ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό

Singular Value Decomposition και Latent Semantic Analysis

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

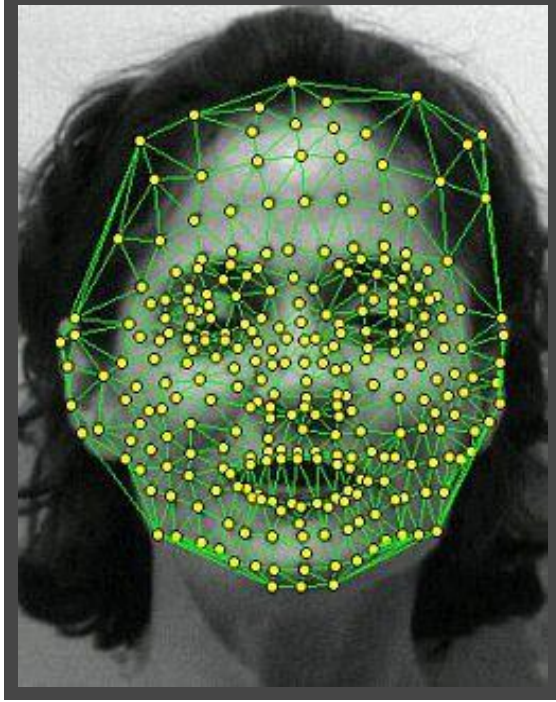
Άδειες Χρήσης

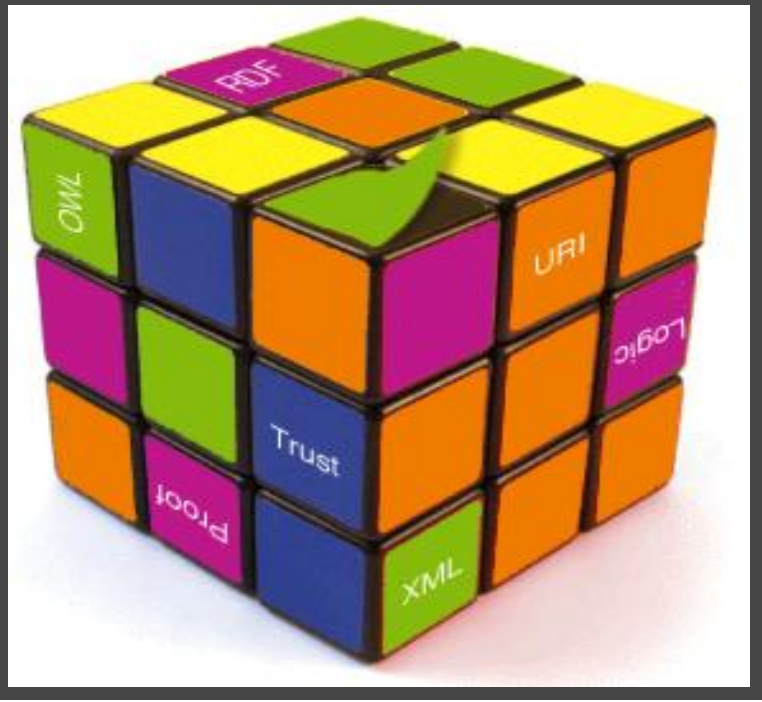
- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης





Παράρτημα 1: Singular Value Decomposition και Latent Semantic Analysis

Εισαγωγή

- **Βασική ιδέα**
  - Αναπαράσταση των δεδομένων σε λιγότερες διαστάσεις
  - Μετατροπή συσχετιζόμενων μεταβλητών σε ένα σύνολο μη-συσχετιζόμενων που αποδίδουν καλύτερα τις συσχετίσεις των δεδομένων
  - Αναγνωρίζουν και ταξινομούν τις διαστάσεις με βάση τις οποίες τα δεδομένα παρουσιάζουν τη μεγαλύτερη ποικιλία
    - Προέκυψαν πολύ ενδιαφέροντα αποτελέσματα στην πορεία

Γραμμική Άλγεβρα

- Πίνακας
- Διάνυσμα

$$\vec{v} = [4, 11, 8, 10]$$

	Doc 1	Doc 2	Doc 3
abbey	2	3	5
spinning	1	0	1
soil	3	4	1
stunned	2	1	3
wrath	1	1	4

