ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

# Ενότητα 8 – Supervised Learning
## Support Vector Machines (*SVMs)*
## *Part 1*
## *Classification*

**Μέθοδοι Μηχανικής Μάθησης στα Χρηματοοικονομικά**

**Αθανάσιος Σάκκας, Επ. Καθηγητής, ΟΠΑ**

# Support Vector Machines (SVMs)

- Ένα πλεονέκτημα των SVMs είναι ότι δουλεύουν καλά όταν υπάρχει μεγάλος αριθμός χαρακτηριστικών.
- Πιο συγκεκριμένα, δύναται ο αριθμός των χαρακτηριστικών να είναι μεγαλύτερος του αριθμού των παρατηρήσεων.
- Το μειονέκτημα είναι ότι κάνει classification των παρατηρήσεων σε θετικές και αρνητικές.
- Σε αντίθεση με τη logistic regression και τα decision tress, δε δίνει πιθανότητες.

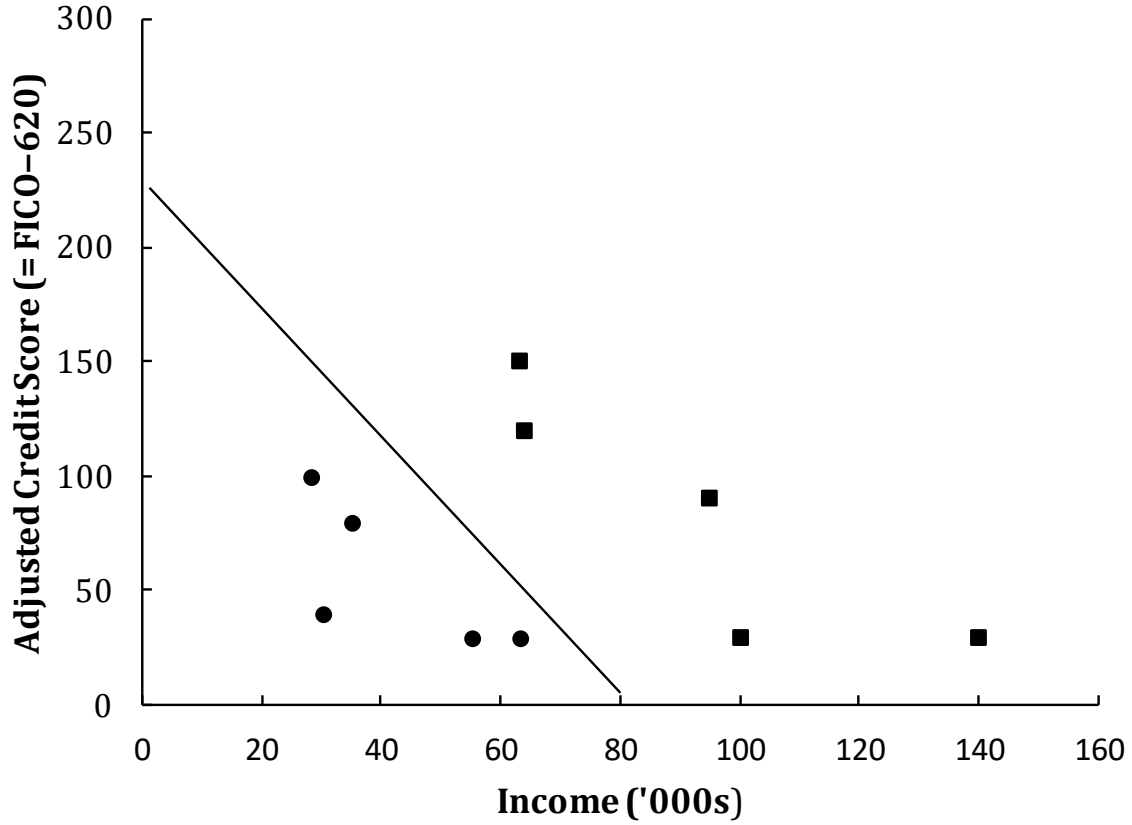
Στο πρώτο μέρος, θα αναλύσουμε τη χρήση των SVMs για classification. Θα χρησιμοποιήσουμε ένα μικρό δείγμα από την **Εφαρμογή LendingClub.**

| Credit score | Adjusted credit score | Income ('000s) | Default =0; good loan=1 |
|---|---|---|---|
| 660 | 40 | 30 | 0 |
| 650 | 30 | 55 | 0 |
| 650 | 30 | 63 | 0 |
| 700 | 80 | 35 | 0 |
| 720 | 100 | 28 | 0 |
| 650 | 30 | 140 | 1 |
| 650 | 30 | 100 | 1 |
| 710 | 90 | 95 | 1 |
| 740 | 120 | 64 | 1 |
| 770 | 150 | 63 | 1 |

