Εργασία 4 - Υπολογιστική Νοημοσύνη

Επίλυση προβλήματος ταξινόμησης με χρήση μοντέλων TSK

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Εργασία του: **Παπαδόπουλου Κωνσταντίνου**

ΑΕΜ: **8677**

Εισαγωγή

Στόχος της εργασίας αυτής είναι να διερευνηθεί η ικανότητα των μοντέλων TSK στην επίλυση προβλημάτων ταξινόμησης (classification), σύμφωνα με τα ζητούμενα της εκφώνησης της Εργασίας 4 του μαθήματος Υπολογιστικής Νοημοσύνης.

Στη συνέχεια, αναλύουμε τη διαδικασία εργασίας πάνω στα ζητούμενα προβλήματα, σχολιάζοντας και τη λογική συγγραφής του κώδικα, όπου αυτό κρίνεται απαραίτητο (υπάρχει σχολιασμός και στον ίδιο τον πηγαίο κώδικα, ο οποίος είναι σε Matlab) και παραθέτουμε τα απαραίτητα διαγράμματα.

1 Εφαρμογή σε απλό dataset

Στην πρώτη φάση της εργασίας, επιλέγεται από το UCI repository το Haberman's Survival[1], το οποίο περιλαμβάνει 306 δείγματα (instances), από 3 χαρακτηριστικά (attributes) το καθένα.

Στη συνέχεια ακολουθούμε τα εξής βήματα:

Ζητούμενα του προβλήματος:

- ✓ *Διαχωρίζουμε τα δεδομένα σε σύνολα εκπαίδευσης-επικύρωσης-ελέγχου*

[1] https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival

Αρχικά, ελέγχουμε ότι στο dataset δεν υπάρχουν missing values (βλέπουμε στο *Abstract: Missing Values?-No*), επομένως ούτε ανάγκη διαχείρισης αυτών.

Όπως και στην Εργασία 3, προχωρούμε σε διαχωρισμό των δεδομένων σε τρία μη επικαλυπτόμενα υποσύνολα D_{trn}, D_{val}, D_{chk} (60%, 20% και 20% των δειγμάτων αντίστοιχα). Φροντίζουμε να συμπεριλάβουμε τις μέγιστες και τις ελάχιστες τιμές του κάθε χαρακτηριστικού στο σύνολο εκπαίδευσης, έτσι ώστε να έχουμε ένα ακόμη πιο αντιπροσωπευτικό σύνολο σε εύρος.

Ελέγχουμε επίσης τη συχνότητα εμφάνισης όμοιων κλάσεων στο dataset και αναλογικά διαμοιράζουμε αυτές σε κάθε ένα από τα τρία υποσύνολα.

- ✓ *Εκπαιδεύουμε TSK μοντέλα με διαφορετικές παραμέτρους*

  Σε αυτό το στάδιο θα εξεταστούν τα διάφορα μοντέλα TSK που δίνονται στην εκφώνηση, τα οποία διαφέρουν ως προς το πλήθος των IF-THEN κανόνων τους .

  Χρησιμοποιούμε τη συνάρτηση *genfis2()* για τη δημιουργία του μοντέλου, η οποία κάνει χρήση του αλγορίθμου Subtractive Clustering (SC).

  Για την εκπαίδευση των μοντέλων χρησιμοποιούμε τη συνάρτηση *anfis()*, η οποία χρησιμοποιεί υβριδική μέθοδο, σύμφωνα με την οποία οι παράμετροι των συναρτήσεων συμμετοχής βελτιστοποιούνται μέσω της μεθόδου οπισθοδιάδοσης (backpropagation algorithm), ενώ οι παράμετροι της πολυωνυμικής συνάρτησης εξόδου βελτιστοποιούνται μέσω της μεθόδου ελαχίστων τετραγώνων (Least Squares Algorithm).

  Στα δύο πρώτα μοντέλα, το SC θα εκτελεστεί για όλα τα δεδομένα του συνόλου εκπαίδευσης (class independent).