- Εσ. Γινόμενο

if $\vec{x} = [1, 6, 7, 4]$ and $\vec{y} = [3, 2, 8, 3]$, then

$$\vec{x} \cdot \vec{y} = 1(3) + 6(2) + 7(8) + 3(4) = 83$$

Γραμμική Άλγεβρα

- Ορθογωνικότητα
  - Εσ. Γινόμενο=0 $[2,1,-2,4]\cdot[3,-6,4,2]=2(3)+1(-6)-2(4)+4(2)=0$
- Μοναδιαίο Διάνυσμα

if $\vec{v}=[2,4,1,2]$, then

$$|\vec{v}|=\sqrt{2^2+4^2+1^2+2^2}=\sqrt{25}=5$$

Then $\vec{u}=[2/5,4/5,1/5,1/5]$ is a normal vector because

$$|\vec{u}|=\sqrt{(2/5)^2+(4/5)^2+(1/5)^2+(2/5)^2}=\sqrt{25/25}=1$$

Γραμμική Άλγεβρα

- Ορθοκανονικά Διανύσματα
  - Μοναδιαία και ορθογώνια μεταξύ τους
  - Τα u και υ είναι τέτοια διανύσματα γιατί….

$$\vec{u} = [2/5, 1/5, -2/5, 4/5]$$

$$\vec{v} = [3/\sqrt{65}, -6/\sqrt{65}, 4/\sqrt{65}, 2/\sqrt{65}]$$

$$|\vec{u}| = \sqrt{(2/5)^2 + (1/5)^2 + (-2/5)^2 + (4/5)^2} = 1$$

$$|\vec{v}| = \sqrt{(3/\sqrt{65}) + (-6/\sqrt{65}) + (4/\sqrt{65}) + (2/\sqrt{65})} = 1$$

$$\vec{u} \cdot \vec{v} = \frac{6}{5\sqrt{65}} - \frac{6}{5\sqrt{65}} - \frac{8}{5\sqrt{65}} + \frac{8}{5\sqrt{65}} = 0$$

Γραμμική Άλγεβρα

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & 0 \\ 2 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

- Μέθοδος Gram-Schmidt
  - Πρώτα normalize την πρώτη στήλη
  - Μετά
    - $\vec{w_2} = \vec{v_2} - \vec{u_1} \cdot \vec{v_2} * \vec{u_1}$

$\vec{v_1} = [1, 0, 2, 1]$:

$$\vec{u_1} = [\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}]$$

$$[2,2,3,1] - [\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}] \cdot [2,2,3,1] * [\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}] = [\frac{1}{2}, 2, 0, \frac{-1}{2}]$$

Γραμμική Άλγεβρα

- Μετά normalize το w2
  - $\vec{u_2} = [\frac{\sqrt{2}}{6}, \frac{2\sqrt{2}}{3}, 0, \frac{-\sqrt{2}}{6}]$
- Για το w3 ισχύει ότι...

$$\vec{w_3} = \vec{v_2} - \vec{u_1} \cdot \vec{v_3} * \vec{u_1} - \vec{u_2} \cdot \vec{v_3} * \vec{u_2} = [\frac{4}{9}, \frac{-2}{9}, 0, \frac{-4}{9}]$$

- Normalize το w3
  - $\vec{u_3} = [\frac{2}{3}, \frac{-1}{3}, 0, \frac{-2}{3}]$

- Τελικά
  - $$A = \begin{bmatrix} \frac{\sqrt{6}}{6} & \frac{\sqrt{2}}{6} & \frac{2}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{-1}{3} \\ \frac{\sqrt{6}}{3} & 0 & 0 \\ \frac{\sqrt{6}}{6} & \frac{-\sqrt{2}}{6} & \frac{-2}{3} \end{bmatrix}$$

Γραμμική Άλγεβρα

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3/5 & -4/5 \\ 0 & 4/5 & 3/5 \end{bmatrix}$$

- Ορθογώνιοι Πίνακες
  - Όταν $AA^T = A^T A = I.$
  - Ο Α είναι ορθογώνιος επειδή:
    - $A^T A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3/5 & -4/5 \\ 0 & 4/5 & 3/5 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3/5 & 4/5 \\ 0 & -4/5 & 3/5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Γραμμική Άλγεβρα