**A.** Το πρώτο βήμα είναι να κάνουμε normalization των δεδομένων μας. Για την κατανόηση του συγκεκριμένου παραδείγματος, πραγματοποιούμε ένα κατά προσέγγιση scaling αφαιρώντας το 620 από το credit score.

**Β.** Το δεύτερο βήμα είναι να θεωρήσουμε ένα **Linear Separation των δεδομένων μας σχεδιάζοντας μια ευθεία γραμμή**, όπως βλέπετε παρακάτω.

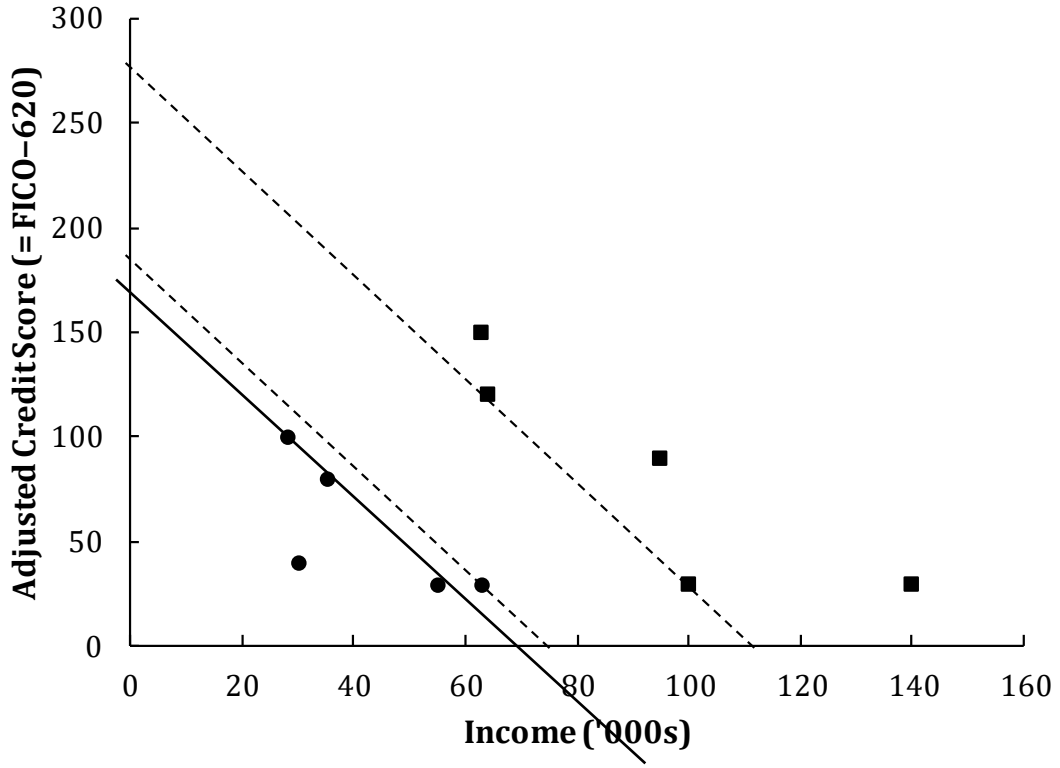
**Σημείωση: Οι κύκλοι είναι** τα κακά δάνεια, ενώ **τα τετράγωνα** τα καλά δάνεια. Επίσης να σημειώσουμε ότι το παράδειγμα αυτό είναι εξιδανικευμένο.