Στα επόμενα τέσσερα μοντέλα (2 κλάσεις επί 2 διαφορετικά radius), θα εξεταστεί ο διαμερισμός του χώρου εισόδου εφαρμόζοντας clustering στα δεδομένα του συνόλου εκπαίδευσης που ανήκουν στην εκάστοτε κλάση ξεχωριστά (class dependent). Σε αυτό το σημείο έπρεπε να βρούμε και τα κέντρα που θα δίναμε στην *genfis2()* στο *radii* argument, κάτι το οποίο κάναμε με την *subclust().*

Παρατηρούμε ότι για μικρό radius, π.χ. 0.2, έχουμε περισσότερα membership functions από ότι για μεγαλύτερο, π.χ. 0.7.

Το dependent παρατηρήθηκε ότι εξάγει καλύτερα αποτελέσματα, μιας και εκπαιδεύεται πιο σωστά ανάλογα με την κλάση.

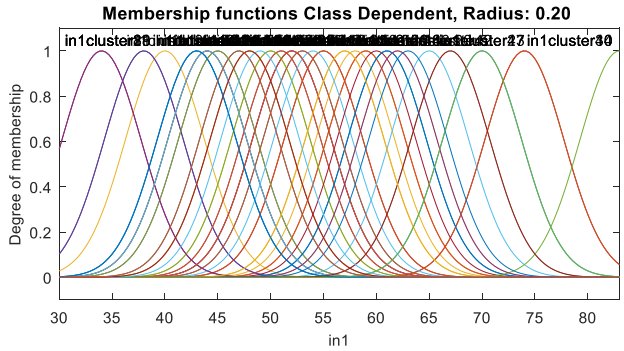
- ✓ *Καθορίζουμε τους δείκτες απόδοσης για την αξιολόγηση των μοντέλων*

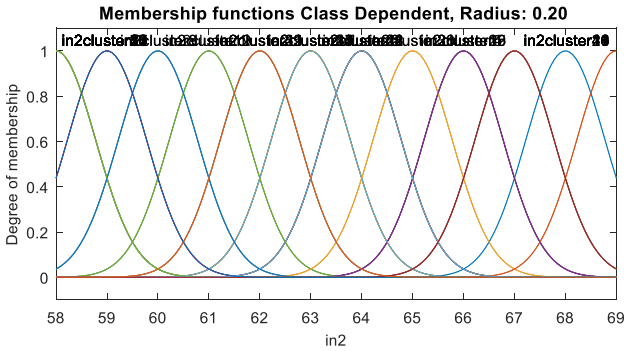
  Στο αρχείο *evaluate_class.m* γράφουμε τις υλοποιήσεις των μετρικών που θα χρησιμοποιήσουμε, δηλαδή των Error Matrix, Overall Accuracy, Producer's Accuracy - User's Accuracy και K-hat.

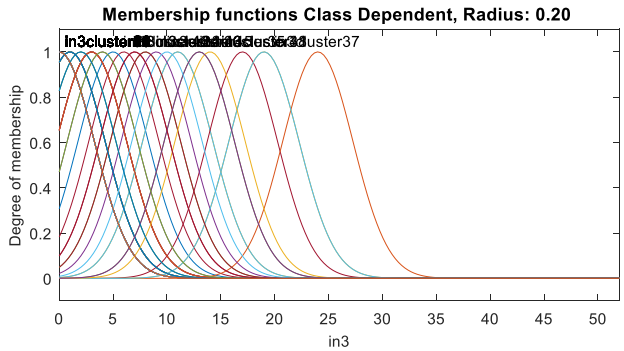
Ζητούμενα του προβλήματος:

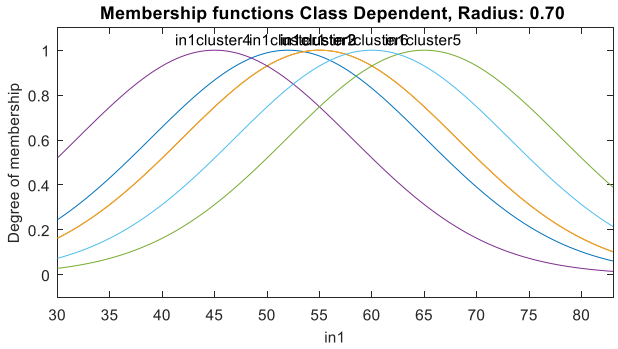
1. Παρατίθενται τα διαγράμματα στα οποία απεικονίζονται οι τελικές μορφές των ασαφών συνόλων που προέκυψαν μέσω της διαδικασίας εκπαίδευσης.

