Αριθμητικές μέθοδοι στην στατιστική

Εισαγωγή

Α. Ν. Γιαννακόπουλος

Ο.Π.Α

Εαρινό Εξάμηνο 2026

1. Εισαγωγή

2. Γιατί;

Εισαγωγή

- Το μάθημα αυτό επικεντρώνεται σε αριθμητικές μεθόδους οι οποίες είναι απαραίτητες στην επιστήμη της στατιστικής, της ανάλυσης δεδομένων και της μηχανικής μάθησης
- Τα περισσότερα μοντέλα που χρησιμοποιούνται καθημερινά και που χρησιμοποιείτε και εσείς, όπως π.χ.
  - ο αλγόριθμός του Netflix ή
  - οι προτάσεις για videos στο Youtube, ή
  - ο τρόπος που ψάχνετε στο Google
  - η ιατρική απεικόνιση
  - η συμπίεση δεδομένων π.χ. videos ή ταινιών στις διάφορες ψηφιακές πλατφόρμες,
  - η ιατρική διάγνωση,
  - οι αλγόριθμοι χορήγησης δανείων απο τις τράπεζες, αλλά και
  - ένα σωρό άλλες εφαρμογές

  έχουν στο υπόβαθρο τους καποια μεθοδολογία η οποία καταλήγει σε ένα πρόβλημα αριθμητικής ανάλυσης και ένα αλγοριθμικό πλαίσιο η ορθή υλοποίηση του οποίου θα μας οδηγήσει στο σωστό αποτέλεσμα

Διαφορετικά μοντέλα μπορεί να χρησιμοποιούν την ίδια μεθοδολογία οι οποία βασίζεται στις ιδιες αριθμητικές μεθόδους οι οποίες εν γένει χωρίζονται σε 3 διαφορετικές κατηγορίες

- Μέθοδοι αριθμητικής γραμμικής άλγεβρας – Η βάση για οποιαδήποτε αριθμητική μέθοδο αλλά και κατηγορίες μεθόδων απο μόνες τους πολύ σημαντικές
  - Μείωση διαστατικότητας (SVD, PCA, Laplacian eigenmaps, Manifold learning κ.α.)
  - Ανάλυση εικόνας
  - και άλλα ...
- Μέθοδοι αριθμητικής βελτιστοποίησης – Τα περισσότερα μοντέλα μηχανικής μάθησης (γραμμική ή μη γραμμική παλινδρόμηση, ridge regression, Lasso, Support Vector Machines κ.α. καταλήγουν σε προβλήματα ελαχιστοποίησης κάποιας συνάρτησης σφάλματος
  - Gradient descent, stochastic gradient descent, Newton and quasinewton methods etc
  - Μέθοδοι δύισμού

- Προσέγγιση συναρτήσεων ή συναρτησιακών δεδομένων – Approximation theory
  - Reproducing Kernel Hilbert Spaces
  - Splines

Σκοπός του μαθήματος είναι μια hands on εισαγωγή στις θεμελιώδεις αριθμητικές μεθόδους που χρειαζόμαστε στην στατιστική και μηχανική μάθηση στις 3 παραπάνω κατηγορίες με σκόπο και την υλοποίηση τους και την εφαρμογή τους.

Γιατί;

Τα περισσότερα αν όχι όλα τα προβλήματα στην στατιστική, την επιστημη δεδομένων και τη μηχανική μάθηση ανάγονται σε προβλήματα αριθμητικής ανάλυσης κυρίως βελτιστοποίησης.

Η επίλυση αυτών των προβλημάτων είναι απαραίτητη για την επίλυση του αρχικού μας προβλήματος.

Τα προβλήματα βελτιστοποίησης στα οποία ανάγονται τα αρχικά μας προβλήματα ειναι κατα κανόνα

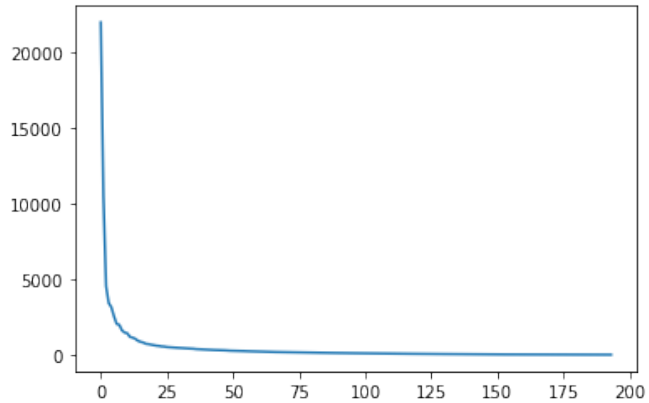
- Υψηλών διαστάσεων
- Δεν επιδέχονται αναλυτικής λύσης και χρειάζονται ειδικές αριθμητικές μεθόδους για την λύση τους.
- Λόγω της πολυπλοκότητας τους και των ιδιαιτεροτήτων τους δεν υπάρχει μια one size fits all τύπου μαύρο κουτί μέθοδος που μπορει να χρησιμοποιηθεί σε οποιοδήποτε πρόβλημα.
- Χρειάζεται ένας καλός συνδυασμός αναλυτικών και αριθμητικών μεθόδων, πολλές φορές taylor made για συγκεκριμένη κατηγορία προβλημάτων.

Διαδικαστικά

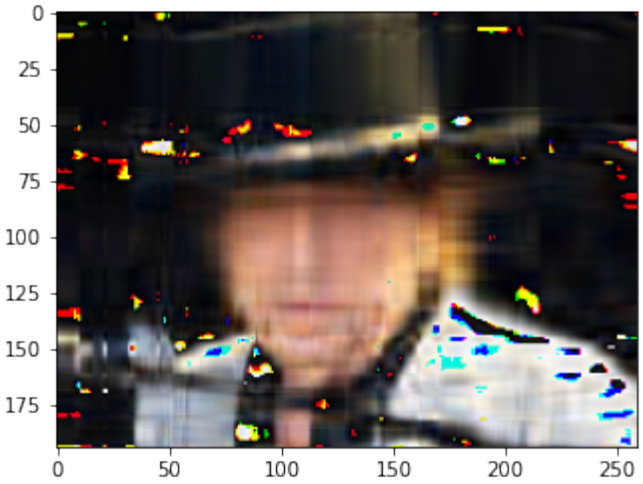
- 4 ωρες την εβδομάδα
  - Διαλέξεις
  - Υπολογιστικές εφαρμογές
- Αξιολόγηση: Ενδιάμεσες και απαλλακτικές εργασίες