# SVM Approach

- Στην προσέγγιση των support vector machine (SVM) approach βρίσκουμε ένα μονοπάτι για να χωρίσουμε τα δεδομένα σε δύο κλάσεις.
- Στην “hard margin” περίπτωση ο τέλειος διαχωρισμός είναι δυνατός (όπως στο παράδειγμά μας).
- Ο αλγόριθμος βρίσκει την **ευρύτερη** δυνατή διαδρομή (για να μεγιστοποιήσει τα οφέλη από το regularization).
- Τα δεδομένα πρέπει να κανονικοποιηθούν (Πραγματοποιούμε κατά προσέγγιση κανονικοποίηση αφαιρώντας 620 από το credit score στο παράδειγμά μας)
- **Τα support vectors είναι οι παρατηρήσεις στην άκρη του μονοπατιού.**

Το καλύτερο μονοπάτι για παράδειγμα. Η σταθερή γραμμή θα χρησιμοποιηθεί για τη διάκριση των καλών και των κακών δανείων



# The Separating hyperplane

- Με δύο μόνο χαρακτηριστικά, ο διαχωρισμός επιτυγχάνεται με μια γραμμή:

$$w_1 x_1 + w_2 x_2 = b$$

- Με $m$ χαρακτηριστικά ο διαχωρισμός επιτυγχάνεται με ένα $(m-1)$ dimensional hyperplane:

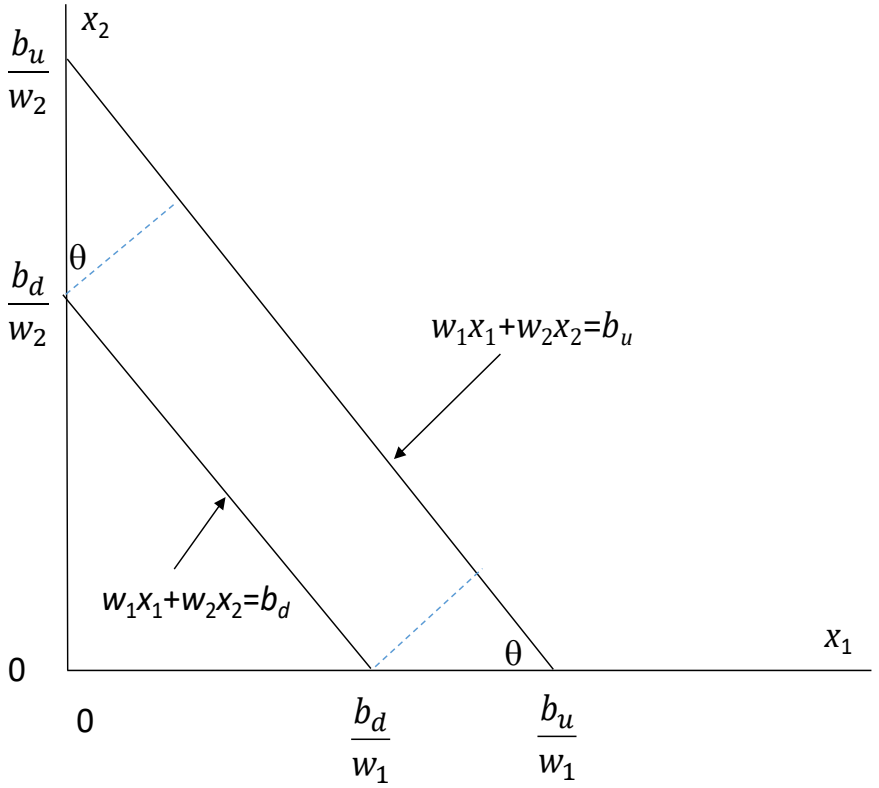
$$\sum_{j=1}^{m} w_j x_j = b$$

Τα $w_j$ είναι σταθερά και αναφέρονται ως weights.

# Regularization

- Το Regularization περιλαμβάνει τη μείωση του μεγέθους των βαρών, $w_j$.
- Αυτό αποφεύγει το overfitting όπου υπάρχουν δύο συσχετιζόμενες μεταβλητές η μία με μεγάλο θετικό βάρος και η άλλη με μεγάλο αρνητικό βάρος.
- **Ελαχιστοποιούμε το άθροισμα των τετραγώνων των βαρών. Αυτό είναι το ίδιο με τη μεγιστοποίηση του πλάτους της διαδρομής (width of the pathway).**

# Notation


$x_2$
$\frac{b_u}{w_2}$
$\frac{b_d}{w_2}$
θ
$w_1x_1+w_2x_2=b_u$
$w_1x_1+w_2x_2=b_d$
θ
$x_1$
0
0
$\frac{b_d}{w_1}$
$\frac{b_u}{w_1}$

