Επεξεργασία Δεδομένων και Αλγόριθμοι Μάθησης

Αναφορά Σετ Ασκήσεων 2

Όνομα: Μακάριος Χρηστάκης

Περιεχόμενα

Πρόβλημα 1

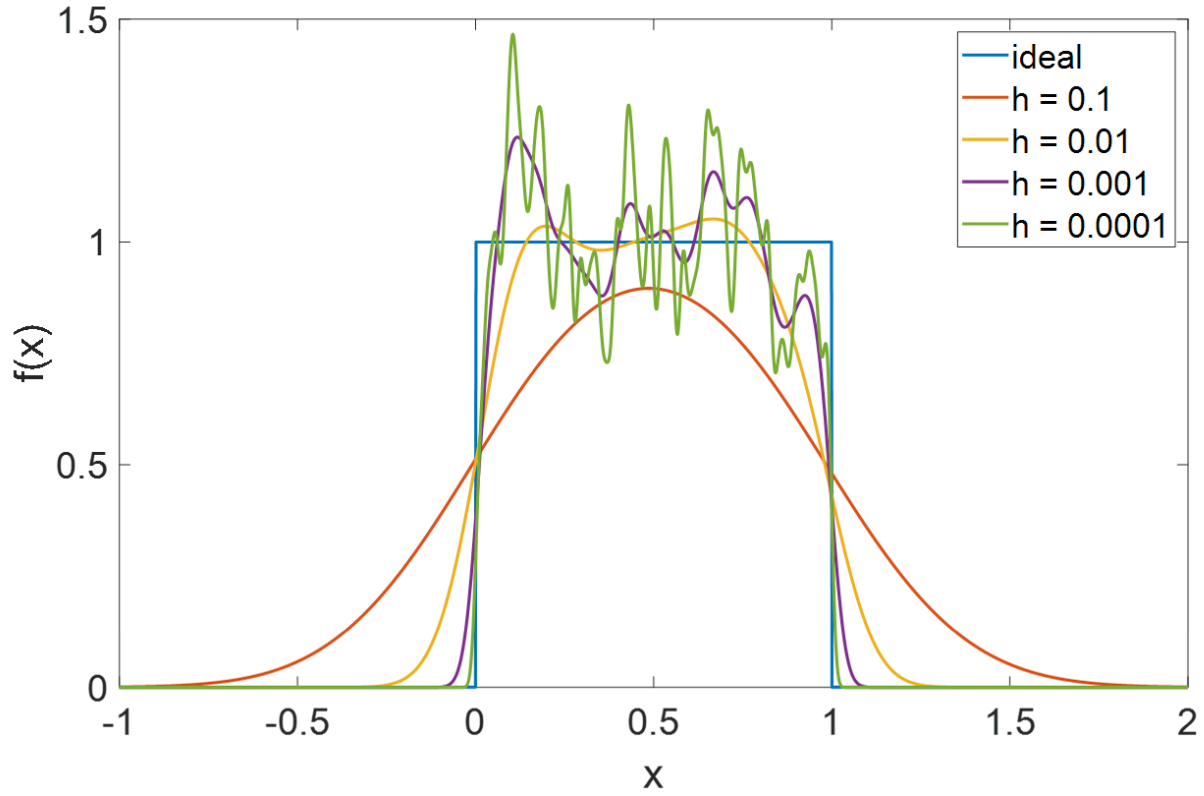
Όπως μάθαμε στην θεωρία για να προσεγγίσουμε την συνάρτηση πυκνότητας πιθανότητας f(x) μιας τυχαίας μεταβλητής στο σημείο x, έχοντας στη διάθεση μας *N* υλοποιήσεις της τυχαίας μεταβλητής αυτής $x_i, i = 1 \dots N$ μπορούμε να χρησιμοποιήσουμε κάποια από τις συναρτήσεις kernel που προσεγγίζουν την κρουστική συνάρτηση και να υπολογίσουμε την παράσταση:

$$\hat{f}(x) = \frac{\sum_{i=1}^{N} K(x - x_i\ , h)}{N} \qquad (1)$$

Για το σκοπό αυτό δημιουργήθηκαν 1000 υλοποιήσεις από ομοιόμορφη κατανομή στο διάστημα [0,1]. Παρακάτω θα μελετηθεί η επίδραση των διαφόρων kernel function, όπως και της παραμέτρου h, στην προσέγγιση αυτής της συνάρτησης πυκνότητας πιθανότητας.