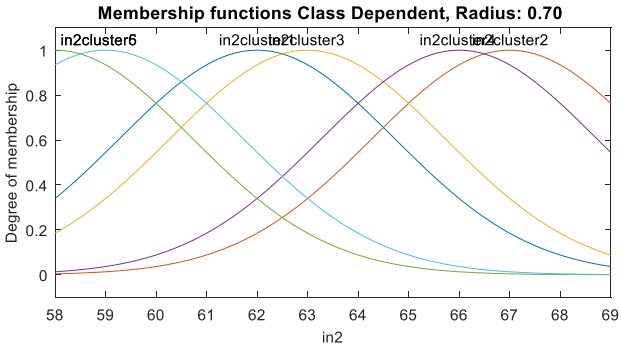
Class Dependent Membership Functions

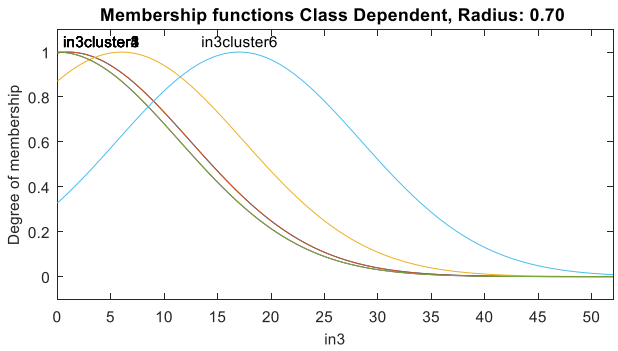




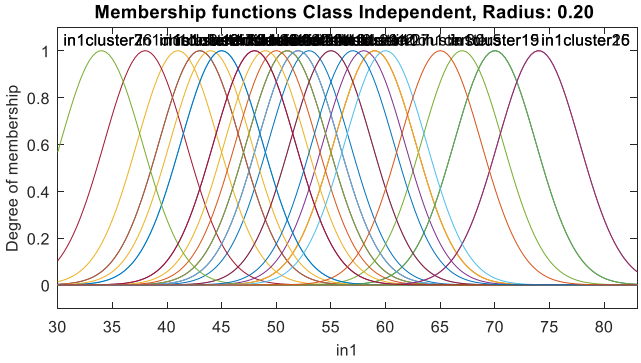


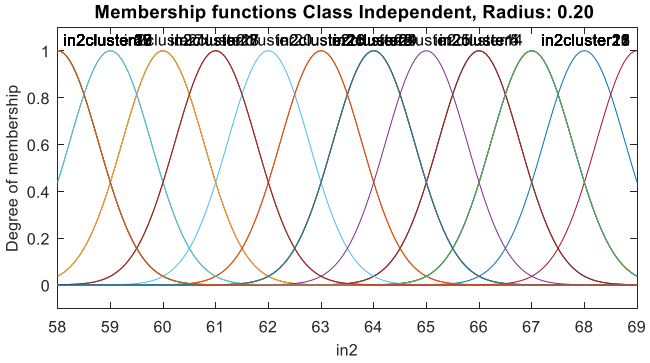


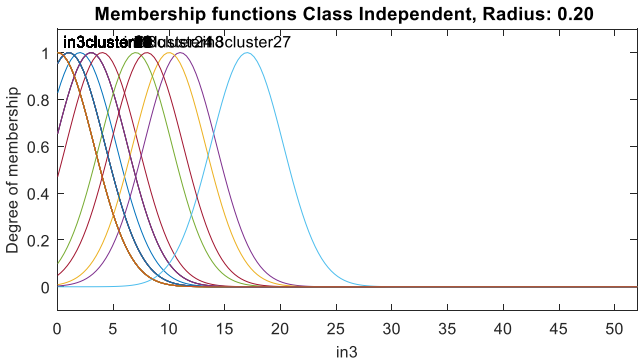


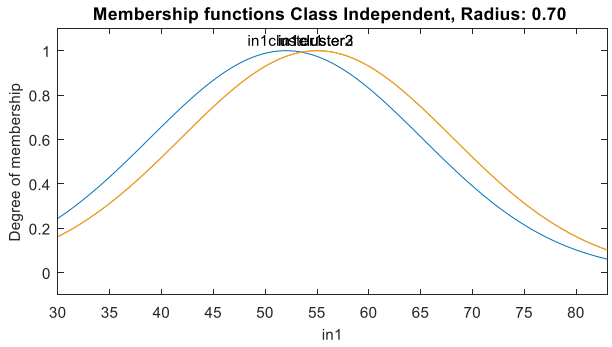


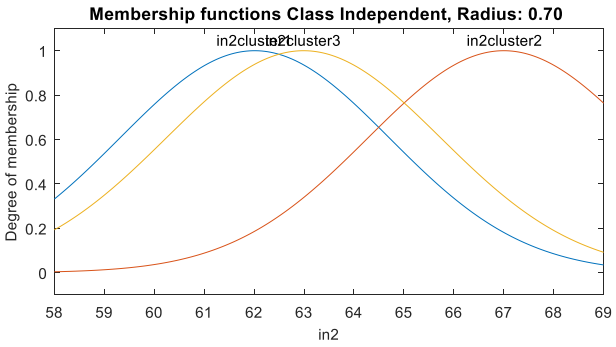
Class Independent Membership Functions

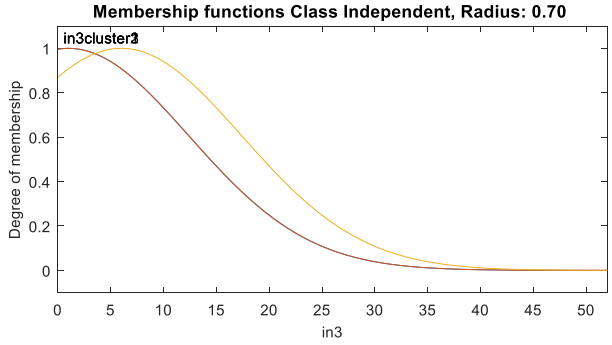




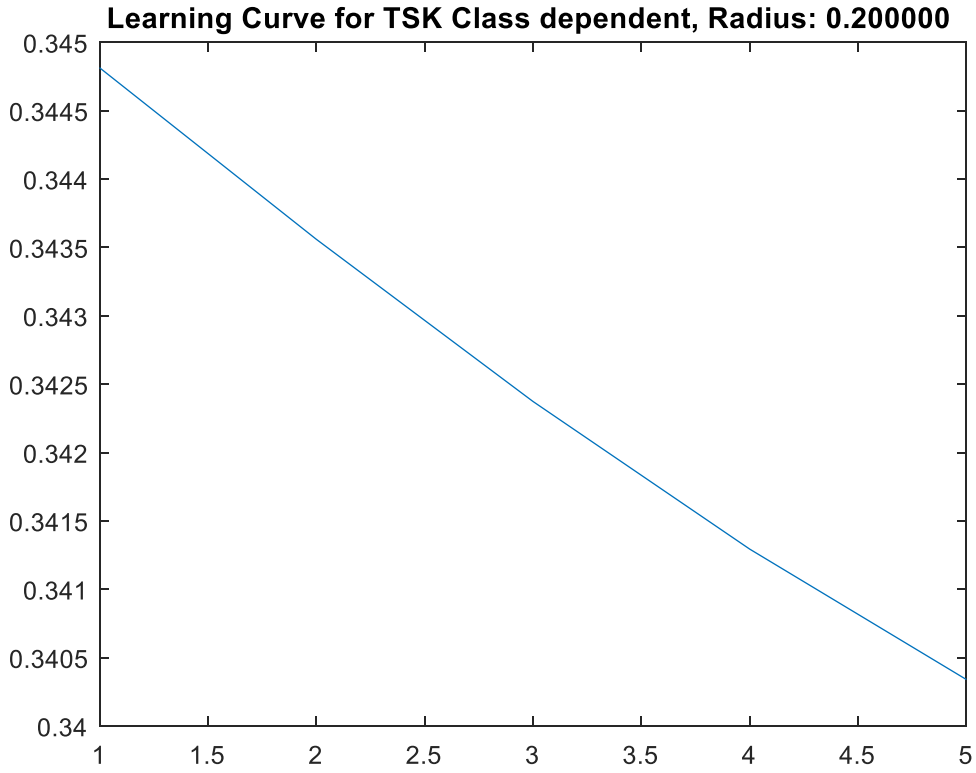


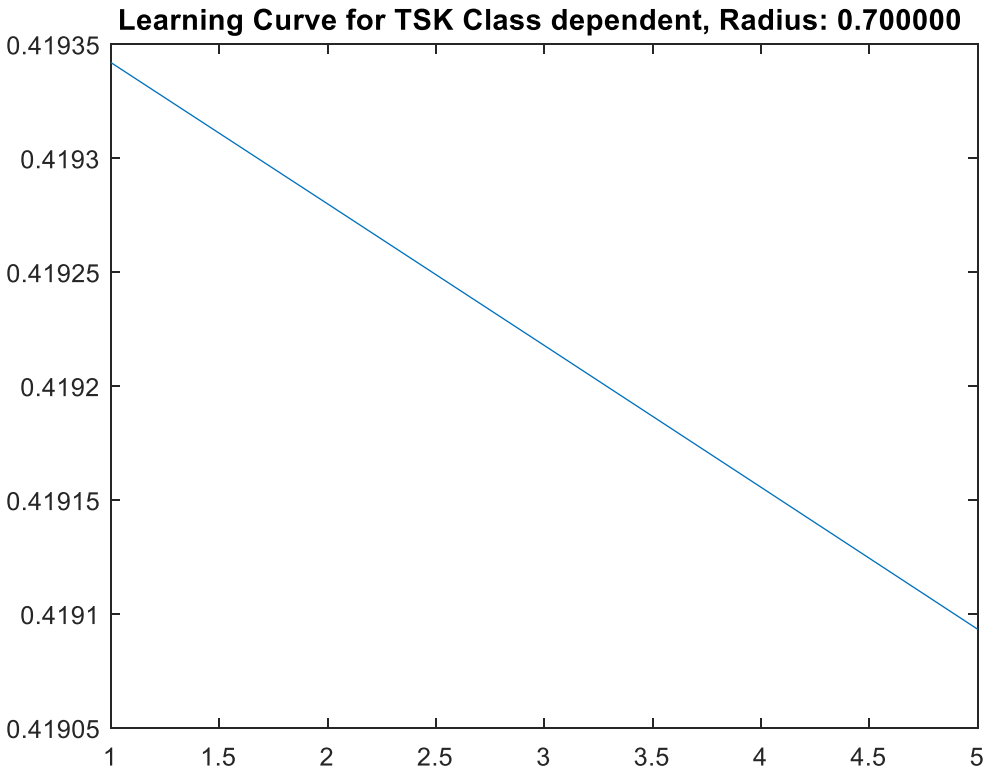


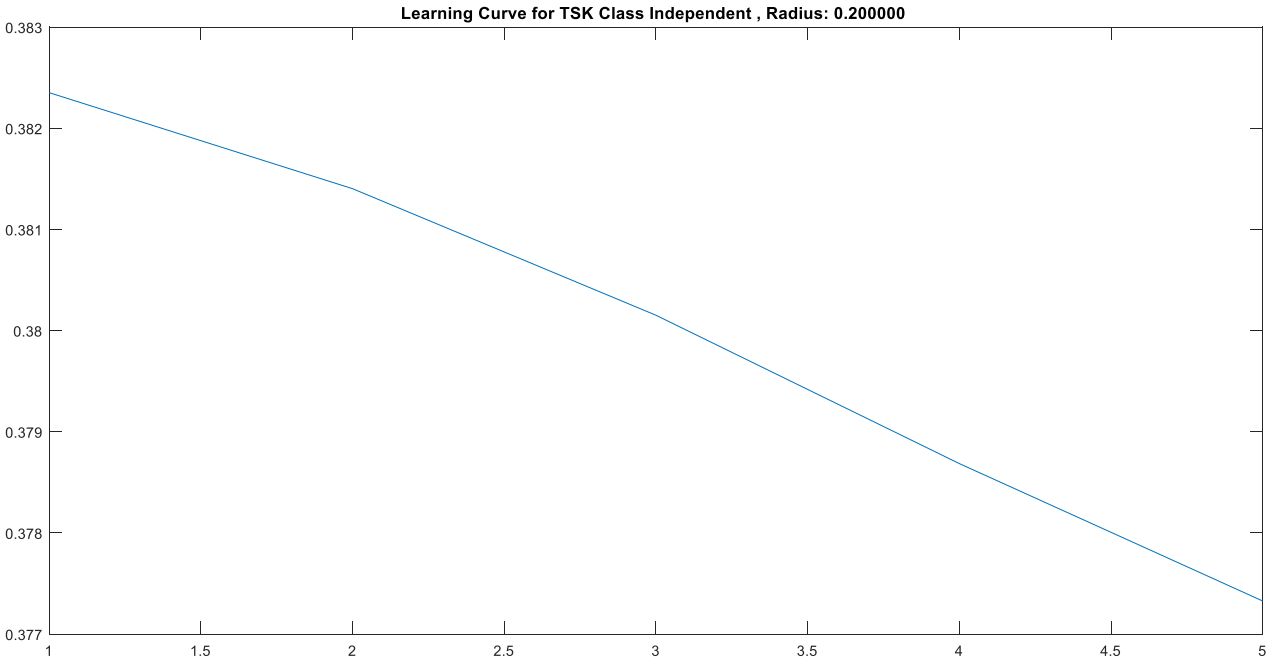