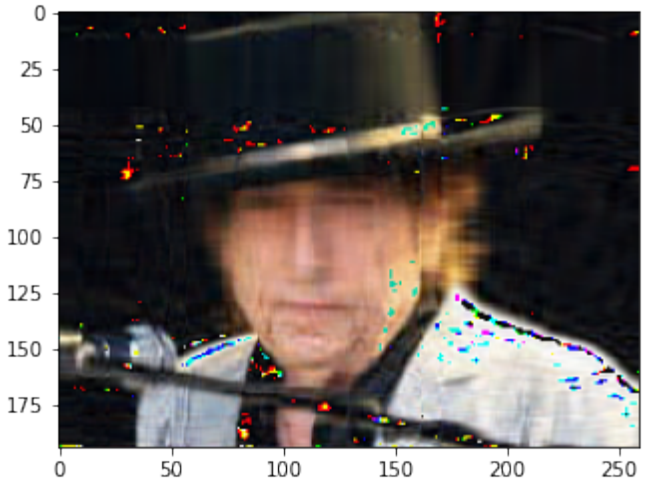
Ανάλυση εικόνας

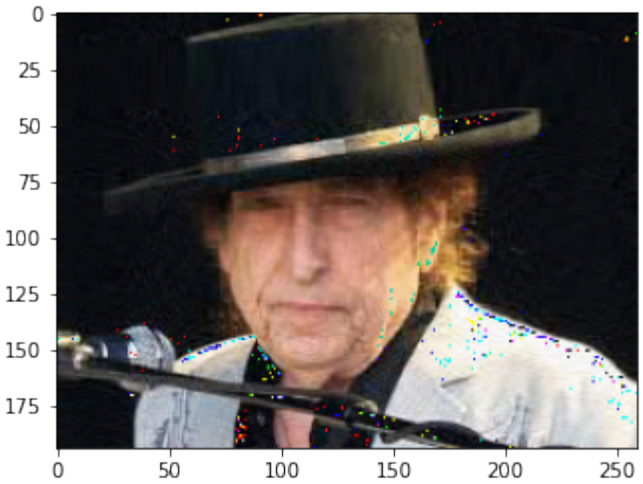
Μια εικόνα μπορεί να περιγραφεί σαν ένας πίνακας



Μια εικόνα 256×256 pixels είναι πχ ενας πινακας $\mathbb{R}^{256\times256}$.


0
25
50
75
100
125
150
175
0
50
100
150
200
250


0
25
50
75
100
125
150
175
0
50
100
150
200
250


0
25
50
75
100
125
150
175
0
50
100
150
200
250

10, 20, 50

Τυπικό πρόβλημα 1: Επιλογή στατιστικού μοντέλου με μεγιστοποίηση πιθανοφάνειας.

Έστω πως έχουμε ένα δείγμα απο παρατηρήσεις απο μια τυχαια μεταβλητή X με τιμές στον $\mathbb{R}^m$,

$$S := \{X_1, \cdots, X_n\} \subset \mathbb{R}^m.$$

Θεωρούμε ότι οι παρατηρήσεις προέρχονται απο μια παραμετρική οικογένεια κατανομών $f(\cdot, \theta)$ όπου $\theta = (\theta_1, \cdots, \theta_d)$, υπο την έννοια οτι

$$\begin{aligned} P(X \in [x, x + dx]) &= P(X \in [x_1, x_1 + dx_1] \times \cdots [x_m, x_m + dx_m) \\ &\simeq f(x; \theta)dx = f(x_1, \cdots, x_m; \theta)dx_1 \cdots dx_m \end{aligned}$$

Λόγω της ανεξαρτησίας του δείγματος η πιθανότητα να παρατηρήσουμε το S είναι ανάλογη της πιθανοφάνειας

$$L(\theta; X_1, \cdots, X_n;) = f(X_1; \theta) \cdots f(X_n; \theta)$$

την οποία και εκλαμβάνουμε σαν μια συνάρτηση της παραμετρου $\theta \in \mathbb{R}^d$.

Αν δεν γνωρίζουμε για ποια τιμή της παραμετρου θ η παραπάνω οικογένεια περιγράφει καλύτερα το δείγμα μπορούμε να επιλέξουμε το $\theta = \theta^*$ έτσι ώστε να μεγιστοποιείται η πιθανοφάνεια του δείγματος

$$\max_{\theta\in\mathbb{R}^d} L(\theta; X_1, \cdots, X_n) \text{ δηλ. } \theta^* \in \arg\max_{\theta\in\mathbb{R}^d} L(\theta; X_1, \cdots, X_n))$$

