ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΒΑΘΙΑ ΜΑΘΗΣΗ
2Η ΕΡΓΑΣΙΑ

Ιωάννης Δεϊρμεντζόγλου 10015

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΕΜ 10015 Email deirmentz@ece.auth.gr

1. Εισαγωγή

Στην πρώτη εργασία του μαθήματος παρουσιάστηκαν τα αποτελέσματα εφαρμογής κατηγοριοποιητών Nearest Centroid, k-Nearest Neighbor, MultiLayer Perceptron και Convolutional Neural Network στη βάση δεδομένων CIFAR-10, όπου η σύγκριση έδειξε ότι το καλύτερο μοντέλο ήταν αυτό ενός CNN.

Στην 2η εργασία ζητούμενο είναι κατηγοριοποίηση με χρήση ενός **SVM** για διαχωρισμό 2 ή περισσοτέρων κλάσεων. Σε αυτήν την εργασία χρησιμοποιήθηκαν πάνω από 1 datasets (δεν θα παρουσιαστούν όλα στην αναφορά) τα οποία προ επεξεργάζονται με παρόμοια διαδικασία :

- **CIFAR-10**. Η CIFAR-10 περιλαμβάνει 60.000 εικόνες σε 10 διαφορετικές κατηγορίες. Κάθε εικόνα ανήκει σε μία από τις κατηγορίες: αεροπλάνα, αυτοκίνητα, πουλιά, γάτες, ελάφια, σκύλοι, βατράχια, ίπποι, πλοία και φορτηγά. Οι εικόνες έχουν διαστάσεις 32x32x3 (3 χρωματικά κανάλια –RGB εικόνες) . Αρχικά , τα δεδομένα που φορτώνονται από την cifar 10 χωρίζονται σε train και test set με διαστάσεις 50.000x32x32x3 (50k εικόνες) και 10.000x32x32x3 αντίστοιχα . Οι εικόνες προκειμένου να κατηγοριοποιηθούν , κανονικοποιουνται μεταξύ 0 και 1 και μετατρέπονται σε 1d πίνακες. Στην συνέχεια ,από τα train και test set επιλέγονται 2 κλάσεις και διαχωρίζονται σε ένα νέο dataset με train set 10.000 δειγμάτων και test set 2.000 δειγμάτων με labels -1 και 1.
- **UCI – Breast Cancer Wisconsin .** Η Breast Cancer Wisconsin περιέχει πληροφορίες για καρκίνο του μαστού από 569 δείγματα κυττάρων και 30 features. Πρόκειται για ένα σύνολο δεδομένων που χρησιμοποιείται συχνά για την ανάπτυξη και εκπαίδευση μοντέλων μηχανικής μάθησης για την ταξινόμηση καρκινικών κυττάρων του μαστού σε κακοήθη-Malignant(M) και καλοήθη- Benign(B). Οι εικόνες προκειμένου να κατηγοριοποιηθούν , κανονικοποιουνται μεταξύ 0 και 1 και τα labels γίνονται -1 και 1.

Σημειώνεται ότι οι περισσότερες δοκιμές έγιναν με το dataset **Breast Cancer Wisconsin** καθώς τα SVM έχει αποδειχθεί ότι δουλεύουν καλύτερα σε δεδομένα με σχετικά λίγα δείγματα και πολλά features και επιπλέον είναι πολύ πιο σύντομα σε χρόνους εκπαίδευσης και testing για τα hyperparameter tuning που θα παρουσιαστούν παρακάτω.

Ανάλυση κυρίων συνιστωσών (PCA)

Με ανάλυση κύριων συνιστωσών μειώνεται ο αριθμός των features τέτοιος ώστε να διατηρείται το 95 % της διακύμανσης των αρχικών δεδομένων . Γενικά , στόχος είναι να δημιουργήσουμε νέα features που προκύπτουν από το αρχικό σύνολο με σκοπό να διατηρηθεί όσο περισσότερη πληροφορία γίνεται σε πολύ μικρότερο αριθμό features .