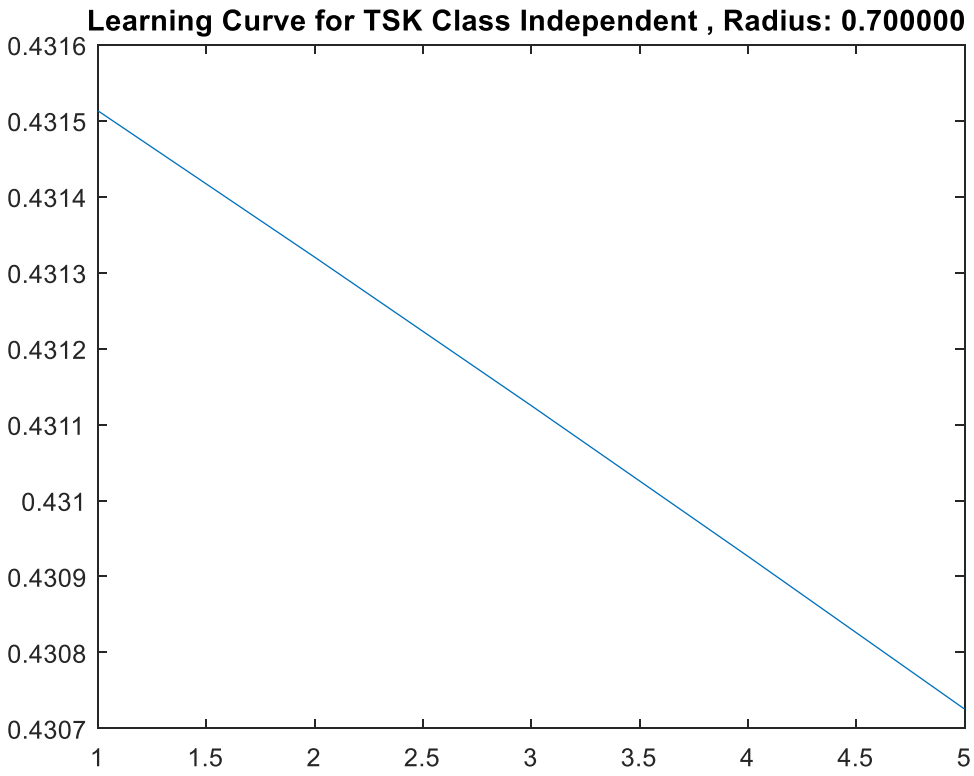


2. Παρατίθενται τα διαγράμματα μάθησης (learning curves) όπου απεικονίζεται το σφάλμα του μοντέλου συναρτήσει του αριθμού των επαναλήψεων (iterations).




Learning Curve for TSK Class dependent, Radius: 0.700000
0.41935
0.4193
0.41925
0.4192
0.41915
0.4191
0.41905
1
1.5
2
2.5
3
3.5
4
4.5
5


Learning Curve for TSK Class Independent , Radius: 0.200000
0.383
0.382
0.381
0.38
0.379
0.378
0.377
1
1.5
2
2.5
3
3.5
4
4.5
5


Learning Curve for TSK Class Independent , Radius: 0.700000
0.4316
0.4315
0.4314
0.4313
0.4312
0.4311