Ερώτημα (α) **Gaussian Kernel**

Χρησιμοποιώντας τον Gaussian Kernel υπολογίζουμε την προσέγγιση της PDF μέσω της σχέσης (1) σε 2000 σημεία στο διάστημα [-1,2] και μεταβάλλουμε την τιμή της παραμέτρου h. Τα αποτελέσματα φαίνονται παρακάτω:

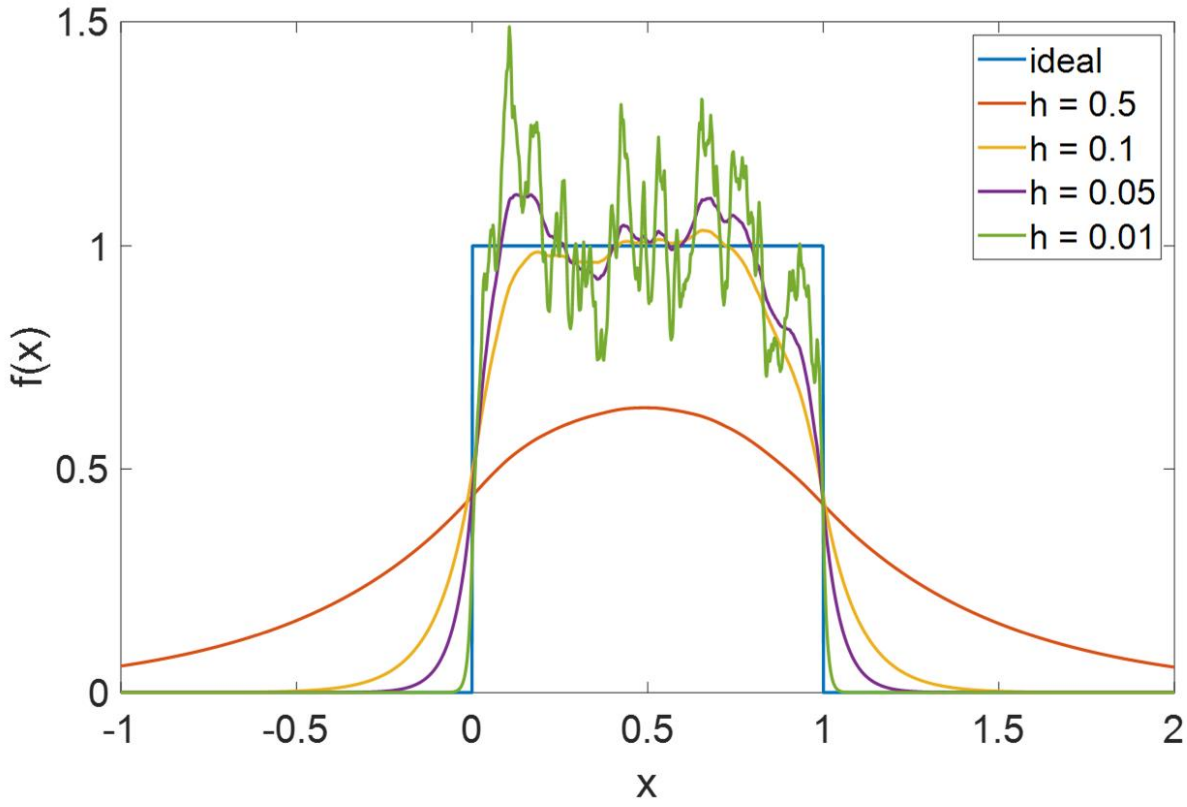


Σχήμα 1: Προσέγγιση με Gaussian Kernel

Όπως φαίνεται στο Σχήμα 1, όσο μειώνουμε την τιμή του h η προσέγγιση πετυχαίνει με όλο και μεγαλύτερη ακρίβεια τα όρια της ιδανικής καμπύλης όπου υπάρχει η ασυνέχεια, ωστόσο εισάγεται θόρυβος στην προσέγγιση της τιμής 1 της ιδανικής καμπύλης. Ο θόρυβος αυτός μειώνεται όσο αυξάνεται το h, αλλά φυσικά χάνονται τα όρια της ιδανικής μεταβολής.

Ερώτημα (β) **Laplacian Kernel**

Για τον Laplacian Kernel υπολογίζοντας στα ίδια 2000 σημεία την προσέγγιση της PDF για διάφορες τιμές του h παίρνουμε:



Σχήμα 2:Προσέγγιση με Laplacian Kernel

Όπως φαίνεται στο Σχήμα 2 και με τον Laplacian Kernel έχουμε το tradeoff μεταξύ ακρίβειας προσέγγισης των ορίων της καμπύλη. Ο Laplacian Kernel ωστόσο αρχίζει να εμφανίζει αυτόν τον θόρυβο σε μεγαλύτερες κατά περίπου μια τάξη μεγέθους τιμές του h απ' ότι ο Gaussian Kernel.

Πρόβλημα 2

Ερώτημα (α)

Η συνάρτηση $\Phi(\boldsymbol{X})$ που θέλουμε να προσεγγίσουμε έγκειται στο χώρο V που ορίζει ο Gaussian Kernel:

$$K\ (\boldsymbol{X}, \boldsymbol{Y}\ ) = e^{\frac{-\|\boldsymbol{X}-\boldsymbol{Y}\|^2}{h}}$$

