7. Aldagai Aukeraketa

Espezialitatea: Konputazioa, hirugarren ikasmaila
Titulazioa: Informatika Ingeniaritzako Gradua
Konputazio Zientzia eta Adimen Artifiziala Saila
Euskal Herriko Unibertsitea

1. Sarrera

Aldagaien azpimultzoaren aukeraketa

- Feature Subset Selection (FSS): Aldagaien azpimultzoaren aukeraketa.
- Eredu sailkatzaileak induzitzeko aldagai iragarleen azpimultzo egokia aukeratzea beharrezkoa eta egokia gertatzen da askotan.
- Aukeraketa egiteko bi filosofia nagusi daude:
  - Filter edo iragazki modukoak
  - Wrapper edo bilgarri modukoak

2. Zergatik egin aldagai iragarleen aukeraketa?

Ez da egia aldagai iragarle gehiago erabiltzeagatik, induzitutako sailkatzailea hobea izango denik

- Aldagai iragarle batzuk ez dira esanguratsuak
  - Aldagai iragarle baten balioa ezagutzeak ez badu *C* klase-aldagaiaren iragarpenean laguntzen, esanguratsua ez dela esaten da
- Aldagai iragarle batzuk beren artean erredundanteak dira
  - Aldagai iragarle bat beste aldagai iragarleetatik ondoriozta badaiteke, erredundantea dela esaten da

- Aldagai-aukeraketa egitearen helburua: esanguratsuak ez diren eta erredundanteak diren aldagaiak identifikatzea
- Datuak ondo adierazten dituzten bi ereduren artean, sinpleena aukeratu

3. Aldagai-aukeraketa egitearen abantailak

Abantailak

- Sailkatzailea hobea izango da, orokortzeko gaitasun handiagoa izango du (*overfitting* ekidingo da)
- Eredu sailkatzailea ulergarriagoa izango da
- Sailkatzailearen indukzio-prozesua azkarragoa izango da
- Eredua induzitzeko informazio gutxiago beharko denez, datuak lortzea gutxiago kostatuko da

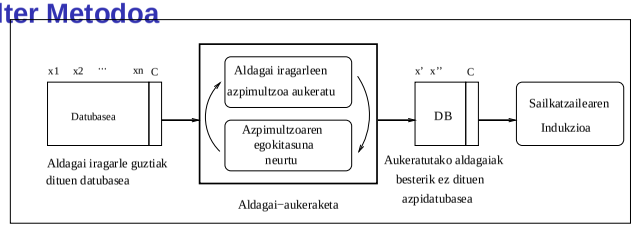
Abantaila horiek kostu konputazional handiagoaren truke lortuko dira; aldagaien azpimultzo optimoa aurkitzearen prezioa!

4. Aldagai-aukeraketarako metodoak

Filter Metodoa: iragazki moduko metodoa

- Aldagai iragarleek *C* klase-aldagaiaren iragarpenerako ematen duten informazioa neurtuko da: elkarrekiko informazioa, irabazi-ratioa, chi karratu...
- Neurri horren arabera aldagai iragarleak ordenatuko dira, esanguratsuena denetik hasi, eta informazio gutxien ematen duenera
- k aldagai iragarle esanguratsuenak aukeratuko dira, eta aukeratutako aldagai iragarle horiekin egingo da eredu sailkatzailearen indukzioa
- Aldagai iragarleak ordenatzeko erabili den neurriak ez du kontuan hartzen zein izango den gerora erabiliko den eredu sailkatzailea: sailkapen-zuhaitza, Naive Bayes...

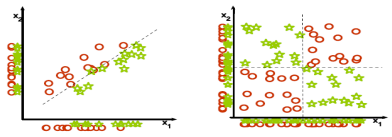
Filter Metodoa



- Abantailak: Konputazionalki azkarra. Gainera, eredu sailkatzaileagandik independentea izateagatik, aldagai-aukeraketa behin bakarrik egingo da, eta aldagai horietan oinarrituz, nahi adina eredu sailkatzaile induzituko dira
- Desabantaila: aldagai-aukeraketa eredu sailkatzaileagandik independente egiteagatik, ez da beraien arteko elkarrekintza kontuan hartzen, eta desegokia izan daiteke. Gainera, aldagaiak banaka aztertzen badira, ez dira aldagaien arteko elkarrekintzak kontuan hartzen. Binaka aztertu? Hirunaka?