2. ΚΝΝ και Nearest Centroid Κατηγοριοποιητές

Παρακάτω παρουσιάζονται τα αποτελέσματα εφαρμογής των κατηγοριοποιητών Nearest Centroid και k-Nearest Neighbor.

Αποτελέσματα

Breast Cancer Wisconsin

Nearest Neighbor with k=1

```
Training time: 0.010 seconds
Testing time: 0.103 seconds
For the 1-NN classifier accuracy is: 0.8333
For the 1-NN classifier, F1 score is: 0.8355
```

Nearest Neighbor with k=3

```
Training time: 0.01030111312866211 seconds
Testing time: 0.064 seconds
For the 3-NN classifier accuracy is: 0.9123
For the 3-NN classifier, F1 score is: 0.9135
```

Nearest Centroid with Euclidean distance

```
Training time: 0.003 seconds
Testing time: 0.007 seconds
For the Nearest Centroid classifier accuracy is: 0.9123
For the Nearest Centroid classifier, F1 score is: 0.9304
```

Nearest Centroid with Manhattan distance

```
Training time: 0.014 seconds
Testing time: 0.019 seconds
For the Nearest Centroid classifier accuracy is: 0.9123
For the Nearest Centroid classifier, F1 score is: 0.8962
```

Για το συγκεκριμένο dataset παρατηρείται ότι όλοι οι παραπάνω ταξινομητες δουλεύουν σε αρκετά ικανοποιητικό βαθμό με τον 1-NN να έχει την χειρότερη απόδοση και τον nearest Centroid με την ευκλείδεια απόσταση ως μετρική την καλύτερη .

3. SVM Implementation

Η κλάση SVM υλοποιεί ένα Support Vector Machine - SVM, χρησιμοποιώντας το εργαλείο cvxopt για την επίλυση ενός προβλήματος βελτιστοποίησης (Quadratic Programming - QP) για την εύρεση των support vectors. Αποτελείται από τις παρακάτω μεταβλητές :

- **kernel (πυρήνας)**: Ο τύπος πυρήνα που χρησιμοποιείται για τη μετασχηματισμό των δεδομένων.
- **C**: Το C είναι μία παράμετρος κανονικοποίησης, η οποία πολλαπλασιάζεται με τις μεταβλητές χαλαρότητας ξi στο τροποποιημένο πρόβλημα του τετραγωνικού προγραμματισμού. Στην ουσία είναι το βάρος τους κόστους των λάθος ταξινομήσεων. Για τιμή C=0, αγνοούμε τελείως τις παραμέτρους χαλαρότητας και οι λάθος ταξινομήσεις δεν μας ενδιαφέρουν καθόλου. Αν πάλι το C λάβει μεγάλη τιμή, δίνεται μεγαλύτερη σημασία στη σωστή ταξινόμηση των προτύπων.
- **gamma**: Η παράμετρος πυρήνα γκαουσιανής συνάρτησης. Επηρεάζει την ευαισθησία του μοντέλου στα σημεία δεδομένων.
- **degree** : Ο βαθμός του πυρήνα πολυωνυμικής συνάρτησης.
- **constant** : Η σταθερά πολυωνυμικού πυρήνα.
- **disp** : Ένα boolean που ελέγχει εάν θέλουμε να εμφανίζονται μηνύματα κατά τη διάρκεια της εκπαίδευσης και των προβλέψεων.

