubasea handian, $N \to \infty$, limitean: $(1 - 1/N)^N \approx e^{-1} \approx 0{,}368$
- D_i-ren asmatze-tasa: $\hat{p}_{i,a}$. $D - D_i$-ren asmatze-tasa: $\hat{p}_{i,0}$
- B bootstrap laginetarako, $\hat{p}_a = 1/B \sum_{i=1}^{B} \hat{p}_{i,a}$, $\hat{p}_0 = 1/B \sum_{i=1}^{B} \hat{p}_{i,0}$

$$\hat{p}_M = 0{,}368\, \hat{p}_a + 0{,}632\, \hat{p}_0$$

0,632 **bootstrapping**



4. Kostuan oinarritutako ebaluazioa: ROC analisia

- Zenbait kasutan, sailkatzaileak egin ditzakeen bi errore motek ez dituzte ondorio berberak eragiten. Kostuen matrizea:

Erroreen matrizea

	$C=1$	$C=0$
$C_M=1$	TP	FP
$C_M=0$	FN	TN

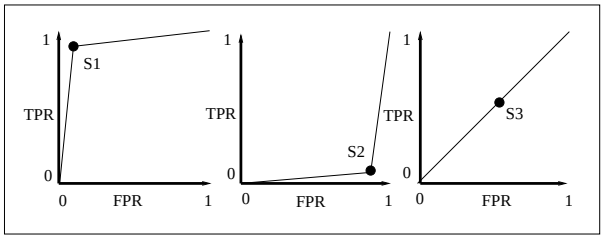
Kostuen matrizea

	$C=1$	$C=0$
$C_M=1$		FP_{cost}
$C_M=0$	FN_{cost}	

- C klaseak bi balio posible dituen problemetan erabili ohi da
- Bi errore mota horiek bereiztuz, sailkatzaileen jokaera ikusi ahal izango dugu grafikoki
- Sailkatzaileen portaera aztertuz, modu egokienean jokatzen dutenak aukeratu ahal izango dira

ROC espazioa. Sailkatzaile onak eta txarrak

Gogoratu: $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$



- S1 ona da: TPR altua eta FPR baxua
- S2 txarra da: TPR baxua eta FPR altua
- S3 txarra da

Bibliografia

Oinarrizko bibliografia

- Liburua: Introducción a la Minería de Datos
  - Capítulo 17: Técnicas de Evaluación
- J. Hernández Orallo, MªJ. Ramírez Quintana, C. Ferri Ramírez
- Pearson Prentice Hall, 2004, ISBN: 84-205-4091-9

Wikipedia

- http://en.wikipedia.org/wiki/Accuracy_and_precision
- http://en.wikipedia.org/wiki/Sensitivity_and_specificity
- http://en.wikipedia.org/wiki/Cross-validation_(statistics)
- http://en.wikipedia.org/wiki/Receiver_operating_characteristic