Η συνάρτηση $\widehat{\Phi}(\boldsymbol{X})$ ανήκει στον γραμμικό υποχώρο Ω του V που ορίζεται από τις συναρτήσεις $K(\boldsymbol{X}, \boldsymbol{X_i})$ όπου X_i τα γνωστά μας παραδείγματα. Συνεπώς από το representer theorem θα έχουμε:

$$< \Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X_i}) > = 0 \Rightarrow < \Phi(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X_i}) > = < \widehat{\Phi}(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X_i}) > \Rightarrow \widehat{\Phi}(\boldsymbol{X_i}) = \Phi(\boldsymbol{X_i})$$

Συνεπώς οι τιμές της ιδανικής, άγνωστης συνάρτησης $\Phi(\mathbf{X})$ στα σημεία $\boldsymbol{X_i}$ που γνωρίζουμε θα ταυτίζονται με τις τιμές της $\widehat{\Phi}(\boldsymbol{X})$, στα σημεία αυτά. Άρα μπορούμε στο κριτήριο κόστους να αντικαταστήσουμε στα πρώτα 2 αθροίσματα την συνάρτηση $\Phi(\boldsymbol{X})$ με $\widehat{\Phi}(\boldsymbol{X})$.

Ερώτημα (β)

Αρχικά μπορούμε να εκφράσουμε την τελευταία νόρμα του κριτηρίου ως:

$$\|\Phi(\boldsymbol{X})\|^2 = \left\|\Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X}) + \widehat{\Phi}(\boldsymbol{X})\right\|^2 \Rightarrow$$

$$\|\Phi(\boldsymbol{X})\|^2 = \left\|\Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X})\right\|^2 + \left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2 + 2 < \Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X}), \widehat{\Phi}(\boldsymbol{X}) > \quad (2)$$

