4η Εργασία στην Υπολογιστική Νοημοσύνη-Classification

Δημήτριος Τικβίνας
ΑΕΜ 9998

Σεπτέμβριος 2023

dtikvina@ece.auth.gr

Εισαγωγή

Σκοπός της εργασίας αυτής είναι η διερεύνηση των ικανοτήτων των μοντέλων TSK στην επίλυση προβλημάτων ταξινόμησης (classification). Η εργασία αποτελείται από δύο μέρη, το *πρώτο* από τα οποία προορίζεται για μια απλή διερεύνηση της διαδικασίας εκπαίδευσης και αξιολόγησης των TSK μοντέλων, ενώ το *δεύτερο* περιλαμβάνει μια πιο συστηματική προσέγγιση στο πρόβλημα της εκμάθησης από δεδομένα, σε συνδυασμό με προεπεξεργαστικά βήματα όπως επιλογή χαρακτηρηστικών (feature selection) και μεθόδους βελτιστοποίησης των μοντέλων μέσω της διασταυρωμένης επικύρωσης (cross validation).

Εφαρμογή σε απλό dataset

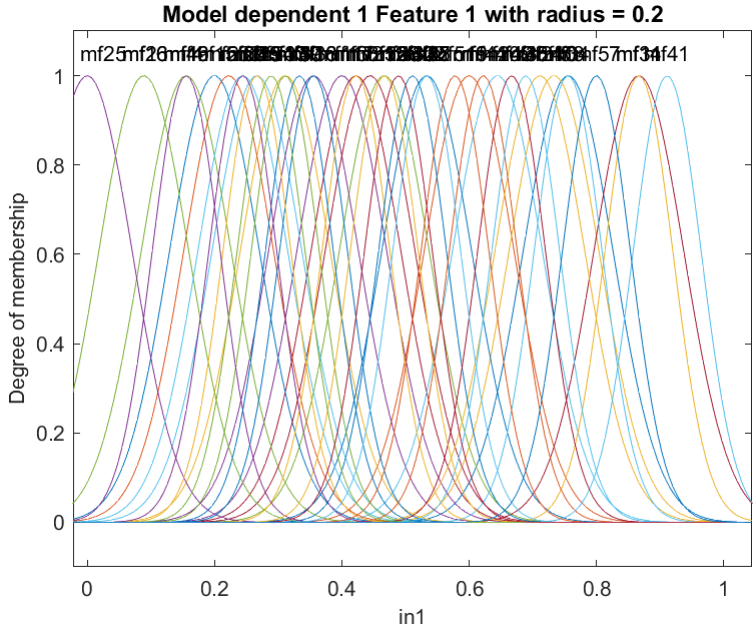
Στην πρώτη φάση της εργασίας, επιλέγεται από το UCI repository το **Haberman's Survival**, το οποίο περιλαμβάνει 306 δείγματα (instances), από 3 χαρακτηριστικά (attributes) το καθένα. Η 4η στήλη αποτελεί την target variable και περιέχει 2 κλάσεις, την 1 και την 2. Επίσης, θα εκπαιδευτούν τέσσερα μοντέλα TSK στα οποία θα μεταβάλλεται το πλήθος των ασαφών κανόνων IF-THEN. Στα μοντέλα 2 και 4, το subtractive clustering θα εκτελεστεί για όλα τα δεδομένα του συνόλου εκπαίδευσης (class independent), ενώ μοντέλα 1 και 3 ο διαμερισμός του χώρου εισόδου θα γίνει εφαρμόζοντας clustering στα δεδομένα του συνόλου εκπαίδευσης που ανήκουν στην εκάστοτε κλάση ξεχωριστά (class dependent). Επιπλέον, η παράμετρος που καθορίζει το μέγεθος των clusters και κατ' επέκταση τον αριθμό των κανόνων, είναι η τιμή της ακτίνας r των clusters. Η τιμή αυτή στα τέσσερα μοντέλα που πρέπει να υλοποιήσουμε θα πάρει 2 ακραίες τιμές όπως είναι το 0.2 και το 0.9 ανά 2 μοντέλα, ώστε τελικά ο αριθμός των κανόνων ανάμεσα στα μοντέλα να εμφανίζει σημαντική διαφορά. Περιληπτικά τα μοντέλα που υλοποιήθηκαν είναι τα παρακάτω:

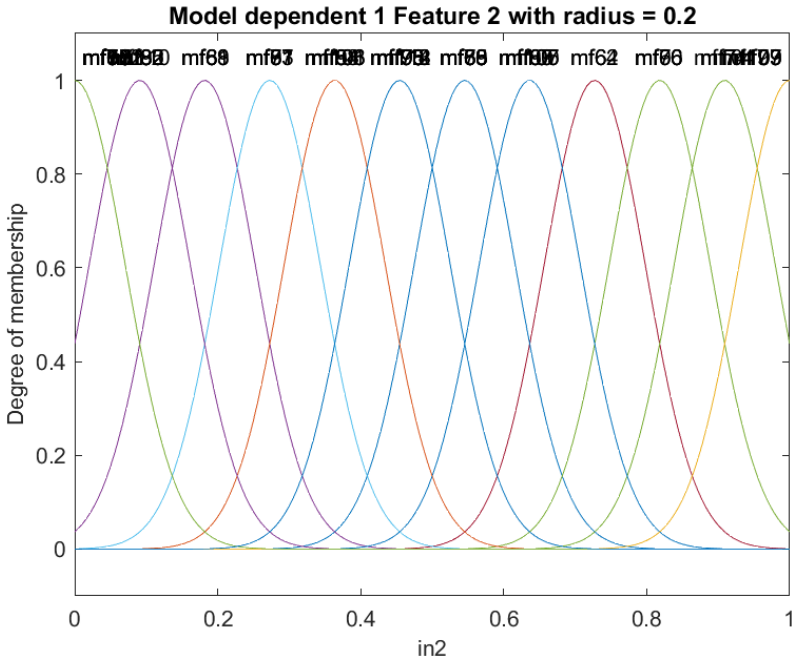
	Cluster's radius	Class (in)dependent
TSK model 1	0.2	class dependent
TSK model 2	0.9	class independent
TSK model 3	0.2	class dependent
TSK model 4	0.9	class independent

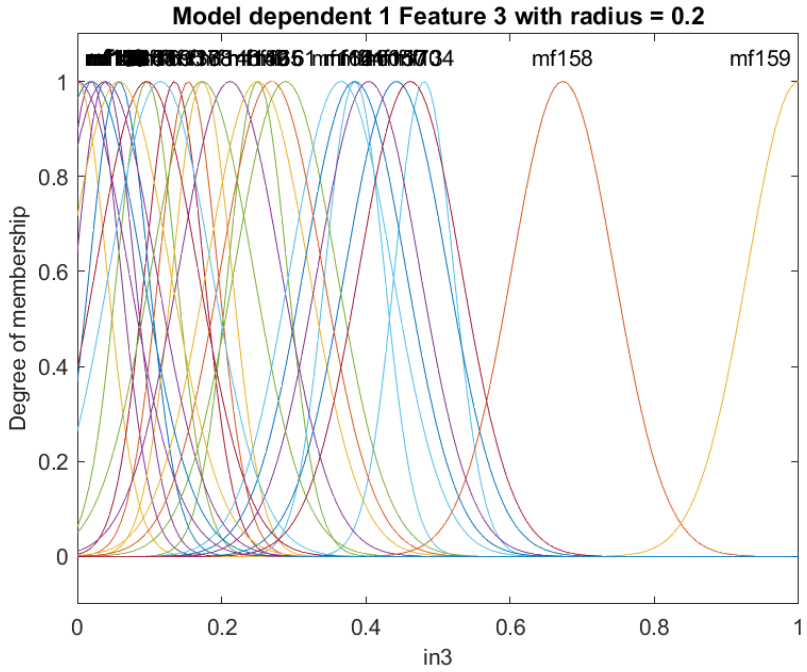
Τα αποτελέσματα και τα ζητούμενα του πρώτου μέρους παρατίθονται παρακάτω:

