ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Παρουσίαση Εργασίας: Bayes, kNN, SVM

Ηλιάνα Κόγια
ΑΕΜ:10090

2023/24

Dataset (μέρη Α,Β,Γ):
3 classes, 2 features

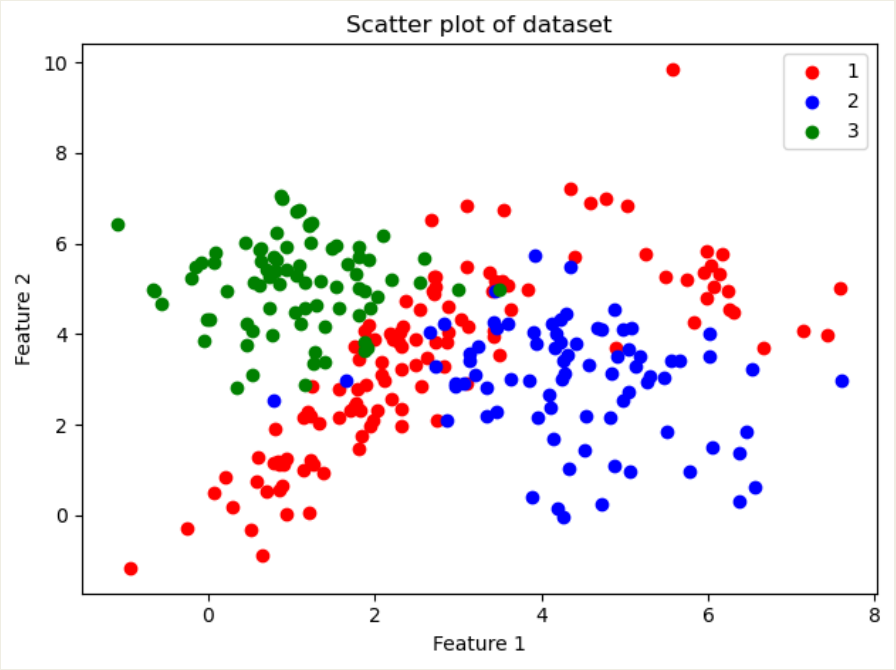
- Scatter plot of data
- Ιστογράμματα των τιμών των 2 feature

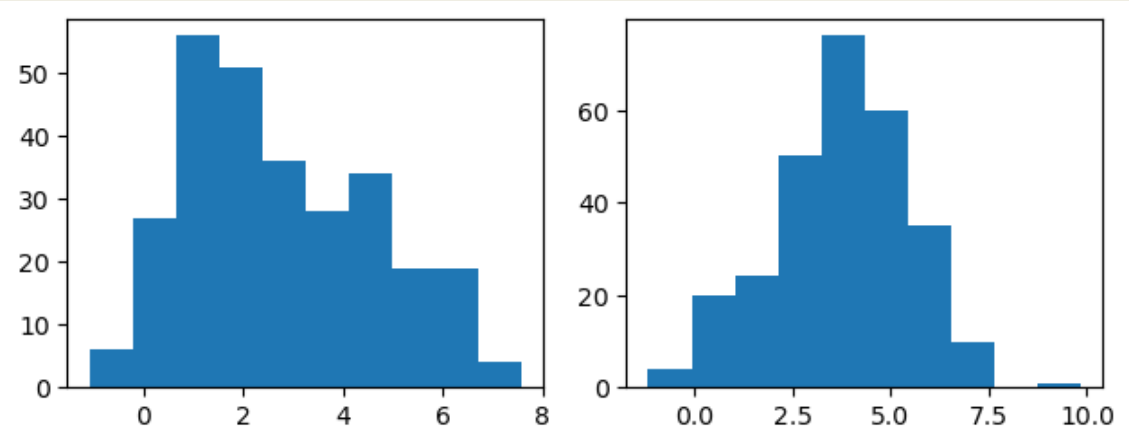
Train data:

Class 1: 53 samples

Class 2: 39 samples

Class 3: 48 samples





Μέρος Α
Bayes Classifier με Maximum Likelihood εκτίμηση παραμέτρων

- Q1 : ίδιος πίνακας συνδιασποράς για όλες τις κλάσεις

Για κάθε κλάση υπολογίζουμε τη discriminant function g για ένα δείγμα **x**, σύμφωνα με τον τύπο:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