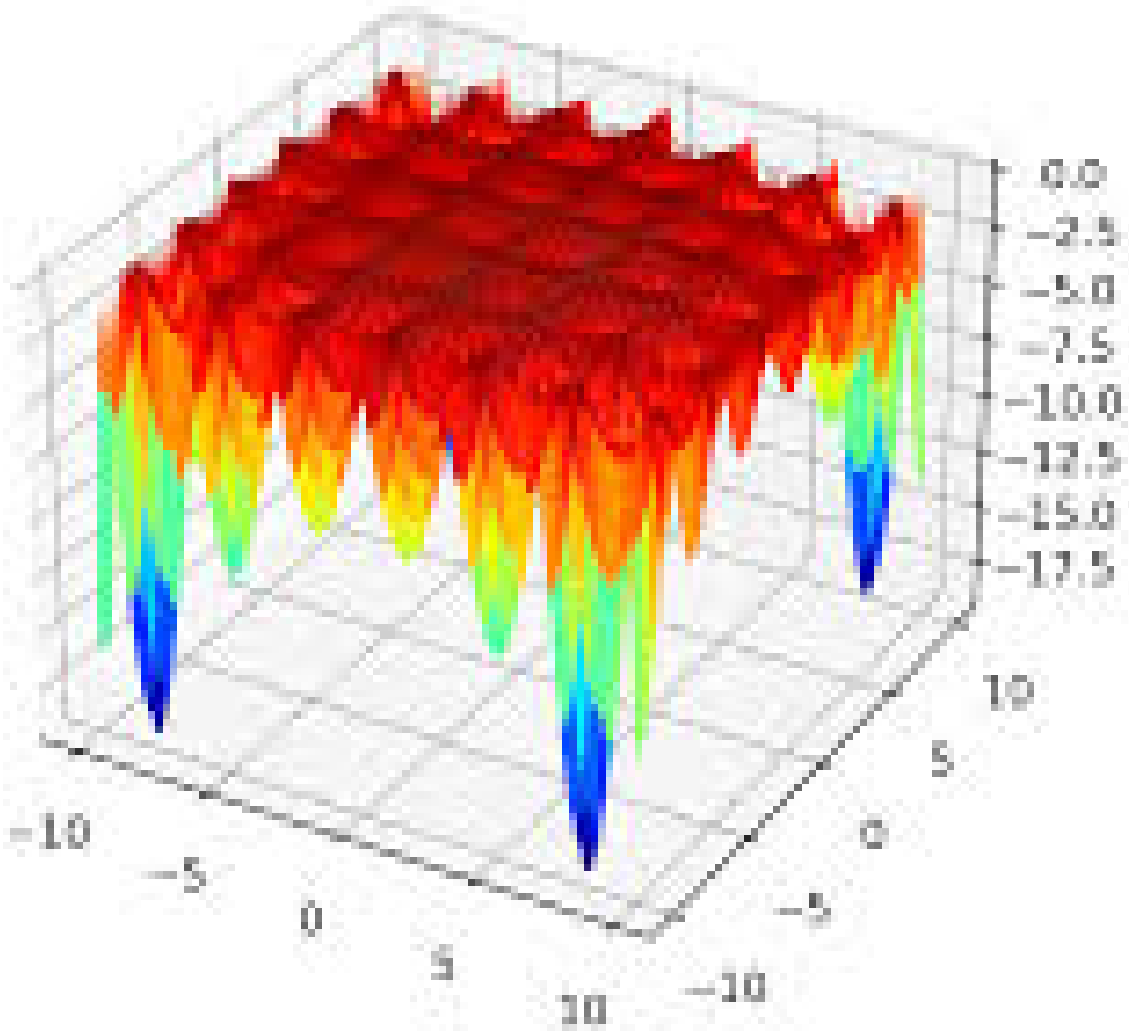
Παίρνουμε την συναρτηση

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; X_1, \cdots, X_n) := -\ln(L(\theta; X_1, \cdots, X_n)),$$

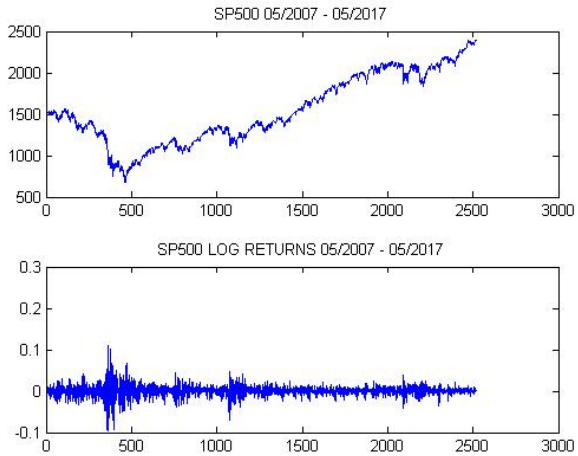
η οποία λόγω μονοτονίας μετατρέπει το πρόβλημα του μεγίστου σε πρόβλημα ελαχιστοποίησης, και το κάνει απο πολλαπλασιαστικό σε προσθετικό, δηλ. στο πρόβλημα

$$\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta; X_1, \cdots, X_n) \text{ δηλ. } \theta^* \in \arg\min_{\theta\in\mathbb{R}^d} \mathcal{L}(\theta; X_1, \cdots, X_n)$$

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; X_1, \cdots, X_n) = -\sum_{i=1}^{n} \ln(f(X_i; \theta))$$


0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
−17.5
−10
−5
0
5
10
10
5
0
−5
−10

Παράδειγμα: Το υπόδειγμα Garch(1,1) για την μεταβλητότητα στις χρηματοοικονομικές αγορές



Ένα πολύ συνηθισμενο υπόδειγμα για τις λογαριθμικές αποδόσεις των χρηματοοικονομικών αγορών είναι το υπόδειγμα Garch(1,1) σύμφωνα με το οποίο

$$x(t) = \ln(r(t))\,|_{r(1),\cdots,r(t-1)} \sim N(0, \sigma^2(t)),$$
$$\sigma^2(t) = \alpha_0 + \alpha_1 x^2(t-1) + \beta_1 \sigma^2(t-1)$$

Τα δεδομένα μας τυπικά είναι τα $x(0) = X_0, x(1) = X_1, \cdots, x(n) = X_n$, ενώ χρειάζεται να προσδιορίσουμε τις μεταβλητότητες $\sigma(0), \sigma(1), \cdots, \sigma(n)$

Αν γνωρίζουμε τις τιμές των

$$\theta = (\sigma(0), \alpha_0, \alpha_1, \beta_1),$$

τότε απο τις παρατηρήσεις $X = \{X_0, X_1, \cdots, X_n\}$ και την εξίσωση εξέλιξης για τα $\sigma^(t)$ μπορούμε να υπολογίσουμε όλα τα $\sigma(1), \cdots, \sigma(n)$.

Η πιθανοφάνεια του δείγματος $X_0, X_1, \cdots, X_n$ είναι

$$L(\theta; X) = \prod_{i=0}^{n} \frac{1}{\sqrt{2\pi\sigma^2(i)}} \exp\left(-\frac{X_i^2}{2\sigma^2(i)}\right),$$

$$\sigma^2(t) = \alpha_0 + \alpha_1 X^2(t-1) + \beta_1 \sigma^2(t-1), \quad t = 1, 2, \cdots, n$$

και η εκτίμηση παραμέτρων καταλήγει στο πρόβλημα ελαχιστοποίησης

$$\min_{\theta \in \mathbb{R}^4} \sum_{i=0}^{n} \left[\frac{X_i^2}{2\sigma^2(i)} - \frac{1}{2}\ln(\sigma^2(i))\right],$$

$$\sigma^2(t) = \alpha_0 + \alpha_1 X^2(t-1) + \beta_1 \sigma^2(t-1), \quad t = 1, 2, \cdots, n$$

υπο τους περιορισμούς

$$\sigma(0) > 0, \ \alpha_0 > 0, \alpha_1 > 0, \beta_1 > 0, \ \alpha_1 + \beta_1 < 1.$$

Τυπικό πρόβλημα 2: Γραμμική παλινδρόμηση και ομαλοποίηση

Ένα τυπικό πρόβλημα στην μηχανική μαθηση ειναι η σύνδεση χαρακτηριστικών (features)

$$\mathbf{X} = (X_1, \cdots, X_d) \in \mathcal{X}$$

με ιδιότητες που περιγράφονται απο τις μεταβλητές

$$\mathbf{Y} = (Y_1, \cdots, Y_m) \in \mathcal{Y}.$$

Το ζητούμενο είναι μια συναρτηση $f : \mathcal{X} \to \mathcal{Y}$ η οποία να περιγράφει ικανοποιητικά την σύνδεση παρατηρήσεων $(\mathbf{X}_i, \mathbf{Y}_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \cdots, n$, υπο την έννοια ότι

$$\mathbf{Y}_i \simeq f(\mathbf{X}_i)$$

Έχοντας καθορίσει την συνάρτηση αυτή μόλις παρατηρήσουμε ένα καινούργιο σετ χαρακτηριστικών $\mathbf{X}' = (X'_1, \cdots, X'_d)$ να 'προβλέψουμε' την απόκριση του $\mathbf{Y}' = f(\mathbf{X}')$.