Όμως αφού η $\widehat{\Phi}(\boldsymbol{X})$ είναι κάθετη προβολή της $\Phi(\boldsymbol{X})$ στον υποχώρο Ω, το σφάλμα $\Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X})$, θα είναι κάθετο στον υποχώρο Ω που βρίσκεται η$\widehat{\Phi}(\boldsymbol{X})$, οπότε το εσωτερικό γινόμενο των δύο διανυσμάτων αυτών θα είναι μηδενικό. Συνεπώς από τη σχέση (2) θα έχουμε:

$$\|\Phi(\boldsymbol{X})\|^2 = \left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2 + \left\|\Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X})\right\|^2 \Rightarrow \left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2 = \|\Phi(\boldsymbol{X})\|^2 - \left\|\Phi(\boldsymbol{X}) - \widehat{\Phi}(\boldsymbol{X})\right\|^2$$

Κι αφού ο τελευταίος όρος είναι μη αρνητικός καταλήγουμε στο:

$$\left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2 \leq \|\Phi(\boldsymbol{X})\|^2$$

Αφού λοιπόν αποδείξαμε πως οι τιμές της συνάρτησης $\widehat{\Phi}(\boldsymbol{X})$ στα σημεία $\boldsymbol{X_i}$ είναι ίδιες με αυτές της συνάρτησης $\Phi(\boldsymbol{X})$ στα σημεία αυτά και η νόρμα της $\widehat{\Phi}(\boldsymbol{X})$ στον τελευταίο όρο του κριτηρίου θα είναι το πολύ ίση με την νόρμα της $\Phi(\boldsymbol{X})$, αντικαθιστώντας την $\Phi(\boldsymbol{X})$ με $\widehat{\Phi}(\boldsymbol{X})$ στο κριτήριο και ελαχιστοποιώντας το ως προς την $\widehat{\Phi}(\boldsymbol{X})$, το Loss που θα προκύψει θα είναι μικρότερο ή ίσο του βέλτιστου που θα πρόκυπτε εάν ελαχιστοποιούσαμε την αρχική παράσταση! Συνεπώς, μπορούμε να αντικαταστήσουμε την $\Phi(\boldsymbol{X})$ με $\widehat{\Phi}(\boldsymbol{X})$ στο κριτήριο αυτό.

Ερώτημα (γ)

Στο κριτήριο προς ελαχιστοποίηση είναι

$$\min_{\widehat{\Phi}\in\Omega} \sum_{\boldsymbol{X_i}\in stars} \left(1-\widehat{\Phi}(\boldsymbol{X_i})\right)^2 + \sum_{\boldsymbol{X_j}\in circles} \left(1+\widehat{\Phi}(\boldsymbol{X_j})\right)^2 + \lambda\left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2 \quad (3)$$

Ορίζοντας $y_i = \pm 1$ η επιθυμητή τιμή της Φ(**X**) για το παράδειγμα $\boldsymbol{X_i}$ και

$$\widehat{\Phi}(\boldsymbol{X}) = \boldsymbol{c}^T\boldsymbol{K}(\boldsymbol{X})$$

όπου

$$\boldsymbol{c} = [a_1 \dots a_N \;\; \beta_1 \dots \beta_M]^T$$

$$\boldsymbol{K}(\boldsymbol{X}) = [K(\boldsymbol{X},\boldsymbol{X_1}) \dots K(\boldsymbol{X},\boldsymbol{X_N})\; K(\boldsymbol{X},\boldsymbol{X_{N+1}}) \dots K(\boldsymbol{X},\boldsymbol{X_{N+M}})]^T$$

με τα γνωστά παραδείγματα $\boldsymbol{X_i} \in \begin{cases} stars, & i = 1 \dots N \\ circles, & i = N+1 \dots N+M \end{cases}$

Η εξίσωση (3) μπορεί να γραφεί ισοδύναμα ως:

$$\min_{\boldsymbol{c}} \sum_{i=1}^{N+M} (1 - y_i\boldsymbol{c}^T\boldsymbol{K}(\boldsymbol{X_i}))^2 + \lambda\left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2$$

Η νόρμα στον τελευταίο όρο μπορεί να γραφεί ως:

$$\left\|\widehat{\Phi}(\boldsymbol{X})\right\|^2 = <\widehat{\Phi}(\boldsymbol{X}), \widehat{\Phi}(\boldsymbol{X})> = \sum_{i=1}^{N+M}\sum_{j=1}^{N+M} c_i c_j K(\boldsymbol{X_i},\boldsymbol{X_j}) = \boldsymbol{c}^T\boldsymbol{\Lambda}\boldsymbol{c}$$