Συναρτήσεις :

- **Fit ():** Η μέθοδος **fit** αναλαμβάνει την εκπαίδευση του μοντέλου. Κατά την εκτέλεσή της, υπολογίζει τον γραμμικό πίνακα (Gram matrix) από τα δεδομένα εκπαίδευσης, χρησιμοποιώντας τον πυρήνα που έχει οριστεί. Στην συνέχεια , δημιουργούνται οι πίνακες που απαιτούνται για το πρόβλημα βελτιστοποίησης Quadratic Programming(P, q, A, b). Έπειτα, ορίζονται οι πίνακες περιορισμών (constraints) ανάλογα με την παράμετρο C. Ο πίνακας G περιέχει τους περιορισμούς του προβλήματος QP και ο πίνακας h περιέχει τις αντίστοιχες τιμές στις συνθήκες περιορισμού. Οι συντελεστές Lagrange υπολογίζονται από τη λύση του QP προβλήματος και με βάση αυτούς προκύπτουν τα support vectors. Τέλος, υπολογίζεται το σταθερό όριο (bias).

- **Predict ():** Η συνάρτηση predict υλοποιεί την πρόβλεψη του μοντέλου SVM ανάλογα με τον τύπο του kernel που χρησιμοποιείται στο test set και επιστρέφει ως έξοδο τις προβλέψεις (-1 ή 1 για κάθε test sample).

Τύποι kernel: Οι τύποι πυρήνων που χρησιμοποιούνται παρουσιάζονται παρακάτω :

- **Linear Kernel (Γραμμικός Πυρήνας)**: Υπολογίζει το εσωτερικό γινόμενο μεταξύ των 2 vectors εισόδου.

- **Gaussian – RBF Kernel**: Υπολογίζει την τιμή του gaussian πυρήνα για τα 2 vectors εισόδου βρίσκοντας αρχικά την ευκλείδεια απόσταση μεταξύ τους και βάζοντας τα στην εκθετική συναρτηση.

- **Polynomial Kernel:** Υπολογίζει την τιμή του πολυωνυμικού πυρήνα για δύο vectors εισόδου βάσει της εκθετικής αύξησης στο εσωτερικό γινόμενο τους και του βαθμού του πολυωνύμου.

- **Sigmoid Kernel**: Υπολογίζει την τιμή του Sigmoid πυρήνα για δύο vectors εισόδου με βάση τη σιγμοειδή συνάρτηση, το εσωτερικό γινόμενο των διανυσμάτων τους και τις παραμέτρους alpha και c.

Μια άλλη σημαντική παράμετρος η οποία επηρεάζει αν είναι διαφορετική για ιδίες παραμέτρους την απόδοση του SVM είναι το threshold για το οποίο ο τελεστής Lagrange ενός δείγματος θεωρείται support vector, καθώς ουσιαστικά καθορίζει το πλήθος των support vectors.

Δοκιμές μοντέλων SVM με το **UCI – Breast Cancer Wisconsin :**

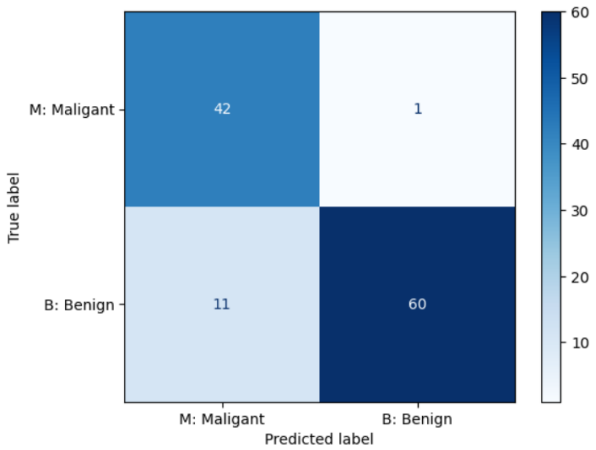
1. **Linear Kernel**
   Για C=1 παρατίθενται τα ακόλουθα αποτελέσματα :

```
Number of samples: 455
Number of features: 30
Gram matrix elapsed time: 1.2065250873565674s

QP solver elapsed time: 0.3557860851287842s

Testing elapsed time: 0.4353342056274414s

Test set accuracy:  0.8947368421052632
```