Το καλύτερα μοντέλο f^* θα καθοριστεί απο μια οικογένεια μοντέλων $\mathcal{F}$, που τυπικά είναι ενα σύνολο συναρτήσεων $\mathcal{F} := \{f : \mathcal{X} \to \mathcal{Y}\}$ και καθορίζεται σαν αυτό το μοντέλο το οποίο ελαχιστοποιεί κάποια συνάρτηση σφάλματος που ποσοτικοποιεί την απόκλιση του μοντέλου μεταξύ των $\mathbf{X}_i$ και των $\mathbf{Y}_i$, $i = 1, \cdots, n$.

$$f^* = \arg\min_{f \in \mathcal{F}} \mathcal{L}(f),$$

$$\mathcal{L}(f) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(f(\mathbf{X}_i), \mathbf{Y}_i)$$

Το σύνολο $\mathcal{F}$ μπορεί είτε να ειναι ένα (απειροδιάστατο) σύνολο συναρτήσεων είτε να περιοριστούμε σε καποιες παραμετρικές μορφές (πεπερασμένης διάστασης παραμετροποιήσεις).

Οι συναρτήσεις $\mathcal{L}_i$ μπορεί να έχουν μορφές που εξαρτώνται απο την φύση των $\mathbf{Y}_i$ (π.χ. αν ειναι πραγματικοί αριθμοί, διανύσματα, πίνακες, διακριτές τιμές - φυσικοί αριθμοί κλπ)

Γραμμική παλινδρόμηση

Το πιο απλό μοντέλο είναι ένα γραμμικό μοντέλο που συνδέει τα $\mathbf{X_i} = (X_{i1}, \cdots, X_{id}) \in \mathbb{R}^d$ με τα $\mathbf{Y}_i = Y_i \in \mathbb{R}$,

$$Y_i \simeq w_1 X_{i1} + \cdots + w_d X_{id}, \quad i = 1, \cdots, n$$

Οι συναρτήσεις f καθορίζονται απο τα διανύσματα $\mathbf{w} = (w_1, \cdots, w_d)^T$.

Μια λογική συνάρτηση σφάλματος είναι η τετραγωνική απόκλιση των Y_i απο τις προβλεπόμενες απο το μοντέλο

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} \frac{1}{2} | Y_i - (w_1 X_{i1} + \cdots + w_d X_{id})|^2$$

Η συνάρτηση σφάλματος μπορεί να γραφεί και σε πιο συμπαγή μορφή χρησιμοποιώντας τον πίνακα δεδομένων

$$D = \underbrace{\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}}_{n\times d} = \underbrace{\begin{pmatrix} X_{11} & \cdots & X_{1d} \\ X_{21} & \cdots & X_{2d} \\ \ddots & \vdots & \ddots \\ X_{n1} & \cdots & X_{nd} \end{pmatrix}}_{n\times d}, \quad Y = \underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1}, \quad w = \underbrace{\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}}_{d\times 1},$$

Η συναρτηση σφάλματος γράφεται ως

$$\mathcal{L}(w) = \frac{1}{2}\|Dw - Y\|^2$$

Το γραμμικό μοντέλο w^* που επιλέγουμε ειναι αυτό που αντιστοιχεί στην λύση του προβλήματος βελτιστοποίησης

$$w^* \in \arg\min_{w\in\mathbb{R}^d} \mathcal{L}(w)$$

το οποίο μπορεί να γραφεί σαν την λύση του γραμμικού προβλήματος

$$D^T Dw = D^T y \implies w = (D^T D)^{-1} D^T y$$

Αν ο πίνακας $D^T D$ δεν είναι αντιστρέψιμος τότε μπορούμε να χρησιμοποιήσουμε στην θέση του $(D^T D)^{-1}$ τον ψευδοαντίστροφο Moore-Penrose $(D^T D)^+$.

Πολλες φορές τα δεδομένα είναι τέτοιας μορφής που ο πίνακας $D^T D$ δεν είναι 'καλός ' πχ. μπορει να είναι ill conditioned κλπ.

Για παράδειγμα μπορεί να εχουμε φαινόμενα συγγραμμικότητας (co-linearity) κλπ.

Επίσης, μπορεί ένα μοντέλο με μεγάλες θετικές και αρνητικές τιμές των διαφόρων συνιστωσών του w, να κάνει μεν καλή προσαρμογή στα δεδομένα αλλά να οδηγεί σε μεγάλες διακυμάνσεις.

Για την επίλυση τέτοιων προβλημάτων πολλες φορές καταφεύγουμε σε τροποποιήσεις του παραπάνω μοντέλου χρησιμοποιώντας όρους ομαλοποίησης (regularization).

Ridge regression - Tikhonov regularization

Στο παραπάνω πλαίσιο τροποποιούμε την συνάρτηση σφάλματος στην

$$\mathcal{L}(w) = \frac{1}{2}\|Dw - Y\|^2 + \frac{\lambda}{2}\|w\|^2, \;\; \lambda > 0.$$

Ο όρος ομαλοποίησης ' ποινικοποιεί ' μοντέλα τα οποία παρουσιάζουν μεγάλες τιμές για τους συντελεστές $w = (w_1, \cdots, w_d)$, μεσω του όρου $\|w\|^2$.

Ο συντελεστής $\lambda > 0$, δίνει την σχετική βαρύτητα των δύο όρων, του $\|Dw - Y\|^2$ που καθοριζει την πιστότητα του μοντέλου, και του $\|w\|^2$, που αποκλείει μοντέλα με μεγάλες τιμες στα w_i.

Η ομαλοποίηση αυτή επιλύει προβλήματα πολυ-συγγραμικότητας και οδηγεί σε πιο συνεπή μοντέλα.

Το μοντέλο w^* που θα επιλέξουμε

$$w^* \in \arg\min_{w \in \mathbb{R}^d} \mathcal{L}(w), \quad \mathcal{L}(w) = \frac{1}{2}\|Dw - Y\|^2 + \frac{\lambda}{2}\|w\|^2, \ \ \lambda > 0.$$

Μπορεί κανείς να δειξει (θα το δούμε λεπτομερώς στο μέλλον) ότι το μοντέλο θα λυνει το γραμμικό πρόβλημα

$$(D^T D + \lambda I)w = D^T y \implies w = (D^T D + \lambda I)^{-1} D^T y.$$

Ο πίνακας $D^T D + \lambda I$ είναι πάντοτε αντιστρέψιμος για οποιοδήποτε $\lambda > 0$.

