Μέρος 1

Το επιλεγμένο σύνολο δεδομένων περιέχει περιπτώσεις από μια μελέτη που διεξήχθη μεταξύ 1958 και 1970 στο Νοσοκομείο Billings του Πανεπιστημίου του Σικάγο σχετικά με την επιβίωση των ασθενών που είχαν υποβληθεί σε χειρουργική επέμβαση για καρκίνο του μαστού. Περιλαμβάνει 306 δείγματα και κάθε ένα από αυτά εμφανίζει τα ακόλουθα 3 χαρακτηριστικά:

1. Ηλικία του ασθενούς κατά τη διάρκεια της επέμβασης (αριθμητική)
2. Έτος λειτουργίας του ασθενούς (έτος - 1900, αριθμητικό)
3. Αριθμός ανιχνευμένων θετικών μασχαλιαίων αδένων

Μετά την αναδιάταξη του συνόλου δεδομένων με τυχαίο τρόπο και το διαμερισμό του σε μη επικαλυπτόμενα σύνολα εκπαίδευσης, επικύρωσης και ελέγχου, προχωράμε στη διαδικασία δημιουργίας των τεσσάρων μοντέλων, όπως αυτά ζητούνται από την εργασία.

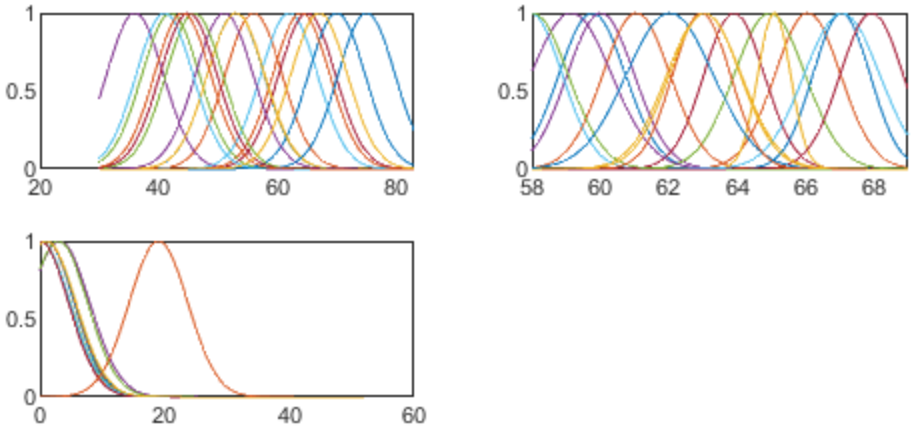
Ζητείται να δημιουργηθούν δύο ζεύγη μοντέλων, όπου κάθε ζεύγος διέπεται από το ίδιο πλήθος κανόνων βάσης, αλλά από διαφορετική μέθοδο διαμέρισης του χώρου εισόδου. Έτσι, σε κάθε ζεύγος μοντέλων, στο ένα μοντέλο, ο χώρος εισόδου διαμερίζεται με τη μέθοδο του subtractive clustering σε όλα τα δεδομένα του συνόλου εκπαίδευσης (class independent), ενώ στο άλλο μοντέλο, ο χώρος εισόδου διαμερίζεται με τη μέθοδο του subtractive clustering, αλλά αυτή τη φορά το clustering εφαρμόζεται στα δεδομένα του συνόλου εκπαίδευσης που ανήκουν στην κάθε κλάση ξεχωριστά (class dependent).

Για τα δύο ζεύγη μοντέλων, όσον αφορά το πλήθος των κανόνων, επιλέγονται οι τιμές 12 και 60 και αντίστοιχα υπολογίζονται οι τιμές των ακτινών των clusters για κάθε μοντέλο. Τα αποτελέσματα της διαδικασίας αυτής παρουσιάζονται στον παρακάτω πίνακα.

