Ασαφή Συστήματα Υπολογιστική Νοημοσύνη

Εργασία 4-Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Παναγιώτης Σαββίδης
8094
11.7.2021

Σκοπός της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στην επίλυση προβλημάτων ταξινόμησης. Η εργασία αυτή έχει 2 κομμάτια. Στο πρώτο (*main_4_1.m*)γίνεται εκπαίδευση 4 TSK μοντέλων με την χρήση ενός μικρού μεγέθους dataset χωρίς κάποια προ επεξεργασία του. Στο δεύτερο (*main_4_2.m*) , το dataset είναι πολύ μεγαλύτερο και γι' αυτό τον λόγο χρειάζεται να χωρίσουμε το dataset για να μειώσουμε τον χρόνο εκτέλεσης. Για να επιτύχω αυτό τον διαχωρισμό, χρησιμοποίησα την συνάρτηση split_scale.m που υπάρχει στο elearning.

Μέρος 1ο :

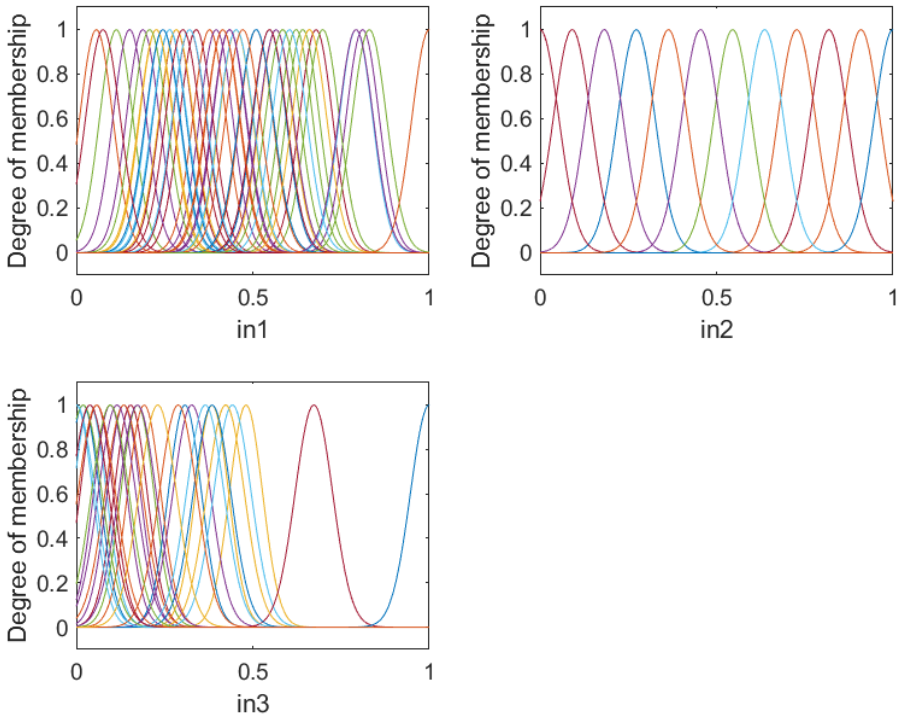
Το dataset που μας δόθηκε για το 1ο κομμάτι της 4ης εργασίας είναι το Haberman's Survival dataset και από αυτό, σύμφωνα με την εκφώνηση, εκπαιδεύτηκαν 4 TSK τα οποία διέφεραν ως προς τον τρόπο διαχωρισμού σε clusters. Συγκεκριμένα τα πρώτα δύο έγιναν με class dependent clustering και τα υπόλοιπα 2 με class independent clustering. Σε όλα τα TSK μοντέλα χρησιμοποίησα 150 εποχές. Οι κανόνες που προέκυψαν για τα μοντέλα μου είναι αντίστοιχα:

Μοντέλο 1 -> 23 Κανόνες
Μοντέλο 2 -> 65 Κανόνες
Μοντέλο 3 -> 4 Κανόνες
Μοντέλο 4 -> 17 Κανόνες