Όπου $\boldsymbol{\Lambda}$ ο (N+M)x(N+M) συμμετρικός πίνακας με στοιχεία $\Lambda_{i,j} = K(\boldsymbol{X_i},\boldsymbol{X_j})$

Οπότε το κριτήριο προς ελαχιστοποίηση είναι:

$$\min_{\boldsymbol{c}} \mathrm{L}(\mathbf{c}) = \min_{\boldsymbol{c}} \sum_{i=1}^{N+M} (1 - y_i\boldsymbol{c}^T\boldsymbol{K}(\boldsymbol{X_i}))^2 + \lambda\, \boldsymbol{c}^T\boldsymbol{\Lambda}\boldsymbol{c}$$

Οπότε για να βρούμε το βέλτιστο διάνυσμα συντελεστών **c**, αφού το κριτήριο $L(\boldsymbol{c})$ είναι τετραγωνικό ως προς το **c** αρκεί να λύσουμε την εξίσωση

$$\frac{\partial L(\boldsymbol{c})}{\partial \boldsymbol{c}} = \mathbf{0}$$

Οπότε έχουμε:

$$\frac{\partial L(\boldsymbol{c})}{\partial \boldsymbol{c}} = \mathbf{0} \Rightarrow -2\sum_{i=1}^{N+M}\left(1 - y_i\boldsymbol{c}^T\boldsymbol{K}(\boldsymbol{X_i})\right)y_i\boldsymbol{K}(\boldsymbol{X_i})^{\boldsymbol{T}} + 2\lambda\boldsymbol{c}^T\boldsymbol{\Lambda} = \mathbf{0}$$

$$\Rightarrow -\sum_{i=1}^{N+M}\left(y_i\boldsymbol{K}(\boldsymbol{X_i})^{\boldsymbol{T}}\right) + \sum_{i=1}^{N+M}\left(y_i^2\boldsymbol{c}^T\boldsymbol{K}(\boldsymbol{X_i})\boldsymbol{K}(\boldsymbol{X_i})^{\boldsymbol{T}}\right) + \lambda\boldsymbol{c}^T\boldsymbol{\Lambda} = \mathbf{0}$$

$$\Rightarrow \boldsymbol{c}^T \sum_{i=1}^{N+M} \left(\boldsymbol{K}(\boldsymbol{X_i})\boldsymbol{K}(\boldsymbol{X_i})^T\right) + \lambda \boldsymbol{c}^T \boldsymbol{\Lambda} = \sum_{i=1}^{N+M} \left(y_i \boldsymbol{K}(\boldsymbol{X_i})^T\right)$$

$$\Rightarrow \boldsymbol{c^T}\left(\boldsymbol{\Lambda^2} + \lambda\, \boldsymbol{\Lambda}\right) = \sum_{i=1}^{N+M} \left(y_i \boldsymbol{K}(\boldsymbol{X_i})^T\right)$$

$$\Rightarrow \boldsymbol{c^T} = \sum_{i=1}^{N+M} \left(y_i \boldsymbol{K}(\boldsymbol{X_i})^T\right)\left(\boldsymbol{\Lambda^2} + \lambda\, \boldsymbol{\Lambda}\right)^{-1} \quad (4)$$

Ερώτημα (δ)