Aldagaien arteko elkarrekintzak baztertuz gero,

Univariate selection may fail



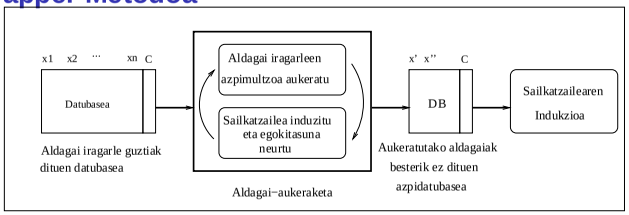
Guyon-Elisseeff, JMLR 2004; Springer 2006

4. Aldagai-aukeraketarako metodoak

Wrapper Metodoa: bilgarri moduko metodoa

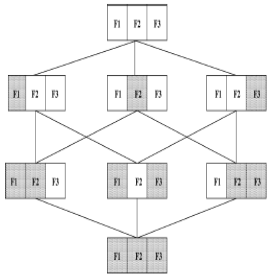
- Aldagai iragarleen azpimultzoak, azpimultzo horretan oinarrituz induzitutako eredu sailkatzailearen bidez ebaluatuak izango dira
- n aldagai iragarle izanik, 2^n azpimultzo sor daitezke, eta horietako onena zein den erabaki beharko da
- Azpimultzo kopurua, aldagai iragarle kopuruarekin batera, esponentzialki hazten da. Ezinezkoa gertatzen da aldagaien azpimultzo guztietarako sailkatzailea entrenatu eta testeatzea, guztien artetik onena aukeratzeko
- Aldagaien azpimultzo optimoaren bilaketa, zenbait algoritmo heuristiko erabiliz egin ohi da, hala nola, algoritmo genetikoak, etab

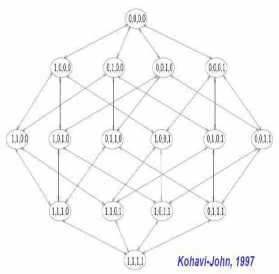
Wrapper Metodoa



- Abantailak: oso onak dira aldagaien azpimultzo optimoaren bilaketaren eta eredu sailkatzailearen arteko elkarrekintza kontuan hartzen. Gainera, aldagai iragarleen arteko mendekotasunak ondo antzematen dituzte
- Desabantailak: Konputazionalki oso garestiak dira. Gainera, entrenamendurako datubasera gehiegi egokitzeko arriskua handiagoa da filter metodoetan baino, sailkatzailearen orokortzeko gaitasuna galduko delarik

Aldagaien azpimultzo optimoaren bilaketa





N features, 2^N possible feature subsets!

Aldagaien azpimultzo optimoaren bilaketa

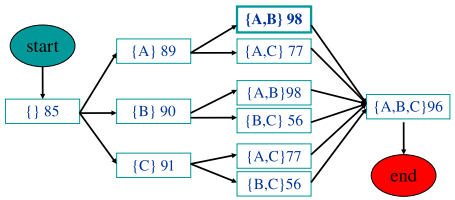
A, B, C aldagai iragarleak eta *M* sailkatzailea *izanik,* klase aldagaia iragarri nahi dugu, {A,B,C} aldagaien azpimultzo txikiena erabiliz, eta asmatze-tasarik handiena lortuz

Aldagai iragarleen azpimultzoak	Eredu sailkatzailea	Asmatze-tasa
{A,B,C}	*M*	**96%**
{A,B}	*M*	**98%**
{A,C}	*M*	77%
{B,C}	*M*	56%
{A}	*M*	89%
{B}	*M*	90%
{C}	*M*	91%
{ }	*M*	85%

Aldagaien azpimultzo optimoaren bilaketa

Aldagai iragarle kopuru handietarako, ez da posible azpimultzo guztiak aztertzea

Azpimultzo optimoaren bilaketarako, algoritmo heuristikoren bat erabili beharko da



Bilaketa heuristikoarekin, agian ez dugu azpimultzo optimoa aurkituko

5. Bestelako metodoak

- Embedded methods. Eredu sailkatzailea eraikitzearekin batera egiten da aldagaien azpimultzo optimoaren bilaketa. Adibidez, sailkapen-zuhaitzak. Wrapper metodoetan bezala, aldagai-aukeraketa eredu sailkatzailearentzat bereziki egiten da, wrapper metodoak bezain garestiak izan gabe
- Aldagai-aukeraketa egin beharrean, aldagai iragarleei pisu bat egoki dakieke
- Datuen dimentsioaren murrizketa egin daiteke aldagai-aukeraketa egin gabe, jatorrizko aldagai iragarleetatik beste aldagai batzuk eratorriz. Teknika desberdinak daude. Adibidez, transformazio linealen bidez, SVD (Singular Value Decomposition)