# Τα Μαθηματικά...

Αν *P είναι το πλάτος του μονοπατιού* (width of pathway)

$$\sin\theta = \frac{Pw_1}{b_u - b_d} \qquad \cos\theta = \frac{Pw_2}{b_u - b_d} \qquad P = \frac{b_u - b_d}{\sqrt{w_1^2 + w_2^2}}$$

Μπορούμε να κάνουμε scale τα $w_1$, $w_2$, $b_u$, and $b_d$ με την ίδια σταθερά χωρίς να αλλάξει το μοντέλο. Μπορούμε επομένως να θέσουμε $b_u$=$b$+1 και $b_d$=$b$−1 ώστε το πλάτος του μονοπατιού (the width of the pathway) να είναι

$$P = \frac{2}{\sqrt{w_1^2 + w_2^2}}$$

Στην περίπτωση του hard margin ο αλγόριθμος ελαχιστοποιηθεί το $w_1^2 + w_2^2$

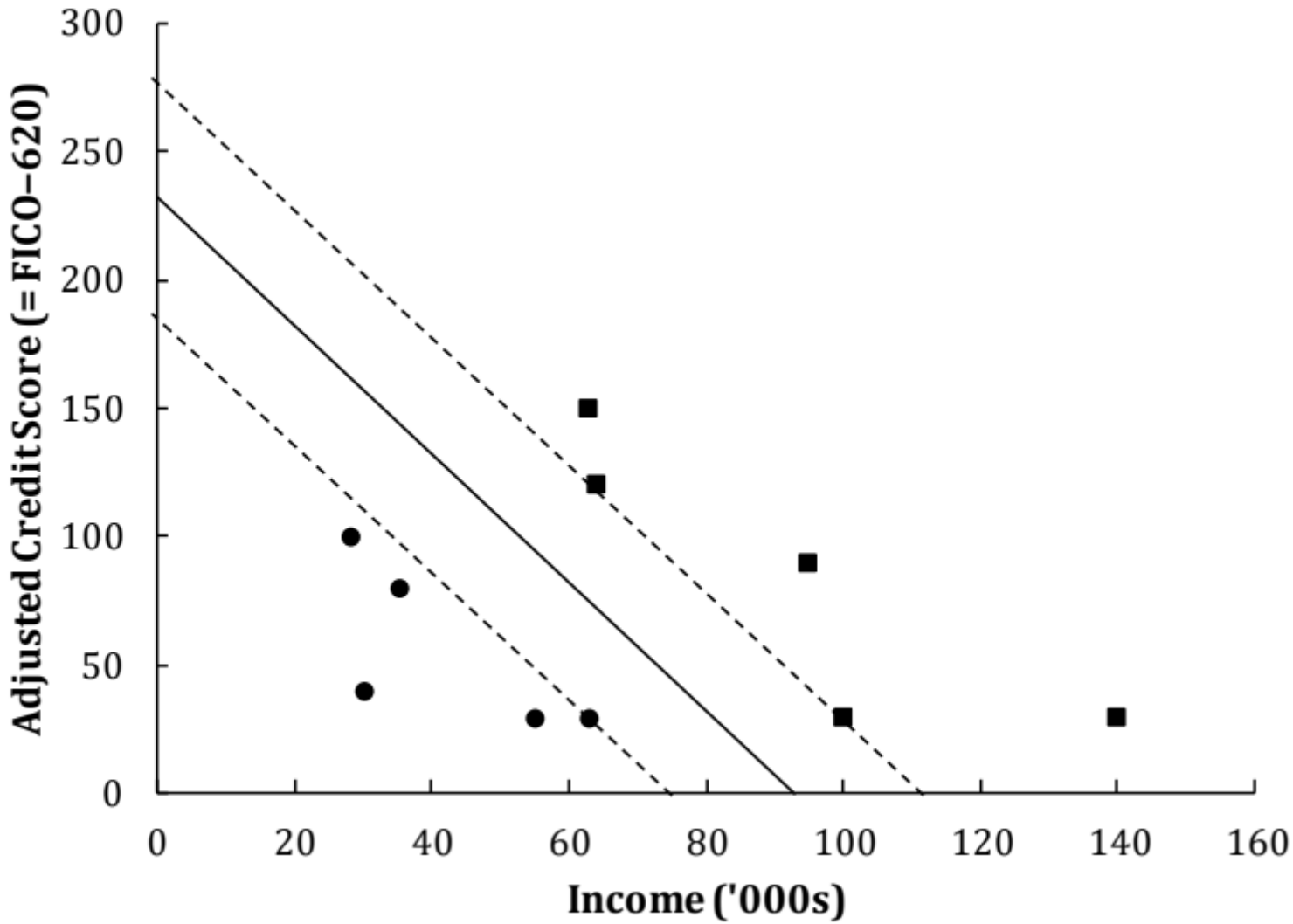
*subject to* τέλειου διαχωρισμού (perfect separation) που επιτυγχάνεται. (Το Regularization προσπαθεί να απλοποιήσει το μοντέλο μειώνοντας το μέγεθος των βαρών (magnitude of weights)).

# Hard margin problem στο παράδειγμά μας (Quadratic programming problem)

Στο παράδειγμά μας ο σκοπός μας είναι να βρούμε τα $b$, $w_1$, and $w_2$ τα οποία θα ελαχιστοποιούν το $w_1^2 + w_2^2$ *subject to*

$$
\begin{aligned}
30w_1 + 40w_2 &\leq b - 1 \\
55w_1 + 30w_2 &\leq b - 1 \\
63w_1 + 30w_2 &\leq b - 1 \\
35w_1 + 80w_2 &\leq b - 1 \\
28w_1 + 100w_2 &\leq b - 1 \\
140w_1 + 30w_2 &\geq b + 1 \\
100w_1 + 30w_2 &\geq b + 1 \\
95w_1 + 90w_2 &\geq b + 1 \\
64w_1 + 120w_2 &\geq b + 1 \\
63w_1 + 150w_2 &\geq b + 1
\end{aligned}
$$

Βρίσκουμε $b$ *=5.054,* $w_1$ *=0.05405* και $w_2$ *=0.02162. Αρα το μέσο του μονοπατιού δίενται από την παρακάτω εξίσωση* $0.05405x_1 + 0.02162x_2 = 5.054$ . Η γραμμή αυτή διαχωρίζει τα καλά από τα κακά δάνεια.Το P= 34.35. Δείτε την επόμενη διαφάνεια.


Adjusted CreditScore (= FICO−620)
300
250
200
150
100
50
0
0
20
40
60
80
100
120
140
160
Income ('000s)

# Το γενικό hard margin problem

- Η αντικειμενική συνάρτηση είναι

$$\sum_{j=1}^{m} w_j^2$$

- Ελαχιστοποιούμε την παραπάνω συνάρτηση για τις τιμές των $w_j$ και $b$ *subject to* την προϋπόθεση να μην υπάρχουν παραβάσεις, δηλαδή:

$$\sum w_j x_j - b > 1 \text{ if loan good}$$

$$\sum w_j x_j - b < -1 \text{ if loan bad}$$

# To Soft Margin Problem

Μετράμε την παραβίαση* μιας παρατήρησης ως τον βαθμό στον οποίο παραβιάζεται η συνθήκη του hard margin condition

Ελαχιστοποιούμε το

$$C \times \sum_{j=1}^{m} z_i \ + \sum_{j=1}^{m} w_j^2$$

Όπου $z_i$ είναι το μέτροπου μία παρατήρηση i παραβιάζει τα hard margin conditions.

$z_i = max\left(b + 1 - \sum_{j=1}^{m} w_j x_{ij}\ , 0\right)$ αν το αποτέλεσμα είναι θετικό.

$z_i = max\left(\sum_{j=1}^{m} w_j x_{ij} - (b-1), 0\right)$ αν το αποτέλεσμα είναι αρνητικό.

Αλλάζοντας το $C$ αλλάζει το trade-off μεταξύ του πλάτους του μονοπατιού (width of the path) και των παραβιάσεων.

Όσο μικραίνει το $C$, το μονοπάτι (pathway) γίνεται πιο πλατύ με περισσότερες παραβιάσεις.

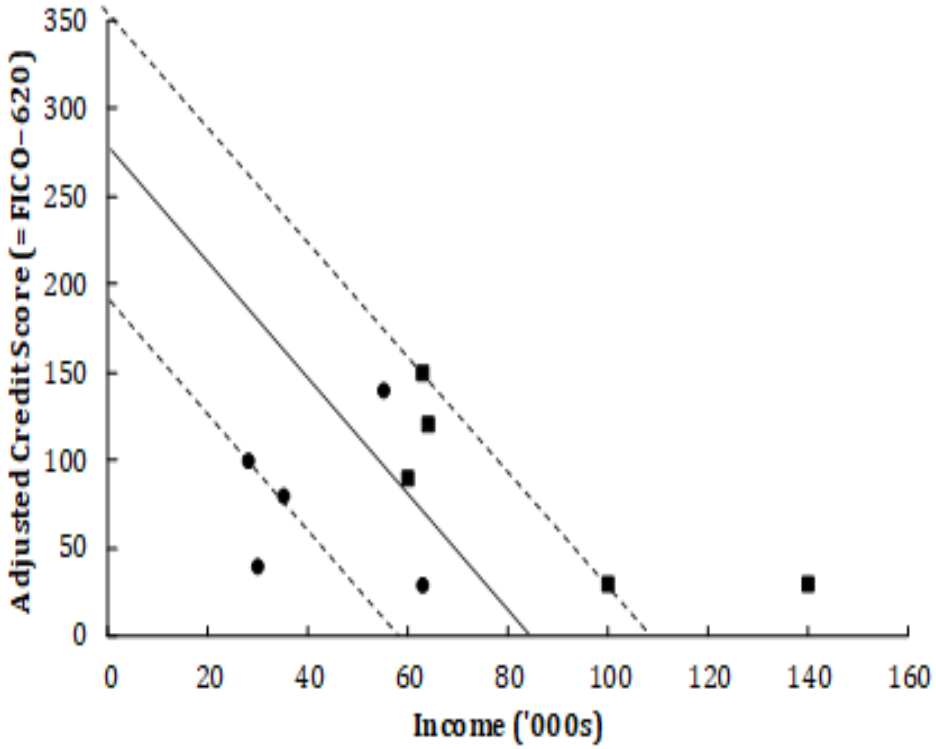
**Με την παραβίαση εννοούμε παρατηρήσεις που βρίσκονται μέσα στο μονοπάτι ή στη λάθος πλευρά του μονοπατιού.*

Αν αλλάξουμε κάποιες τιμές στο παράδειγμά μας. Έστω

| Credit score | Adjusted credit score | Income (‘000s) | Default =0; good loan=1 |
|---|---|---|---|
| 660 | 40 | 30 | 0 |
| 650 | **140** | 55 | 0 |
| 650 | 30 | 63 | 0 |
| 700 | 80 | 35 | 0 |
| 720 | 100 | 28 | 0 |
| 650 | 30 | 140 | 1 |
| 650 | 30 | 100 | 1 |
| 710 | 90 | **60** | 1 |
| 740 | 120 | 64 | 1 |
| 770 | 150 | 63 | 1 |