1ο Μοντέλο TSK

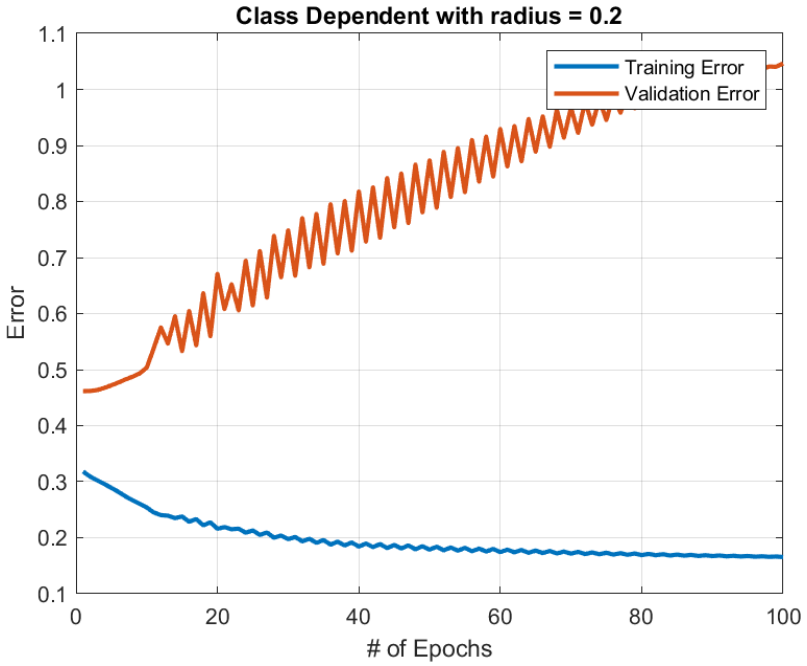
- **Συναρτήσεις συμμετοχής (membership functions)**





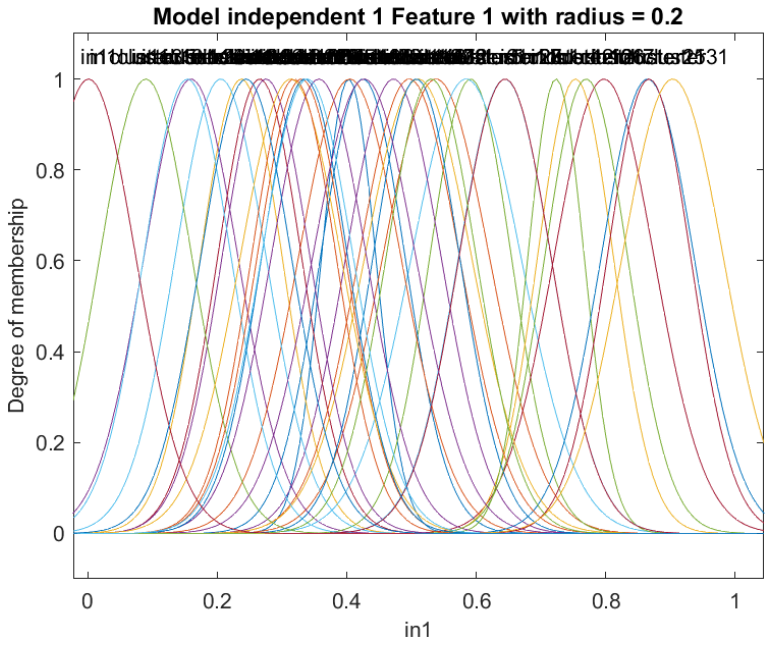


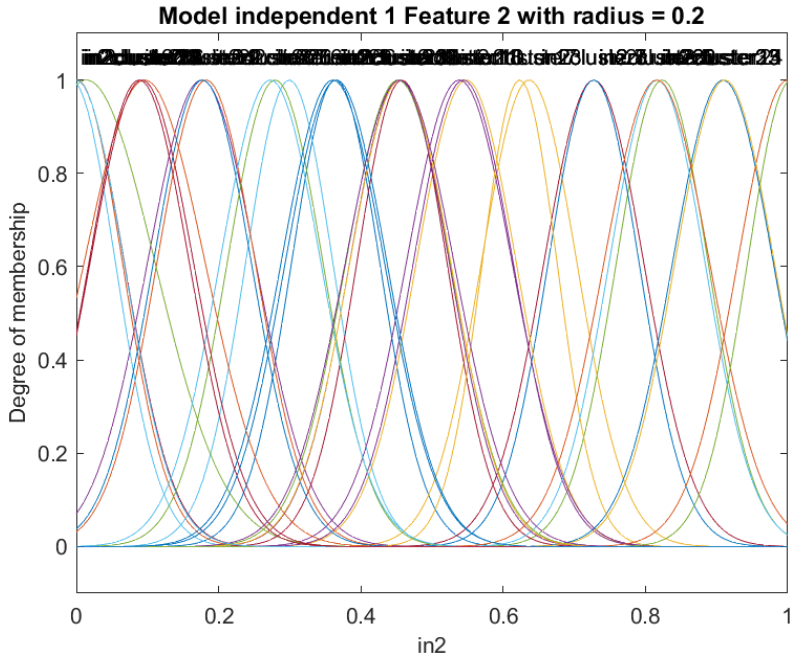
- **Καμπύλες εκμάθησης (learning curves)**