Το ridge regression βελτιώνει το σφάλμα της πρόβλεψης συρρικνώνοντας το άθροισμα των τετραγώνων των συντελεστών του μοντέλου έτσι ώστε να είναι μικρότερο απο κάποια τιμή (που συνδέεται με την παράμετρο λ) έτσι ώστε να αποφεύγονται φαινόμενα overfitting.

Least absolute shrinkage and selection operator (Lasso)

Το μοντέλο ridge regression επιλέγει ενα μοντέλο με χαμηλό $\|w\| < t$, όμως στην ουσία το μοντέλο αυτό εμπλέκει και τις d μεταβλητές $(X_1, \cdots, X_d)$ που περιγράφουν τα χαρακτηριστικά (features) δημιουργώντας έτσι προβλήματα ερμηνείας.

Πολλες φορές ειναι επιθυμητή η επιλογή ενός μοντέλου με το μικρότερο δυνατό πλήθος χαρακτηριστικών απο τα $X_1, \cdots, X_d$) (επιλογή χαρακτηριστικών - feature selection)

Στο παραπάνω πλαίσιο τροποποιούμε την συνάρτηση σφάλματος στην

$$\mathcal{L}(w) = \frac{1}{2}\|Dw - Y\|^2 + \lambda\|w\|_1, \;\; \lambda > 0,$$

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|.$$

Η επιλογη του μοντέλου $w^* \in \arg\min_{w\in\mathbb{R}^d} \mathcal{L}(w)$ έχει σαν αποτέλεσμα την επιλογή ενός μοντέλου για το οποιό ορισμένες απο τις παραμέτρους w_i μηδενίζονται (variable selection)

Το πρόβλημα βελτιστοποίησης

$$\min_{w\in\mathbb{R}^d} \mathcal{L}(w), \quad \mathcal{L}(w) = \frac{1}{2}\|Dw - Y\|^2 + \lambda\|w\|_1, \ \ \lambda > 0,$$

είναι ποιοτικά διαφορετικό και σαφώς δυσκολότερο απο τα προηγούμενα.

Η συνάρτηση $w \mapsto \mathcal{L}(w)$ δεν είναι διαφορίσιμη (λεία) σε οποιοδήποτε $w = (w_1, \cdots, w_d)$ με κάποιο απο τα $w_i = 0$.

Η επίλυση τέτοιων προβλημάτων απαιτεί διαφορετικές αναλυτικές και αριθμητικές τεχνικές απο την μη λεία βελτιστοποίηση (non smooth optimization) όπως π.χ. τεχνικές που συνδέονται με την έννοια του υποδιαφορικού (subgradient)

Elastic net

Το μοντέλο Lasso αντιμετωπίζει προβλήματα όταν π.χ. d μεγάλο αλλά ο αριθμός των δεδομένων n είναι μικρός, ή τα δεδομένα είναι ισχυρά συσχετισμένα.

Στο παραπάνω πλαίσιο τροποποιούμε την συνάρτηση σφάλματος στην

$$\mathcal{L}(w) = \frac{1}{2}\|Dw - Y\|^2 + \lambda_1\|w\|_1 + \frac{1}{2}\lambda_2\|w\|_2^2, \quad \lambda_1, \lambda_2 > 0.$$

Στο πρόβλημα αυτό η επιλογή του μοντέλου επιβάλλει την επίλυση ενός προβλήματος βελτιστοποίησης που αποτελείται απο ένα λείο κομμάτι και ένα μη λείο κομμάτι.

Η μίξη των δύο αυτών κομματιών μας επιτρέπει την κατασκευή ειδικών αλγορίθμων που εκμεταλλεύονται την λειότητα μερους του προβλήματος για την επιτάχυνση του μη λείου αλγορίθμου (forward-backward methods)

Τυπικό πρόβλημα 3: Μη γραμμική παλινδρόμηση

Τα προβλήματα της προηγούμενης κατηγορίας μπορεί να γενικευθούν και για συναρτήσεις $f : \mathcal{X} \to \mathcal{Y} = \mathbb{R}$ οι οποίες δεν είναι γραμμικές.

Τυπικά παραδείγματα είναι συναρτήσεις οι οποίες μπορεί να εκφραστούν σαν αναπτύγματα μιας βάσης συναρτήσεων $\{\phi_i : \mathcal{X} \to \mathbb{R}, \ \ i \in \mathcal{I}\}$, όπου $\mathcal{I}$ ένα σύνολο δεικτών,

$$f(X) = \sum_{i=1}^{|\mathcal{I}|} c_i \phi_i(X) = \sum_{i=1}^{m} c_i \phi_i(X),$$

όπου $c = (c_1, \cdots, c_m) \in \mathbb{R}^m$ παράμετροι που πρέπει να καθοριστούν.

Το μοντέλο παίρνει την μορφή

$$Y_i \simeq \sum_{i=1}^{m} c_i \phi_i(X_i)$$

και για μια κατάλληλα επιλεγμένη συνάρτηση σφάλματος

$$\mathcal{L}(c) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i \left(\sum_{i=1}^{m} c_i \phi_i(X_i), Y_i \right)$$

επιλέγουμε το μοντέλο c^*,

$$c^* \in \arg \min_{c \in \mathbb{R}^m} \mathcal{L}(c).$$

Υπάρχει σχετική ελευθερία σχετικά με την επιλογή των συναρτήσεων f, και αυτό οδηγεί σε μια μεγάλη ποικιλία μοντέλων (πολύ χρήσιμα στις εφαρμογές) που όλα όμως καταλήγουν σε κάποιο πρόβλημα βελτιστοποίησης!

Οι συναρτήσεις βάσης μπορεί να είναι προκαθορισμένες, π.χ.