# C=0.001 Αποτελέσματα

Δείτε το *8.1 SVM Example.xlsx και το 8. SVM_Example_Python.ipynb*



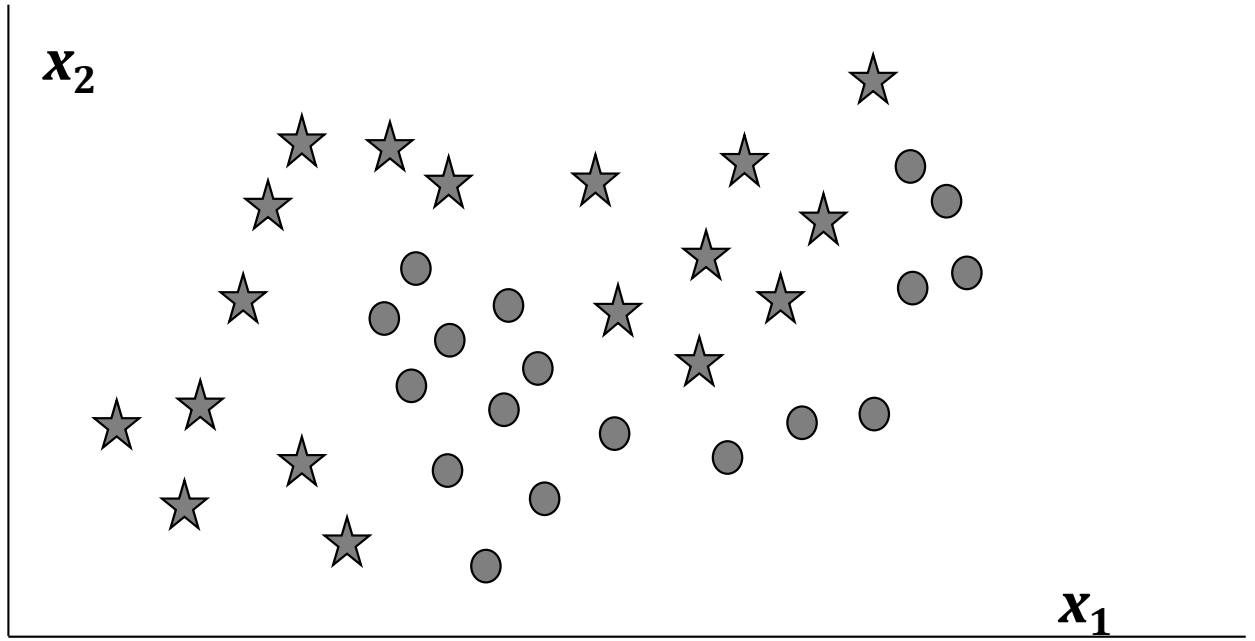
# Η επίδραση του C στο παράδειγμά μας

| $C$ | $w_1$ | $w_2$ | $b$ | Loans mis-classified by centerline | Width of pathway | Total pathway violations |
|---|---|---|---|---|---|---|
| 0.01 | 0.054 | 0.022 | 5.05 | 10% | 34.4 | 2.81 |
| 0.001 | 0.040 | 0.012 | 3.33 | 10% | 48.2 | 3.31 |
| 0.0005 | 0.026 | 0.010 | 2.46 | 10% | 70.6 | 4.79 |
| 0.0003 | 0.019 | 0.006 | 1.79 | 20% | 102.2 | 5.79 |
| 0.0002 | 0.018 | 0.003 | 1.69 | 30% | 106.6 | 5.91 |

# Συμπεράσματα

- Οι παραβιάσεις υπολογίζονται σε σχέση με τα άκρα του μονοπατιού.
- Τα misclassifications υπολογίζονται σε σχέση με το κέντρο του μονοπατιού.
- Το training set χρησιμοποιείται για να αναπτυχθούν εναλλακτικά SVMs, δηλαδή εναλλακτικά boundaries. Και μετά ακολουθούμε τη διαδικασία για τα ML.

# Non-linear separation



# Non-linear classification

- Μέχρι τώρα θεωρήσαμε το μονοπάτι να είναι γραμμική συνάρτηση των χαρακτηριστικών. Τώρα θα εξετάσουμε κατά πόσο η υπόθεση δύναται να «χαλαρώσει»
- Ο στόχος είναι να δημιουργηθούν νέα χαρακτηριστικά έτσι ώστε το όριο να γίνει γραμμικό.
- Ας υποθέσουμε ότι υπάρχει ένα μόνο χαρακτηριστικό (age) και βρίσκουμε ότι οι χαμηλές και οι υψηλές τιμές του χαρακτηριστικού τείνουν να δίνουν ένα αποτέλεσμα ενώ οι ενδιάμεσες τιμές δίνουν ένα άλλο αποτέλεσμα.
- Θα μπορούσαμε να δημιουργήσουμε ένα νέο χαρακτηριστικό ώστε $(v\text{-}m)^2$ όπου $v$ είναι η τιμή χαρακτηριστικού και $m$ η μέση τιμή του.

# Δημιουργώντας νέα χαρακτηριστικά

- Μπορούμε να προσθέσουμε δυνάμεις για κάθε χαρακτηριστικό ως νέο χαρακτηριστικό.

- Εναλλακτικά, μπορούμε να επιλέξουμε συγκεκριμένα ορόσημα και να δημιουργήσουμε νέα χαρακτηριστικά χρησιμοποιώντας τη Gaussian Radial Basis Function (similarity function). Εάν οι τιμές των χαρακτηριστικών σε ένα ορόσημο είναι $\ell_1$, $\ell_2$, ……, $\ell_m$, οι νέες τιμές χαρακτηριστικών υπολογίζονται ως

$$\exp\left(-\gamma \sum_{j=1}^{m} (x_j - \ell_j)^2\right)$$

- Καθώς η παράμετρος γ αυξάνεται, το εύρος επιρροής ενός ορόσημου μειώνεται και το όριο γίνεται λιγότερο ομαλό.