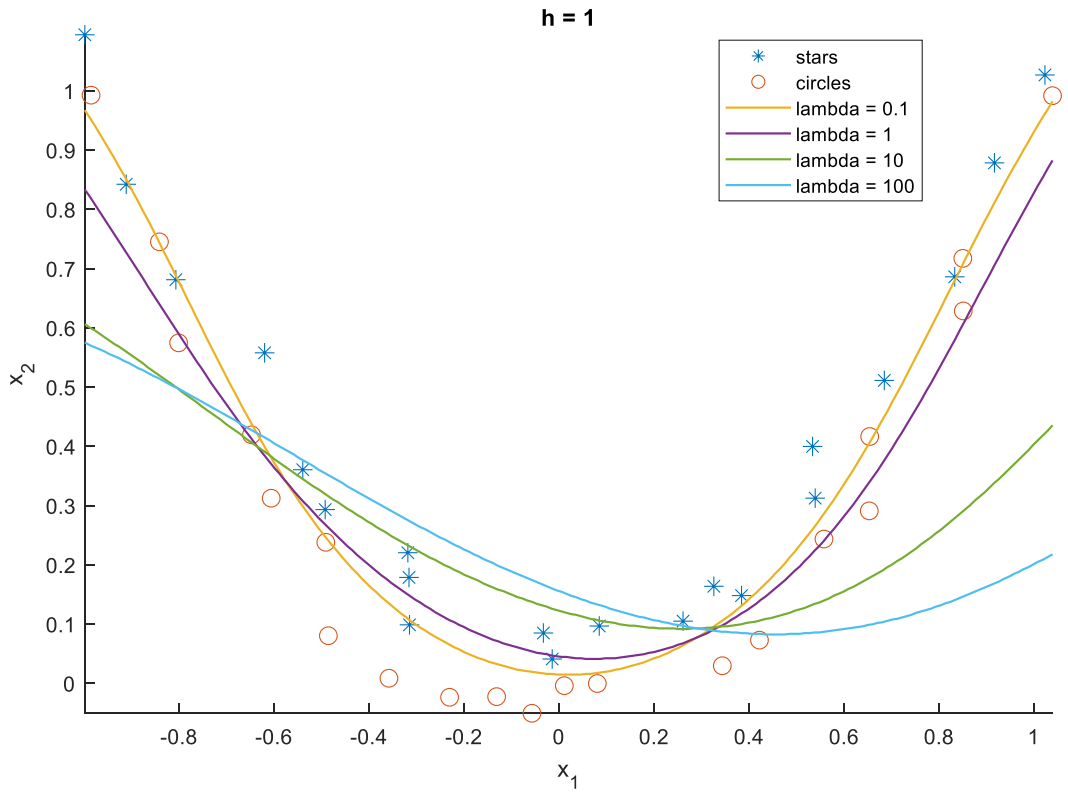
Αφού προσδιορίσαμε μέσω της εξίσωσης (4) τους βέλτιστους συντελεστές, έχουμε ορίσει πλήρως την συνάρτηση $\widehat{\Phi}(\boldsymbol{X})$ η οποία μας δίνει μια προσέγγιση του ιδανικού κατηγοριοποιητή $\Phi(\boldsymbol{X})$. Συνεπώς για να κατηγοριοποιήσουμε ένα νέο διάνυσμα δεδομένων $\boldsymbol{X_{new}}$ αρκεί να κοιτάξουμε την τιμή της $\widehat{\Phi}(\boldsymbol{X_{new}})$ και να δούμε εάν είναι κοντινότερα στο 1 ή στο -1, δηλαδή ο κανόνας κατηγοριοποίησης θα είναι:

$$class\{\boldsymbol{X_{new}}\} = sign\left(\widehat{\Phi}(\boldsymbol{X_{new}})\right)$$

Και το διαχωριστικό σύνορο φυσικά θα δίνεται από την εξίσωση $\widehat{\Phi}(\boldsymbol{X}) = 0$.