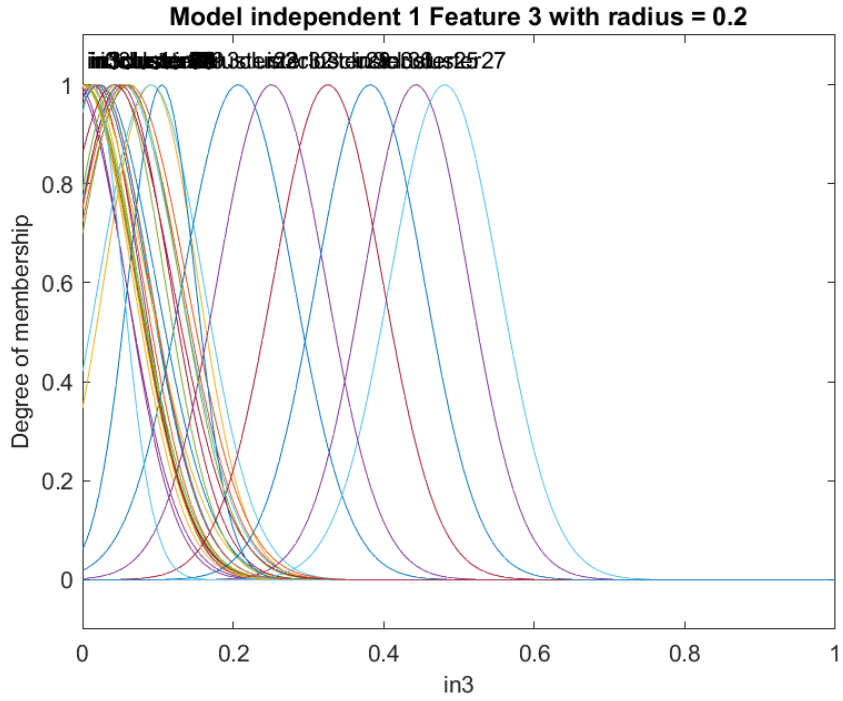


2ο Μοντέλο TSK

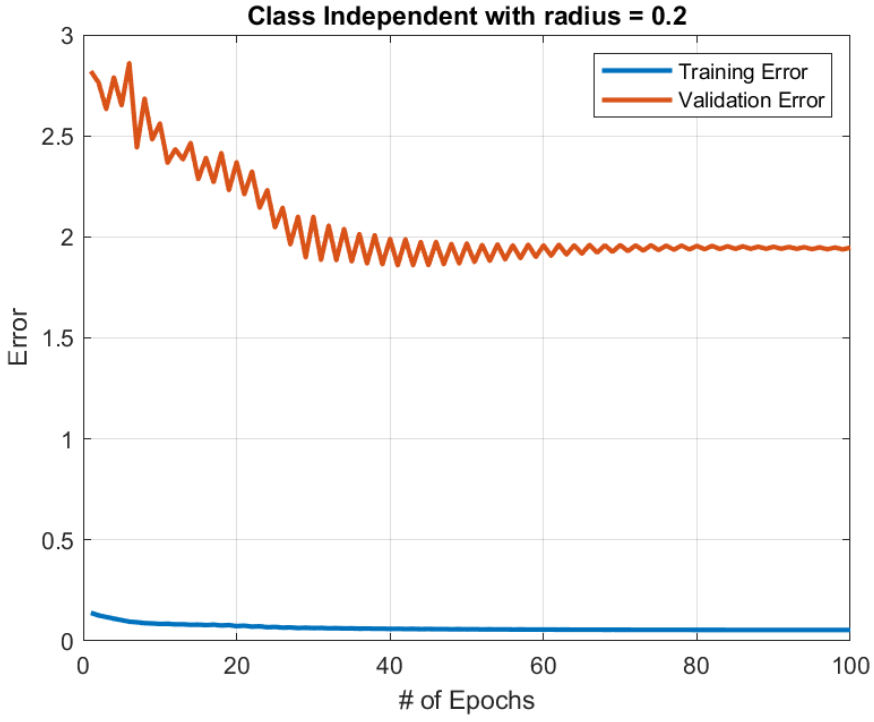
- **Συναρτήσεις συμμετοχής (membership functions)**





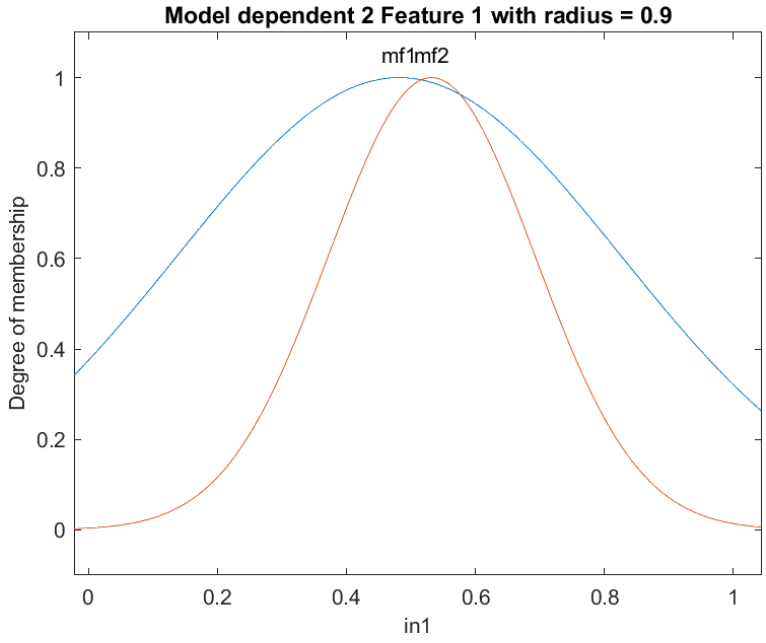


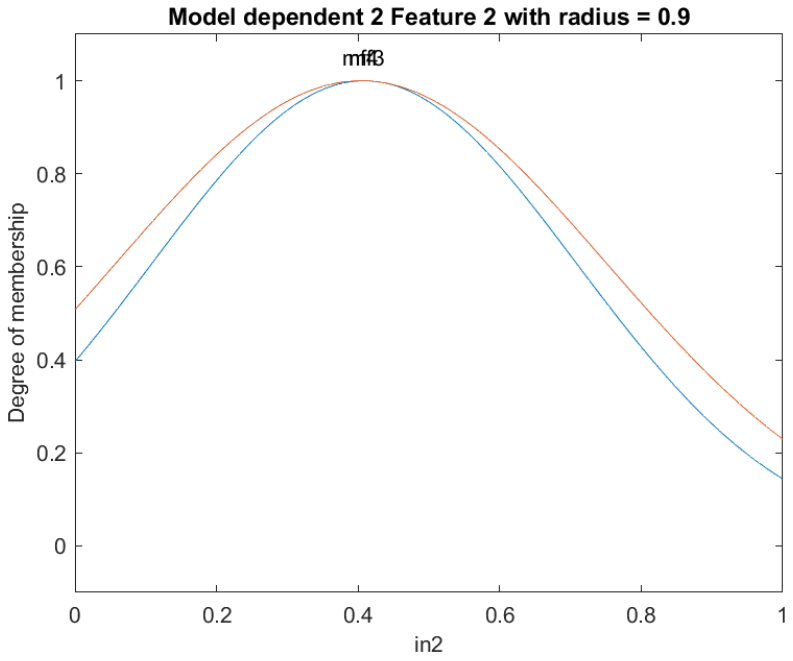
- **Καμπύλες εκμάθησης (learning curves)**

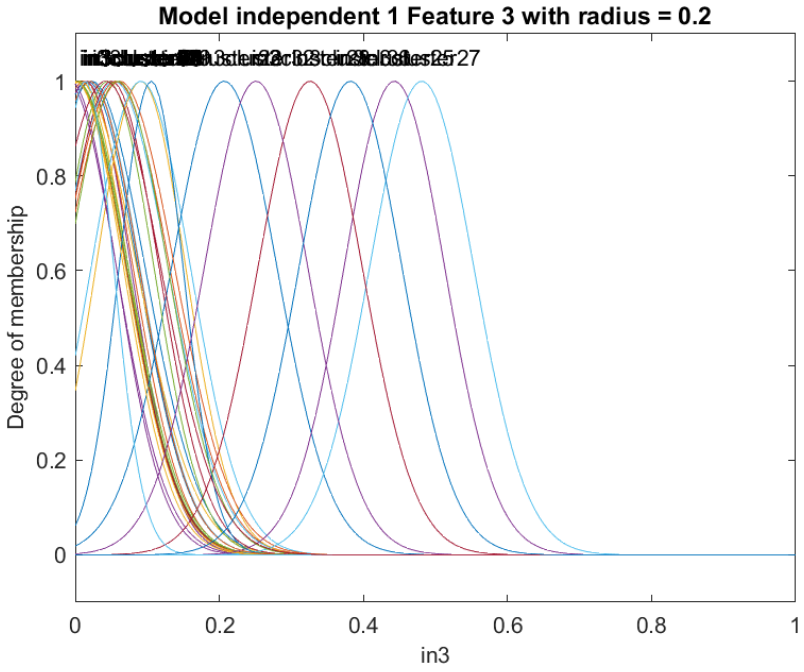


3ο Μοντέλο TSK

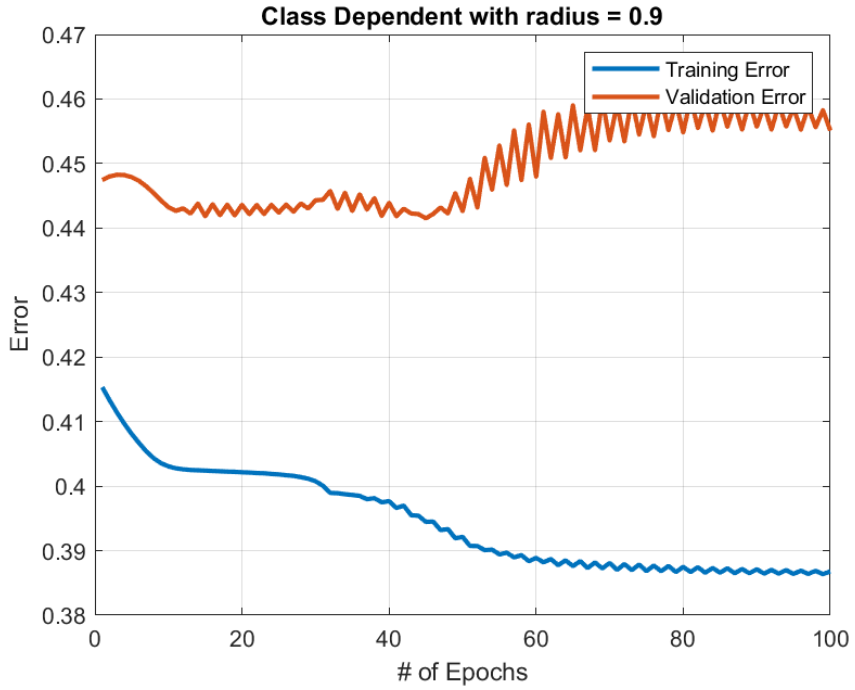
- **Συναρτήσεις συμμετοχής (membership functions)**





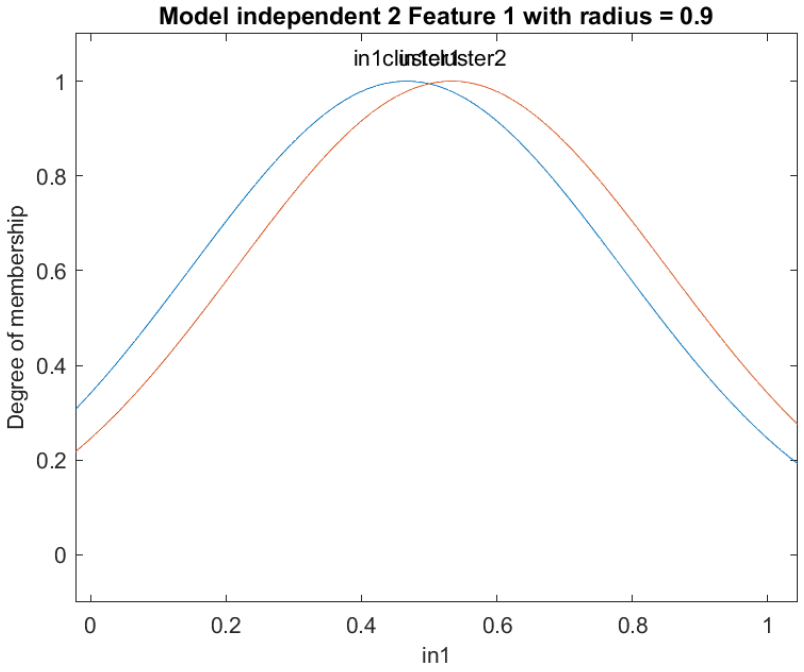


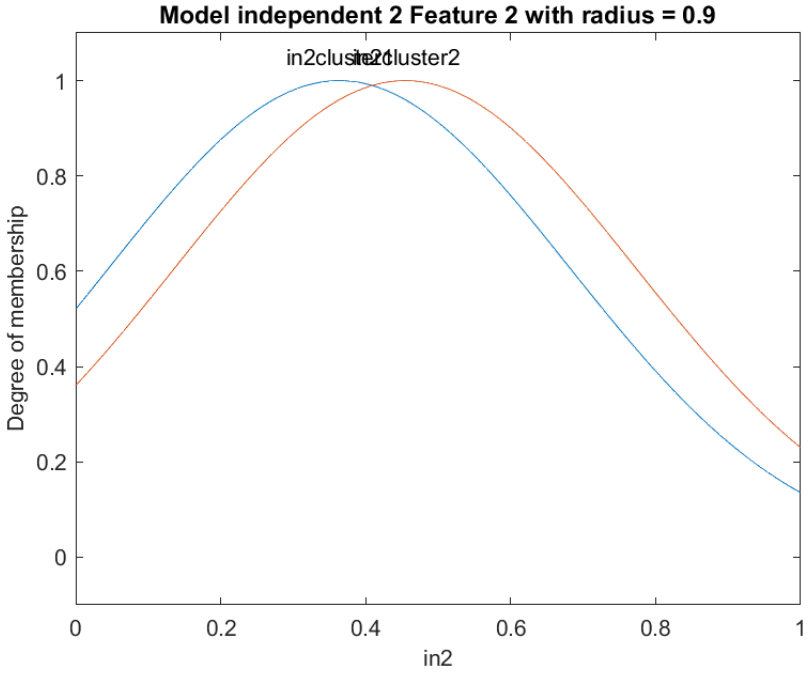
- **Καμπύλες εκμάθησης (learning curves)**

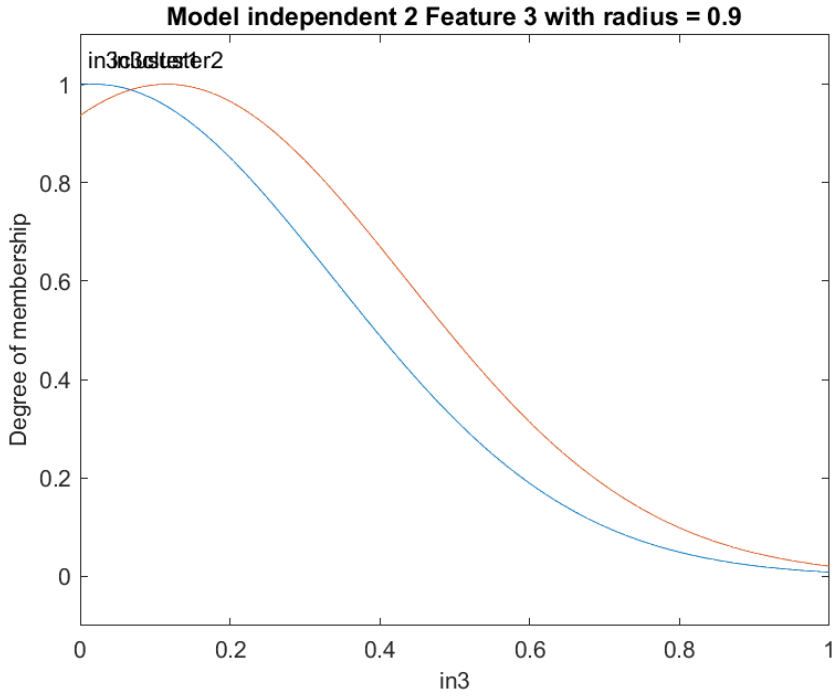


4ο Μοντέλο TSK

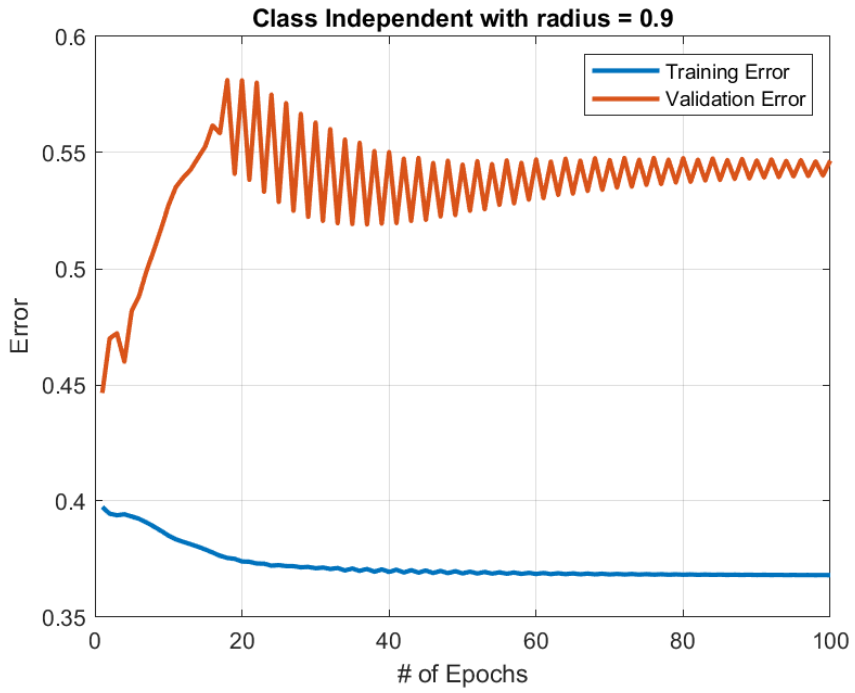
- **Συναρτήσεις συμμετοχής (membership functions)**







- **Καμπύλες εκμάθησης (learning curves)**



Μετρικές των παραπάνω μοντέλων και πίνακες σφαλμάτων

- Μετρικές

	OA	PA 1	PA 2	UA 1	UA 2	k
Model 1	0.6230	0.7674	0.2778	0.7174	0.3333	0.0475
Model 2	0.5574	0.6512	0.3333	0.7000	0.2857	0.0148
Model 3	0.7049	0.8837	0.2778	0.7451	0.5000	0.1855
Model 4	0.6885	0.9302	0.1111	0.7143	0.4000	0.0523

- **Πίνακες σφάλματος**

Model 1	**Predicted Class**	
True Class	**33**	**10**
	13	**5**

Model 2	**Predicted Class**	
True Class	**28**	**15**
	12	**6**

Model 3	**Predicted Class**	
True Class	**38**	**5**
	13	**5**

Model 4	**Predicted Class**	
True Class	**40**	**3**
	16	**2**

- Οι παραπάνω υπολογισμοί πραγματοποιήθηκαν με βάση τον παρακάτω πρότυπο πίνακα για μέτρησης σφαλμάτων:

Model	**Predicted Class**	
True Class	**TP**	**FN**
	FP	**TN**

Όπου:

OA = (TP + TN) / (TP + FP + TN + FN)

PA1 = TP / (FN + TP)

PA2 = TN / (FP + TN)

UA1 = TP / (TP + FP)

UA2 = TN / (TN + FN)

- Αριθμός των κανόνων για κάθε μοντέλο:

	Αριθμός κανόνων
Μοντέλο 1	57
Μοντέλο 2	36
Μοντέλο 3	2
Μοντέλο 4	2

Συμπεράσματα

Από την παραπάνω ανάλυση των επιδόσεων του κάθε μοντέλου πάνω στο dataset μας, καταλήγουμε στο ότι, συνολικά, το μοντέλο **3** είναι το καλύτερο. Έχει τα υψηλότερα OA, UA1, UA2 και K και καλές τιμές PA1 και PA2 . Γενικότερα, καταδεικνύεται από τα αποτελέσματα η λειτουργικότητα του διαμερισμού του χώρου εισόδου εφαρμόζοντας clustering στα δεδομένα του συνόλου εκπαίδευσης που ανήκουν στην κάθε κλάση ξεχωριστά (class dependent). Κρατώντας την ακτίνα των clusters σταθερή, βλέπουμε ότι το μοντέλο 1 και το μοντέλο 3 έχουν υψηλότερο OA από ότι των μοντέλων 2 και 4 αντίστοιχα. Επιπλέον, βλέπουμε ότι δεν υπάρχει κάποια εξάρτηση της απόδοσης των μοντέλων από το πλήθος των κανόνων που έχουν. Παρατηρούμε, επιπρόσθετα, συγκρίνοντας ανά δύο τα μοντέλα 1-3 και 2-4, ότι όσο μειώνεται η ακτίνα των clusters, τόσο περισσότερους κανόνες έχει το μοντέλο TSK. Όσους περισσότερους κανόνες έχει, τόσο πιο περίπλοκο γίνεται και τόσο πιο εύκολα βαίνει σε overtraining. Η αιτία αυτού βρίσκεται στην υπεραριθμία των συναρτήσεων συμμετοχής, οι οποίες μεγιστοποιούν την ακρίβεια στο training set και έτσι χάνεται η γενίκευση. Τέλος, βλέπουμε ότι η υπερεπικάλυψη των

ασαφών συνόλων από τα μοντέλα 1 και 2 με την αφθονία αυτή σε συναρτήσεις συμμετοχής δρα αρνητικά στην OA, συγκριτικά με τα μοντέλα 3 και 4.

Μία μέθοδος που προτείνεται για την βελτίωση του τμήματος υπόθεσης είναι η σταδιακή αφαίρεση mfs που δέχονται υψηλή επικάλυψη, μετά την υλοποίηση και την εκπαίδευση του μοντέλου, με σκοπό την απλοποίση του μοντέλου και την επίτευξη της επιθυμητής του γενίκευσης.

Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στο δεύτερο μέρος της εργασίας θα χρησιμοποιηθεί ένα dataset με υψηλότερο βαθμό διαστασιμότητας. Το dataset αυτό είναι από το UCI repository το Epileptic Seizure Recognition dataset, το οποίο περιέχει 11500 δείγματα, καθένα από τα οποία περιγράφεται από 178 features με την στήλη No. 179 να είναι το Target column, έχοντας τη κλάση του κάθε δείγματος. Ο μεγάλος αριθμός μεταβλητών καθιστά αναγκαία τη χρήση μεθόδων μείωσης της διαστασιμότητας καθώς και του αριθμού των IF – THEN κανόνων. Συνεπώς, θα κάνουμε αναζήτηση πλέγματος για την εύρεση των βέλτιστων τιμών των παραμέτρων kept_features και radius, που είναι ο αριθμός των κρατημένων attributes και της ακτίνας r των clusters αντίστοιχα. Οι δοκιμές που κάναμε ήταν για 5, 7, 9 και 11 κρατημένα features και για ακτίνα r ήταν 0.2, 0.4, 0.6, 0.8 και 1. Οι επιλογές αυτές των τιμών έγιναν ώστε να που μπορούν πρακτικά να υπολογιστούν τα αποτελέσματα τους, σε έναν ικανοποιητικά βραχύ χρόνο. Τελικά, η μετρική που χρησιμοποιήθηκε για να γίνει η αξιολόγηση των δοκιμαστικών μοντέλων με τις παραπάνω τιμές των παραμέτρων, ήταν η μετρική του Overall accuracy.

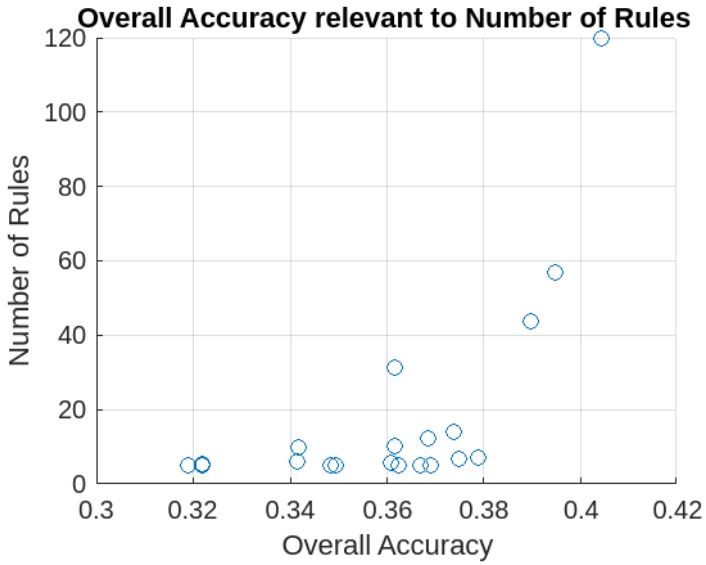
Παρακάτω, παρουσιάζονται τα αποτελέσματα και τα ζητούμενα της παραπάνω διαδικασίας.