SVD. Testuen sailkapenerako adibidea

Testuen sailkapena

- Ohiko aplikazioak: Posta elektronikoan zaborra diren mezuak antzematea, internet bidezko dokumentuen bilaketak, hitzen adieren desanbiguazioa, etab
- Bag of Words bidezko adierazpena. Dokumentua hiztegiaren tamainako bektorearen bidez adierazten da, dokumentuan hitz bakoitzak duen agerpenarekin
- Ohikoa da hiztegiak ehundaka milaka hitz izatea; sarritasun txikieneko eta handieneko hitzak hasieratik kenduta, 15000 ingurura jaitsi daiteke hiztegiaren tamaina
- 5000-800000 dokumentuko datubaseak izaten dira erabilgarri ikerketarako

SVD. Adibidea.

Dokumentuak

- c1-c5: “gizaki-konputagailu arteko interakzioa”
- m1-m4: “grafoen teoria”

c1: Human machine interface for Lab ABC computer applications.
c2: A survey of user opinion of computer system response time.
c3: The EPS user interface management system.
c4: System and human system engineering testing of EPS.
c5: Relation of user-perceived response time to error measurement.
m1: The generation of random, binary, unordered trees.
m2: The intersection graph of paths in trees.
m3: Graph minors IV: Widths of trees and well-quasi-ordering.
m4: Graph minors: A survey.

SVD adibidea: Dokumentuen adierazpena

M matrizea: terminoak x dokumentuak

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} \rightarrow \text{human} \\ \rightarrow \text{interface} \\ \rightarrow \text{computer} \\ \rightarrow \text{user} \\ \rightarrow \text{system} \\ \rightarrow \text{response} \\ \rightarrow \text{time} \\ \rightarrow \text{EPS} \\ \rightarrow \text{survey} \\ \rightarrow \text{trees} \\ \rightarrow \text{graph} \\ \rightarrow \text{minors} \end{matrix}$$

$$\begin{matrix} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ c1 & c2 & c3 & c4 & c5 & m1 & m2 & m3 & m4 \end{matrix}$$

SVD: Singular Value Decomposition

M=U Σ V

$$\mathbf{U} = \begin{pmatrix} 0{,}22 & -0{,}11 & 0{,}29 & -0{,}41 & -0{,}11 & -0{,}34 & 0{,}52 & -0{,}06 & -0{,}41 \\ 0{,}20 & -0{,}07 & 0{,}14 & -0{,}55 & 0{,}28 & 0{,}50 & -0{,}07 & -0{,}01 & -0{,}11 \\ 0{,}24 & 0{,}04 & -0{,}16 & -0{,}59 & -0{,}11 & -0{,}25 & -0{,}30 & 0{,}06 & 0{,}49 \\ 0{,}40 & 0{,}06 & -0{,}34 & 0{,}10 & 0{,}33 & 0{,}38 & 0{,}00 & 0{,}00 & 0{,}01 \\ 0{,}64 & -0{,}17 & 0{,}36 & 0{,}33 & -0{,}16 & -0{,}21 & -0{,}17 & 0{,}03 & 0{,}27 \\ 0{,}27 & 0{,}11 & -0{,}43 & 0{,}07 & 0{,}08 & -0{,}17 & 0{,}28 & -0{,}02 & -0{,}05 \\ 0{,}27 & 0{,}11 & -0{,}43 & 0{,}07 & 0{,}08 & -0{,}17 & 0{,}28 & -0{,}02 & -0{,}05 \\ 0{,}30 & -0{,}14 & 0{,}33 & 0{,}19 & 0{,}11 & 0{,}27 & 0{,}03 & -0{,}02 & -0{,}17 \\ 0{,}21 & 0{,}27 & -0{,}18 & -0{,}03 & -0{,}54 & 0{,}08 & -0{,}47 & -0{,}04 & -0{,}58 \\ 0{,}01 & 0{,}49 & 0{,}23 & 0{,}03 & 0{,}59 & -0{,}39 & -0{,}29 & 0{,}25 & -0{,}23 \\ 0{,}04 & 0{,}62 & 0{,}22 & 0{,}00 & -0{,}07 & 0{,}11 & 0{,}16 & -0{,}68 & 0{,}23 \\ 0{,}03 & 0{,}45 & 0{,}14 & -0{,}01 & -0{,}30 & 0{,}28 & 0{,}34 & 0{,}68 & 0{,}18 \end{pmatrix}$$

SVD: Singular Value Decomposition

M=U Σ V

$$\Sigma = \begin{pmatrix} 3{,}34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2{,}54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2{,}35 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1{,}64 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1{,}50 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1{,}31 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}85 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}56 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}36 \end{pmatrix}$$

SVD: Singular Value Decomposition

M=U Σ V

$$\mathbf{V} = \begin{pmatrix} 0,20 & -0,06 & 0,11 & -0,95 & 0,05 & -0,08 & 0,18 & -0,01 & -0,06 \\ 0,61 & 0,17 & -0,50 & -0,03 & -0,21 & -0,26 & -0,43 & 0,05 & 0,24 \\ 0,46 & -0,13 & 0,21 & 0,04 & 0,38 & 0,72 & -0,24 & 0,01 & 0,02 \\ 0,54 & -0,23 & 0,57 & 0,27 & -0,21 & -0,37 & 0,26 & -0,02 & -0,08 \\ 0,28 & 0,11 & -0,51 & 0,15 & 0,33 & 0,03 & 0,67 & -0,06 & -0,26 \\ 0,00 & 0,19 & 0,10 & 0,02 & 0,39 & -0,30 & -0,34 & 0,45 & -0,62 \\ 0,01 & 0,44 & 0,19 & 0,02 & 0,35 & -0,21 & -0,15 & -0,76 & 0,02 \\ 0,02 & 0,62 & 0,25 & 0,01 & 0,15 & 0,00 & 0,25 & 0,45 & 0,52 \\ 0,08 & 0,53 & 0,08 & -0,03 & -0,60 & 0,36 & 0,04 & -0,07 & -0,45 \end{pmatrix}$$

SVD: Espazio murriztua

2 dimentsiotara murriztuz, $M \rightarrow M_2$

$$\mathbf{M}_2 = \mathbf{U}_2\Sigma_2\mathbf{V}_2^T = \sum_{i=1}^{2} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^T$$

$$\mathbf{M}_2 = \begin{pmatrix} 0,22 & -0,11 \\ 0,20 & -0,07 \\ 0,24 & 0,04 \\ 0,40 & 0,06 \\ 0,64 & -0,17 \\ 0,27 & 0,11 \\ 0,27 & 0,11 \\ 0,30 & -0,14 \\ 0,21 & 0,27 \\ 0,01 & 0,49 \\ 0,04 & 0,62 \\ 0,03 & 0,45 \end{pmatrix} \begin{pmatrix} 3,34 & 0 \\ 0 & 2,54 \end{pmatrix} \begin{pmatrix} 0,20 & 0,61 & 0,46 & 0,54 & 0,28 & 0,00 & 0,02 & 0,02 & 0,08 \\ -0,06 & 0,17 & -0,13 & -0,23 & 0,11 & 0,19 & 0,44 & 0,62 & 0,53 \end{pmatrix}$$

SVD: 2 dimentsioko espazio murriztua

Dokumentuen adierazpen grafikoa planoan

d dokumentu bakoitzaren proiekzioa planoan

$$\mathbf{d}_2 = \mathbf{d}^T\mathbf{U}_2\Sigma_2^{-1}$$

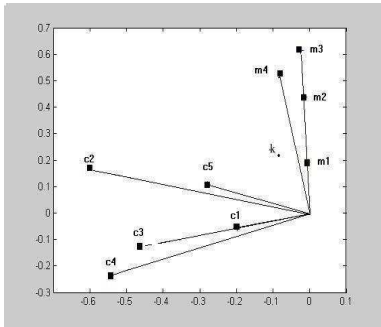
Adibideko 9 dokumentuen proiekzioak

$$\mathbf{D}_2 = \begin{pmatrix} -0.20 & -0.61 & -0.46 & -0.54 & -0.28 & -0.01 & -0.01 & -0.02 & -0.08 \\ -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53 \end{pmatrix}$$

↓	↓	↓	↓	↓	↓	↓	↓	↓
c1	c2	c3	c4	c5	m1	m2	m3	m4

Dokumentuen proiekzioen koordenatuak planoan.

SVD: 2 dimentsioko espazio murriztua



Bibliografia

- http://en.wikipedia.org/wiki/Feature_selection
- "An Introduction to Variable and Feature Selection".
  Isabelle Guyon, André Elisseeff.
  Journal of Machine Learning Research 3, (2003), pp. 1157-1182
- "A Review of Feature Selection Techniques in Bioinformatics".
  Yvan Saeys, Iñaki Inza, Pedro Larrañaga.
  Bioinformatics, Review. Vol.23, n.19, (2007), pp. 2507-2517