$\{\phi_1(x), \cdots, \phi_m(x)\} = \{1, x, \cdots, x^{m-1}\}$, πολυωνυμική παλινδρόμηση,

ή ακόμα και να καθορίζονται απο τα δεδομένα $\mathbf{X}_1, \cdots, \mathbf{X}_n$, μέσω κάποιας συνάρτησης πυρήνα (kernel function) $K : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}$ που έχουμε προεπιλέξει δηλ.

$$\{\phi_1(x) = K(x, X_1), \phi_2(x) = K(x, X_2), \cdots, \phi_n(x) = K(x, X_n)\}$$

οδηγώντας σε αναπαραστάσεις της μορφής

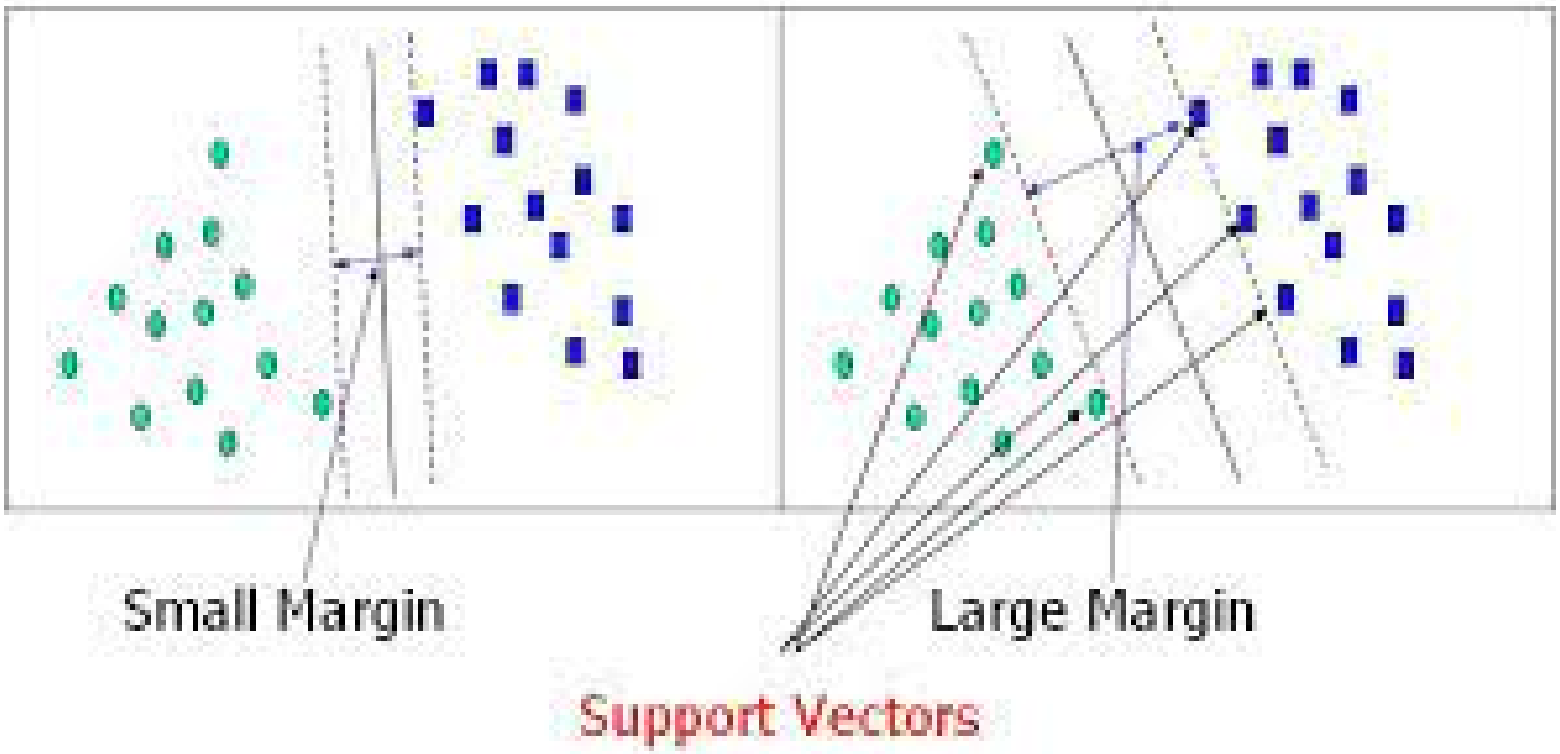
$$f(X) = \sum_{i=1}^{n} c_i K(X, X_i), \quad RKHS$$

για κατάλληλα $c_i \in \mathbb{R}$.

Τυπικό πρόβλημα 4: Support Vector Machines (SVM)

Τα SVM είναι πολύ χρησιμη μεθοδολογία για την κατάταξη δεδομένων σε διακριτές κλάσεις.

Ας υποθέσουμε οτι έχουμε χαρακτηριστικά (features) $\mathbf{X}_i = (X_{i1}, \cdots, X_{id}) \in \mathbb{R}^d$, $i = 1, \cdots, n$, τα οποία θέλουμε να κατατάξουμε σε 2 διακριτές κλάσεις, την A και την B.



Μπορούμε να ορισουμε τις μεταβλητές

$$Y_i = \begin{cases} 1 & \text{αν} \quad \mathbf{X}_i \in A, \\ -1 & \text{αν} \quad \mathbf{X}_i \in B \end{cases}$$

Ιδανικά θέλουμε να βρούμε ένα γραμμικό κανόνα διαχωρισμού δηλ. ένα διάνυσμα $w = (w_1, \cdots, w_d)$ και ένα πραγματικό αριθμό w_0, τέτοια ώστε

$$\langle w, \mathbf{X}_i \rangle - w_0 = \sum_{j=1}^{d} w_j X_{ij} - w_0 > 0 \iff sign(\langle w, \mathbf{X}_i \rangle) = 1 \text{ αν } \mathbf{X}_i \in A,$$

$$\langle w, \mathbf{X}_i \rangle - w_0 = \sum_{j=1}^{d} w_j X_{ij} - w_0 < 0 \iff sign(\langle w, \mathbf{X}_i \rangle) = -1 \text{ αν } \mathbf{X}_i \in B.$$

Γεωμετρικά, το συνολο

$$H = \{z \in \mathbb{R}^d \ : \ \langle w, z \rangle = \sum_{i=1}^{d} w_i z_i - w_0 = 0\} \subset \mathbb{R}^d$$

αντιστοιχεί σε ένα $d-1$-διάστασης υπερεπίπεδο το οποίο διαχωρίζει τα δεδομένα που τα όπτικοποιούμε΄ ως σημεία στον $\mathbb{R}^d$, σε σημεία τυπου A (απο την μία πλευρά) και σημεία τύπου B.