- Πίνακας OA

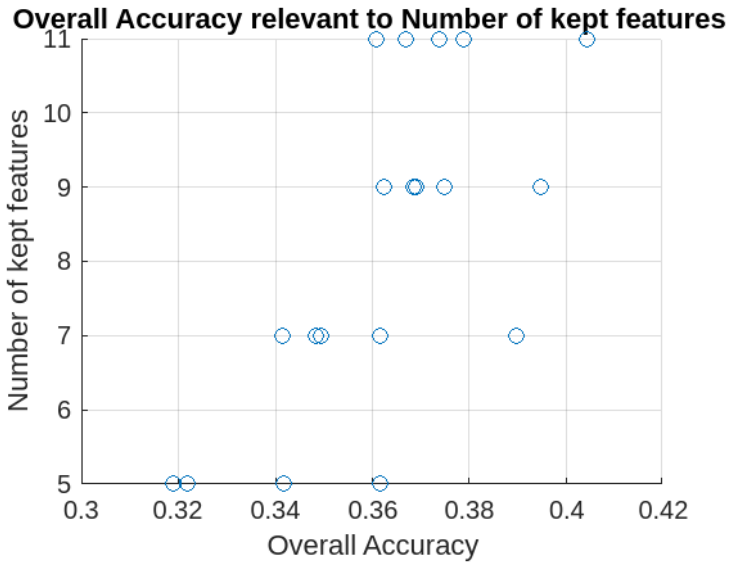
	0.2	**0.4**	**0.6**	**0.8**	**1**
5	0.3615	0.3417	0.3219	0.3219	0.3190
7	0.3897	0.3617	0.3414	0.3484	0.3495
9	0.3946	0.3686	0.3749	0.3691	0.3624
11	0.4043	0.3737	0.3788	0.3607	0.3670

(γραμμές: πλήθος κρατημένων features, στήλες: ακτίνα των clusters)

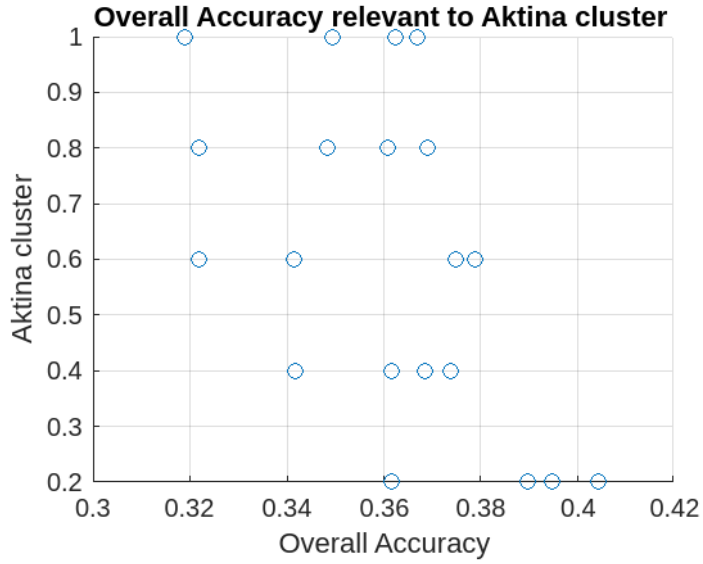
- OA συναρτήσει του πλήθους των κανόνων



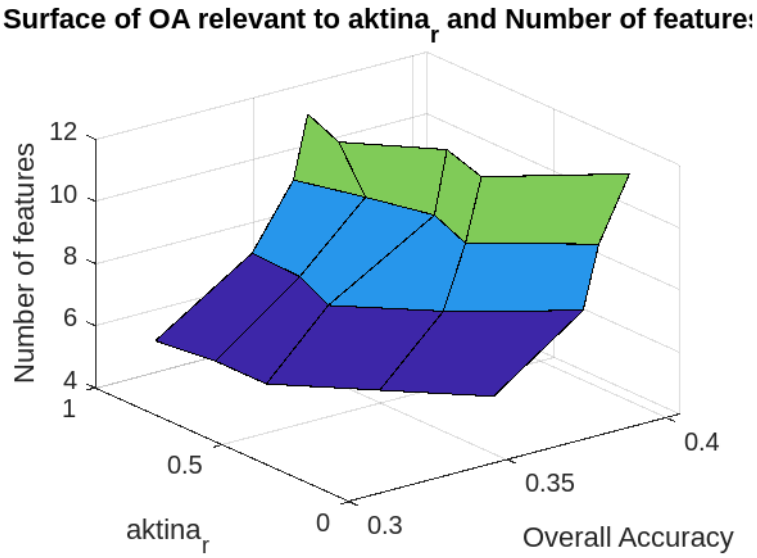
- ΟΑ συναρτήσει του features που κρατήθηκαν



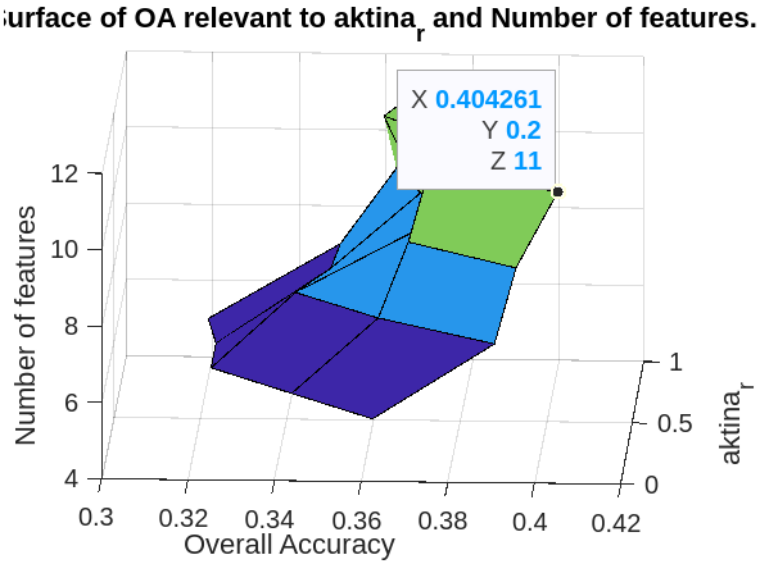
- ΟΑ συναρτήσει της ακτίνας των clusters



- Επιφάνεια ΟΑ συναρτήσει του kept_features και του radius



για τον εντοπισμό των παραμέτρων εκείνων που μεγιστοποιούν την ΟΑ:



- Πίνακας με τον αριθμό των κανόνων μετά το cross validation

	0.2	**0.4**	**0.6**	**0.8**	**1**
5	31.2	9.8	5.4	5	5
7	43.8	10.2	6	5	5
9	56.8	12.4	6.6	5	5
11	119.8	13.8	7	5.6	5

Συμπεράσματα

Από τα παραπάνω αποτελέσματα συμπεραίνουμε ότι όσο αυξάνουμε το πλήθος των features που κρατάμε και όσο μείωνουμε την ακτίνα επιρροής των clusters, τόσο και αυξάνεται η μετρική αξιολόγησης μας, η Overall Accuracy. Αυτό βέβαια δεν σημαίνει απαραίτητα πως η συνεχόμενη αναπροσαρμογή αυτών των δύο

παραμέτρων οδηγεί και σε καλύτερο μοντέλο, καθώς κινδυνεύουμε να οδηγηθούμε σε overfitting στο training set. Αυτό φαίνεται και από την εκθετική αύξηση του συνόλου των κανόνων στον πίνακα μετά το cross validation, όπου περισσότεροι κανόνες σημαίνει μεγαλύτερη προσαρμογή στο σύνολο εκπαίδευσης, κάτι γενικά ανεπιθύμητο λόγω εξάλειψης της γενικότητας. Από το surf plot, καταλήγουμε ότι την βέλτιστη τιμή του OA την επιτυγχάνουμε όταν kept_features = 11 και radius = 0.2, με βάση τις τιμές που επιλέχτηκαν για τις προσομοιώσεις.