Ερώτημα (ε)

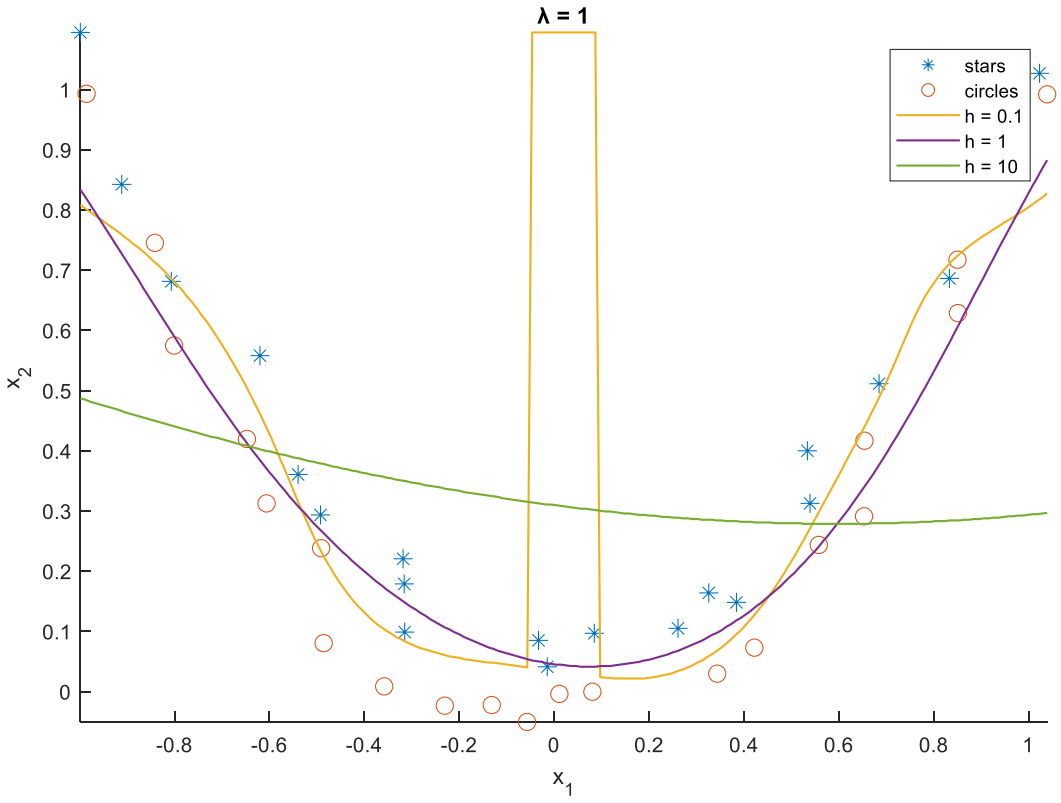
Υπολογίζοντας αριθμητικά το όριο διαχωρισμού για h=1 Και διάφορες τιμές της παραμέτρου λ έχουμε:



Σχήμα 3: Διαχωριστικό σύνορο μεταβάλλοντας την παράμετρο λ

Μπορούμε να παρατηρήσουμε πως όσο μικρότερο είναι το λ τόσο πιο κοντά στα δεδομένα εκπαίδευσης βρίσκεται το διαχωριστικό σύνορο, και εάν το μειώσουμε επαρκώς μπορεί να διαχωρίσει πλήρως τα δεδομένα αυτά. Ωστόσο αυτό δεν είναι επιθυμητό διότι θέλουμε ο κατηγοριοποιητής να μπορεί να διακρίνει και δεδομένα πέραν αυτών που χρησιμοποιήθηκαν για την εκπαίδευση. Το φαινόμενο αυτό ονομάζεται overfitting και πρέπει να αποφεύγεται όταν σχεδιάζουμε κατηγοριοποιητές.

Κρατώντας τώρα σταθερή την παράμετρο λ=1 και μεταβάλλοντας το h παίρνουμε τις παρακάτω καμπύλες:



Σχήμα 4: Διαχωριστικό σύνορο για μεταβλητό h

Μικραίνοντας το h φαίνεται η καμπύλη να μετατοπίζεται όλο και πιο κοντά στα δεδομένα εκπαίδευσης, αλλά εμφανίζονται περίεργες συμπεριφορές για μικρά h, όπως φαίνεται στην κίτρινη καμπύλη γύρω από το 0. Αυτό μπορεί ενδεχομένως να οφείλεται σε ανακρίβειες κατά την αντιστροφή του πίνακα για τον υπολογισμό των βέλτιστων συντελεστών.