Βέβαια δεν είναι απαραίτητο οτι τα σημεία A και B μπορεί να διαχωριστούν πλήρως οπότε θα επιλέξουμε αυτό το υπερεπίπεδο το οποίο μας ελαχιστοποιεί το σφάλμα του διαχωρισμού, όπως αυτό π.χ μπορεί να ποσοτικοποιηθεί απο την συνάρτηση σφάλματος

$$\mathcal{L}(w) = \sum_{i=1}^{n} \max\{0, (1 - Y_i \langle w, \mathbf{X}_i \rangle - w_0\}, \quad \text{hinge loss}$$

ή την

$$\mathcal{L}(w) = \frac{1}{2} \sum_{i=1}^{n} \max\{0, (1 - Y_i \langle w, \mathbf{X}_i \rangle - w_0\}^2, \quad L^2\text{-hinge loss}$$

Παρατηρείστε ότι η συνάρτηση

$$\Phi(w) = \begin{cases} \max\{0, (1 - Y_i \langle w, \mathbf{X}_i \rangle - w_0\} = (1 - Y_i \langle w, \mathbf{X}_i \rangle)^+ = 0 & \\ \text{αν} & \langle w, \mathbf{X}_i \rangle - w_0 > 1 \text{ και } Y_i = +1 \\ \text{ή αν} & \langle w, \mathbf{X}_i \rangle - w_0 < -1 \text{ και } Y_i = -1 \\ & \text{καλή κατάταξη} \\ \max\{0, (1 - Y_i \langle w, \mathbf{X}_i \rangle\} = (1 - Y_i \langle w, \mathbf{X}_i \rangle - w_0)^+ > 1, & \text{αλλιώς, κακή κατάταξη} \end{cases}$$

Η επιλογή του SVM γράφεται με την μορφή των προβλημάτων βελτιστοποίησης

$$\min_{w\in\mathbb{R}^d, w_0\in\mathbb{R}} \mathcal{L}(w) = \sum_{i=1}^{n} \max\{0, (1 - Y_i\langle w, \mathbf{X}_i\rangle - w_0\} + \frac{\lambda}{2}\|w\|^2,$$

hinge loss Μη λείο

ή το

$$\min_{w\in\mathbb{R}^d, w_0\in\mathbb{R}} \mathcal{L}(w) = \frac{1}{2}\sum_{i=1}^{n} \max\{0, (1 - Y_i\langle w, \mathbf{X}_i\rangle - w_0\}^2 + \frac{\lambda}{2}\|w\|^2,$$

L^2-hinge loss Λείο

Ο όρος $\frac{\lambda}{2}\|w\|^2$ αντιστοιχεί σε ένα όρο ομαλοποίησης.

Ο καλός διαχωρισμός προυποθέτει ένα ασφαλές περιθώριο M (margin) μεταξύ των 2 κλάσεων.

Αυτό μπορεί να περιγραφεί μαθηματικά ζητώντας

$$\begin{aligned} \langle w, \mathbf{X}_i \rangle - w_0 > M, \quad & \text{αν} \quad Y_i = +1, \\ \langle w, \mathbf{X}_i \rangle - w_0 < -M, \quad & \text{αν} \quad Y_i = -1, \end{aligned}$$

ή σε πιό συμπαγή μορφή

$$Y_i(\langle w, \mathbf{X}_i \rangle - w_0) > M, \;\; i = 1, \cdots, n.$$

Θα μπορούσαμε να ζητήσουμε να επιλέξουμε αυτό το SVM που μεγιστοποιεί το περιθώριο M, δηλ. να ζητήσουμε την λύση του προβλήματος βελτιστοποίησης

$$\max_{w \in \mathbb{R}^d, w_0 \in \mathbb{R}} M,$$

υπο τον περιορισμό

$$Y_i(\langle w, \mathbf{X}_i \rangle - w_0) > M, \;\; i = 1, \cdots, n.$$

Οι περιορισμοί μπορεί να γραφούν ισοδύναμα ως

$$Y_i(\langle w, \mathbf{X}_i\rangle - w_0) > M, \quad i = 1, \cdots, n, \iff$$
$$Y_i(\langle \frac{w}{\|w\|^2}, \mathbf{X}_i\rangle - \frac{w_0}{\|w\|^2}) > \frac{M}{\|w\|^2}, \quad i = 1, \cdots, n$$

συνεπώς το περιθώριο είναι αντιστρόφως ανάλογο το $\|w\|^2$, οπότε αν επιλέξουμε το υπερεπίπεδο με το μικρότερο $\|w\|^2$ θα είναι σαν να έχουμε επιλέξει αυτό το SVM που δίνει το μεγαλύτερο περιθώριο.

Με αυτό το σκεπτικό αντικαθιστούμε το αρχικό πρόβλημα με το πρόβλημα

$$\min_{w\in\mathbb{R}^d, w_0\in\mathbb{R}} \|w\|^2,$$

υπο τον περιορισμό

$$Y_i(\langle w, \mathbf{X}_i\rangle - w_0) > 1, \quad i = 1, \cdots, n.$$

Η λύση προβλημάτων όπως το παραπάνω προύποθέτει μια θεωρητική μεθοδολογία σχετικα με την βελτιστοποίηση κάτω απο περιορισμούς

Κυρτή βελτιστοποίηση - Πολλαπλασιαστές Lagrange - Συνθήκες Karush - Kuhn -Tucker (KKT)

Επίσης πολύ χρήσιμη μπορεί να είναι και η δυϊκή οπτική στα προβλήματα βελτιστοποίησης (duality theory) που μας επιτρέπει να λύσουμε το αρχικό μας πρόβλημα (primal) αντικαθιστώντας με ένα άλλο, το δυικό του (dual) το οποίο μπορεί πολλές φορές να είναι ευκολότερο να αντιμετωπιστεί.

Π.χ. οι παραπάνω θεωρίες μας εξασφαλίζουν ότι

$$w = \sum_{i=1}^{n} a_i Y_i \mathbf{X}_i,$$

για κατάλληλα $a_i > 0$, δηλ. το w είναι ένα διάνυσμα που 'υποστηρίζεται' απο τα δεδομένα για τα χαρακτηριστικά (support vector) με τα α_i να δίνονται απο το δυικό πρόβλημα

$$\max_{a_i \geq 0, i=1,\cdots,n} \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} Y_i Y_j \langle \mathbf{X}_i, \mathbf{X}_j \rangle a_i a_j$$

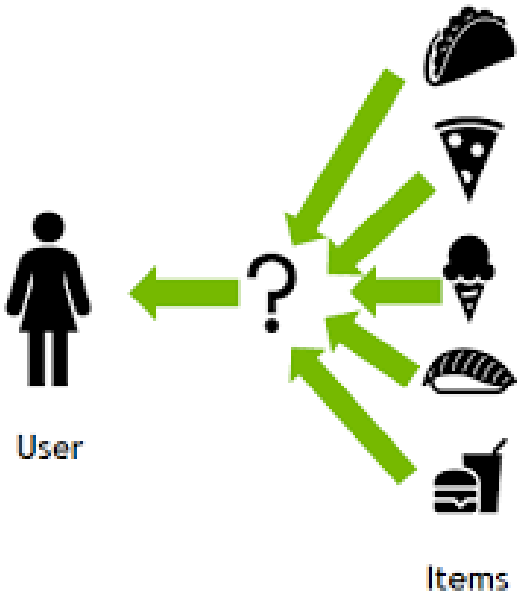
Τυπικό πρόβλημα 5: Παραγοντοποίηση πινάκων

Προβλήματα παραγοντοποίησης πινάκων είναι πολύ διαδεδομένα σε μια μεγάλη γκάμα εφαρμογών στην στατιστική και μηχανική μάθηση.