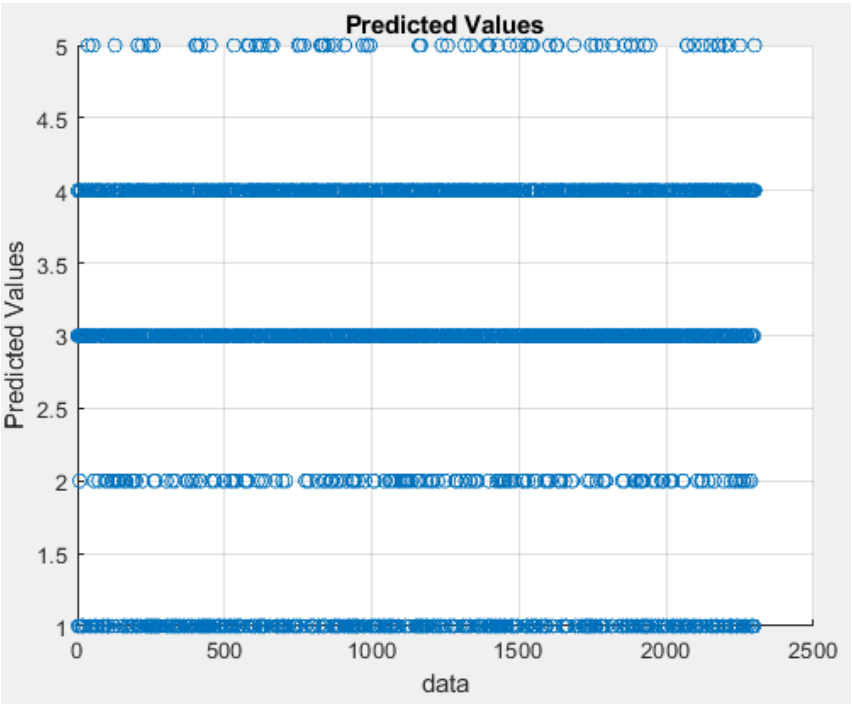
Βέλτιστο Μοντέλο TSM με kept_features = 11 και radius = 0.2

Τα αποτελέσματα και τα ζητούμενα της διαδικασίας αυτής παρατίθονται παρακάτω

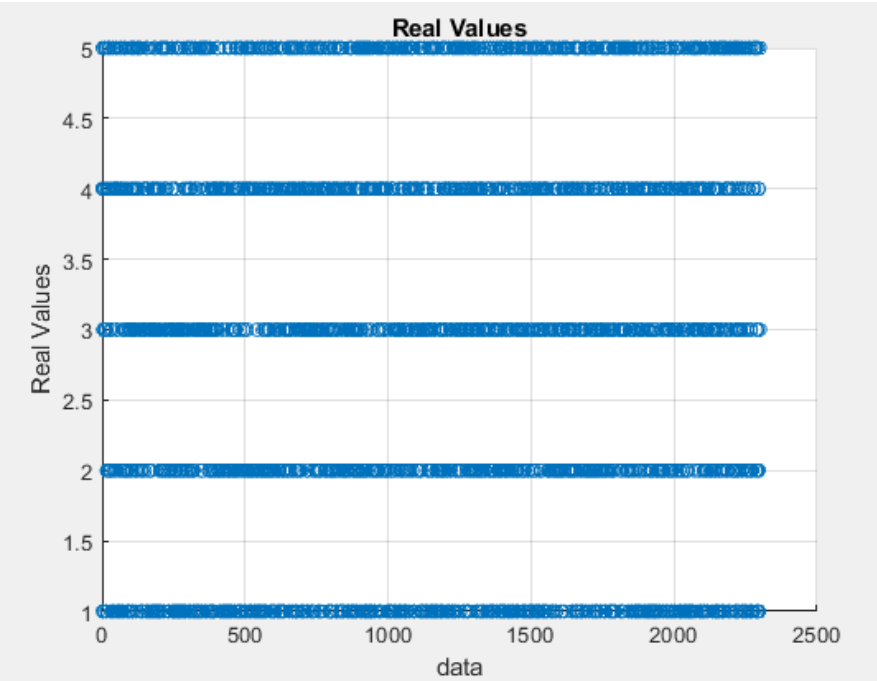
- Μέσοι δείκτες απόδοσης του βέλτιστου μοντέλου

	OA	PA1	PA2	PA3	PA4	PA5	UA1	UA2	UA3	UA4	UA5	K	rule
έλτιστο οντέλο	0.4016	0.7552	0.0996	0.5922	0.5104	0.0504	0.9192	0.2430	0.3007	0.2994	0.5257	0.2520	107.