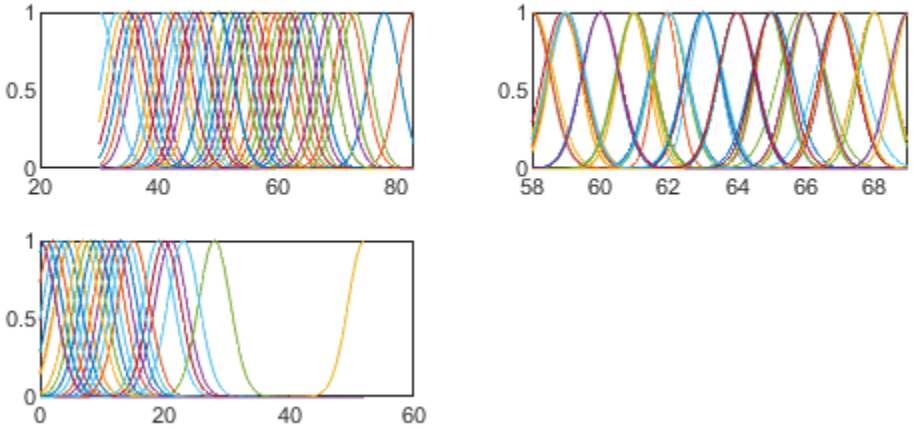
	Μοντέλα με class independent clustering		**Μοντέλα με class dependent clustering**	
	Μοντέλο 1	**Μοντέλο 2**	**Μοντέλο 3**	**Μοντέλο 4**
πλήθος κανόνων	12	60	12	60
ακτίνα clusters	0.255	0.138	0.414	0.202

Μετά την εκπαίδευση των μοντέλων με τη χρήση μια υβριδικής μεθόδου, κατά την οποία οι παράμετροι των συναρτήσεων συμμετοχής και της πολυωνυμικής συνάρτησης βελτιστοποιούνται με τις μεθόδους του back propagation και των ελαχίστων τετραγώνων εξόδου αντίστοιχα, παρουσιάζονται οι συναρτήσεις συμμετοχής των εισόδων που προέκυψαν στα παρακάτω σχήματα.

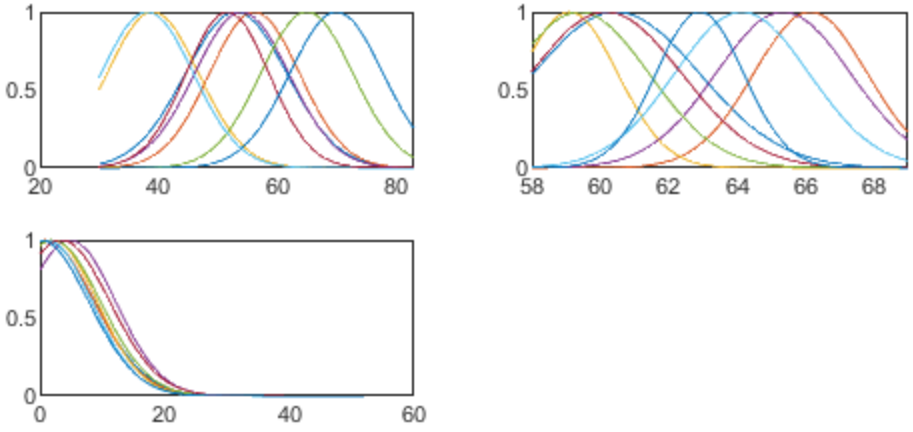
TSK model 1: Trained membership functions

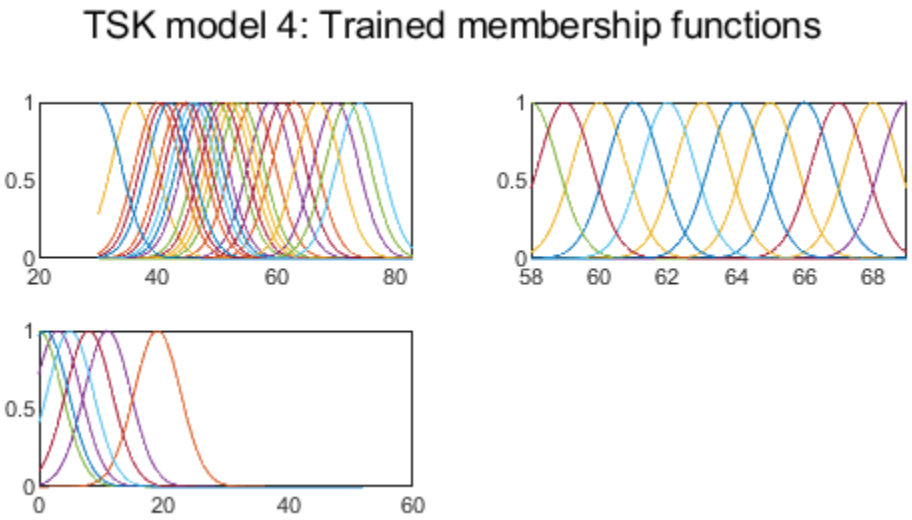


TSK model 2: Trained membership functions



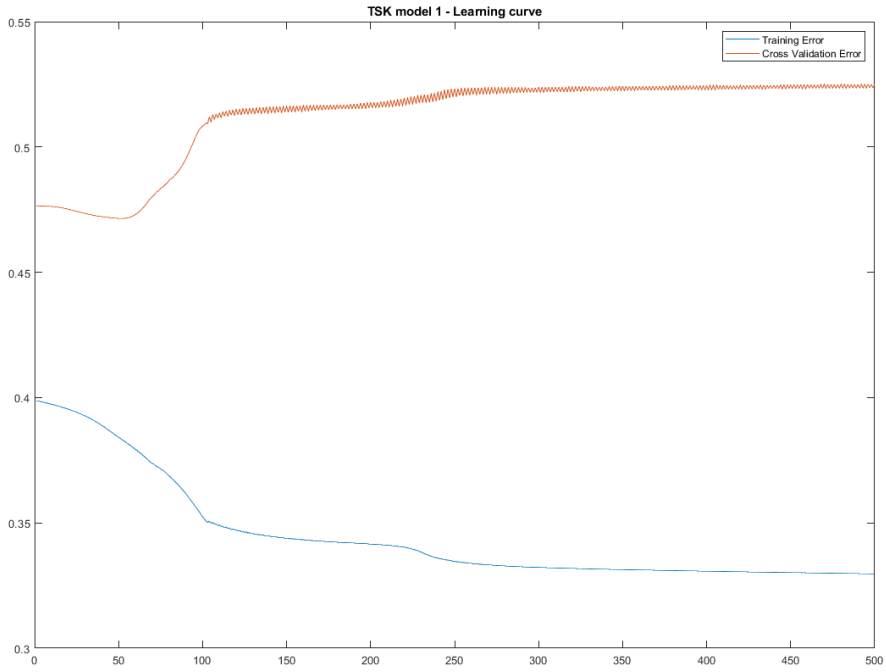
TSK model 3: Trained membership functions

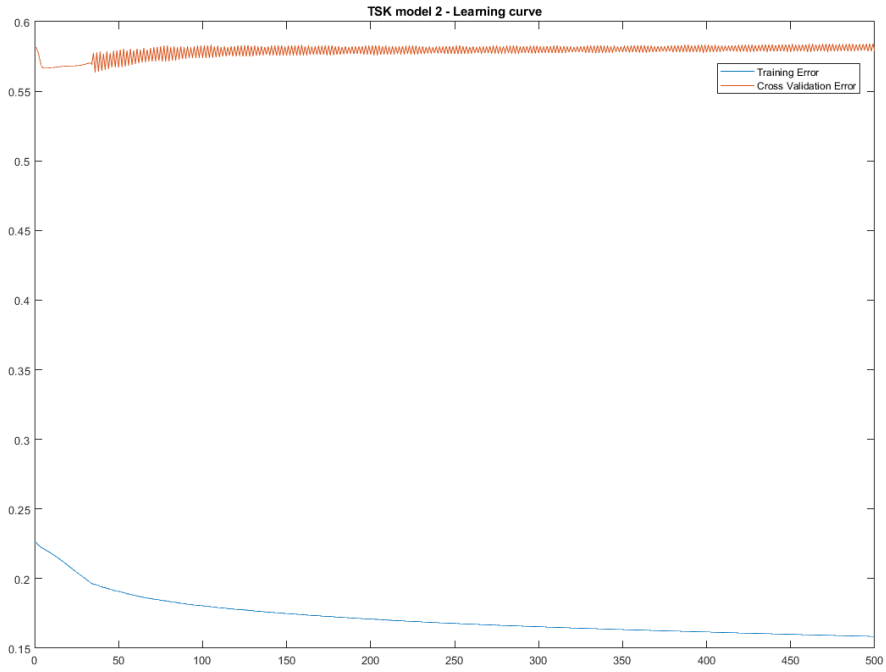


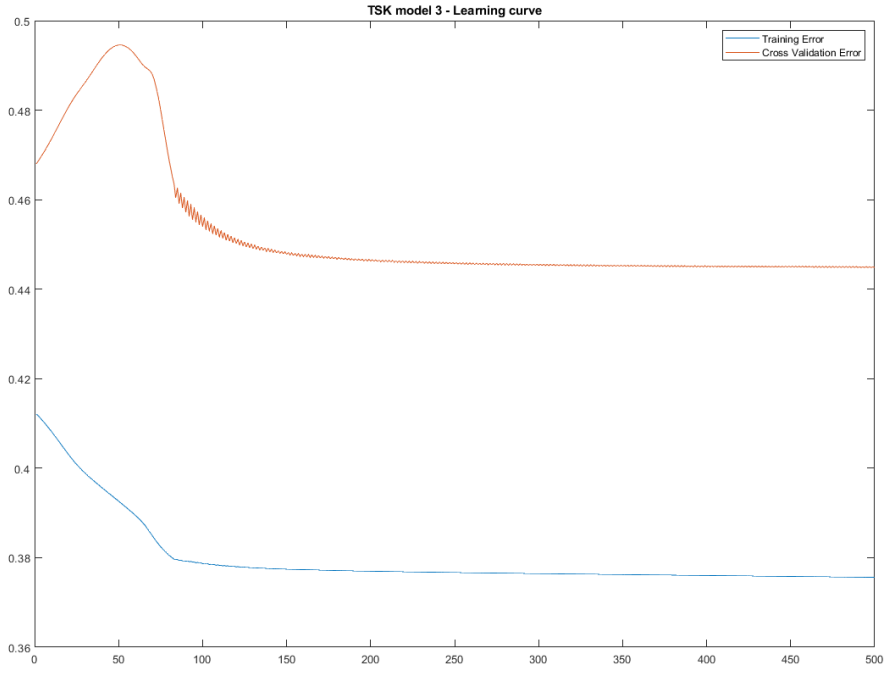


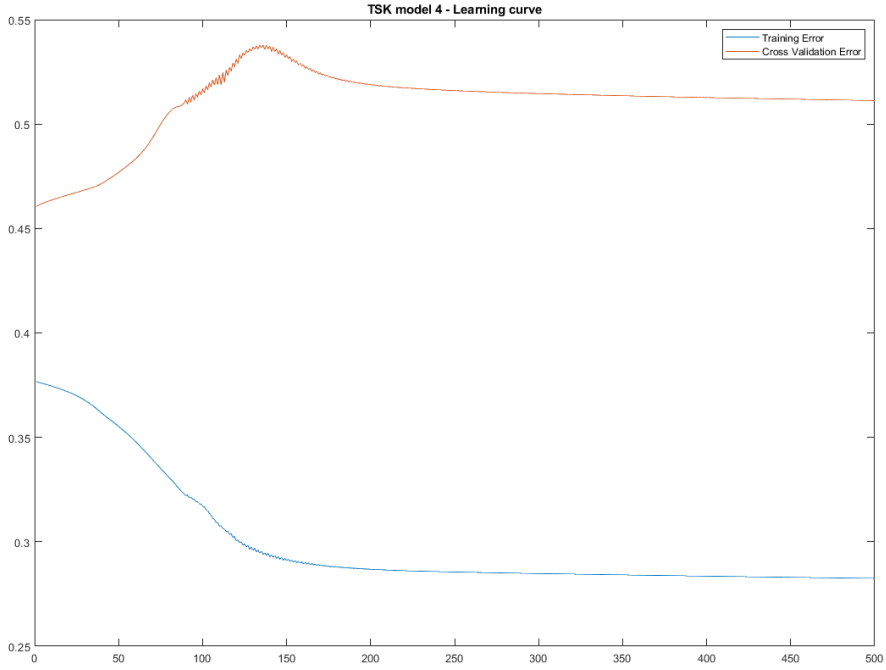
Από τα διαγράμματα των συναρτήσεων συμμετοχής των μοντέλων φαίνεται πως για ίδιο αριθμό κανόνων, η διαμέριση του χώρου εισόδου με τη μέθοδο του subtractive clustering στα δεδομένα του συνόλου εκπαίδευσης που ανήκουν στην κάθε κλάση ξεχωριστά (class dependent) οδηγεί σε συναρτήσεις συμμετοχής που είναι αρκετά σαφέστερα ορισμένες και δεν έχουν τόσο μεγάλη επικάλυψη όσο το αντίστοιχο μοντέλο, όπου για τη διαμέριση του χώρου εισόδου εφαρμόζεται subtractive clustering στο σύνολο των δεδομένων του συνόλου εκπαίδευσης.

Παρακάτω παρουσιάζονται και οι καμπύλες εκμάθησης των τεσσάρων μοντέλων.




TSK model 2 - Learning curve
Training Error
Cross Validation Error
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0
50
100
150
200
250
300
350
400
450
500


TSK model 3 - Learning curve
Training Error
Cross Validation Error
0.5
0.48
0.46
0.44
0.42
0.4
0.38
0.36
0
50
100
150
200
250
300
350
400
450
500



Τέλος, παρουσιάζεται ο πίνακας σφαλμάτων ταξινόμησης και ο πίνακας των δεικτών απόδοσης για κάθε μοντέλο.

Μοντέλο 1	
38	13
7	4

Μοντέλο 2	
31	10
14	7

Μοντέλο 3	
39	11
6	6

Μοντέλο 4	
40	13
5	4

Μοντέλο 1	
OA	0.6774
PA	0.8444
	0.2353
UA	0.7451
	0.3636
$\hat{k}$	0.4641

Μοντέλο 2	
OA	0.6129
PA	0.6889
	0.4118
UA	0.7561
	0.3333
$\hat{k}$	0.357

Μοντέλο 3	
OA	0.7258
PA	0.8667
	0.3529
UA	0.78
	0.5
$\hat{k}$	0.5445

Μοντέλο 4	
OA	0.7097
PA	0.8889
	0.2353
UA	0.7547
	0.4444
$\hat{k}$	0.5177

Μέρος 2

Το επιλεγμένο σύνολο δεδομένων είναι το Superconductivity dataset, το οποίο περιλαμβάνει 21263 δείγματα, καθένα από τα οποία περιγράφεται από 81 χαρακτηριστικά. Λόγω της μεγάλης διαστασημότητας του συνόλου δεδομένων. αρχικά εφαρμόζεται ο αλγόριθμος Relief για την βέλτιστη ταξινόμηση των χαρακτηριστικών, ώστε όταν στη συνέχεια να επιλεγεί μόνο ένας συγκεκριμένος αριθμός χαρακτηριστικών, να χρησιμοποιηθούν εκείνα τα χαρακτηριστικά που συμβάλλουν περισσότερο στην εξαγωγή χρήσιμων συμπερασμάτων.

Πέραν του αριθμού των χαρακτηριστικών που θα επιλεγούν, η δεύτερη ανεξάρτητη μεταβλητή του συγκεκριμένου προβλήματος μοντελοποίησης είναι η ακτίνα των clusters που θα προκύψουν, κάτι που επηρεάζει καθοριστικά και το πλήθος των συναρτήσεων συμμετοχής για το κάθε μοντέλο.

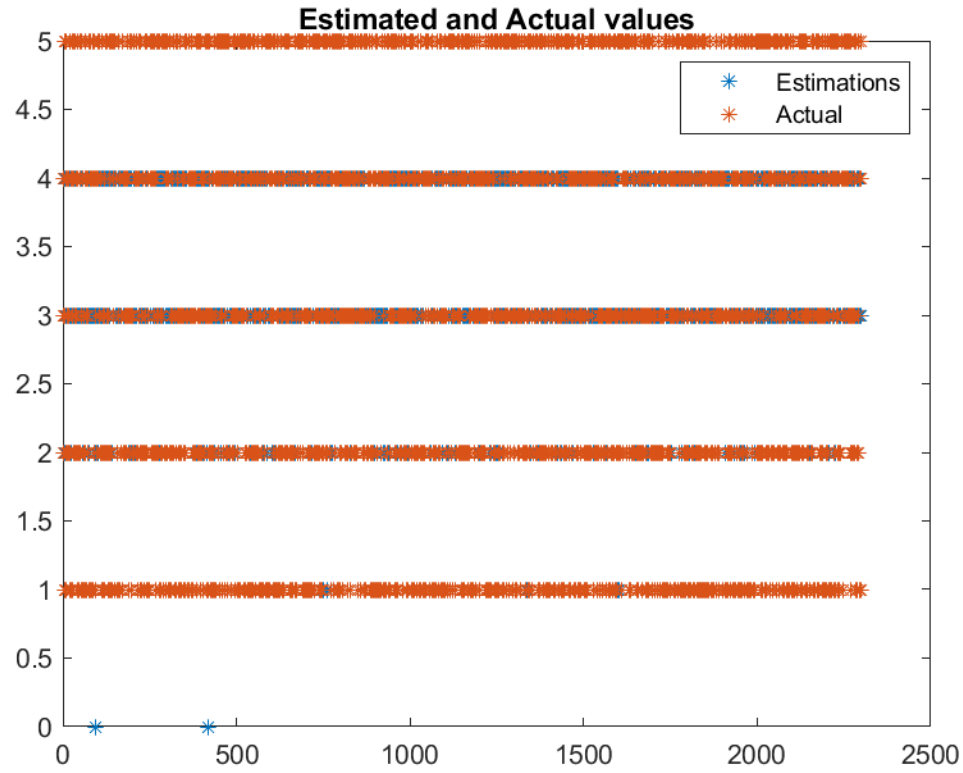
Παράλληλα, γίνεται εφαρμογή subtractive clustering ως μέθοδος ομαδοποίησης για τη δημιουργία των IF-THEN κανόνων.

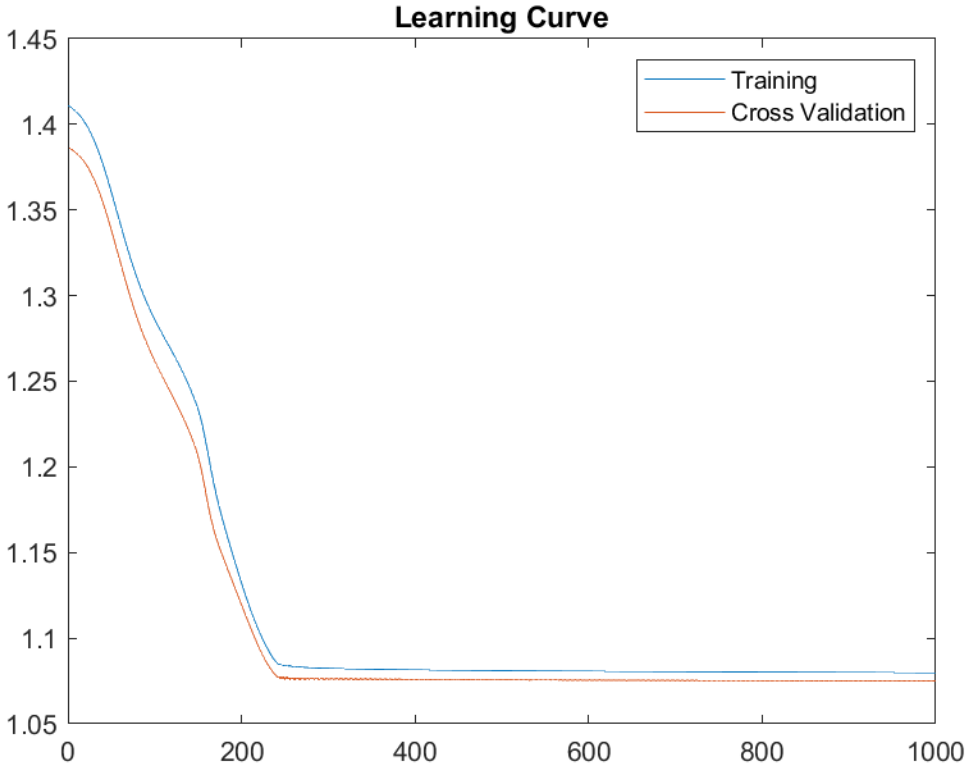
Στη συνέχεια εφαρμόζεται πενταπτυχής διασταυρωμένη επικύρωση σε κάθε φορά τυχαία διαχωρισμένα σύνολα εκπαίδευσης και επαλήθευσης για κάθε μια από τις τιμές ακτινών που έχουν επιλεγεί. Τα μοντέλα που προκύπτουν στη συνέχεια χρησιμοποιούν τη μέθοδο Fuzzy C-means για να ομαδοποιήσουν τα δεδομένα. Τα αποτελέσματα των μέσων όρων σφαλμάτων που προκύπτουν για τους συνδυασμούς των ανεξάρτητων μεταβλητών που επιλέξαμε παρουσιάζονται στον παρακάτω πίνακα.

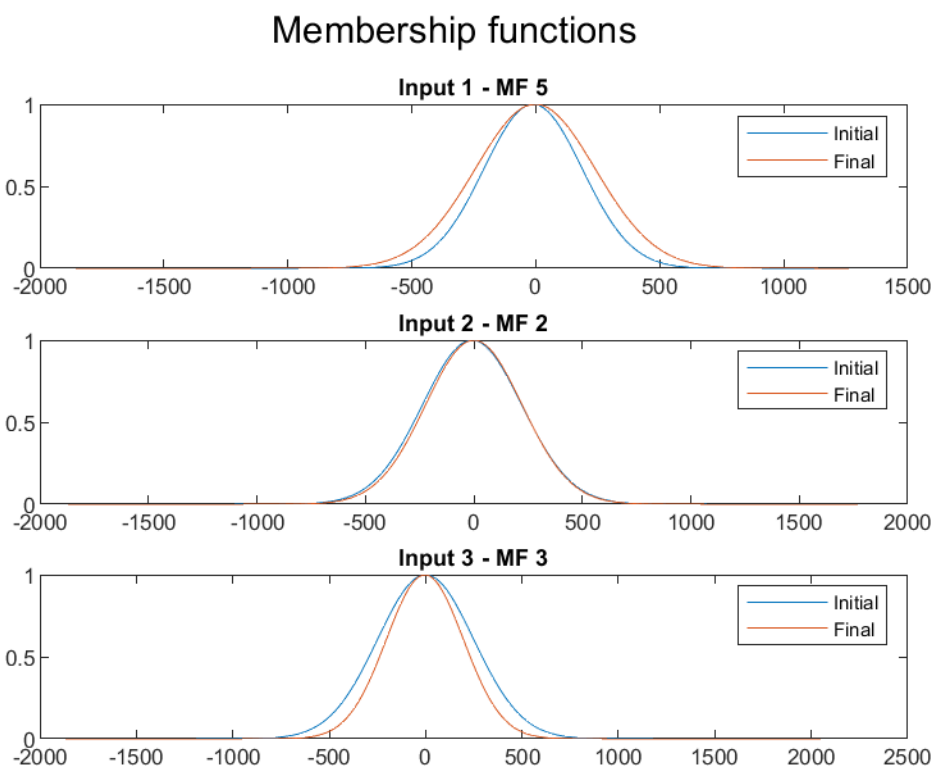
		Number of Features				
		3	6	9	12	15
Clusters Radius	0.18	1.3216	1.3058	1.2885	1.3237	1.3058
	0.22	1.3177	1.3173	1.2933	1.3527	1.3293
	0.26	1.3217	1.3214	1.3022	1.3688	1.344
	0.3	1.3248	1.321	1.3049	1.3711	1.3568
	0.34	1.3246	1.3228	1.3279	1.3692	1.3661
	0.38	1.3254	1.3473	1.3646	1.3752	1.3762

Ο βέλτιστος συνδυασμός αριθμού χαρακτηριστικών και αριθμού συναρτήσεων συμμετοχής είναι εκείνος που επιφέρει το μικρότερο μέσο όρο σφάλματος. Από τον παραπάνω πίνακα είναι προφανές ότι ο συνδυασμός αυτός είναι ο [Number of features, Clusters Radius] = [9,0.18].

Μετά την επιλογή των βέλτιστων παραμέτρων ακολουθεί η εκπαίδευση του μοντέλου και προκύπτουν έτσι η καμπύλη εκμάθησης του μοντέλου, η σύγκριση πραγματικών τιμών και εκτιμήσεων, καθώς και η σύγκριση ορισμένων τυχαίων συναρτήσεων συμμετοχής πριν και μετά τη διαδικασία της εκπαίδευσης και αποτυπώνονται στα παρακάτω σχήματα.


Estimated and Actual values
Estimations
Actual
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
500
1000
1500
2000
2500


Learning Curve
Training
Cross Validation
1.45
1.4
1.35
1.3
1.25
1.2
1.15
1.1
1.05
0
200
400
600
800
1000



Τέλος, παρουσιάζεται ο πίνακας σφαλμάτων ταξινόμησης και ο πίνακας των δεικτών απόδοσης για κάθε μοντέλο.

Πίνακας σφαλμάτων ταξινόμησης				
308	16	1	0	0
94	24	11	22	0
64	199	197	285	174
2	213	221	159	308
0	0	0	0	0

OA	0.2991

PA	0.6581	0.0531	0.4581	0.3412	0

UA	0.9477	0.1589	0.2144	0.1761	NaN

$\hat{k}$	0.1244