Το τυπικό πρόβλημα παραγοντοποίηση πινάκων είναι η προσέγγιση ενός πίνακα $D \in \mathbb{R}^{n \times d}$ σαν γινόμενο του πίνακα $U \in \mathbb{R}^{n \times k}$ και $V \in \mathbb{R}^{k \times d}$ με τους U, V συγκεκριμένης μορφης

$$\underbrace{D}_{n \times d} \simeq \underbrace{U}_{n \times k} \underbrace{V}_{k \times d}$$

Παράδειγμα: Recommender systems



Rating Matrix R

n

m

x	4.5	2.0	x
4.0	x	3.5	x
x	5.0	x	2.0
x	3.5	4.0	1.0

Ας πάρουμε n χρήστες που αξιολογούν d αντικείμενα βαθμολογώντας τα με τον βαθμό s_{ij}.

Π.χ. οι χρήστες αξιολογούν ταινίες (Netflix)

Ο αξιολογήσεις αυτές μπορεί να μαζευτούν σε ένα πίνακα

$$D = \begin{pmatrix} s_{11} & \cdots & s_{1d} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nd} \end{pmatrix}$$

που περιέχει τις αξιολογήσεις όλων των χρηστών για τα αντικείμενα.

Βέβαια στην πράξη, ο κάθε χρήστης δεν αξιολογεί όλα τα αντικείμενα αλλά μόνο ένα υποσύνολο απο αυτά, οπότε ο πίνακας D δεν είναι ποτέ εξ όλοκλήρου γνωστός!

Αυτό που είναι γνωστό ειναι ορισμενα απο τα στοιχεία του πίνακα D, δηλ. αυτά τα οποία αντιστοιχούν στα αντικείμενα που έχει πραγματικα αξιολογήσει ο κάθε χρήστης - ας ονομάσουμε

$$\mathcal{O} = \{(i,j) \ : \ s_{ij} \text{ έχει παρατηρηθεί}\}$$

το συνολο των δεικτών του πίνακα που αντιστοιχούν στα στοιχεια του πίνακα που έχουν παρατηρηθεί.

Το ερώτημα που θέλουμε να απαντήσουμε είναι να προβλέψουμε τα υπόλοιπα στοιχεία του πίνακα D απο τα γνωστά στοιχεία του δηλ. απο τον πίνακα $D_{obs} = (s_{ij})_{(i,j)\in\mathcal{O}}$.

Κατ' αυτόν τον τρόπο θα μπορούμε να υποδείξουμε στον κάθε χρήστη καινούργια αντικείμενα τα οποία δεν έχει ήδη δοκιμάσει και αξιολογήσει, επιλέγοντας αυτά που προβλέπουμε ότι θα αξιολογούσε με υψηλή βαθμολογία άρα και θα προτιμούσε.

Το πρόβλημα αυτό μπορεί να γραφεί σαν ένα πρόβλημα παραγοντοποίησης πινάκων.

Έστω οτι μπορούσαμε να βρούμε

- $U \in \mathbb{R}^{n\times k}$ ο πίνακας των παραγόντων των χρηστών (user factor matrix)
- $V \in \mathbb{R}^{d\times K}$ ο πίνακας των παραγόντων των αντικειμένων (item factor matrix)

τέτοιους ώστε

$$D = UV^T$$

Αν οι πίνακες U, V καθοριστούν τότε θα μπορούσαμε να προβλέψουμε τα στοιχεία του D τα οποία δεν έχουν παρατηρηθεί απο την σχέση

$$\hat{s}_{ij} = \sum_{\ell=1}^{k} u_{i\ell} v_{j\ell} \text{ recommender system}$$

Θα προσπαθήσουμε να κατασκευάσουμε τους U, V μονο απο την γνώση των παρατηρημένων στοιχείων δηλ. βάσει της πιστότητας της προτεινόμενης αναπαράστασης ως προς παρατηρήσεις,όπως περιγράφεται απο την συνάρτηση σφάλματος

$$\mathcal{L}(U, V) = \sum_{(i,j)\in\mathcal{O}} |s_{ij} - \sum_{\ell=1}^{k} u_{i\ell} v_{j\ell}|^2$$

Το πρόβλημα της κατασκευής του συστήματος προτάσεων (recommender system) μπορεί λοιπόν να γραφεί σαν ενα πρόβλημα βελτιστοποίησης της μορφής

$$\min_{U\in\mathbb{R}^{n\times k},V\in\mathbb{R}^{d\times k}} \frac{1}{2}\sum_{(i,j)\in\mathcal{O}} |s_{ij}-\sum_{\ell=1}^{k} u_{i\ell}v_{j\ell}|^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{k} u_{i\ell}^2 + \frac{\lambda}{2}\sum_{j=1}^{d}\sum_{\ell=1}^{k} v_{j\ell}^2$$

με τους όρους

$$\frac{\lambda}{2}\sum_{i=1}^{n}\sum_{\ell=1}^{k} u_{i\ell}^2 =: \frac{\lambda}{2}\|U\|_F^2 \text{ νόρμα Frobenious}$$

$$\frac{\lambda}{2}\sum_{j=1}^{d}\sum_{\ell=1}^{k} v_{j\ell}^2 =: \frac{\lambda}{2}\|U\|_F^2 \text{ νόρμα Frobenious}$$

να παίζουν τον ρόλο των όρων ομαλοποίησης (regularization terms).

Παράδειγμα: Συσταδοποίηση k-means

Το πρόβλημα της συσταδοποίησης (k -means clustering) συνίσταται στο να βρούμε k κεντροιδή επιλεγμένα έτσι ώστε το άθροισμα όλων των στοιχείων ενος $n \times k$ πίνακα δεδομένων απο το κοντινότερο κεντρο που αντιστοιχεί στο καθενα απο αυτά να είναι όσο το δυνατόν ελάχιστο.

Το πρόβλημα αυτό μπορει να δειξει κανείς ότι είναι ένα ισοδύναμο με ένα πρόβλημα βελτιστοποίησης (παραγοντοποίηση πινάκων υπο περιορισμούς) της μορφής

$$\min_{U \in \mathbb{R}^{n \times k}, V \in \mathbb{R}^{d \times k}} \|D - UV^T\|_F^2,$$

υπο τους περιορισμούς

U ορθογώνιος

$u_{ij} \in \{0, 1\}$

Κάθε γραμμή i του πίνακα U θα πρέπει να περιέχει ακριβώς ένα στοιχείο 1, που θα δείχνει και το cluster j στο οποίο θα ανήκει το σημειο i.

Οι στήλες του πίνακα V καθορίζουν την θέση των κέντρων.