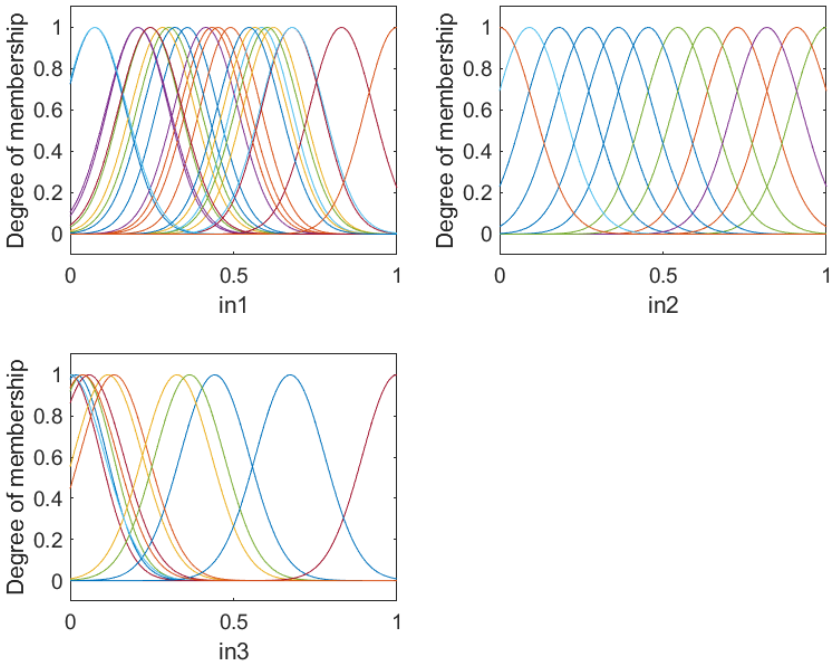
Τα αποτελέσματα της εκτέλεσης φαίνονται στα παρακάτω διαγράμματα:

Membership Functions:

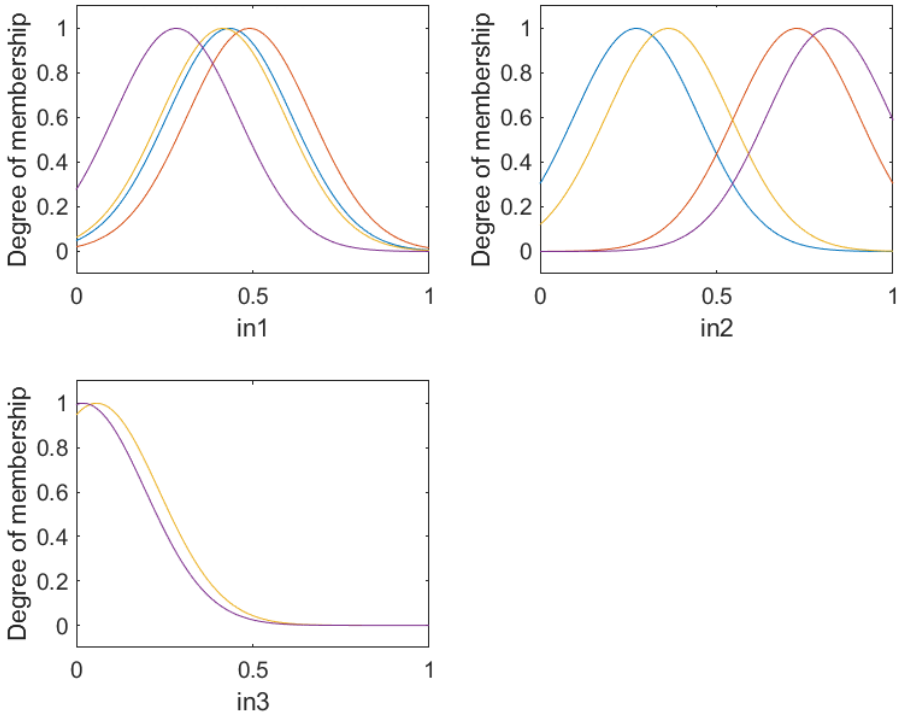
TSK model 1



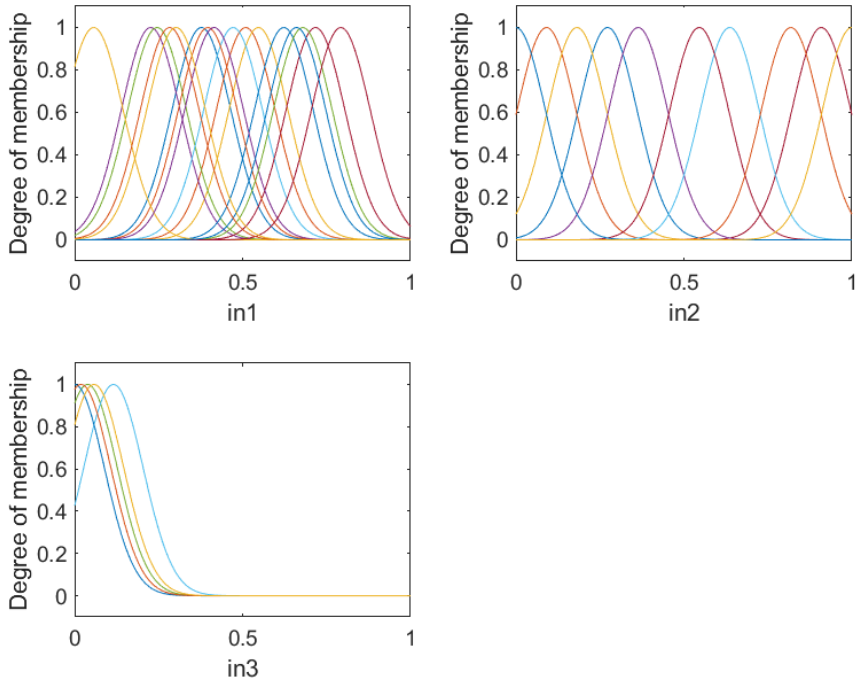
TSK model 2



TSK model 3

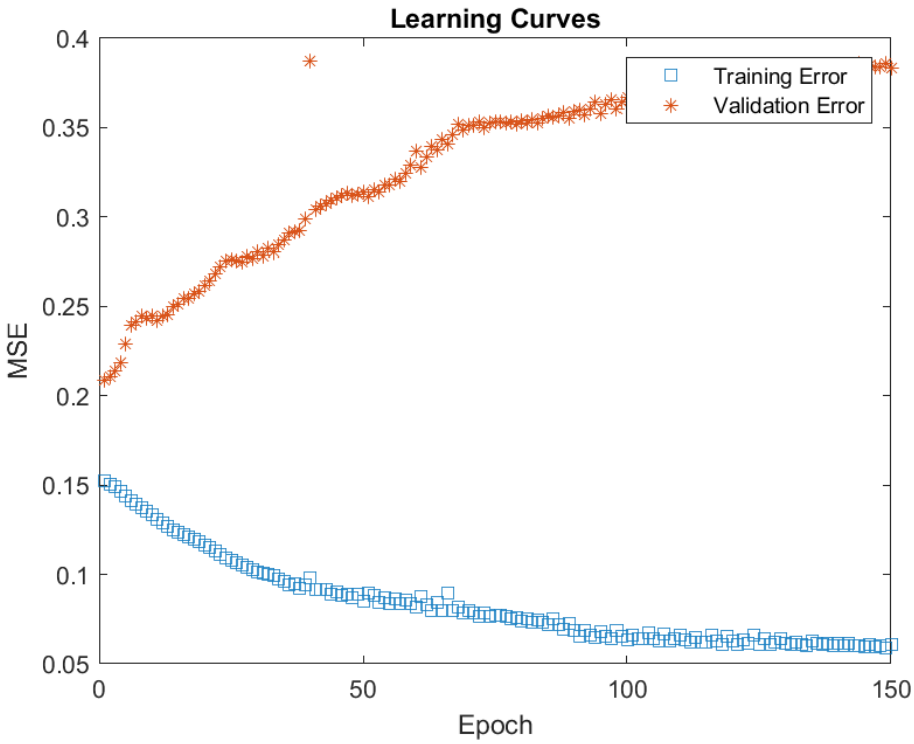


TSK model 4

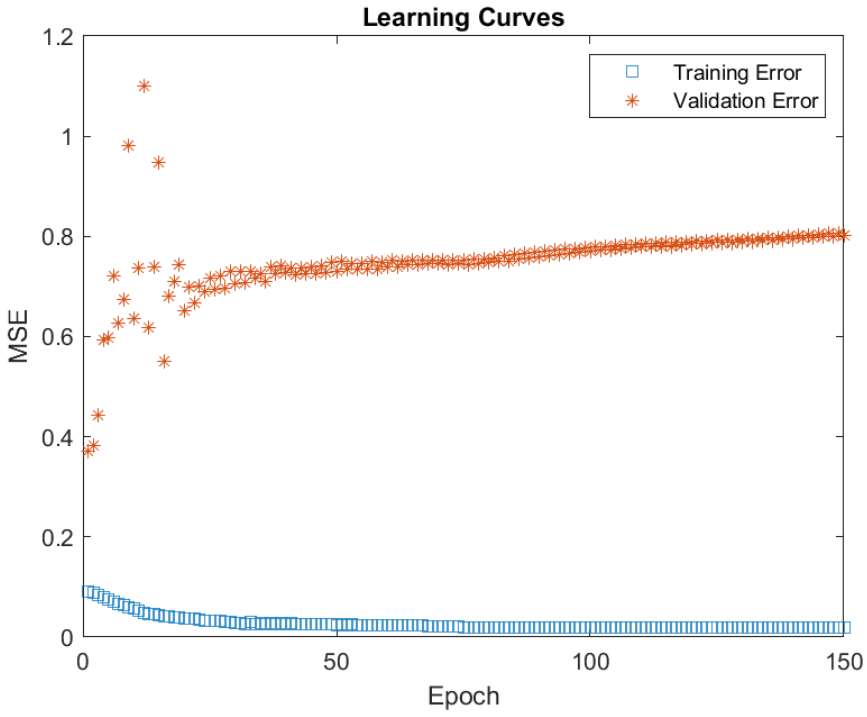


Learning Curves:

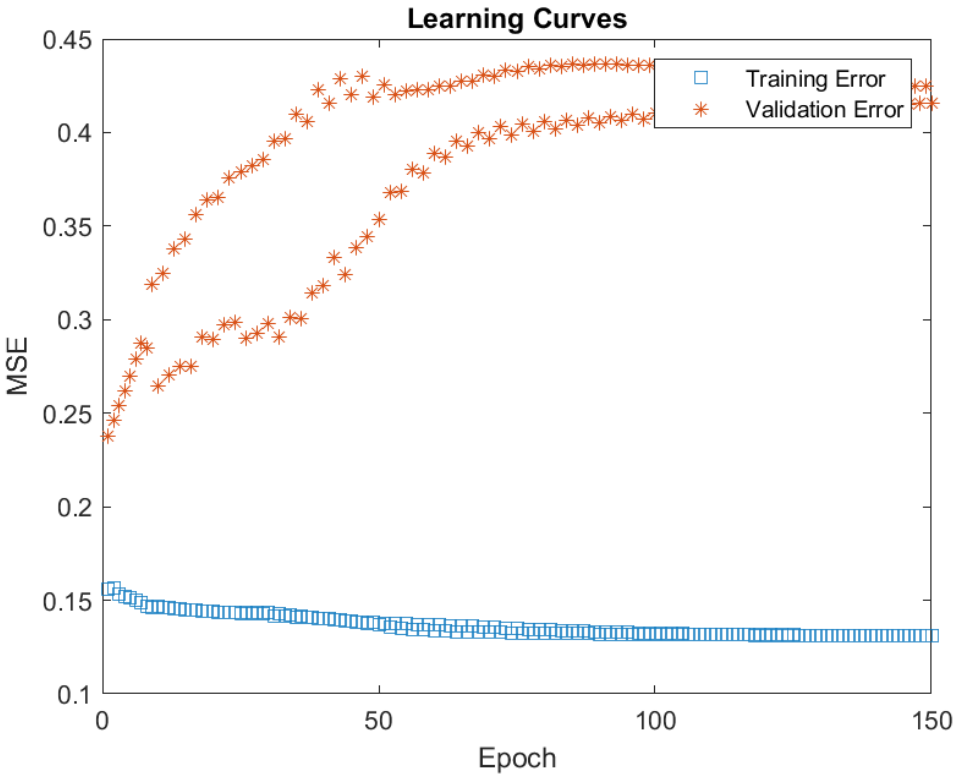
TSK model 1



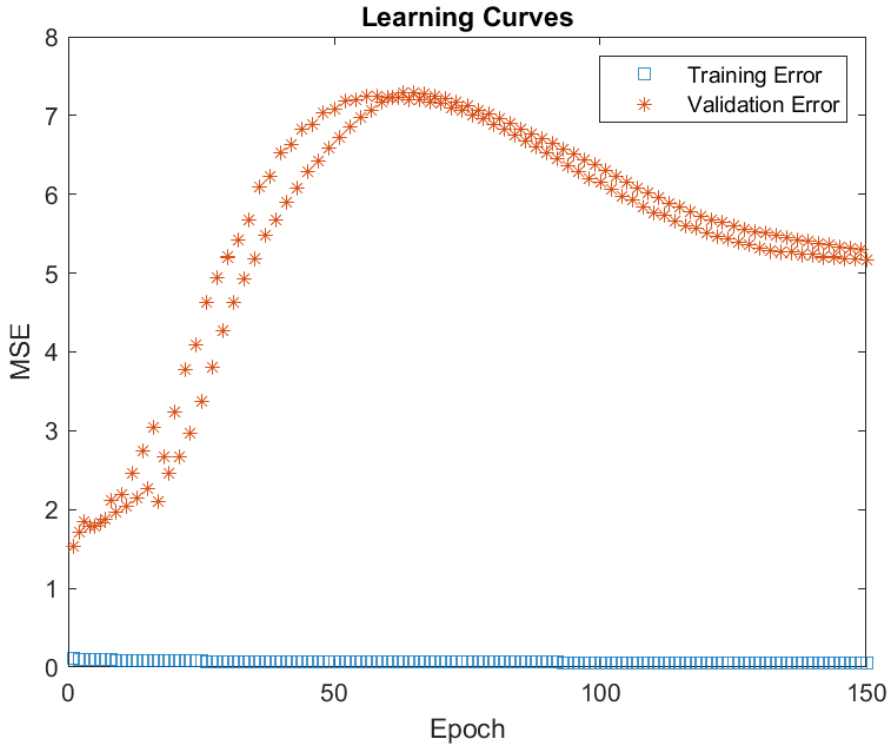
TSK model 2



TSK model 3

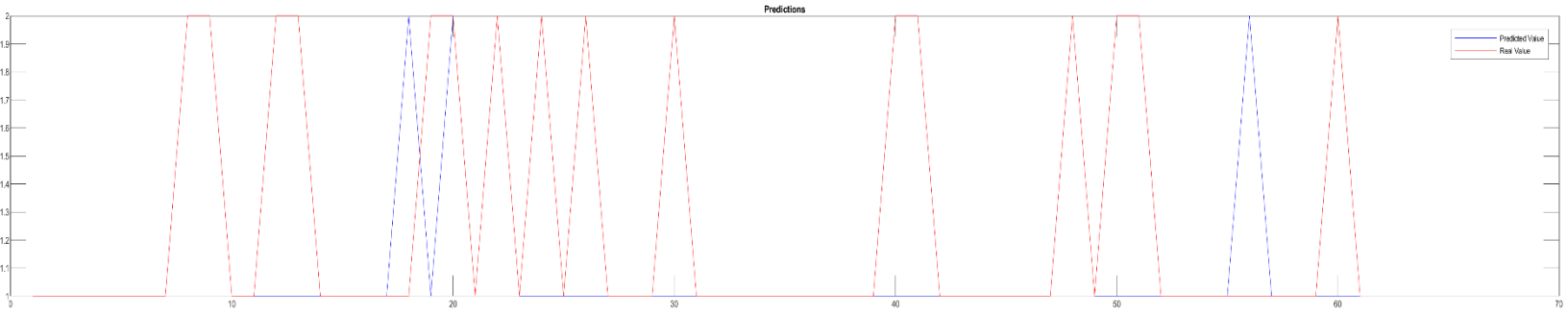


TSK model 4

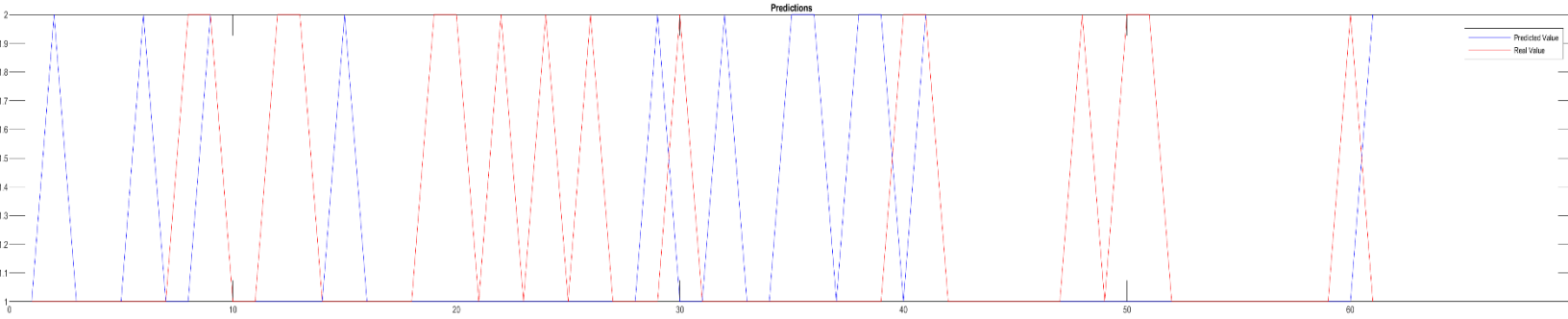


Prediction Errors:

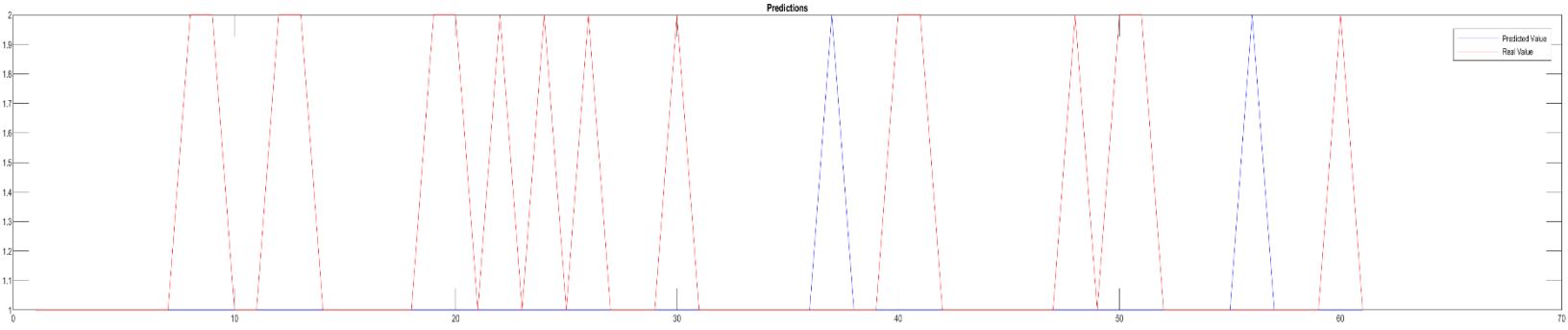
TSK model 1



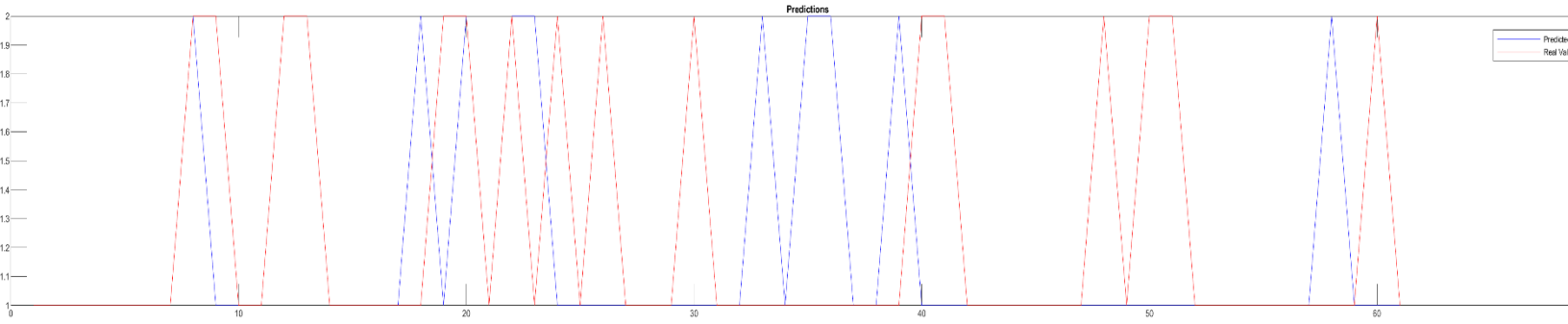
TSK model 2



TSK model 3



TSK model 4



Πίνακας Λανθασμένης Ταξινόμησης

23 Rules	Actual		65 Rules	Actual	
Predicted	C_1	C_2	Predicted	C_1	C_2
C_1	43	11	C_1	35	14
C_2	2	5	C_2	10	2

4 Rules	Actual		17 Rules	Actual	
Predicted	C_1	C_2	Predicted	C_1	C_2
C_1	43	16	C_1	38	11
C_2	2	0	C_2	7	5

		Class Dependent		Class Independent	
Number of Rules		23	65	4	17
Overall Accuracy		0.78689	0.60656	0.70492	0.70492
K hat		0.3274	-0.10574	-0.061896	0.17069
Producer's Accuracy	1	0.95556	0.77778	0.95556	0.84444
	2	0.3125	0.125	0	0.3125
User's Accuracy	1	0.7963	0.71429	0.72881	0.77551
	2	0.71429	0.16667	0	0.41667

Αρχικά από τις membership functions γίνεται αμέσως αντιληπτό ότι στα class dependent μοντέλα υπάρχουν ανάμεικτα αποτελέσματα όσο αναφορά τον διαμερισμό. Στο μοντέλο 1 έχουμε καλό διαμερισμό στο 2^{ο} χαρακτηριστικό αλλά στα άλλα 2 υπάρχει επικάλυψη. Στο μοντέλο 2, υπάρχει μεγάλη επικάλυψη και στα 3 χαρακτηριστικά.

Στα class independent μοντέλα υπάρχει πάλι μεγάλη επικάλυψη με χειρότερη περίπτωση του 3^{ου} χαρακτηριστικού του 3^{ου} μοντέλου. Ωστόσο μια λίγο καλύτερη περίπτωση βλέπουμε στο μοντέλο 4^{ο} στο 2^{ο} χαρακτηριστικό, όπου υπάρχει λίγο καλύτερος διαμερισμός.

Σαν γενικό συμπέρασμα, οι κανόνες δεν συνδέονται άμεσα με τον διαμερισμό των μοντέλων.

Οι Learning Curves δείχνουν ένα σχετικά σταθερό MSE, αν και σε κάποιες περιπτώσεις αυξάνεται. Βέβαια οι τιμές του είναι αρκετά χαμηλές.

Ένα άλλο συμπέρασμα που αξίζει να σημειωθεί είναι ότι οι κανόνες δεν είναι ανάλογοι με καλύτερο μοντέλο. Αν δούμε το βέλτιστο Overall Accuracy το πετυχαίνουμε στους 23 κανόνες, ενώ τόσο σε ψηλότερες όσο και χαμηλότερες τιμές κανόνων δεν υπάρχει βελτίωση. Μάλιστα για τους 4 και 17 κανόνες η τιμή του Accuracy είναι ακριβώς ίδια.

Αξίζει να σημειωθεί το 0 που υπάρχει στο Producer's και User's Accuracy για τους 4 Κανόνες. Πιθανότατα πρόκειται για κάποια ακραία περίπτωση που αν ξανατρέχαμε τον αλγόριθμο να μην επαναλαμβανόταν (κάτι το οποίο δοκίμασα, αλλά δεν υπήρχε χρόνος για να αλλάξω όλο το Report).

Όλες οι υπόλοιπες τιμές των 23 κανόνων είναι οι βέλτιστες, άρα μπορούμε με σιγουριά να πούμε ότι οι 23 κανόνες είναι το βέλτιστο μοντέλο.

Μέρος 2ο :

Το dataset που έπρεπε να χρησιμοποιήσω στο μέρος 2 της 4ης εργασίας είναι το Epileptic Seizure Recognition dataset (το οποίο έχει γίνει Removed από το Repository του UCI και έπρεπε να το βρω από το THMMY), το οποίο έχει πολύ περισσότερα δείγματα και με πολλά περισσότερα χαρακτηριστικά στο καθένα, από το αντίστοιχο dataset του πρώτου μέρους. Όπως μπορεί να αντιληφθεί κανείς, το να δουλέψουμε όπως στο 1ο μέρος της εργασία είναι αδύνατο, καθώς οι κανόνες που θα προκύψουν θα ήταν πάρα πολλοί κάνοντας τον χρόνο εκτέλεσης απαγορευτικά μεγάλο.

Έτσι, για να μειώσω τον χρόνο εκτέλεσης, έτρεξα τον αλγόριθμο relief ο οποίος επιλέγει την βαρύτητα κάθε χαρακτηριστικού και στην συνέχεια επιλέγω τον αριθμό των χαρακτηριστικών που θα επιλεγούν και την ακτίνα των διαφορετικών cluster. Όπως φαίνεται και στην συνάρτηση main_4_2.m οι προδιαγραφές που επέλεξα είναι οι εξής:

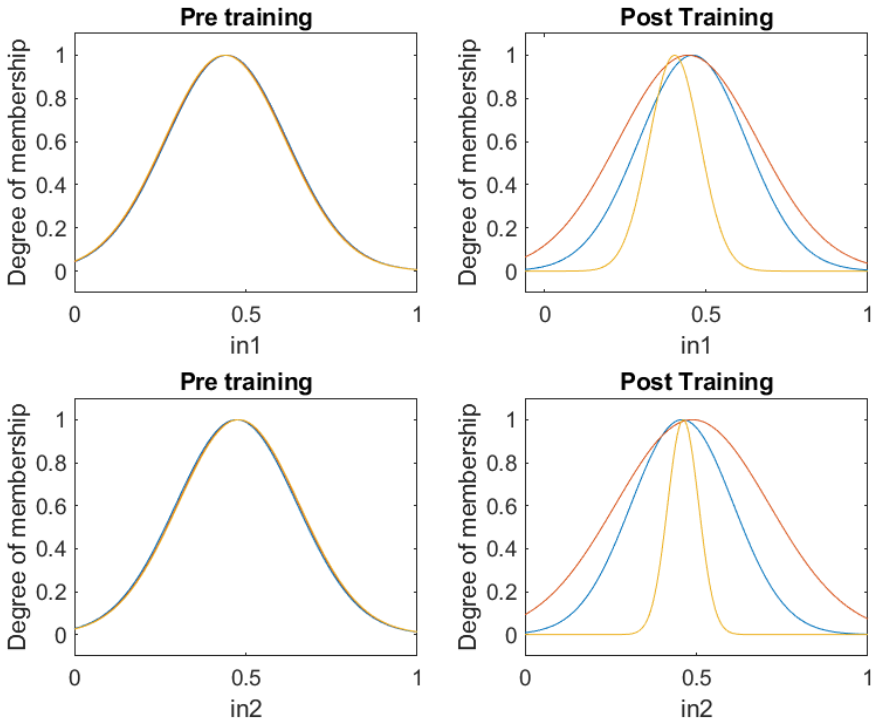
```
%Random Selected characteristics
characteristics = [3 6 12 15];
%Rads
clust_rad = [0.5 0.35 0.2 0.05];
```

Οι αριθμοί αυτοί προέκυψαν από δοκιμές και είναι αυθαίρετοι αλλά παράλληλα και συγκεκριμένοι για να ελέγξουμε τις «δυνατότητες» του μοντέλου μας. Δεν έπρεπε να «ζορίσω» το μοντέλο με μεγάλο αριθμό χαρακτηριστικών καθώς θα είχε αντίκρισμα στον χρόνο εκτέλεσης. Επίσης η πολύ μικρή ακτίνα του cluster ήταν μια δοκιμή η οποία δεν έδειξε κάποιο συγκεκριμένο αποτέλεσμα.

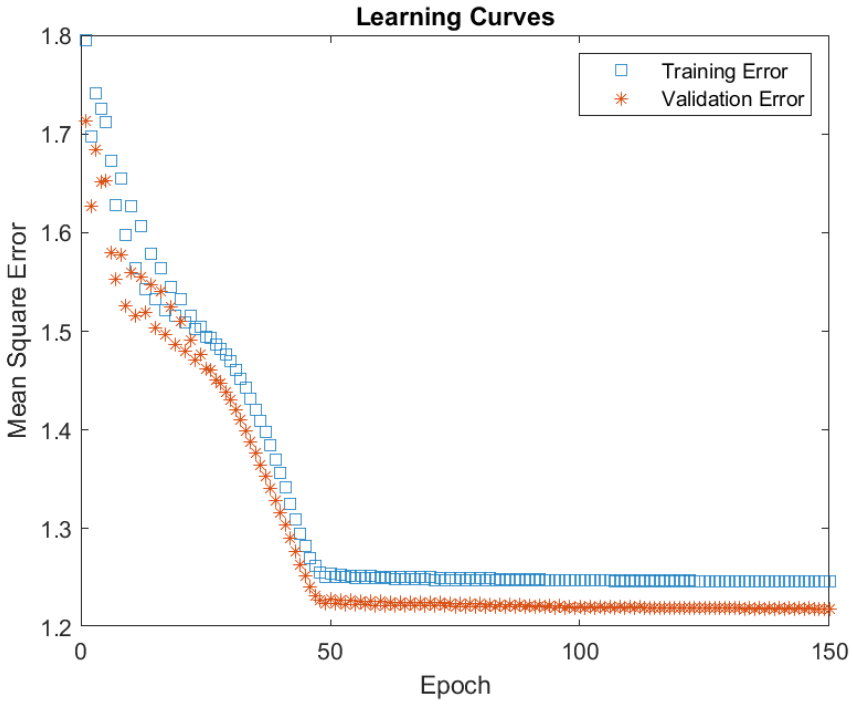
Αφού τελείωσα την προ επεξεργασία των δεδομένων, έπρεπε να εφαρμόσω 5-fold cross validation για 50 εποχές. Ο σκοπός μου είναι να βρω ποιος συνδυασμός από αριθμό χαρακτηριστικών και ακτίνας οδηγεί στο μικρότερο σφάλμα και με αυτό τον συνδυασμό τιμών να εκπαιδεύσω το τελικό μου μοντέλο. Όμως όλα τα σφάλματα που έπαιρνα είχαν την τιμή NaN κάτι το οποίο οδήγησε στον εναλλακτικό υπολογισμό του σφάλματος, που ουσιαστικά αφαιρούμε όλες τις NaN τιμές.

Μετά από αυτή την αλλαγή τα αποτελέσματα που πήρα συνοψίζονται στα παρακάτω διαγράμματα όπου προέκυψαν με 145 κανόνες.

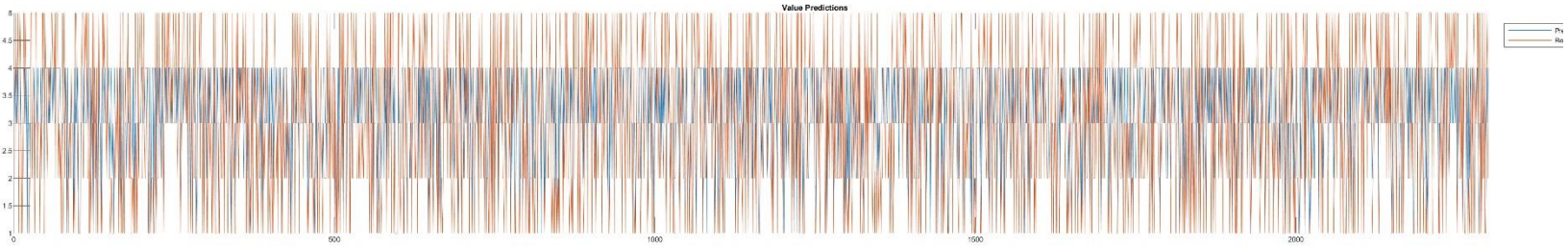
Membership functions:



Learning Curves:



Prediction Errors:



Error Matrix

		Actual				
	Classes	C_1	C_2	C_3	C_4	C_5
Predicted	C_1	285	14	1	1	0
	C_2	102	27	14	26	0
	C_3	64	179	222	270	202
	C_4	5	239	221	165	263
	C_5	0	0	0	0	0

Accuracies

	Overall Accuracy	Producer's Accuracy					User's Accuracy					K hat
Model	0.30391	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5	0.13019
		0.625	0.0588	0.485	0.357	0	0.947	0.160	0.237	0.185	NaN	

Από τις Membership Functions παρατηρούμε ότι υπάρχει πλήρης επικάλυψη πριν γίνει το Training, αλλά μετά το Training υπάρχει μια μικρή βελτίωση. Αντιθέτως από την Learning Curve παρατηρούμε ότι τα Errors είναι αρκετά μικρά και σταθεροποιούνται πλήρως μετά τις 50 περίπου εποχές.

Το σημαντικότερο συμπέρασμα προκύπτει από τους πίνακες των Accuracies και του Error Matrix. Το μοντέλο μας, παρόλο το Training και το clustering δεν κατάφερε να ταξινομήσει κανένα δείγμα στην τάξη πέντε. Κατά τ ’άλλα το μοντέλο μας τα πήγε «αρκετά καλά» αφού τα accuracies έχουν μεγάλες τιμές, εκτός από την κλάση δύο και πέντε.

Δυστυχώς η πίεση του χρόνου δεν μου επέτρεψε να πειράξω τις αρχικές τιμές ώστε να αποφευχθεί το παραπάνω πρόβλημα.