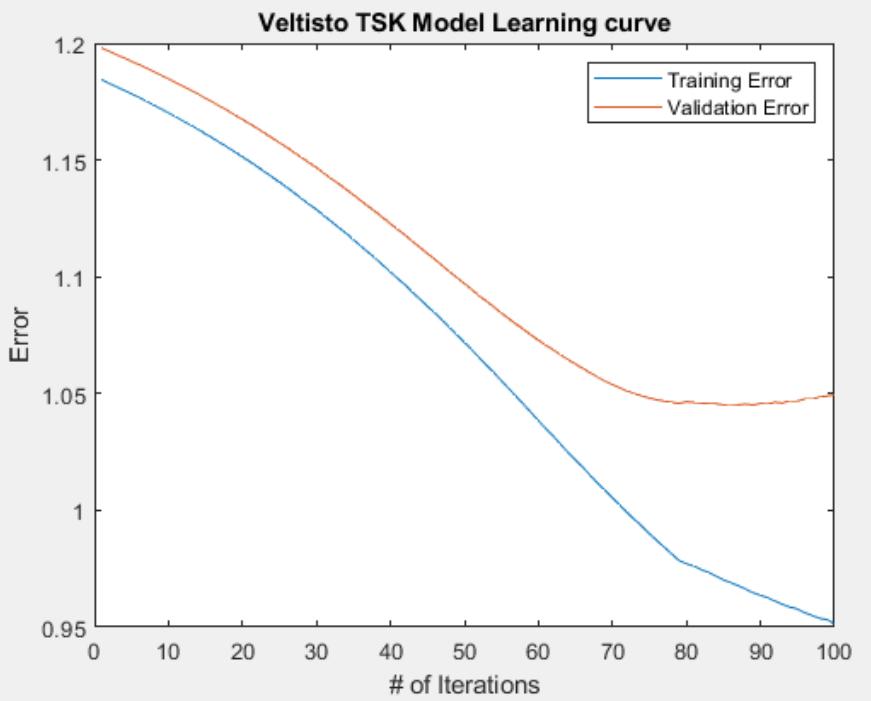
- Προβλέψεις τελικού μοντέλου



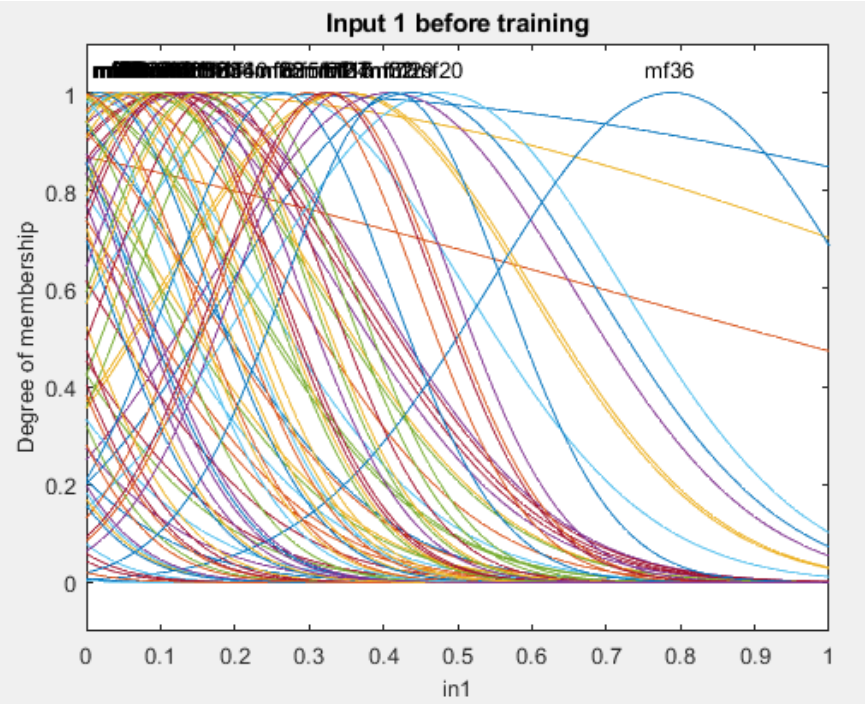
- Πραγματικές τιμές του dataset

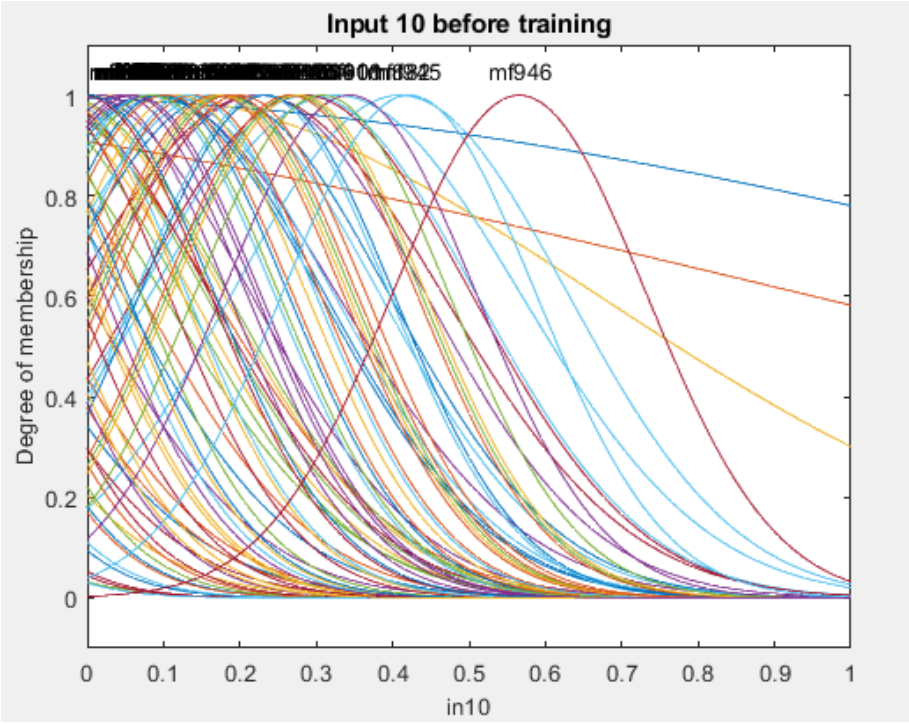


- Learning curve

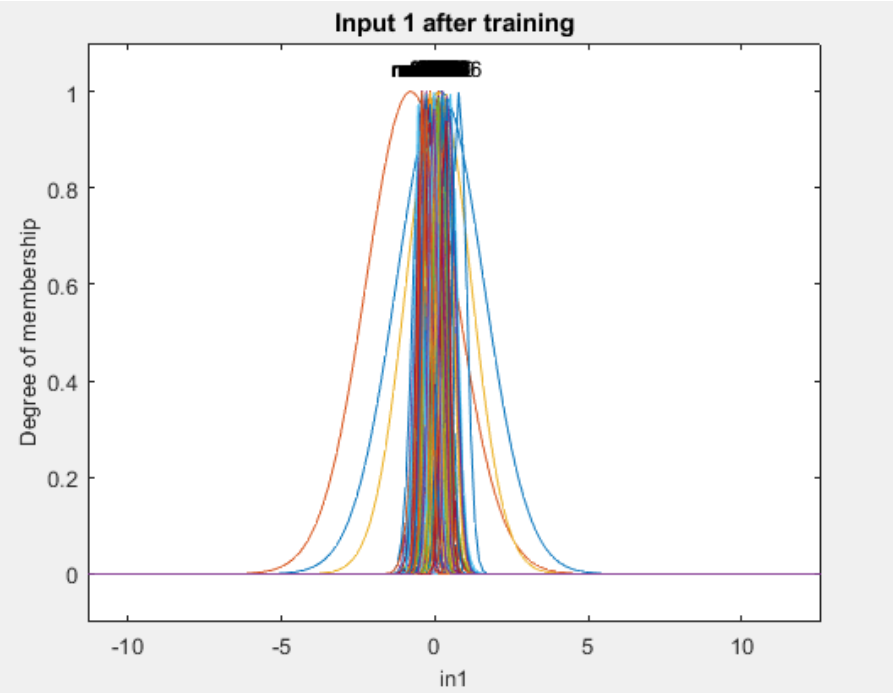


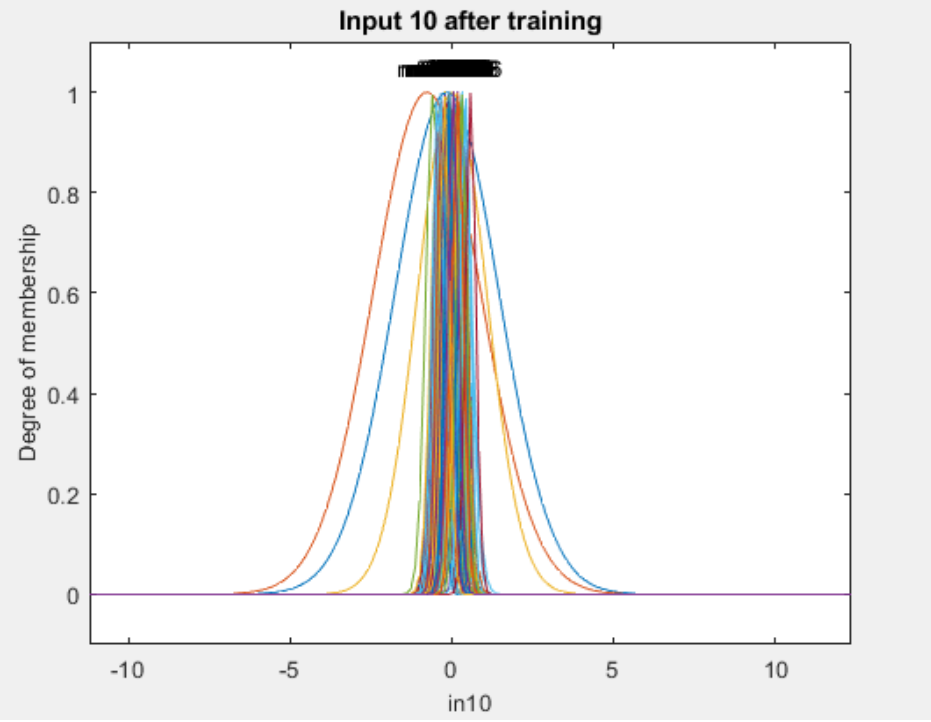
- Ασαφή σύνολα πριν την εκπαίδευση





- Αντίστοιχα ασαφή σύνολα μετά την εκπαίδευση





- Μέσος πίνακας σφαλμάτων classification

True Class	Predicted class				
	337	59	49	15	0
	38	51	247	122	2
	1	37	291	124	7
	2	39	164	247	8
	0	12	159	2559	30

Συμπεράσματα για το Βέλτιστο Μοντέλο

Όπως αναφέραμε και πριν, το βέλτιστο μοντέλο που βρέθηκε από την αναζήτηση πλέγματος (grid search), ήταν για τις τιμές των παραμέτρων kept_features = 11 και radius = 0.2. Από τον πίνακα με τους μέσους δείκτες απόδοσεις μπορούμε να εξάγουμε τα εξής συμπεράσματα:

1. Αν ένα δείγμα ανήκει στην κλάση 1, τότε έχει 75.52% πιθανότητα να ταξινομηθεί σωστά. Ομοίως και για τις υπόλοιπες 4 κλάσεις.
2. Η πιθανότητα να ανήκει ενα δείγμα στην κλάση 1, εφόσον έχει ταξινομηθεί από το βέλτιστο μοντέλο στην κλάση 1 είναι 91.92%. Ομοίως για τις υπόλοιπες 4 κλάσεις.

Με βάση τις δύο αυτές μετρικές (PA και UA), το βέλτιστο μοντέλο προβλέπειαρκετά καλά τις κλάσεις 1 και 3, μέτρια την κλάση 4, και αρκετά άσχημα τις κλάσεις 2 και 5. Επίσης, ο μέσος αριθμός ασαφών κανόνων είναι 107.2 ~ 108, ένα αριθμός σημαντικά μικρότερος από αυτόν που θα προέκυπτε με grid partitioning με δύο ή τρία ασαφή σύνολα ανά είσοδο, δηλαδή 2^{11} ή 3^{11} κανόνες. Καταδεικνύεται έτσι η απαραίτητη επιλογή του subtractive clustering έναντι του grid partitioning. Τέλος, παρατηρούμε μεγάλη επικάλυψη ανάμεσα στην πληθώρα των ασαφών κανόνων, κάτι που σημαίνει ότι δεν είναι αναγκαία.