Και το classification Report

	precision	recall	f1-score	support
-1.0	0.79	0.98	0.88	43
1.0	0.98	0.85	0.91	71
accuracy			0.89	114
macro avg	0.89	0.91	0.89	114
weighted avg	0.91	0.89	0.90	114

Ενώ παρατίθεται και ο confusion matrix



Παρατηρείται ότι γενικά κατηγοριοποιεί σωστά το svm με linear kernel . Από τον πίνακα σύγχυσης παρατηρείται ότι η κλάση Maligant (κακοήθης) κατηγοριοποιείται 42 φορές σωστά ενώ μόλις μια αναγνωρίζεται λάθος ως Benign ενώ από την άλλη η κλάση Benign λειτουργεί λίγο χειροτέρα καθώς 60 φορές κατηγοριοποιείται σωστά ενώ 11 φορές αναγνωρίζεται ως Malignant. Γενικά, είναι προτιμότερο να αναγνωρίζεται κάποιος που δεν έχει καρκίνο λάθος (δηλαδή Malignant ενώ είναι Benign) πάρα το αντίθετο . Επομένως , μπορούμε να είμαστε σχετικά ικανοποιημένοι καθώς αυτό γίνεται μια στις 43 φορές .

2. Gaussian-RBF Kernel

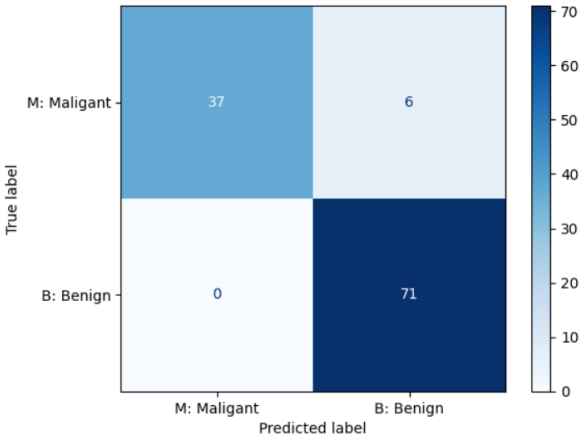
Για C = 1 και gamma = 0.1 παρατίθενται τα ακόλουθα αποτελέσματα :

```
Number of samples: 455
Number of features: 30
Gram matrix elapsed time: 3.9026596546173096s

QP solver elapsed time: 0.5097439289093018s

Testing elapsed time: 1.5202796459197998s

Test set accuracy:  0.9473684210526315
```

Και το classification Report

	precision	recall	f1-score	support
-1.0	1.00	0.86	0.92	43
1.0	0.92	1.00	0.96	71
accuracy			0.95	114
macro avg	0.96	0.93	0.94	114
weighted avg	0.95	0.95	0.95	114

Ενώ παρατίθεται και ο confusion matrix



Παρατηρείται ότι γενικά κατηγοριοποιεί σωστά το svm με RBF kernel . Από τον πίνακα σύγχυσης παρατηρείται ότι η κλάση Malignant (κακοήθης) κατηγοριοποιείται 37 φορές σωστά ενώ 6 φορές αναγνωρίζεται λάθος ως Benign ενώ από την άλλη η κλάση Benign λειτουργεί άριστα καθώς 71 φορές κατηγοριοποιείται σωστά ενώ καμία αναγνωρίζεται ως Malignant. Γενικά, είναι προτιμότερο να αναγνωρίζεται κάποιος που δεν έχει καρκίνο λάθος (δηλαδή Malignant ενώ είναι Benign) πάρα το αντίθετο . Επομένως , δεν μπορούμε να είμαστε ικανοποιημένοι καθώς 6 ασθενείς που έχουν καρκίνο δεν έχουν την σωστή διάγνωση κάτι που κάνει προβληματικό το συγκεκριμένο μοντέλο πάρα το γεγονός ότι το accuracy έχει αυξηθεί σε σχέση με το γραμμικό kernel .

3. Polynomial Kernel

Για C = 1 και degree = 2 (βαθμός πολυωνύμου) παρατίθενται τα ακόλουθα αποτελέσματα :

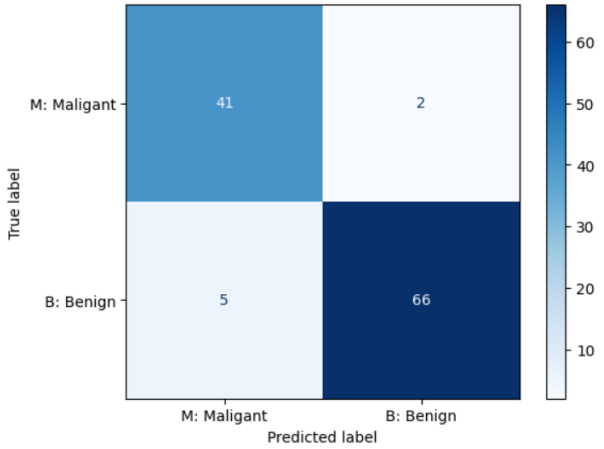
```
Number of samples: 455
Number of features: 30
Gram matrix elapsed time: 3.012148857116699s

QP solver elapsed time: 0.35548996925354004s

Testing elapsed time: 0.8083102703094482s

Test set accuracy:  0.9385964912280702
```

Παρατηρείται ότι το πιο χρονοβόρο μοντέλο είναι αυτό που χρησιμοποιεί τον πολυωνυμικού πυρήνα .

Και το classification report :

	precision	recall	f1-score	support
-1.0	0.89	0.95	0.92	43
1.0	0.97	0.93	0.95	71
accuracy			0.94	114
macro avg	0.93	0.94	0.94	114
weighted avg	0.94	0.94	0.94	114

Ενώ παρατίθεται και ο confusion matrix



Παρατηρείται ότι γενικά κατηγοριοποιεί σωστά το svm με polynomial kernel . Από τον πίνακα σύγχυσης παρατηρείται ότι η κλάση Malignant (κακοήθης) κατηγοριοποιείται 41 φορές σωστά ενώ 2 φορές αναγνωρίζεται λάθος ως Benign ενώ από την άλλη η κλάση Benign λειτουργεί καλά καθώς 66 φορές κατηγοριοποιείται σωστά ενώ 5 φορές αναγνωρίζεται ως Malignant. Γενικά, είναι προτιμότερο να αναγνωρίζεται κάποιος που δεν έχει καρκίνο λάθος (δηλαδή Malignant ενώ είναι Benign) πάρα το αντίθετο . Επομένως μπορούμε να είμαστε σχετικά ικανοποιημένοι καθώς 2 ασθενείς που έχουν καρκίνο δεν έχουν την σωστή διάγνωση κάτι που κάνει το συγκεκριμένο μοντέλο καλύτερο σε σχέση με αυτό που χρησιμοποιεί RBF Kernel και κοντά σε αυτό με το linear kernel καθώς ενώ το accuracy έχει αυξηθεί σε σχέση με το γραμμικό kernel, 1 παραπάνω ασθενής δεν έχει σωστή διάγνωση .

4. Hyperparameter Tuning

Σε αυτό το στάδιο της εργασίας και μετά από τις δοκιμές για κάποιες υπερπαραμετρους θα γίνει προσπάθεια εύρεσης των καταλληλότερων για το επιλεγμένο dataset. Αρχικά , ενώνονται τα train και test sets σε ένα ενιαίο , προκειμένου να χρησιμοποιηθεί η τεχνική cross validation. Για κάθε τύπο kernel θα γίνει grid search για τις αντίστοιχες υπερπαραμετρους . Με for loops για κάθε παράμετρο που εξετάζεται επιλέγεται το βέλτιστο μοντέλο και ελέγχονται όλοι οι δυνατοί συνδυασμοί, με 5-fold Cross validation για να έχει αξιοπιστία η τελική επιλογή και προκειμένου να επιτευχθούν πιο ισχυρές εκτιμήσεις της ακρίβειας κάθε μοντέλου Τέλος, επιλέγεται το μοντέλο με την καλύτερη ακρίβεια. Το dataset που θα χρησιμοποιηθεί είναι το breast cancer Wisconsin. Χρησιμοποιήθηκε το συγκεκριμένο dataset ώστε να γίνει σύγκριση μεταξύ πολλών μοντέλων και να χρησιμοποιηθούν όλες οι παράμετροι και να σχηματιστούν όλοι οι δυνατοί συνδυασμοί . Ένα θέμα το οποίο , παρατήρησα είναι ότι λόγω του μικρού dataset (αρά και μικρό test set) υπήρχε διαφοροποίηση στο accuracy για τις ιδίες παραμέτρους στο ήδη χωρισμένο dataset , στο οποίο έγιναν δοκιμές αρχικά .

A. Linear Kernel SVM

Η μοναδική παράμετρος είναι το C και το grid search για την εύρεση του βέλτιστου μοντέλου θα γίνει με τις εξής τιμές : **C = [0.1, 1, 10, 100, 1000 , 10e3]**

Αποτελέσματα :

Model	C	Accuracy
1	0.1	0.0
2	1	0.8945
3	10	0.9069
4	100	0.9157
5	1000	0.9245
6	10e3	0.9145

Παρατηρείται ότι :

```
Optimal parameters for the model:
C =  1000.0
Accuracy: 0.9245
```

Το μοντέλο με χρήση γραμμικού πυρήνα παρατηρείται ότι λειτουργεί ικανοποιητικά και διαχωρίζει τις 2 κλασεις . Θα μπορούσε να χαρακτηριστεί γραμμικά διαχωρίσιμο καθώς το accuracy είναι αρκετά υψηλό και παράλληλα με αύξηση του C έως έναν βαθμό αυξάνεται.

B. Gaussian Kernel SVM

Οι παράμετροι για grid search είναι το C και το gamma . Το grid search για την εύρεση του βέλτιστου μοντέλου γίνεται με 2 παραμέτρους σε αυτήν την περιπτωση με το **C = [1, 10, 100, 1000, 10e3]** και **gamma = [0.01, 0.1,**

1, 10, 100]. Τα μοντέλα- συνδυασμοί που προκύπτουν παρουσιάζονται παρακάτω :

Model	C	Gamma	Accuracy
1	1	0.01	0.8260
2	1	0.1	0.8699
3	1	1	0.8804
4	1	10	0.8700
5	1	100	0.9156
6	10	0.01	0.8699
7	10	0.1	0.8823
8	10	1	0.8734
9	10	10	0.8770
10	10	100	0.8331
11	100	0.01	0.8822
12	100	0.1	0.8752
13	100	1	0.7188
14	100	10	0.8294
15	100	100	0.7239
16	1000	0.01	0.8699
17	1000	0.1	0.6836
18	1000	1	0.6537
19	1000	10	0.6747
20	1000	100	0.8066
21	10e3	0.01	0.6749
22	10e3	0.1	0.6274
23	10e3	1	0.6309
24	10e3	10	0.6291
25	10e3	100	0.8472

Παρατηρείται ότι :

```
Optimal parameters for the model:
C =  1.0
gamma =  100.0
Accuracy:  0.915680794907623
```

Γενικότερα , παρατηρείται ότι σε αρκετές περιπτώσεις- συνδυασμούς υπάρχει σημαντική πτώση στο accuracy . Χαρακτηριστικό παράδειγμα είναι τα μοντέλα 17 με 24 οπού το accuracy μειώνεται αναμεσά στο 60-65%. Το βέλτιστο μοντέλο με RBF Kernel λειτουργεί λίγο χειροτέρα από το αντίστοιχο γραμμικού πυρήνα. Επίσης , παρατηρείται ότι η παράμετρος C είναι πολύ μικρότερη σε σχέση με τον γραμμικό πυρήνα ενώ το γράμμα αρκετά μεγάλο.

C. Polynomial Kernel SVM

Οι παράμετροι για το grid search είναι το C , ο βαθμός του πολυωνύμου degree και η παράμετρος constant. Οι τιμές για τις οποίες γίνεται το grid search είναι : **C = [0.1 , 1, 10, 100, 1000] , degree = [2, 3, 4] , constant = [0 , 0.1, 1, 2] .** Τα αποτελέσματα που προκύπτουν παρατίθενται παρακάτω :

Model	C	degree	constant	Accuracy
1	0.1	2	0	0
2	0.1	2	0.1	0
3	0.1	2	1	0
4	0.1	2	2	0
5	0.1	3	0	0
6	0.1	3	0.1	0
7	0.1	3	1	0
8	0.1	3	2	0
9	0.1	4	0	0
10	0.1	4	0.1	0
11	0.1	4	1	0
12	0.1	4	2	0
13	1	2	0	0.8787
14	1	2	0.1	0.8927
15	1	2	1	0.8998
16	1	2	2	0.9015
17	1	3	0	0.8664
18	1	3	0.1	0.8734
19	1	3	1	0.9033
20	1	3	2	0.9051
21	1	4	0	0.8241
22	1	4	0.1	0.8630
23	1	4	1	0.9033
24	1	4	2	0.9174
25	10	2	0	0.9015
26	10	2	0.1	0.9015
27	10	2	1	0.9122
28	10	2	2	0.9156
29	10	3	0	0.8822
30	10	3	0.1	0.8927
31	10	3	1	0.9156
32	10	3	2	0.9173
33	10	4	0	0.8699
34	10	4	0.1	0.8822
35	10	4	1	0.9156
36	10	4	2	0.9244
37	100	2	0	0.9103
38	100	2	0.1	0.9190
39	100	2	1	0.9192

40	100	2	2	0.9261
41	100	3	0	0.8998
42	100	3	0.1	0.9050
43	100	3	1	0.9191
44	100	3	2	0.9262
45	100	4	0	0.8823
46	100	4	0.1	0.9016
47	100	4	1	0.9226
48	100	4	2	0.9226
49	100	2	0	0.9156
50	1000	2	0.1	0.9191
51	1000	2	1	0.9279
52	1000	2	2	0.9261
53	1000	3	0	0.9086
54	1000	3	0.1	0.9156
55	1000	3	1	0.9297
56	1000	3	2	0.9279
57	1000	4	0	0.8946
58	1000	4	0.1	0.8946
59	1000	4	1	0.9156
60	1000	4	2	0.9279

Παρατηρείται ότι :

```
Optimal parameters for the model:
C =  1000.0
Degree =  3
Constant =  1.0
Accuracy:  0.9297469337059463
```

Γενικά , παρατηρείται ότι σε κάθε μοντέλο – συνδυασμό που δοκιμάστηκε τα αποτελέσματα είναι πολύ ικανοποιητικά (πάνω από 82% accuracy) σε αντίθεση με τους άλλους 2 τύπους kernel . Επιστής, ήταν και η πιο χρονοβόρα εκτέλεση.

Σύνοψη

Στον παρακάτω πίνακα συνοψίζονται τα accuracy για τους KNN, NN και SVM κατηγοριοποιητες

1 – Nearest Neighbor	83,55 %
3 – Nearest Neighbors	91,35 %
Nearest Centroid (Euclidean distance)	93,04 %
SVM with linear kernel	92,45 %
SVM with RBF kernel	91,56 %
SVM with Polynomial kernel	**92,97 %**