0.431
0.4309
0.4308
0.4307
1
1.5
2
2.5
3
3.5
4
4.5
5

3. Παρουσιάζονται οι τιμές των δεικτών απόδοσης OA, PA, UA και K-hat.

```
OA =          UA =          PA =          Khat =

    0.7419        1.0000        0.7377
                  0.0588        1.0000        0.0832
```

Παρατηρούμε ότι όσο μεγαλύτερος είναι ο αριθμός των κανόνων τόσο καλύτερη είναι η ακρίβεια του ταξινομητή με βάση των δείκτη της συνολικής ακρίβειας. Επίσης με βάση τους δείκτες PA , UA παρατηρούμε ότι όσο αυξάνει ο αριθμός των κανόνων τείνουν να είναι και αυτές μικρότερες. Σε σχετικά μικρούς αριθμούς κανόνων παρατηρείται μεγάλη συγκέντρωση συναρτήσεων συμμετοχής σε ορισμένες περιοχές.

2 Εφαρμογή σε dataset με υψηλή διαστασιμότητα

Στο δεύτερο μέρος της εργασίας, επιλέγεται από το UCI repository το Epileptic Seizure Recognition dataset[2], το οποίο περιλαμβάνει 11500 δείγματα, καθένα από τα οποία περιγράφεται από 179 μεταβλητές/χαρακτηριστικά.

Λόγω του μεγέθους των διαστάσεων των δειγμάτων είναι φανερό ότι μια απλή εφαρμογή του TSK μοντέλου καθίσταται αρκετά δύσκολη, ένα πρόβλημα γνωστό και ως κατάρα των διαστάσεων (curse of dimensionality). Για αυτό το λόγο θα χρησιμοποιήσουμε μεθόδους ομαδοποίησης των δεδομένων και ουσιαστικά μείωσης της διασπασιμότητας, καθώς και του αριθμού των IF-THEN κανόνων. Αυτό θα γίνει μέσω της επιλογής χαρακτηριστικών και της ασαφούς ομαδοποίησης.

Λόγω έλλειψης χρόνου, καθώς το πρόγραμμα δεν ολοκλήρωσε το τρέξιμό του, ενώ έχει γραφτεί ο κώδικας, δεν ήταν δυνατή η παράθεση των διαγραμμάτων. Ισχύουν και σε αυτήν την περίπτωση παρόμοια σχόλια με αυτά που έγιναν στο προηγούμενο μέρος.

[2] https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition