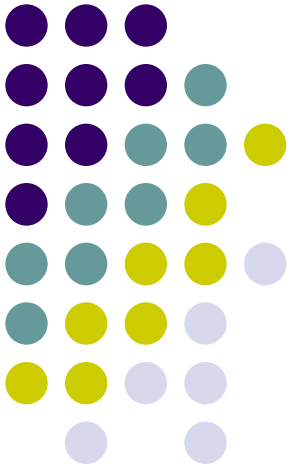
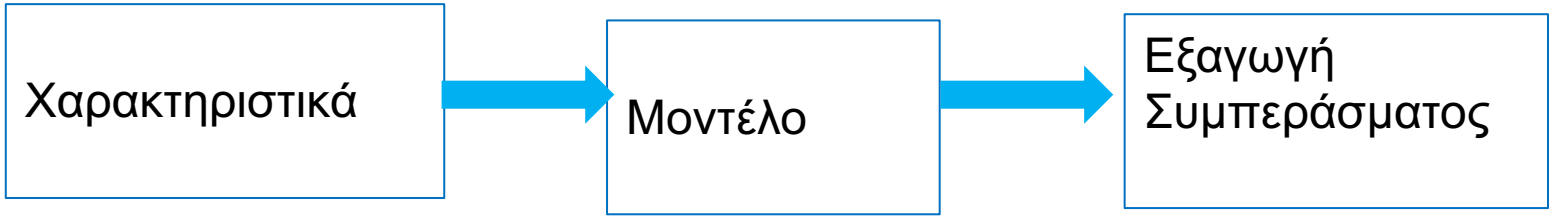
Θεωρία Αποφάσεων

Σ. Λυκοθανάσης, Καθηγητής
Δ. Κοσμόπουλος

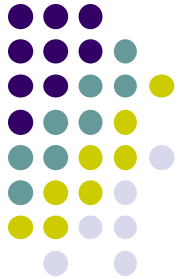
Τμήμα Μηχανικών Η/Υ & Πληροφορικής - Εργαστήριο Αναγνώρισης Προτύπων
Διευθυντής: Σ. Λυκοθανάσης, Καθηγητής



Γενικό σχήμα μάθησης

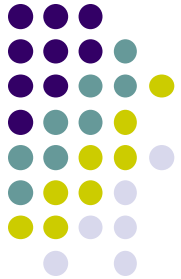


Μείωση διαστάσεων χαρακτηριστικών



- Γιατί μείωση διαστάσεων ;
- Τρεις βασικοί λόγοι:
  - Περισσότερο Επεξηγήσιμα Αποτελέσματα
  - Μείωση Υπολογιστικού Κόστους
  - Καλύτερη Γενίκευση (Αποφυγή Υπερκεπαίδευσης)

Μείωση διαστάσεων χαρακτηριστικών



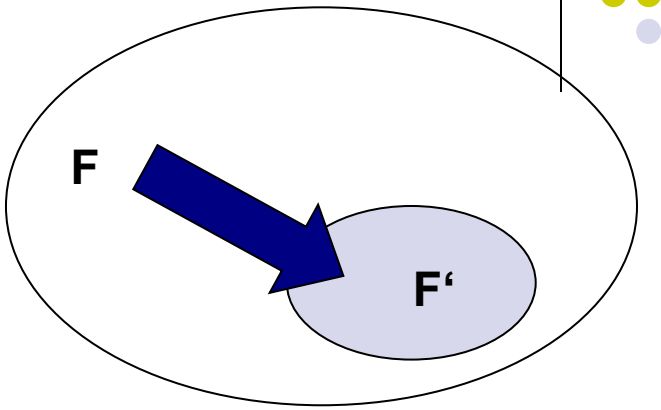
- Μπορεί να επιτευχθεί με δύο μεθόδους:

  - **Επιλογή χαρακτηριστικών (Feature Selection):** Επιλογή ενός βέλτιστου υποσυνόλου χαρακτηριστικών μεγέθους m, από ένα σύνολο n χαρακτηριστικών (ή τουλάχιστον ενός «καλού» υποσυνόλου...)

  - **Εξαγωγη/ Μετασχηματισμός χαρακτηριστικών** (Feature Extraction):. Μετασχηματισμός των χαρακτηριστικών που δημιουργεί ένα νέο σύνολο χαρακτηριστικών, λιγότερων διαστάσεων από το αρχικό.

Επιλογή Χαρακτηριστικών – Εξαγωγή Χαρακτηριστικών

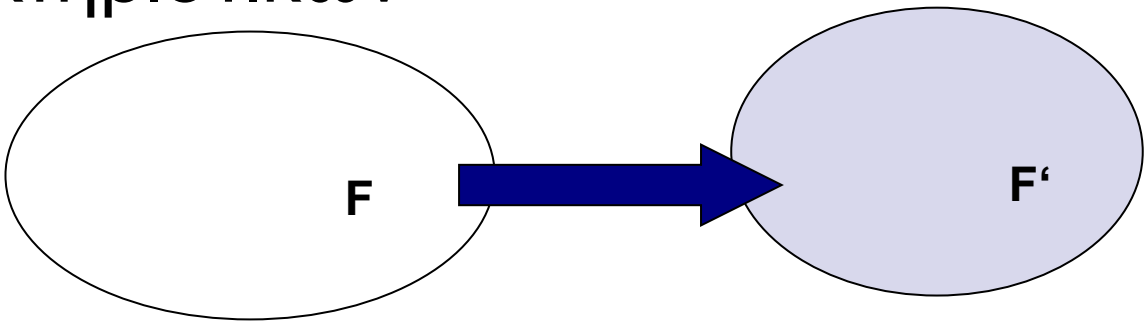
- Επιλογή Χαρακτηριστικών:



$$\{f_1,\ldots,f_i,\ldots,f_n\} \xrightarrow{f.selection} \{f_{i_1},\ldots,f_{i_j},\ldots,f_{i_m}\}$$

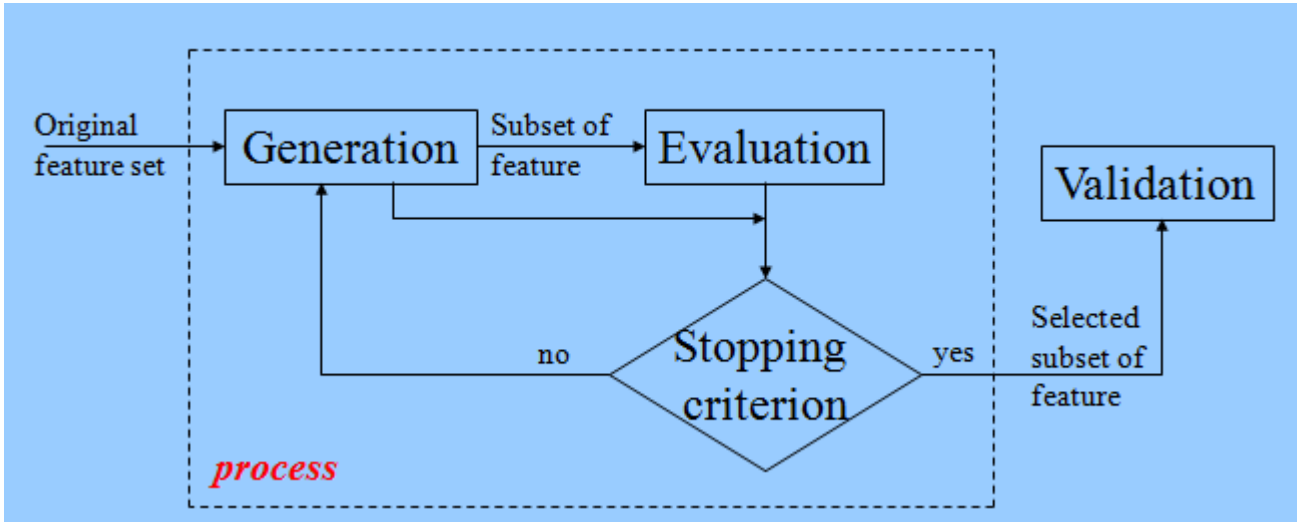
$$i_j \in \{1,\ldots,n\};\ j=1,\ldots,m$$
$$i_a = i_b \Rightarrow a=b;\ \ a,b \in \{1,\ldots,m\}$$

- Εξαγωγή Χαρακτηριστικών



$$\{f_1,\ldots,f_i,\ldots,f_n\} \xrightarrow{f.extraction} \{g_1(f_1,\ldots,f_n),\ldots,g_j(f_1,\ldots,f_n),\ldots,g_m(f_1,\ldots,f_n)\}$$

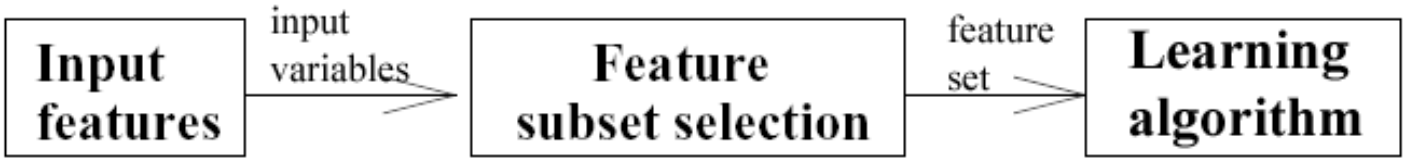
Επιλογή Χαρακτηριστικών
Feature Selection



- Αντικειμενική Συνάρτηση
  - Filters (Φίλτρα)
  - Wrappers (Περιτυλίγματα)
  - Embedded
- Στρατηγικές Αναζήτησης
  - Σειριακοί αλγόριθμοι
  - Εκθετικοί αλγόριθμοι
  - Στοχαστικοί αλγόριθμοι

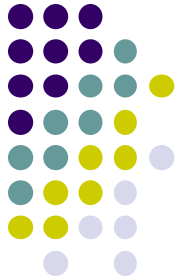
Μέθοδοι Φιλτραρίσματος

- Μέθοδοι Φιλτραρίσματος (ένα σκορ συσχέτισης παράγεται για κάθε χαρακτηριστικό από τα δεδομένα και τα χαρακτηριστικά με το χαμηλότερο σκορ φιλτράρονται)
  - Univariate (κάθε χαρακτηριστικό εξετάζεται ξεχωριστά από τα υπόλοιπα)
  - Multivariate (λαμβάνονται υπ' όψιν οι εξαρτήσεις των χαρακτηριστικών)



Τα χαρακτηριστικά επιλέγονται ανεξάρτητα από τον ταξινομητή που θα χρησιμοποιηθεί!!

Univariate Μέθοδοι Φιλτραρίσματος



- T-test – Anova:
  - Συγκρίνουν τις μέσες τιμές που παίρνει ένα χαρακτηριστικό όταν το παράδειγμα ανήκει στην μία κλάση με αυτές που παίρνει όταν ανήκει στην άλλη
  - Υποθέτει κανονική κατανομή και ίδια διασπορά
- Gamma test:
  - Αξιολογεί την εξάρτηση κάθε χαρακτηριστικού με την κλάση ταξινόμησης υποθέτοντας ότι κοντινά σημεία ανήκουν στην ίδια κλάση (μοιάζει με κοντινότερους γείτονες)
- Wilcoxon Rank Test:
  - Συγκρίνει τις διαφορές μεταξύ των παραδειγμάτων εκπαίδευσης των δύο κλάσεων χωρίς να υποθέτει κάποια κατανομή

T-test

- Χρησιμοποιείται για να συγκρίνουμε 2 κατανομές
- Εξετάζουμε ένα μόνο χαρακτηριστικό ανεξάρτητα από τα υπόλοιπα
- Προϋποθέσεις
  - Κανονικές κατανομές
  - Στατιστική ανεξαρτησία
  - Ίσες τυπικές αποκλίσεις

T-test

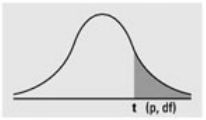
- Ορισμοί
  - Null hypothesis (h0): οι κατανομές πιθανότητας των χαρακτηριστικών ταυτίζονται
  - h1: οι κατανομές διαφέρουν (επιθυμητό)
- Το *t* σχετίζεται με πιθανότητα να ισχύει η h0 (τιμές κοντά στο μηδέν ενισχύουν την h0)
- σ_i: διασπορές, n_i: πλήθος δειγμάτων

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

T-test – εξαγωγή πιθανότητας

- Έστω n = 10 => df=n-1=9
- Αν π.χ. t = 2.43 από τον πίνακα η πιθανότητα της h0 είναι μεταξύ 0.025 και 0.01 (ακριβέστερα με παρεμβολή μεταξύ τιμών 2.26216 και 2.82144)

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460

ANOVA

- Σύγκριση περισσότερων κατανομών
- Επιλέγουμε ένα χαρακτηριστικό και εξετάζουμε για κάθε κλάση
- Υποθέσεις
  - Κανονικές κατανομές
  - Ίδια διασπορά
  - Ανεξάρτητα δείγματα

ANOVA

- Null hypothesis (h0): για όλες τις κλάσεις το χαρακτηριστικό έχει την ίδια μέση τιμή
- h1: τουλάχιστο για μια κλάση η μέση τιμή διαφέρει (επιθυμητό)

		Τιμές χαρακτηριστικού					
		δείγμα 1	δείγμα 2	δείγμα 3	δείγμα 4	δείγμα 5	δείγμα 6
	κλάση 1	7	8	15	9	10	11
	κλάση 2	12	17	13	19	15	18
	κλάση 3	14	18	19	16	18	17
	κλάση 4	19	25	22	18	20	23

ANOVA

$$\bar{y}_{..} = \frac{1}{n}\sum_{i=1}^{C}\sum_{j=1}^{n} y_{ij}$$

συνολική μέση τιμή

(*n* δείγματα / κλάση, *C* κλάσεις)

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

μ.τ. για την κλάση *i*

$$SS_{\kappa\lambda} = n\sum_{i=1}^{C}(\bar{y}_{i.} - \bar{y}_{..})^2$$

$$SS = n\sum_{i=1}^{C}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

$$F = \frac{SS_{\kappa\lambda}/(C-1)}{SS/C(C-1)}$$

ANOVA

$$F = \frac{SS_{\kappa\lambda}/(C-1)}{SS/C(C-1)}$$

- Αριθμητής: αμερόληπτος εκτιμητής σ^2 υπό την προϋπόθεση ότι ισχύει η h0
- Παρονομαστής: αμερόληπτος εκτιμητής σ^2 ανεξάρτητα αν ισχύει η h0
- F: ποσοτικοποιεί κατά πόσο ισχύει η h0

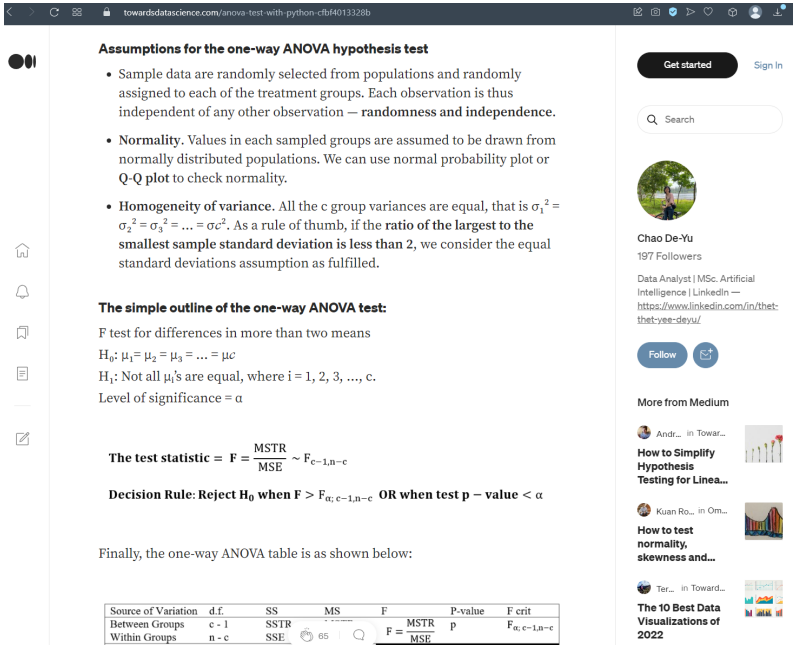
ANOVA

	Τιμές χαρακτηριστικού						
	δείγμα 1	δείγμα 2	δείγμα 3	δείγμα 4	δείγμα 5	δείγμα 6	$\bar{y}_{i.}$
κλάση 1	7	8	15	9	10	11	10
κλάση 2	12	17	13	19	15	18	15,66667
κλάση 3	14	18	19	16	18	17	17
κλάση 4	19	25	22	18	20	23	21,16667
						$\bar{y}_{..}$	15,95833
		n=6				$SS_{\kappa\lambda}$	382.79
		C=4				SS	130.17
						F	19.60

$p(f>19.60) = 3.59 \times 10^{-6}$

Η πιθανότητα με την οποία ισχύει η h0

ANOVA



towardsdatascience.com/anova-test-with-python-cfbf4013328b

Assumptions for the one-way ANOVA hypothesis test

- Sample data are randomly selected from populations and randomly assigned to each of the treatment groups. Each observation is thus independent of any other observation — **randomness and independence**.
- **Normality**. Values in each sampled groups are assumed to be drawn from normally distributed populations. We can use normal probability plot or **Q-Q plot** to check normality.
- **Homogeneity of variance**. All the c group variances are equal, that is $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma c^2$. As a rule of thumb, if the **ratio of the largest to the smallest sample standard deviation is less than 2**, we consider the equal standard deviations assumption as fulfilled.

The simple outline of the one-way ANOVA test:

F test for differences in more than two means

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu c$

H_1: Not all μ_i's are equal, where i = 1, 2, 3, ..., c.

Level of significance = α

The test statistic $= F = \frac{MSTR}{MSE} \sim F_{c-1,n-c}$

Decision Rule: **Reject** H_0 **when** $F > F_{\alpha;\, c-1,n-c}$ **OR when test** $p - value < \alpha$

Finally, the one-way ANOVA table is as shown below:

Source of Variation	d.f.	SS	MS	F	P-value	F crit
Between Groups	c - 1	SSTR		$F = \frac{MSTR}{MSE}$	p	$F_{\alpha;\, c-1,n-c}$
Within Groups	n - c	SSE				

Get started Sign In

Search

Chao De-Yu

197 Followers

Data Analyst | MSc. Artificial Intelligence | LinkedIn — https://www.linkedin.com/in/thet-thet-yee-deyu/

Follow

More from Medium

Andr... in Towar...
How to Simplify Hypothesis Testing for Linea...

Kuan Ro... in Om...
How to test normality, skewness and...

Ter... in Toward...
The 10 Best Data Visualizations of 2022

$\bar{y}_{i.}$

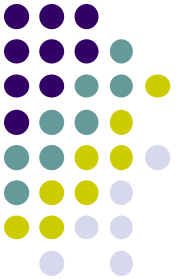
$\bar{y}_{..}$

$SS_{\kappa\lambda}$

Παράδειγμα κώδικα:

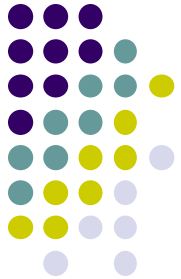
https://towardsdatascience.com/anova-test-with-python-cfbf4013328b

Multivariate Μέθοδοι Φιλτραρίσματος



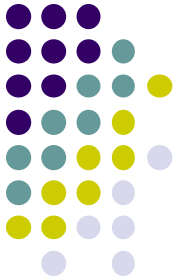
- Bivariate methods:
  - Ταξινόμησε τα ζευγάρια χρησιμοποιώντας την μέθοδο Diagonal Linear Discriminant
  - Επίλεξε ζευγάρια
- Minimum redundancy - Maximum relevance (MRMR):
  - Ευρετική μέθοδος για εύρεση βέλτιστου υποσυνόλου με μεγιστοποίηση της συσχέτισης των χαρακτηριστικών με την έξοδο και ελαχιστοποίηση της αμοιβαίας πληροφορίας.
- Correlation based Feature Selection (CFS)

Correlation-based Feature Selection



- Η ιδέα είναι να επιλέξετε χαρακτηριστικά που συσχετίζονται έντονα με την κλάση - στόχο αλλά έχουν χαμηλή συσχέτιση μεταξύ τους.

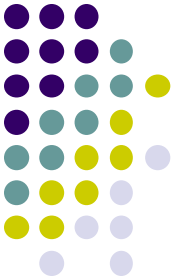
Correlation-based Feature Selection



$$r_{xy} = \frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y}$$

- Η συσχέτιση (correlation) είναι μια στατιστική μέτρηση που περιγράφει τον βαθμό σχέσης μεταξύ δύο μεταβλητών.
- Η τιμή της συσχέτισης κυμαίνεται από -1 έως 1.
  - συσχέτιση 1 υποδηλώνει τέλεια θετική σχέση,
  - -1 τέλεια αρνητική σχέση και
  - 0 καμία σχέση (αυτό που αναζητούμε).

Correlation-based Feature Selection



$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}.$$

- $\overline{r_{ff}}$: average feature-feature correlation
- $\overline{r_{cf}}$: average feature-class correlation
- k: number of features of that subset

Feature-feature correlation for features *x*,*y*

S: feature subset, c: target class

Feature-class correlation

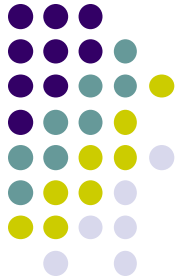
$$r_{xy} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

M_1,M_0 μέσες τιμές χαρακτηριστικού για δείγματα που ανήκουν και δεν ανήκουν στην κλαση αντίστοιχα

n_1, n_0 αριθμός δειγμάτων που ανήκουν και δεν ανήκουν στην κλαση αντίστοιχα, n συνολικός αριθμός δειγμάτων, σ_n συνολική τυπική απόκλιση

$$r_{cf} = \frac{M_1 - M_0}{\sigma_n}\sqrt{\frac{n_1 n_0}{n^2}}$$

Correlation-based Feature Selection



Πώς γίνεται η επιλογή;

1. Βρές το χαρακτηριστικό που μεγιστοποιεί το μέτρο που ορίσαμε (k=1) και πρόσθεσε το στο σύνολο επιλογής.

2. Έλεγξε το συνδυασμό του με όλα τα υπόλοιπα και διάλεξε το ζεύγος με το πρώτο που μεγιστοποιεί το μέτρο που ορίσαμε.

3. Συνέχισε μέχρι να φτάσεις ορισμένο αριθμό επαναλήψεων

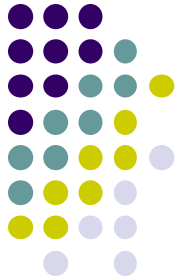
4. Αν έχεις και άλλες επαναλήψεις δοκίμασε άλλους συνδυασμούς που δεν έχουν δοκιμαστεί με τον ίδιο τρόπο

[Ευρετική μέθοδος που δεν εγγυάται βέλτιστο αποτέλεσμα]

Παράδειγμα κώδικα:

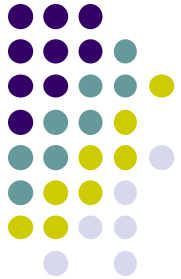
https://johfischer.com/2021/08/06/correlation-based-feature-selection-in-python-from-scratch/

Correlation-based Feature Selection



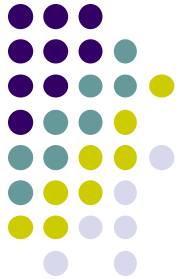
- Πλεονεκτήματα:
  - Μειώνει την υπερεκπαίδευση του μοντέλου, Βελτιώνει την απόδοση του μοντέλου, Μειώνει το υπολογιστικό κόστος
- Μειονεκτήματα:
  - Μπορεί να παραβλέψει χαρακτηριστικά που δεν συσχετίζονται γραμμικά αλλά παρέχουν σημαντικές πληροφορίες
  - Η συσχέτιση δεν συνεπάγεται αιτιότητα

Σύνοψη για Μεθόδους Φιλτραρίσματος



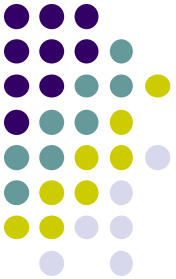
- Συνήθως γρήγορες μέθοδοι
- Επειδή η επιλογή είναι ανεξάρτητη από τον ταξινομητή είναι πιο γενική
- Από την άλλη αυτό μπορεί να είναι και μειονέκτημα. (Το υποσύνολο χαρακτηριστικών δεν είναι το βέλτιστο για τον ταξινομητή που θα χρησιμοποιηθεί)
- Κάποιες φορές χρησιμοποιούνται ως βήμα προεπεξεργασίας για άλλες μεθόδους επιλογής/εξαγωγής χαρακτηριστικών

Επιλογή Χαρακτηριστικών Περιτυλίγματος (Wrapper Feature selection)



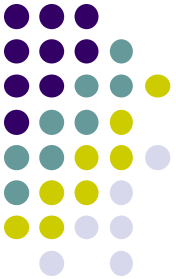
- Οι μέθοδοι περιτυλίγματος θεωρούν την επιλογή ενός συνόλου χαρακτηριστικών ως πρόβλημα αναζήτησης.
- Προετοιμάζονται, αξιολογούνται και συγκρίνονται διαφορετικοί συνδυασμοί χαρακτηριστικών.
- Για κάθε επιλογή χαρακτηριστικών εκπαιδεύεται ένα μοντέλο, το οποίο αξιολογείται με βάση την ακρίβεια του και στο τέλος κρατάμε το βέλτιστο.

Επιλογή Χαρακτηριστικών Περιτυλίγματος (Wrapper Feature selection)



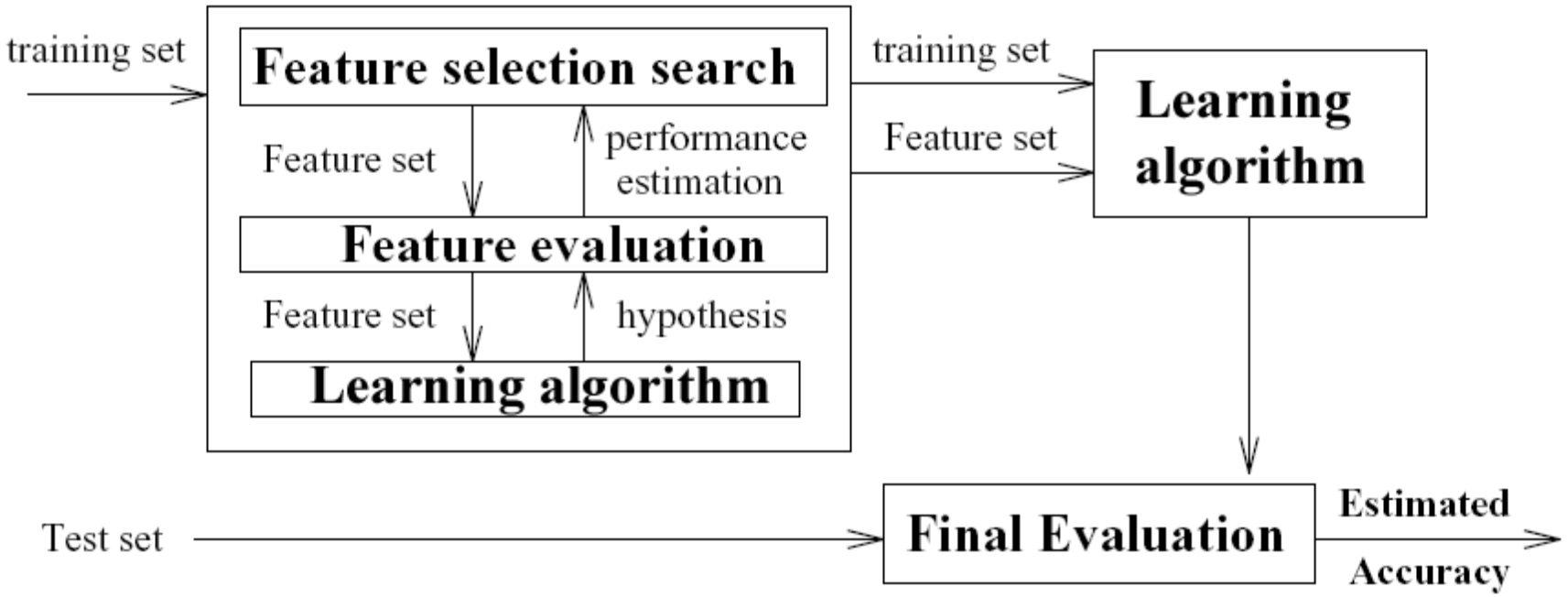
- Ο ταξινομητής θεωρείται «μαύρο κουτί»
- Η «διεπαφή» του μαύρου κουτιού χρησιμοποιείται για να δίνει ένα σκορ στα υποσύνολα των χαρακτηριστικών σύμφωνα με την απόδοση του ταξινομητή (όταν χρησιμοποιεί τα δεδομένα υποσύνολα χαρακτηριστικών).
- Τα αποτελέσματα είναι συνήθως διαφορετικά για διαφορετικούς ταξινομητές.
- Πρέπει να καθορίσει κανείς:
  - Πώς να αναζητήσει των χώρο λύσεων που ορίζεται από όλα τα πιθανά υποσύνολα χαρακτηριστικών.
  - Πώς να υπολογίσει την απόδοση του ταξινομητή.

Επιλογή Χαρακτηριστικών Περιτυλίγματος (Wrapper Feature selection)

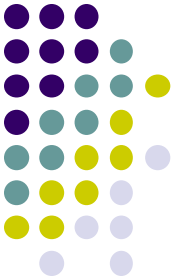


- Το πρόβλημα εύρεσης του βέλτιστου υποσυνόλου είναι NP-hard!
- Μεγάλο εύρος ευρετικών τεχνικών μπορούν να χρησιμοποιηθούν.
  Δύο διαφορετικές κατηγορίες:
  - Προς-τα-εμπρός επιλογή (Forward selection)
    (ξεκινάει με κενό σύνολο χαρακτηριστικών και προσθέτει ένα χαρακτηριστικό σε κάθε βήμα)
  - Προς-τα-πίσω απαλοιφή (Backward elimination)
    (ξεκινάει με όλα τα χαρακτηριστικά και απαλείφει ένα χαρακτηριστικό σε κάθε βήμα)
- Η απόδοση συνήθως αποτιμάται σε ένα σύνολο επικύρωσης (validation set) ή μέσω cross-validation
- Μειονέκτημα: Μεγάλο υπολογιστικό κόστος διότι για κάθε επιλογή χαρακτηριστικών γίνεται εκπαίδευση νέου μοντέλου!

Επιλογή Χαρακτηριστικών Περιτυλίγματος (Wrapper Feature selection)

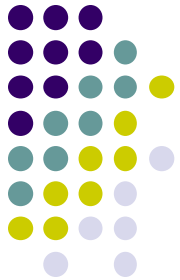


Ενσωματωμένη Επιλογή Χαρακτηριστικών Embedded (Embedded Feature selection)



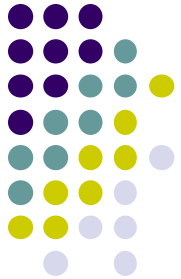
- Αναζητούν το βέλτιστο υποσύνολο χαρακτηριστικών σε συνεργασία με τον ταξινομητή (δεν χρησιμοποιούν απλά την έξοδο του ταξινομητή αλλά επιλέγουν χαρακτηριστικά ως μέρος της εκπαίδευσης).
- Βέλτιστο υποσύνολο χαρακτηριστικών μόνο για τον συγκεκριμένο ταξινομητή
- Μικρότερο υπολογιστικό κόστος σε σχέση με τις μεθόδους περιτυλίγματος διότι δεν γίνεται εκπαίδευση για κάθε συνδυασμό χαρακτηριστικών παρά μόνο μία φορά

Ακολουθιακή Προς-τα εμπρός Επιλογή Χαρακτηριστικών (Sequential Forward Feature Selection)



- Υπο-βέλτιστη μέθοδος (δεν εξετάζονται όλα τα δυνατά υποσύνολα χαρακτηριστικών)
- Μεγάλη μείωση υπολογιστικού κόστους
- Εφαρμόζεται τόσο σε wrapper όσο και σε embedded μεθόδους

Αλγόριθμος Ακολουθιακής Προς-τα εμπρός Επιλογής Χαρακτηριστικών



- Σε κάθε βήμα της αναζήτησης το υποσύνολο χαρακτηριστικών επεκτείνεται κατά ένα «βέλτιστο» χαρακτηριστικό

- Κατά την πρώτη επανάληψη κάθε χαρακτηρικό αξιολογείται βάσει ενός κριτηρίου και επιλέγεται το βέλτιστο χαρακτηριστικό x^*

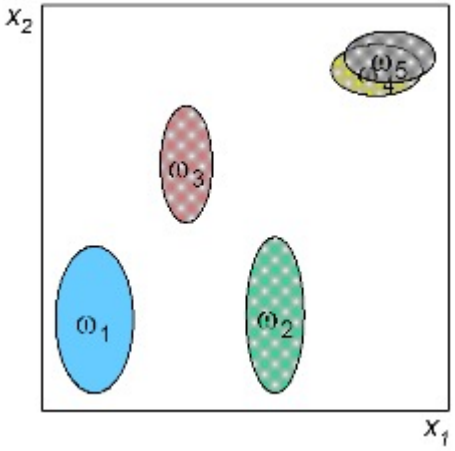
- Κατά την δεύτερη επανάληψη το κριτήριο υπολογίζεται για όλα τα ζεύγη (x^*,x_n) και επιλέγεται το καλύτερο υποσύνολο με δύο χαρακτηριστικά, κ.τ.λ.

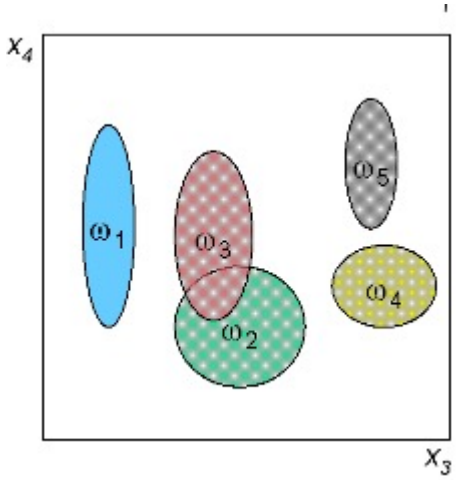
- Συνήθως είτε προκαθορίζεται ο αριθμός m των χαρακτηριστικών του βέλτιστου υποσυνόλου, ή η επέκταση σταματάει όταν κανένα από τα υποσύνολα - παιδιά δεν οδηγεί σε βελτίωση της απόδοσης.

- Για αποφυγή όμως πρόωρου σταματήματος ορίζεται κάποιες φορές μια πιο χαλαρή συνθήκη τερματισμού, που προβλέπει την συνέχιση των επεκτάσεων εφόσον υπάρχει κάποιο υποσύνολο-παιδί που οδηγεί σε απόδοση το ίδιο καλή με την έως τώρα καλύτερη και τερματισμό αν η απόδοση δεν βελτιωθεί μετά από N διαδοχικές επεκτάσεις

Αλγόριθμος Ακολουθιακής Προς-τα εμπρός Επιλογής Χαρακτηριστικών

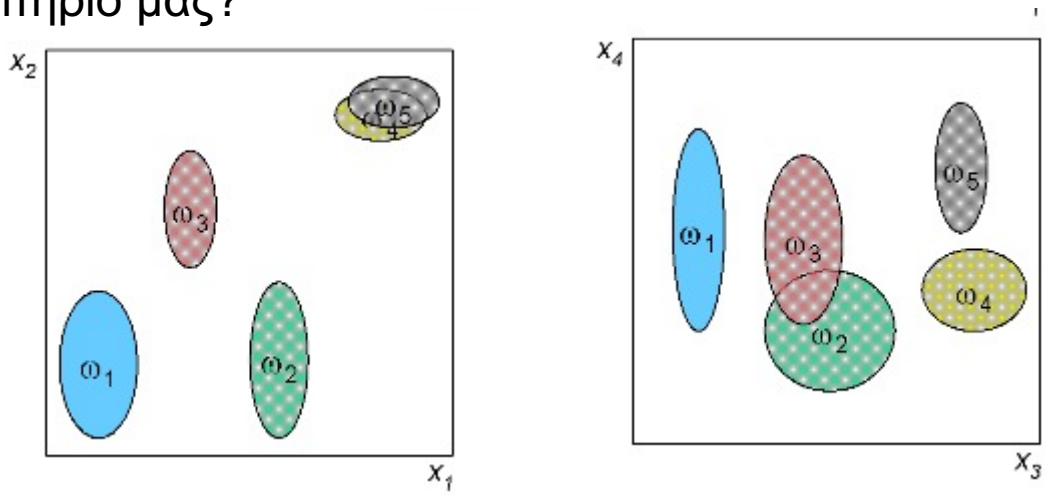
Γιατί να μην επιλέξουμε απλά τα πρώτα *m* χαρακτηριστικά σύμφωνα με το κριτήριο μας ?





Αλγόριθμος Ακολουθιακής Προς-τα εμπρός Επιλογής Χαρακτηριστικών

Γιατί να μην επιλέξουμε απλά τα πρώτα *m* χαρακτηριστικά σύμφωνα με το κριτήριο μας?



Οι περισσότερες μέθοδοι επιλογής θα ταξινομούσαν τα χαρακτηριστικά σύμφωνα με τη σειρά:

$$J(x_1) > J(x_2) \approx J(x_3) > J(x_4)$$

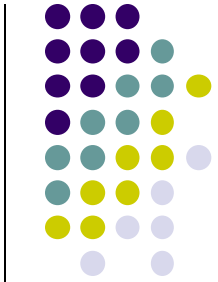
Διαχωρίζουν τα δεδομένα σε 3 κλάσεις

Διαχωρίζει μόνο την ω_4 από την ω_5 αν γνωρίζουμε τις υπόλοιπες

Διαχωρίζει τις κλάσεις ω_1, ω_2, ω_3 και το συνδυασμό $\{\omega_4,\omega_5\}$

Ο ιδανικός συνδυασμός είναι το x_1 και x_4!! Αν τα επιλέγαμε απλά τα πρώτα χαρακτηριστικά θα επιλέγαμε το x_1 και είτε το x_2 είτε το x_3 και οι κλάσεις ω_4 και ω_5 δεν θα διαχωριζόταν!

Προς-τα εμπρός Επιλογή Χαρακτηριστικών



1.Start with No Features: Initially, the model starts with no features.

2.Iteratively Add Features:

1. In each iteration, the algorithm tries out each of the remaining features (i.e., those not already selected) and evaluates the model performance by adding each feature to the set of selected features.
2. The feature whose addition gives the best improvement in performance (e.g., accuracy, F1-score) is permanently added to the set of selected features.

3.Performance Evaluation:

1. The performance of the model is typically evaluated using cross-validation or a holdout validation set.
2. The metric for performance evaluation depends on the problem type (classification, regression, etc.) and specific requirements (accuracy, precision, recall, etc.).

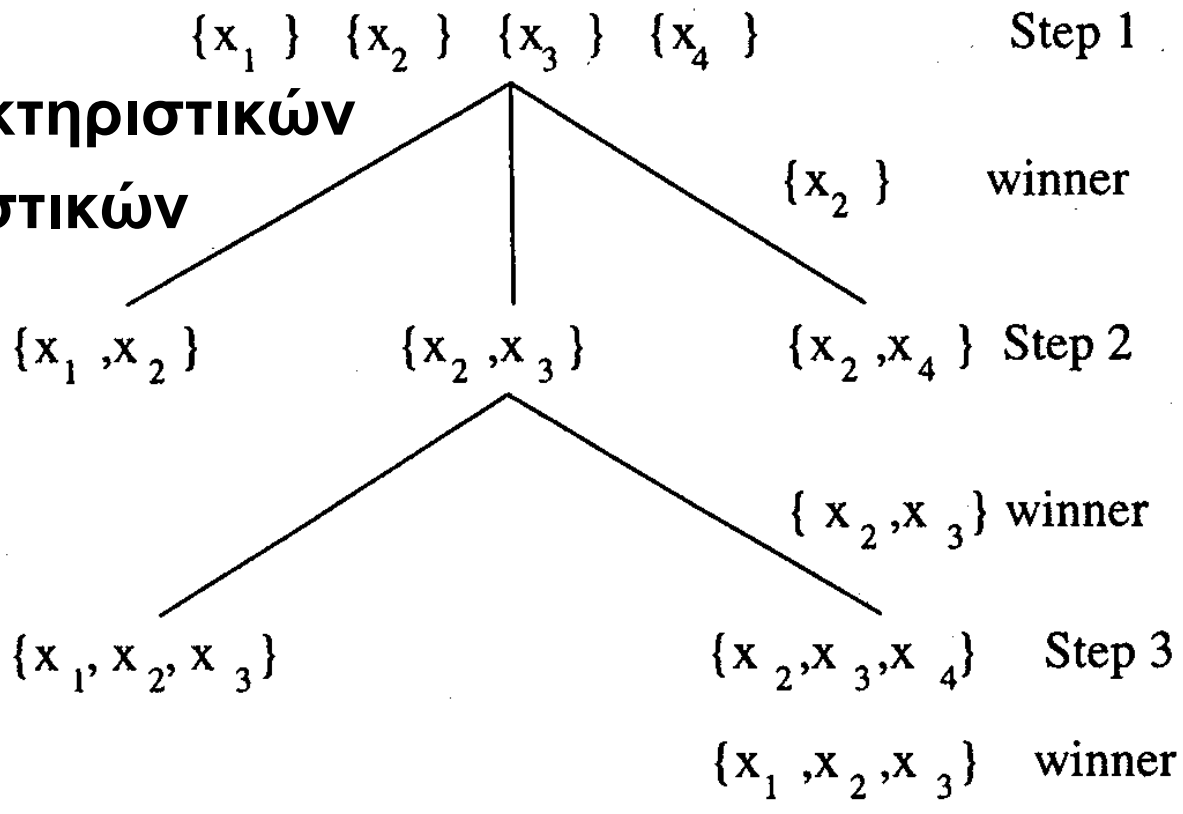
4.Stopping Criterion:

1. The process stops when adding new features does not improve the model performance beyond a certain threshold or when all features have been exhausted.
2. Sometimes, a predefined number of features to select is set as a stopping criterion.

Ακολουθιακή Προς-τα εμπρός Επιλογή Χαρακτηριστικών

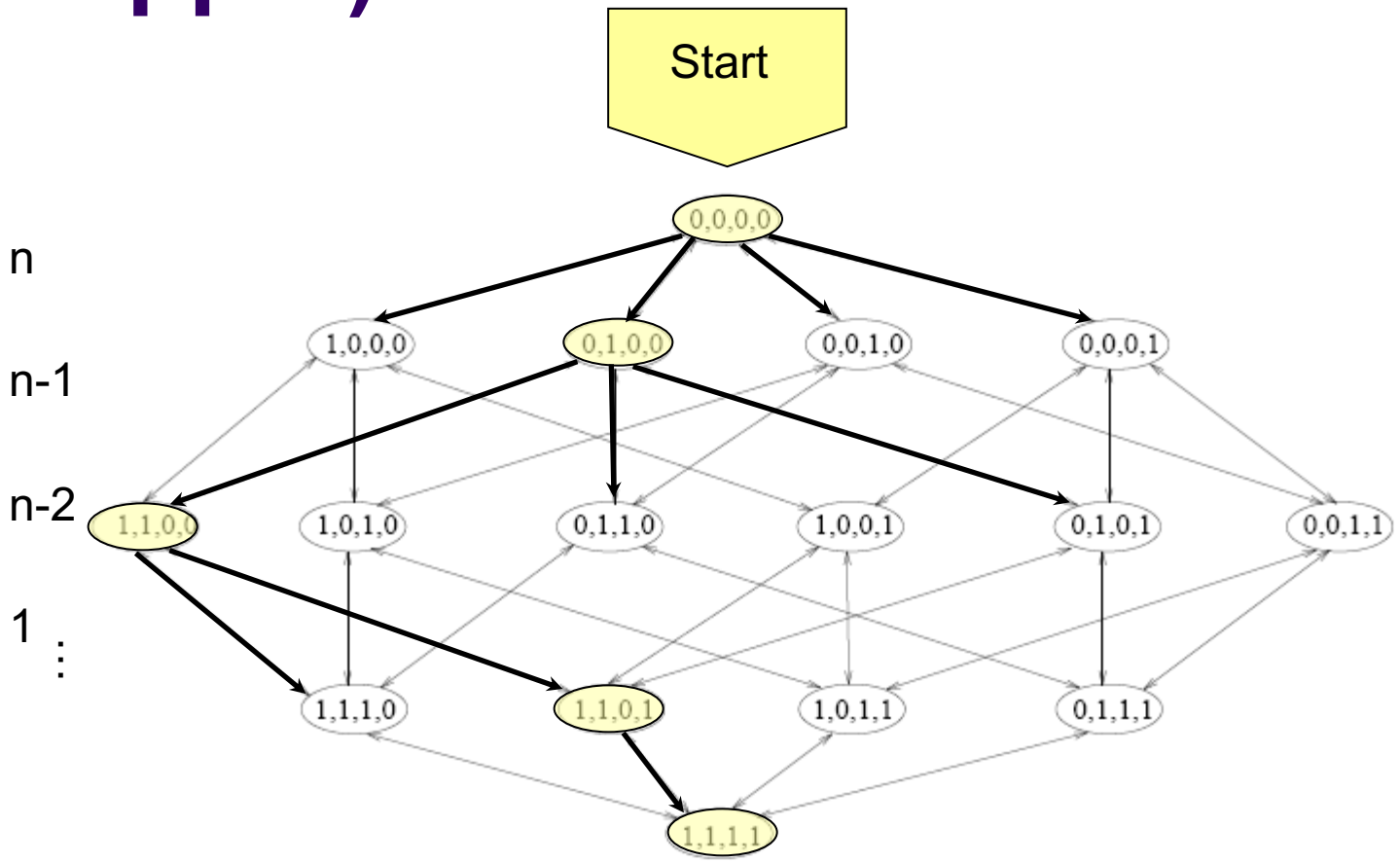
Παράδειγμα:

Επιλογή m=3 χαρακτηριστικών από n=4 χαρακτηριστικών

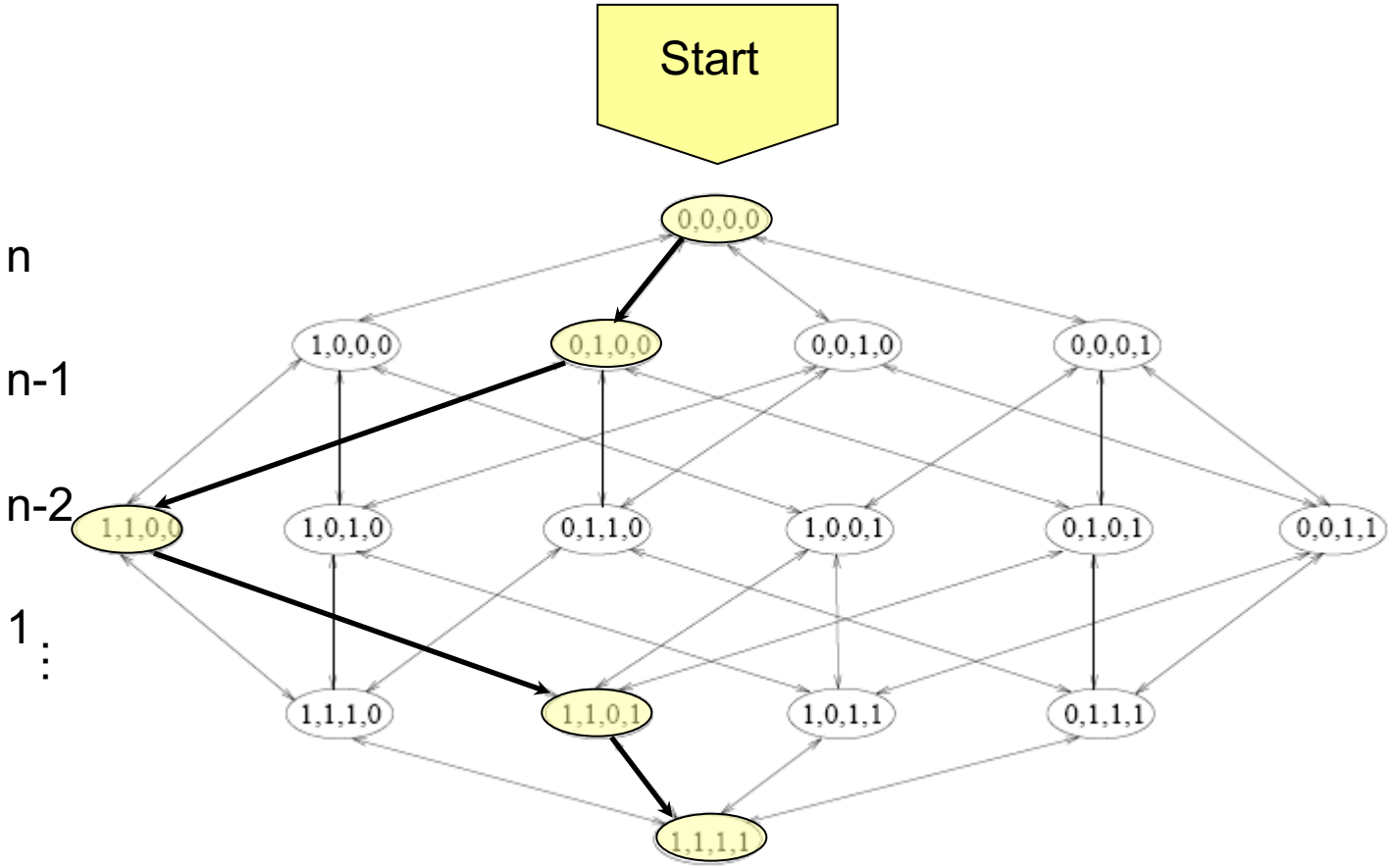


Sequential forward feature selection search

Προς-τα-εμπρος επιλογή (wrapper)



Προς-τα-εμπρος επιλογή (embedded)

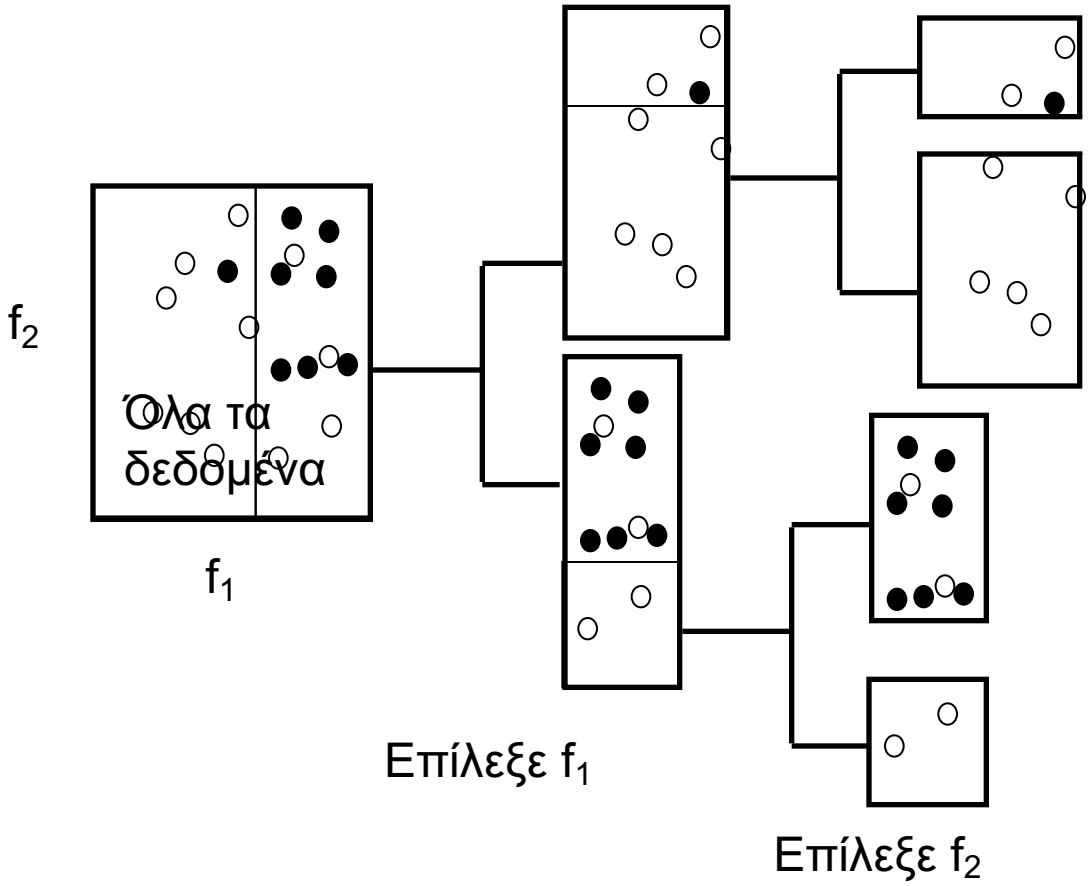


Καθοδηγούμενη αναζήτηση: δεν λαμβάνουμε υπόψη εναλλακτικά μονοπάτια

Προς-τα-εμπρος Επιλογή με Δένδρα Απόφασης (embedded)

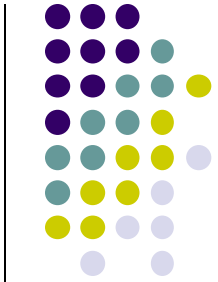
- Δένδρα Απόφασης όπως τα

  ICART *(Breiman, 1984)* ή *C4.5 (Quinlan, 1993)*



Σε κάθε βήμα επίλεξε το χαρακτηριστικό που ελαχιστοποιεί περισσότερο την εντροπία (των διάφορετικών κλάσεων). Προσπαθούμε δηλαδή να οδηγηθούμε σε φύλλα που θα έχουν δείγματα μόνο μίας κλάσης

Προς-τα-πίσω Απαλοιφή
Backward Elimination



1.Start with All Features: Initially, the model includes all available features.

2.Iteratively Remove Features:

1. In each iteration, the algorithm temporarily removes each feature one at a time and evaluates the model performance without that feature.
2. The feature whose removal causes the least deterioration in performance (or sometimes even improves it) is permanently eliminated from the set.

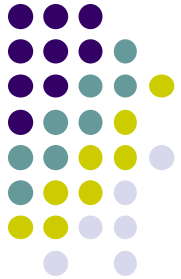
3.Performance Evaluation:

1. Model performance is typically assessed using cross-validation or a holdout validation set, similar to forward feature selection.
2. The evaluation metric depends on the specific problem type and requirements.

4.Stopping Criterion:

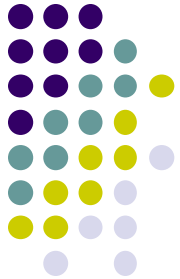
1. This process continues until the removal of additional features leads to a significant decrease in model performance.
2. Alternatively, the process can stop when a predefined number of features is reached.

Προς-τα-πίσω Απαλοιφή Backward Elimination



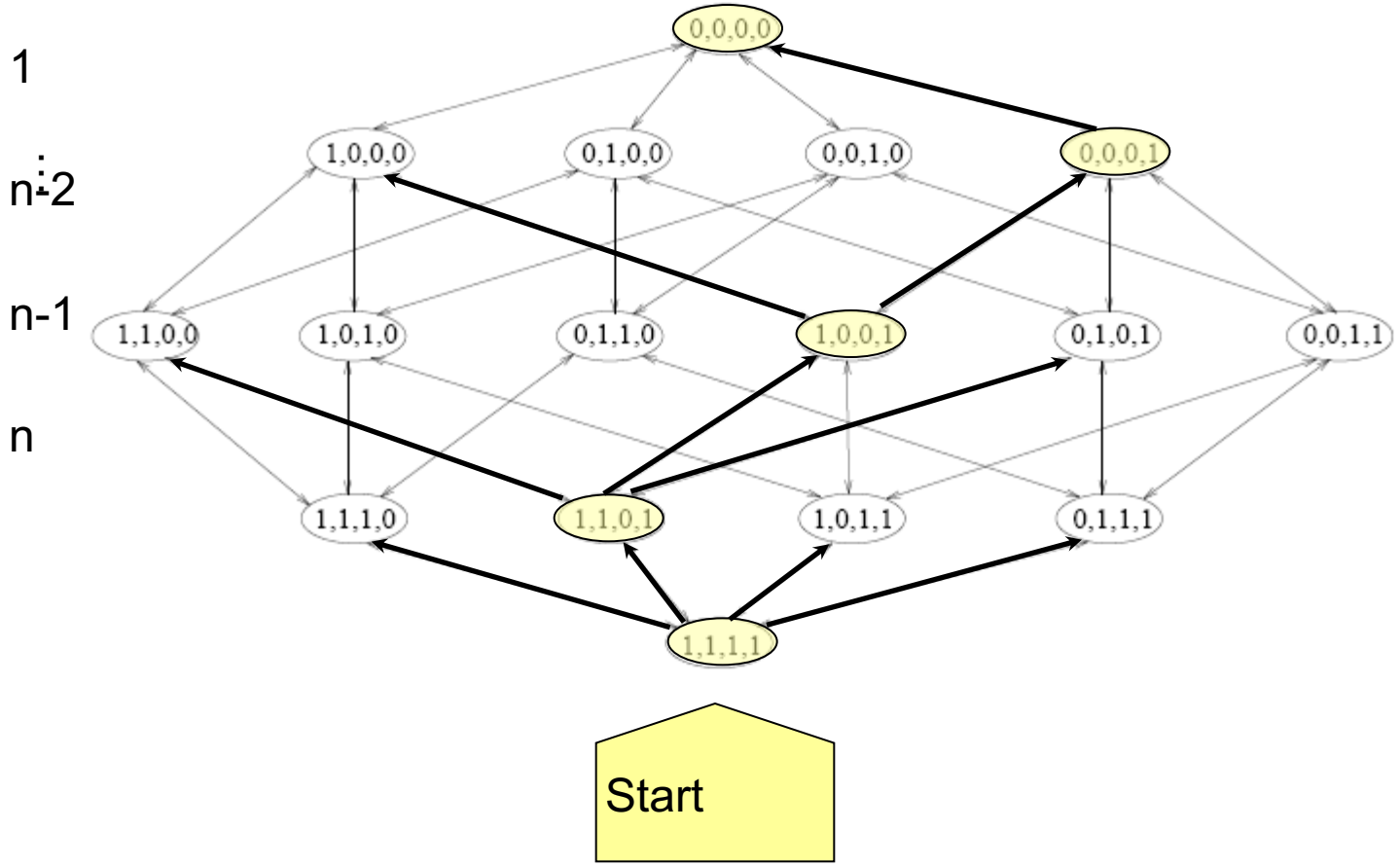
- Αρχικά λιγότερο αποδοτική από την forward μέθοδο, επειδή ξεκινά με όλα τα χαρακτηριστικά, αλλά γίνεται πιο αποτελεσματική καθώς εξαλείφονται τα χαρακτηριστικά.
- Καλύτερο για χειρισμό αλληλεπιδράσεων χαρακτηριστικών: Είναι πιο πιθανό να διατηρήσει σημαντικές αλληλεπιδράσεις χαρακτηριστικών σε σύγκριση με την forward μέθοδο, καθώς αξιολογεί την απόδοση με ένα μεγαλύτερο σύνολο χαρακτηριστικών αρχικά.
- Εξάρτηση μοντέλου: Η αποτελεσματικότητα της μεθόδου επηρεάζεται από την επιλογή του μοντέλου που χρησιμοποιείται για την αξιολόγηση.

Προς-τα-πίσω Απαλοιφή Backward Elimination

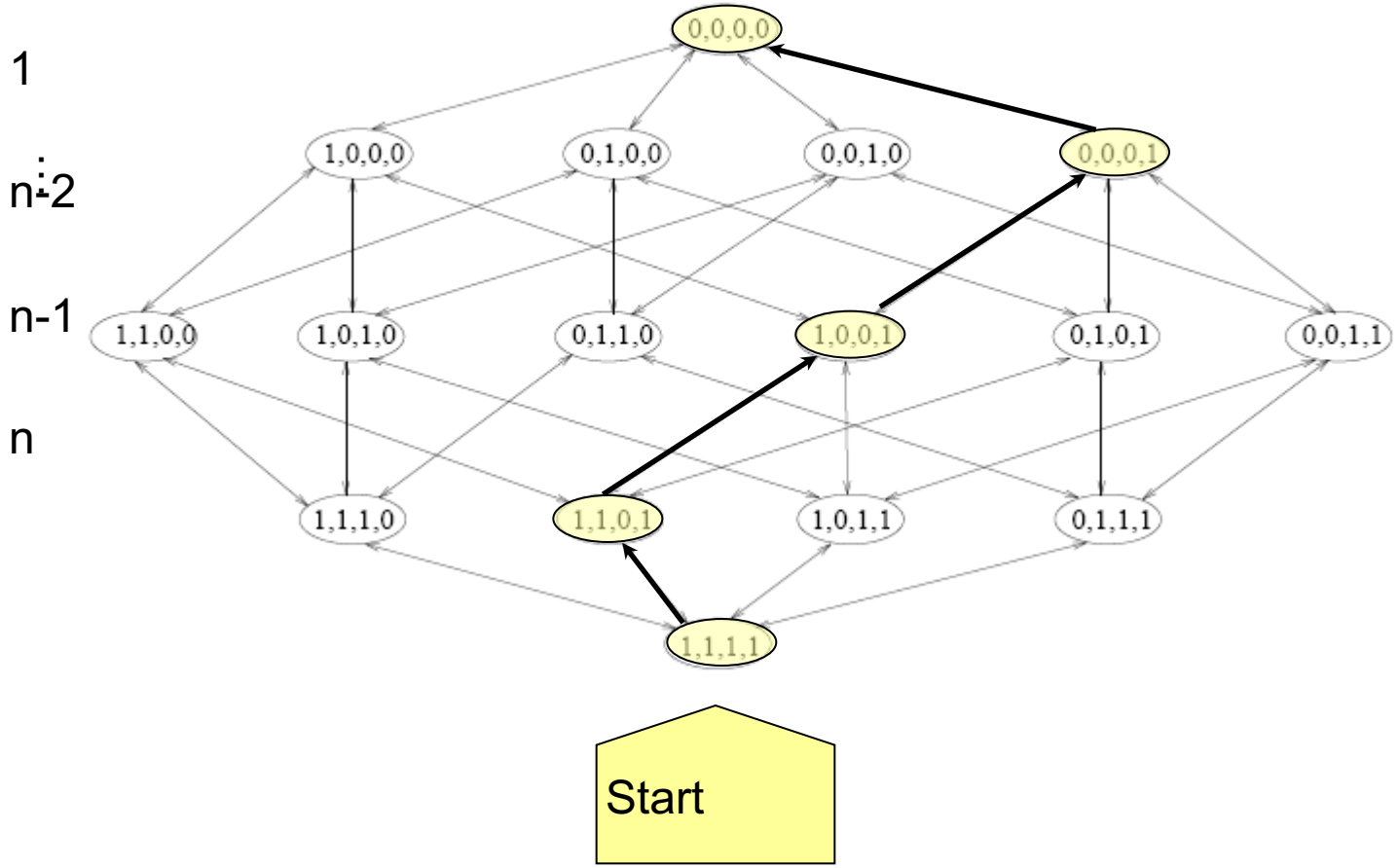


- Περιπτώσεις χρήσης:
  - Σύνολα δεδομένων με λιγότερα χαρακτηριστικά: Ιδιαίτερα χρήσιμο όταν ο αριθμός των χαρακτηριστικών δεν είναι υπερβολικά μεγάλος, καθώς η έναρξη με όλα τα χαρακτηριστικά μπορεί να είναι υπολογιστικά ακριβή.
  - Σύνθετα μοντέλα: Συχνά χρησιμοποιούνται σε σενάρια όπου το μοντέλο πρέπει να απλοποιηθεί για καλύτερη ερμηνεία ή για μείωση της υπερβολικής προσαρμογής

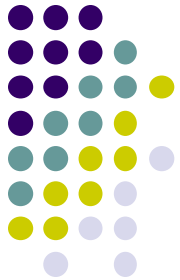
Προς-τα-πίσω Απαλοιφή
Backward Elimination (wrapper)



Προς-τα-πίσω Απαλοιφή
Backward Elimination (embedded)



L-Προς-τα-Εμπρός-R-Προς-τα Πίσω Επιλογή Χαρακτηριστικών (L-Forward R-Backward Feature Selection)

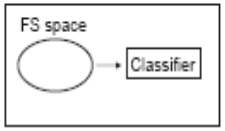
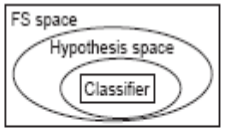
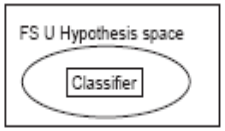


- L βήματα Προς-τα-Εμπρός Επιλογή ακολουθούνται από R βήματα Προς-τα-Πίσω Επιλογής

  - Αποφυγή του μειονεκτήματος της Προς-τα-Εμπρός Επιλογης κάποιο χαρακτηριστικό που επιλέχθηκε αρχικά γιατί έδωσε καλή απόδοση μεμονωμένα να μην συνδυάζεται καλά με τα χαρακτηριστικά που επιλέχθηκαν αργότερα
  - Αποφυγή του μειονεκτήματος της Προς-τα-Πίσω Επιλογης κάποιο χαρακτηριστικό που απαλείφθηκε αρχικά γιατί έδωσε κακή απόδοση μεμονωμένα να συνδυάζεται καλά με τα χαρακτηριστικά που επιλέχθηκαν αργότερα

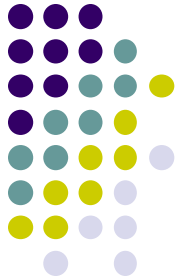
Feature Selection techniques in a nutshell

Table 1. A taxonomy of feature selection techniques. For each feature selection type, we highlight a set of characteristics which can guide the choice for a technique suited to the goals and resources of practitioners in the field.

	Model search	Advantages		Disadvantages	Examples
Filter		Univariate	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	Chi-square Euclidean distance t-test Information gain, Gain ratio [6]
		Multivariate	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation based feature selection (CFS) [45] Markov blanket filter (MBF) [62] Fast correlation based feature selection (FCBF) [136]
Wrapper		Deterministic	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) [60] Sequential backward elimination (SBE) [60] Plus q take-away r [33] Beam search [106]
		Randomized	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing [110] Genetic algorithms [50] Estimation of distribution algorithms [52]
Embedded		Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies		Classifier dependent selection	Decision trees Weighted naive Bayes [28] Feature selection using the weight vector of SVM [44, 125]

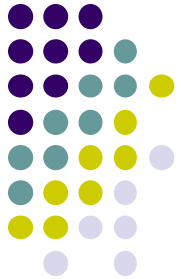
Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics Bioinformatics. 2007 Oct 1;23(19):2507-17

Εξαγωγή Χαρακτηριστικών



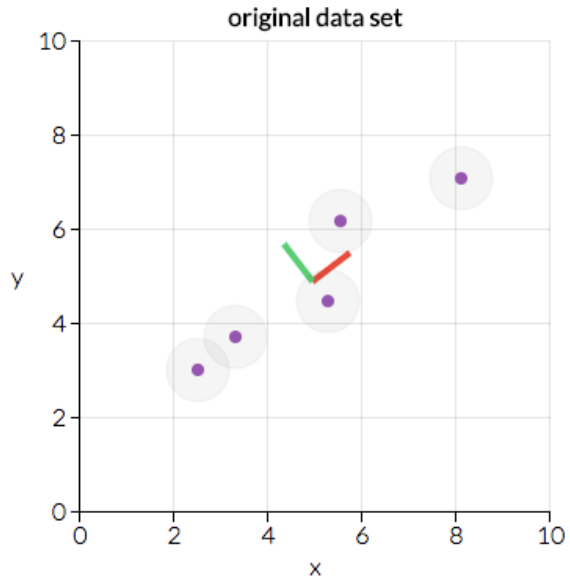
- Η πιο δημοφιλής τεχνική εξαγωγής χαρακτηριστικών είναι η μέθοδος Ανάλυσης Κυρίων Συνιστωσών (Principal Component Analysis - PCA).
- Αποτελεί μια μη επιβλεπόμενη τεχνική και ανήκει στην κατηγορία των μεθόδων γραμμικού μετασχηματισμού δεδομένων σε ένα χώρο χαμηλότερης διάστασης

Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)

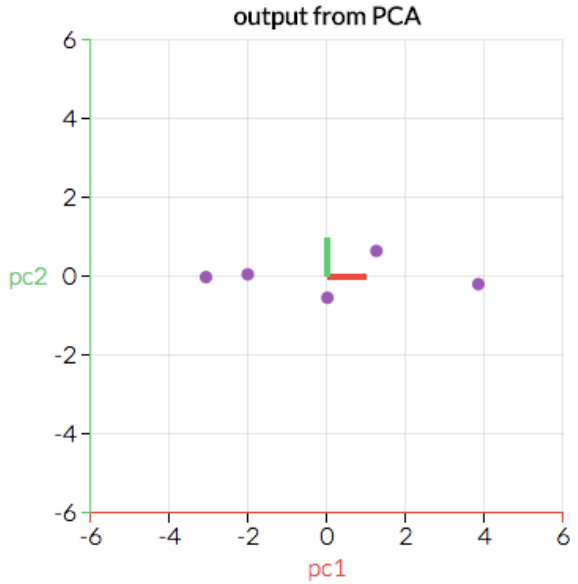


- Πότε χρησιμοποιούμε PCA
  - Θέλουμε να μειώσουμε τη διάσταση αλλά δεν ξέρουμε ποιες μεταβλητές να αφαιρέσουμε
  - Θέλουμε ανεξαρτησία των μεταβλητών
  - Δεν μας νοιάζει αν χάσουμε σε επεξηγηματικότητα

Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)



Αρχικά δεδομένα



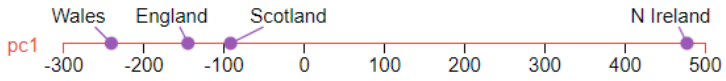
Μετασχηματισμός με PCA

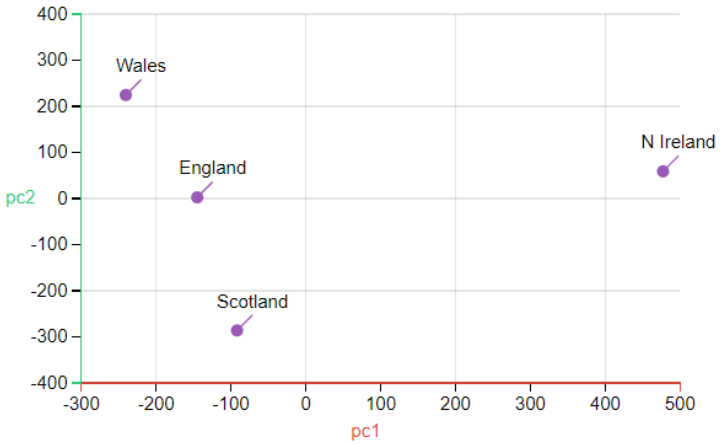
Άξονες: κατεύθυνση μέγιστης μεταβλητότητας

Πηγή: https://setosa.io/ev/principal-component-analysis/

Παράδειγμα: διατροφικές συνήθειες στο HB

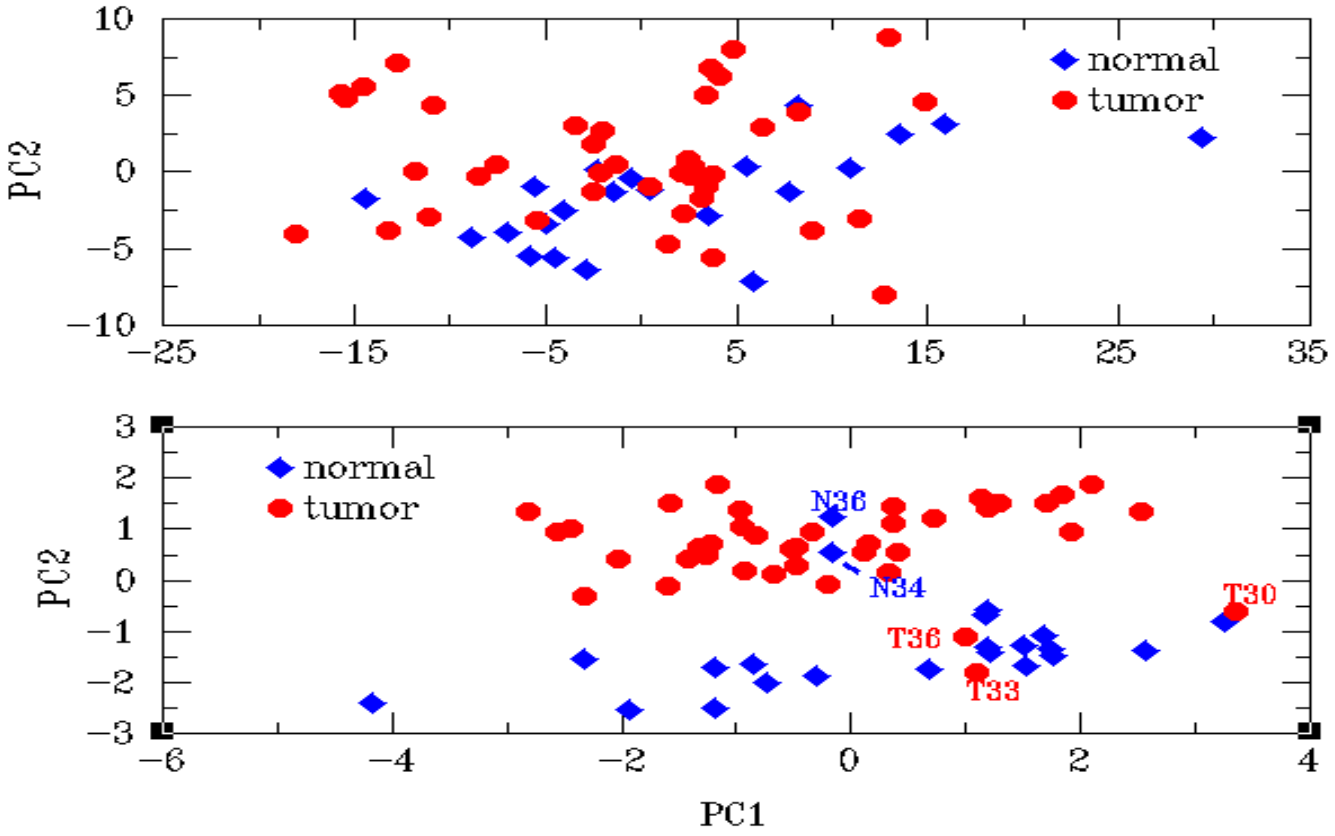
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175





Πηγή: https://setosa.io/ev/principal-component-analysis/

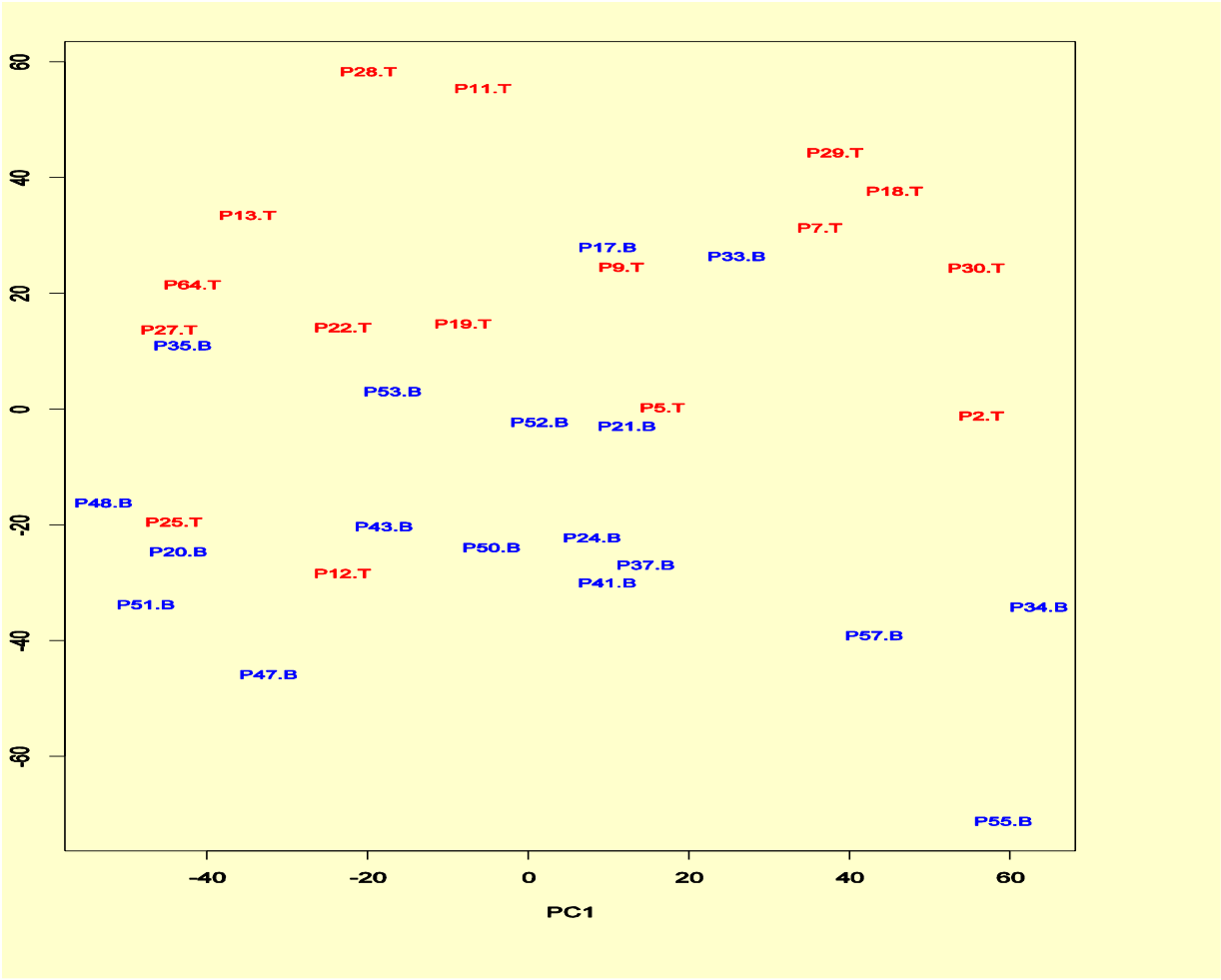
PCA



Εφαρμογή της PCA σε 2.000 γονίδια (πάνω) και στα 50 πιο σημαντικά γονίδια (κάτω) από colon tissue data
(http://dir.niehs.nih.gov/microarray/datamining/public_html/colon.html)

PCA σε όλα τα γονίδια
Leukemia data, precursor B and T

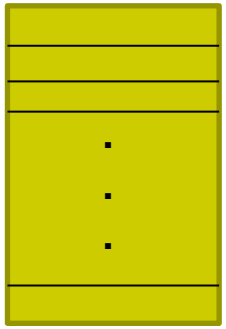
- 34 ασθενείς, 8973 διαστάσεις (γονίδια) μειώθηκαν στις 2



Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)

A1

$\mathbf{X}$

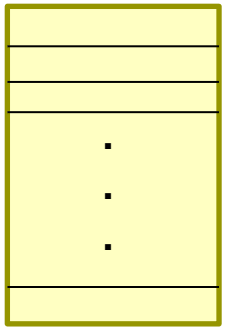


$(n \times p)$

(n: δείγματα, p: χαρακτηριστικά)

A2 Κανονικοποίηση, $(\mu,\sigma)=(0,1)$

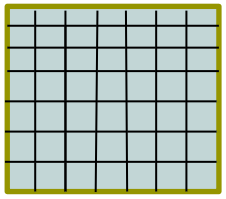
$$\mathbf{Z} = \frac{\mathbf{X} - \bar{\mathbf{X}}}{std(\mathbf{X})}$$



$(n \times p)$

A3 Πίνακας συμμεταβλητότητας

$\mathbf{Z}^T\mathbf{Z}$

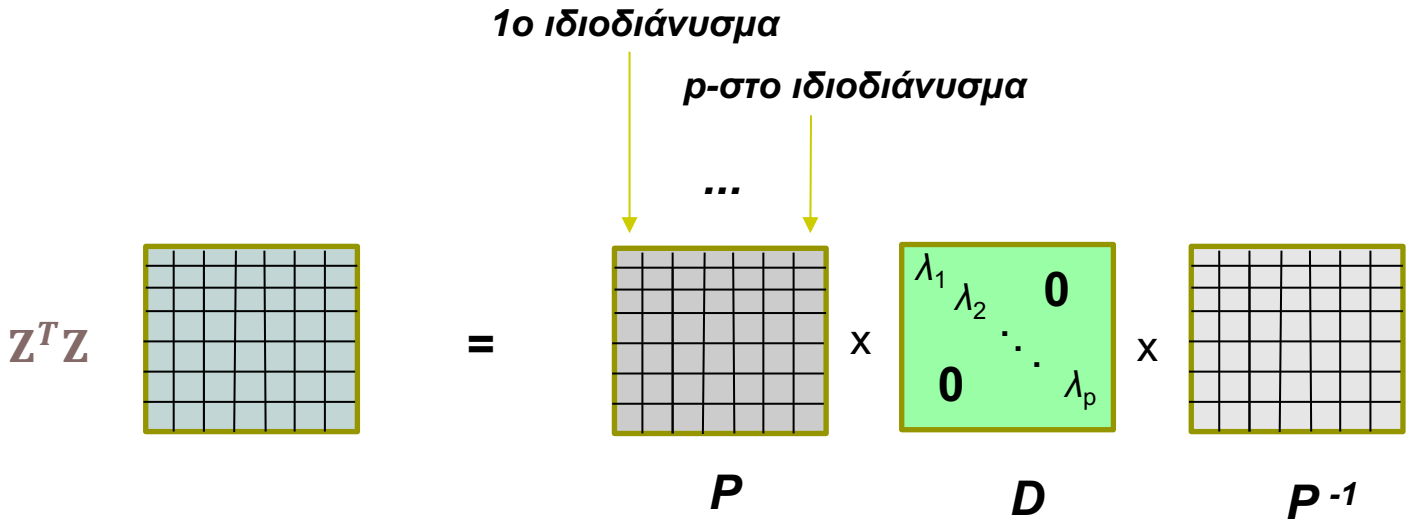


$(p \times p)$

Δείχνει πώς εξαρτώνται οι μεταβλητές μεταξύ τους (πολύ σημαντική πληροφορία).

Συμμετρικός, θετικά ημιορισμένος, άρα προχωρούμε στο επόμενο βήμα.

Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)



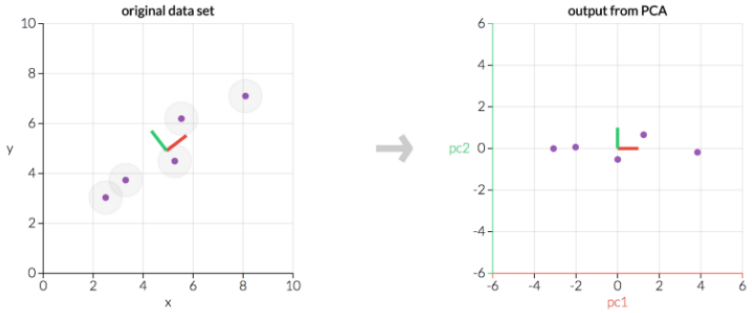
A4 Ανάλυση σε: $Z^TZ = PDP^{-1}$ (λογισμικό)

Όπου ***D*** = diag(λ_1, λ_2, … , λ_p), $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$

και ***P*** *(p*x*p)* ο αντίστοιχος πίνακας ιδιοδιανυσμάτων

Τα ιδιοδιανύσματα (στήλες) είναι ανεξάρτητα (ορθογώνια)

Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)



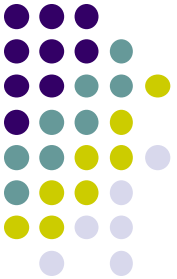
$\mathbf{Z}$ $\mathbf{Z}^* = \mathbf{ZP}$

B1 **Z*** μια μετασχηματισμένη εκδοχή του **Z**

Κάθε στήλη του **Z*** είναι γραμμικά ανεξάρτητη λόγω του ότι ο **P** έχει γραμμικά ανεξάρτητες στήλες

Μέχρι στιγμής δεν κάναμε μείωση διάστασης

Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)



A5 Πόσα από τα *p* χαρακτηριστικά (*p'*) θα κρατήσουμε τελικά χωρίς να χάσουμε χρήσιμη πληροφορία;

1. Αφελής τρόπος: διάλεξε *p'* αυθαίρετα
2. Πιο σωστά: Διάλεξε ποσοστό μεταβλητότητας που θέλουμε να κρατήσουμε χρησιμοποιώντας τις ιδιοτιμές:

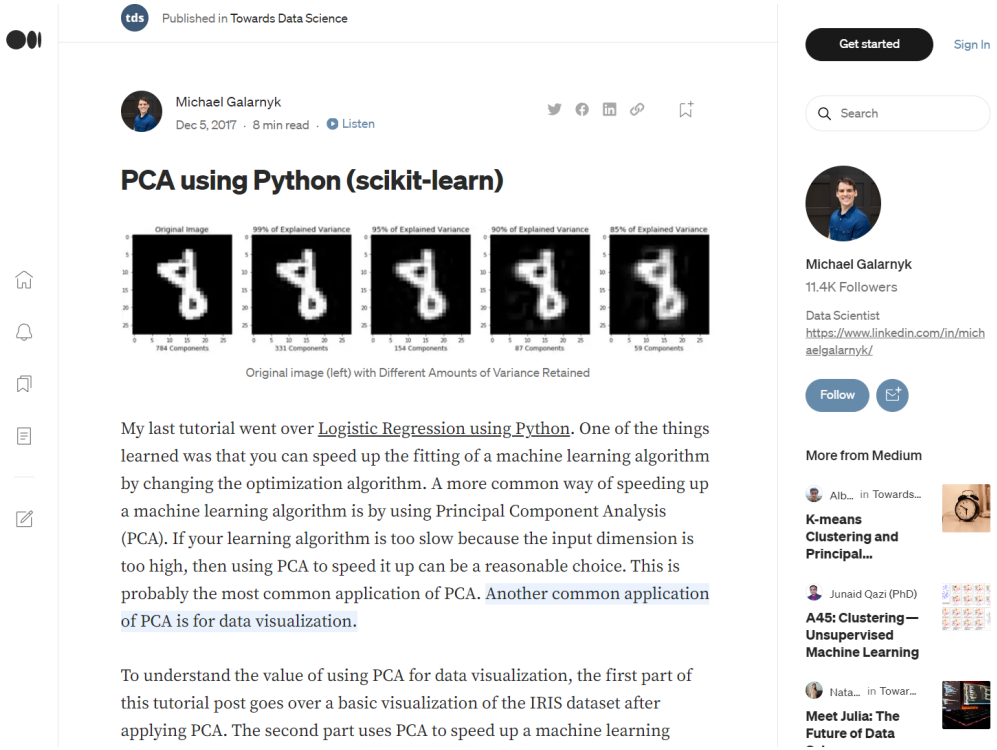
Διάλεξε *K* χαρακτηριστικά έτσι ώστε η παρακάτω ποσότητα να είναι μεγαλύτερη από το ζητούμενο κατώφλι μεταβλητότητας *T* που θέλουμε να καλύψουμε:

$$\sum_{i=1}^{K} \lambda_i \; / \sum_{i=1}^{p} \lambda_i \; > T$$

B2 Αφαιρούμε τα αντίστοιχα ιδιοδιανύσματα (στήλες) από τον **P** για να πάρουμε τον $\mathbf{P}^r$ (pxp') και κάνουμε τον μετασχηματισμό:

$\mathbf{Z}^{*r} = \mathbf{Z}\,\mathbf{P}^r$ $\quad$ $\mathbf{Z}^{*r}$ (nxp')

Μέθοδος Ανάλυσης Κυρίων Συνιστωσών (PCA)



Published in Towards Data Science

Get started Sign In

Michael Galarnyk
Dec 5, 2017 · 8 min read · Listen

Search

PCA using Python (scikit-learn)

Original image (left) with Different Amounts of Variance Retained

My last tutorial went over Logistic Regression using Python. One of the things learned was that you can speed up the fitting of a machine learning algorithm by changing the optimization algorithm. A more common way of speeding up a machine learning algorithm is by using Principal Component Analysis (PCA). If your learning algorithm is too slow because the input dimension is too high, then using PCA to speed it up can be a reasonable choice. This is probably the most common application of PCA. Another common application of PCA is for data visualization.

To understand the value of using PCA for data visualization, the first part of this tutorial post goes over a basic visualization of the IRIS dataset after applying PCA. The second part uses PCA to speed up a machine learning

Michael Galarnyk
11.4K Followers

Data Scientist
https://www.linkedin.com/in/michaelgalarnyk/

Follow

More from Medium

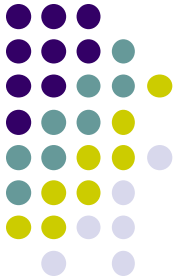
Alb... in Towards...
K-means Clustering and Principal...

Junaid Qazi (PhD)
A45: Clustering — Unsupervised Machine Learning

Nata... in Towar...
Meet Julia: The Future of Data

Παράδειγμα κώδικα:
https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

Άλλες Τεχνικές Εξαγωγής Χαρακτηριστικών



- Linear Discriminant Analysis (LDA)
- Independent Component Analysis (ICA)
- Partial Least Square Analysis (PLS)
- ….