- Ορίζουσα
  - Μια συνάρτηση που μειώνει ένα τετραγωνικό πίνακα σε ένα αριθμό
- Det(A)
- |A|

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

$$A = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

$$|A| = \begin{vmatrix} 4 & 1 \\ 1 & 2 \end{vmatrix} = 4(2) - 1(1) = 7.$$

Γραμμική Άλγεβρα

- Ορίζουσα
  - Μέθοδος expansion by row

$$\begin{vmatrix} -1 & 4 & 3 \\ 2 & 6 & 4 \\ 3 & -2 & 8 \end{vmatrix} = (-1)\begin{vmatrix} 6 & 4 \\ -2 & 8 \end{vmatrix} - (4)\begin{vmatrix} 2 & 4 \\ 3 & 8 \end{vmatrix} + (3)\begin{vmatrix} 2 & 6 \\ 3 & -2 \end{vmatrix} =$$

$$-1(6 \cdot 8 - 4 \cdot -2) - 4(2 \cdot 8 - 4 \cdot 3) + 3(2 \cdot -2 - 3 \cdot 6) =$$

$$-56 - 16 - 66 = -138$$

$\lambda = 1$ is $[1, -1]$.

$\lambda = 3$ is $[1, 1]$

Γραμμική Άλγεβρα

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

- Ιδιοδιανύσματα και Ιδιοτιμές

$$A\vec{v} = \lambda\vec{v}$$

$$A\vec{v} = \lambda\vec{v} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

πράξεις

$$2x_1 + x_2 = \lambda x_1$$

$$x_1 + 2x_2 = \lambda x_2$$

$$(2-\lambda)x_1 + x_2 = 0$$

$$x_1 + (2-\lambda)x_2 = 0$$

Συνθήκη: ορίζουσα=0

$$\begin{vmatrix} (2-\lambda) & 1 \\ 1 & (2-\lambda) \end{vmatrix} = 0$$

$$(2-\lambda)(2-\lambda) - 1 \cdot 1 = 0$$

$$\lambda^2 - 4\lambda + 3 = 0$$

$$(\lambda - 3)(\lambda - 1) = 0$$

Singular Value Decomposition

- **Βασική ιδέα**
  - Αναπαράσταση των δεδομένων σε λιγότερες διαστάσεις
  - Μετατροπή συσχετιζόμενων μεταβλητών σε ένα σύνολο μη-συσχετιζόμενων που αποδίδουν καλύτερα τις συσχετίσεις των δεδομένων
  - Αναγνωρίζουν και ταξινομούν τις διαστάσεις με βάση τις οποίες τα δεδομένα παρουσιάζουν τη μεγαλύτερη ποικιλία

Full Singular Value Decomposition

- $A_{mn} = U_{mm} S_{mn} V_{nn}^T$
  - $U^T U = I, V^T V = I$
  - οι στήλες του U είναι τα ορθογώνια ιδιοδιανύσματα του AA^T
  - οι στήλες του V είναι τα ορθογώνια ιδιοδιανύσματα του A^TA
  - Ο S είναι διαγώνιος πίνακας με την τετρ. Ρίζα των ιδιοτιμών του U με φθίνουσα σειρά

Παράδειγμα Full SVD

- Έστω πίνακας Α,

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

- Ο ανάστροφος…

$$A^T = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}$$

- Υπολογισμός του U:

$$AA^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$$

$$\begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- Ιδιοτιμές και ιδιοδιανύσματα

$$\begin{vmatrix} (11-\lambda) & 1 \\ 1 & (11-\lambda) \end{vmatrix} = 0$$

$$(11-\lambda)(11-\lambda) - 1 \cdot 1 = 0$$

$$(\lambda - 10)(\lambda - 12) = 0$$

$$\lambda = 10, \lambda = 12$$

$$(11-\lambda)x_1 + x_2 = 0$$

$$x_1 + (11-\lambda)x_2 = 0$$

Παράδειγμα Full SVD

- Για λ=10 → $(11-10)x_1 + x_2 = 0$

  $x_1 = -x_2$

- Για λ=12 → $(11-12)x_1 + x_2 = 0$

  $x_1 = x_2$

- Έστω x1=1 άρα

λ=12 λ=10

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Παράδειγμα Full SVD

- Μετά Gram-Schmidt $\vec{u_1} = \frac{\vec{v_1}}{|\vec{v_1}|} = \frac{[1,1]}{\sqrt{1^2+1^2}} = \frac{[1,1]}{\sqrt{2}} = [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$
  - $$\vec{w_2} = \vec{v_2} - \vec{u_1} \cdot \vec{v_2} * \vec{u_1} =$$
    $$[1,-1] - [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] \cdot [1,-1] * [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] =$$
    $$[1,-1] - 0 * [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}] = [1,-1] - [0,0] = [1,-1]$$

$$\vec{u_2} = \frac{\vec{w_2}}{|\vec{w_2}|} = [\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}]$$

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

Παράδειγμα Full SVD

$$V^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

- Υπολογισμός του V:
- $A^TA = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$
- Ιδιοτιμές κτλ....

$$\begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$(10-\lambda)x_1 + 2x_3 = 0$$
$$(10-\lambda)x_2 + 4x_3 = 0$$
$$2x_1 + 4x_2 + (2-\lambda)x_3 = 0$$

$$\begin{vmatrix} (10-\lambda) & 0 & 2 \\ 0 & (10-\lambda) & 4 \\ 2 & 4 & (2-\lambda) \end{vmatrix} = 0$$

Πράξεις....

λ=12 λ=10 λ=0

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & -1 & 2 \\ 1 & 0 & 5 \end{bmatrix}$$

Gram-schmidt

Παράδειγμα Full SVD

$$A_{mn} = U_{mm} S_{mn} V^T_{nn} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

Latent Semantic Indexing

- Για κείμενα-έγγραφα όταν SVD όχι full
  - Δεν παίρνουμε όλες τις συνιστώσες
- Παράδειγμα
  - The Neatest Little Guide to Stock Market Investing
  - Investing For Dummies, 4th Edition
  - The Little Book of Common Sense Investing: The Only Way to Guarantee Your Fair Share of Stock Market Returns
  - The Little Book of Value Investing
  - Value Investing: From Graham to Buffett and Beyond
  - Rich Dad's Guide to Investing: What the Rich Invest in, That the Poor and the Middle Class Do Not!
  - Investing in Real Estate, 5th Edition
  - Stock Investing For Dummies
  - Rich Dad's Advisors: The ABC's of Real Estate Investing: The Secrets of Finding Hidden Profits Most Investors Miss

LSI-παράδειγμα

1η διάσταση: εμφάνιση συχνότητας λέξεων

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

SVD

book	0.15	-0.27	0.04
dads	0.24	0.38	-0.09
dummies	0.13	-0.17	0.07
estate	0.18	0.19	0.45
guide	0.22	0.09	-0.46
investing	0.74	-0.21	0.21
market	0.18	-0.30	-0.28
real	0.18	0.19	0.45
rich	0.36	0.59	-0.34
stock	0.25	-0.42	-0.28
value	0.12	-0.14	0.23

*

3.91	0	0
0	2.61	0
0	0	2.00

*

T1	T2	T3	T4	T5	T6	T7	T8	T9
0.35	0.22	0.34	0.26	0.22	0.49	0.28	0.29	0.44
-0.32	-0.15	-0.46	-0.24	-0.14	0.55	0.07	-0.31	0.44
-0.41	0.14	-0.16	0.25	0.22	-0.51	0.55	0.00	0.34

1η διάσταση: εμφάνιση μήκους κειμένου

LSI-παράδειγμα

1) The Neatest Little Guide to Stock Market Investing
2) Investing For Dummies, 4th Edition
3) The Little Book of Common Sense Investing: The Only Way to Guarantee Your Fair Share of Stock Market Returns
4) The Little Book of Value Investing
5) Value Investing: From Graham to Buffett and Beyond
6) Rich Dad's Guide to Investing: What the Rich Invest in, That the Poor and the Middle Class Do Not!
7) Investing in Real Estate, 5th Edition
8) Stock Investing For Dummies
9) Rich Dad's Advisors: The ABC's of Real Estate Investing: The Secrets of Finding Hidden Profits Most Investors Miss