Όπου η μέση τιμή και ο πίνακας συνδιασποράς Σ που είναι κοινός για κάθε κλάση βρίσκεται από τους παρακάτω τύπους, οι οποίοι προκύπτουν αν θεωρήσουμε Gaussian κατανομή και χρησιμοποιήσουμε Maximum Likelihood τεχνική:

Για τη μέση τιμή των δειγμάτων που ανήκουν σε μία κλάση:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

Για τον πίνακα συνδιασποράς των δειγμάτων που ανήκουν σε μία κλάση χρησιμοποιείται ο τύπος:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

Ωστόσο, για την πρώτη περίπτωση που θεωρούμε ίδιο Σ, χρησιμοποιούμε τον σταθμισμένο άθροισμα των τριών πινάκων συνδιασποράς:

$$\hat{\Sigma} = \frac{n_A}{n}\hat{\Sigma}_A + \frac{n_B}{n}\hat{\Sigma}_B + \frac{n_C}{n}\hat{\Sigma}_C$$

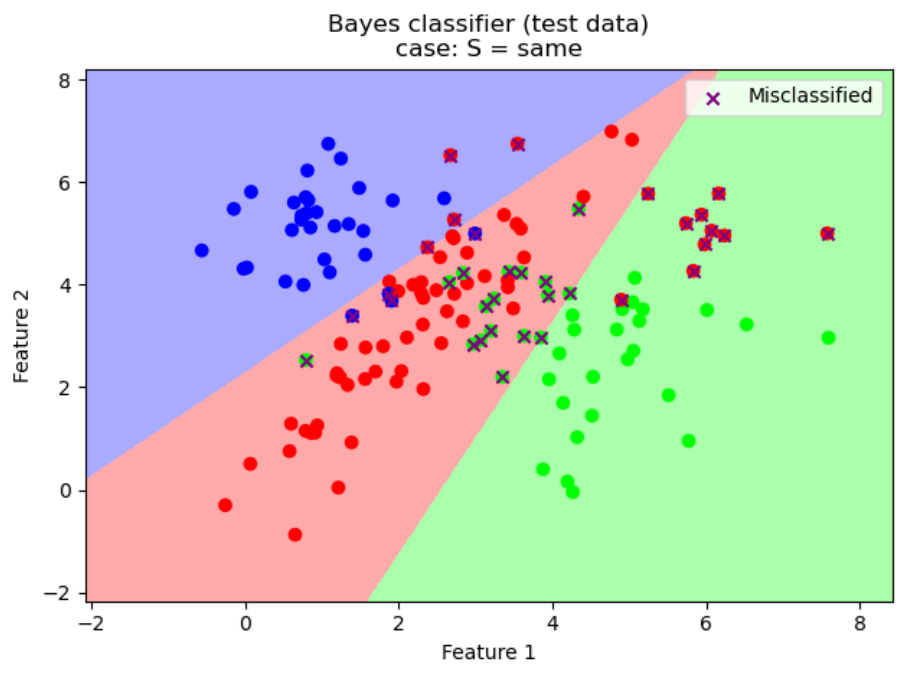
- Q2 : για διαφορετικό πίνακα συνδιασποράς για κάθε κλάση χρησιμοποιείται ο τύπος για την εύρεση των discriminant function:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

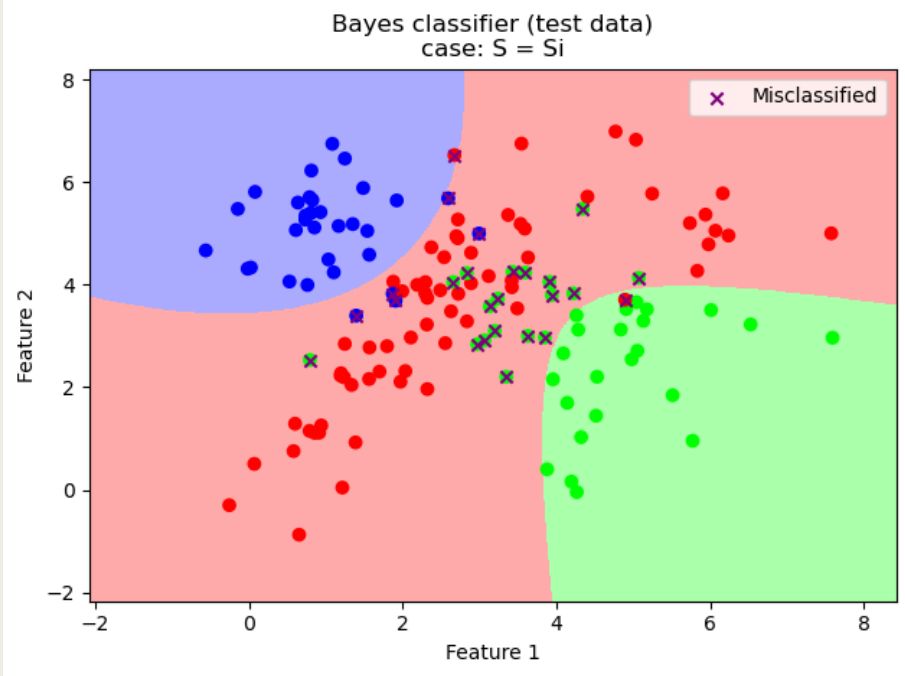
Με $\hat{\Sigma}_i = \hat{\Sigma}_A$, $\hat{\Sigma}_B$, $\hat{\Sigma}_C$ αντίστοιχα για κάθε μια από τις 3 κλάσεις

Εκπαιδεύουμε τα δεδομένα με Bayes ταξινομητή και προκύπτουν τα παρακάτω αποτελέσματα, από τα οποία παρατηρούμε ότι η υπόθεση για διαφορετικούς πίνακες Σ δίνει το καλύτερο αποτέλεσμα στο συγκεκριμένο πρόβλημα ταξινόμησης:

Case 1:
Accuracy in test set: 74.28%
Classification error: 25.72%
Γραμμικά όρια των περιοχών απόφασης

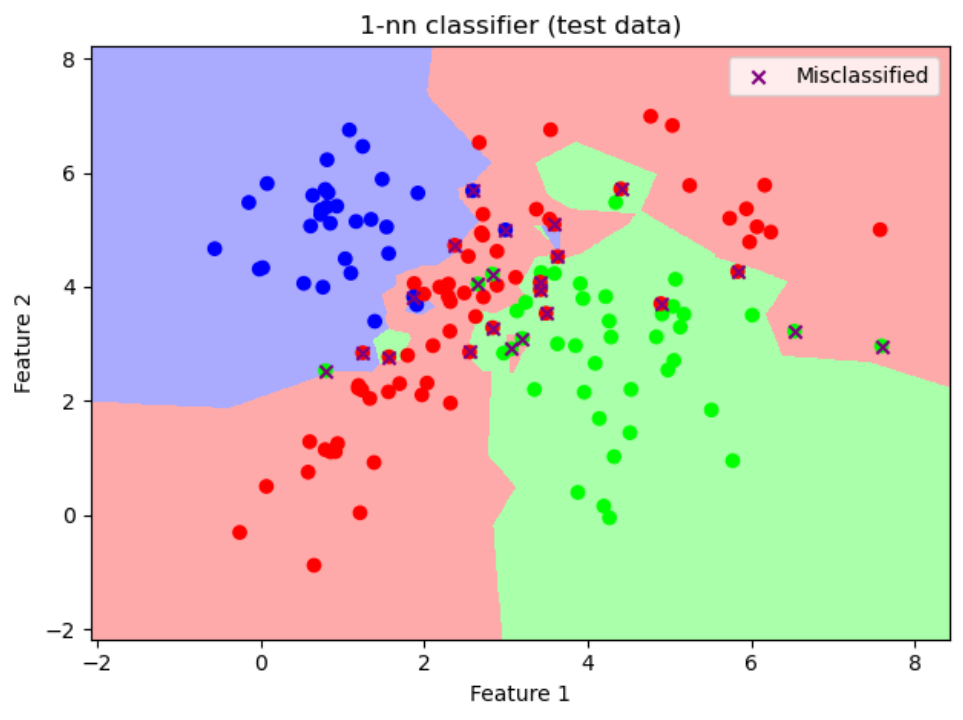


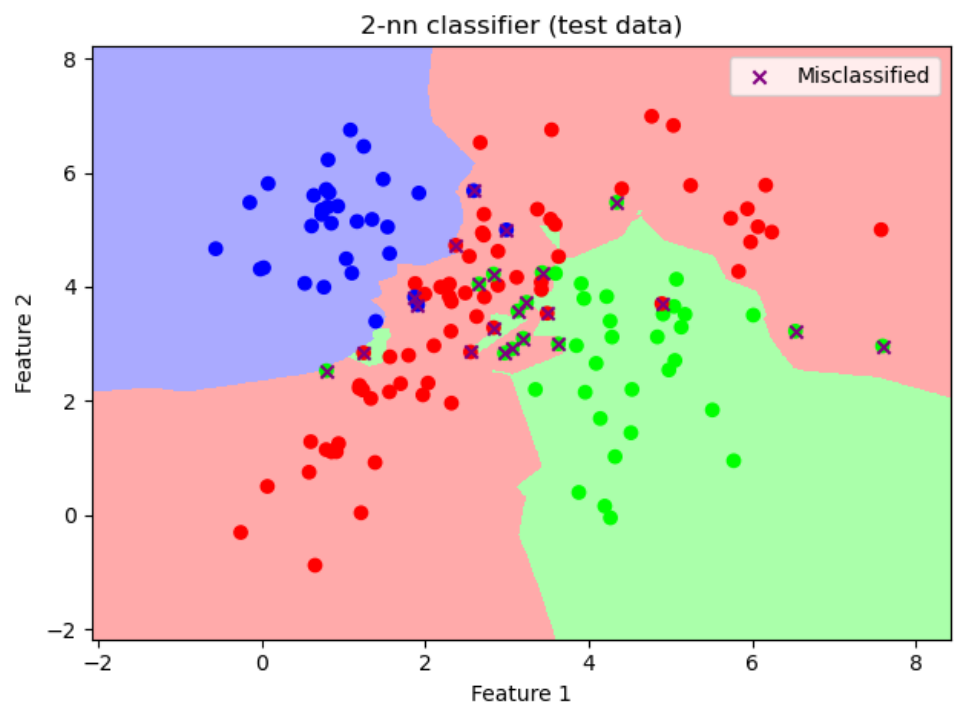
Case 2:
Accuracy in test set: 82.14%
Classification error: 17.84%
Μη γραμμικά όρια των περιοχών απόφασης

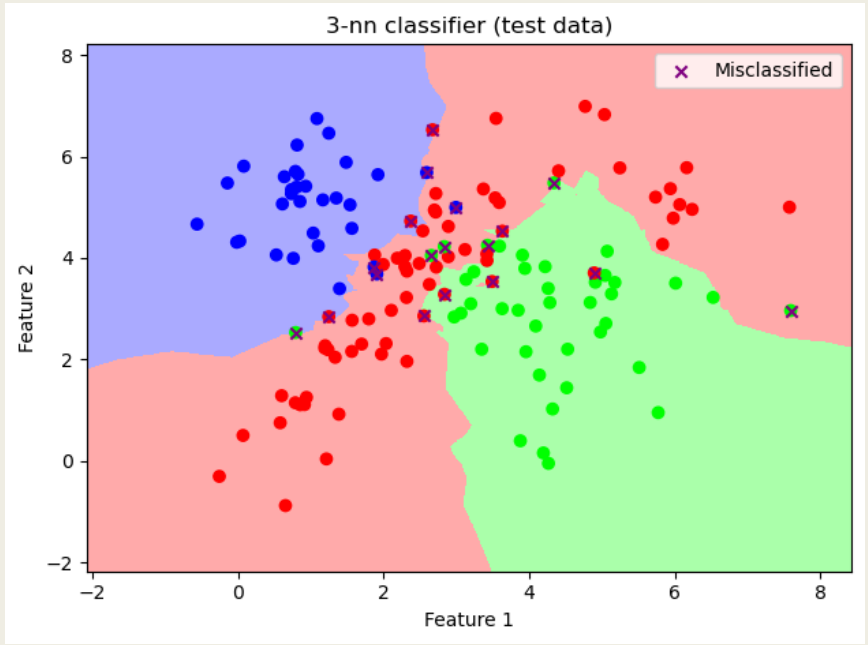


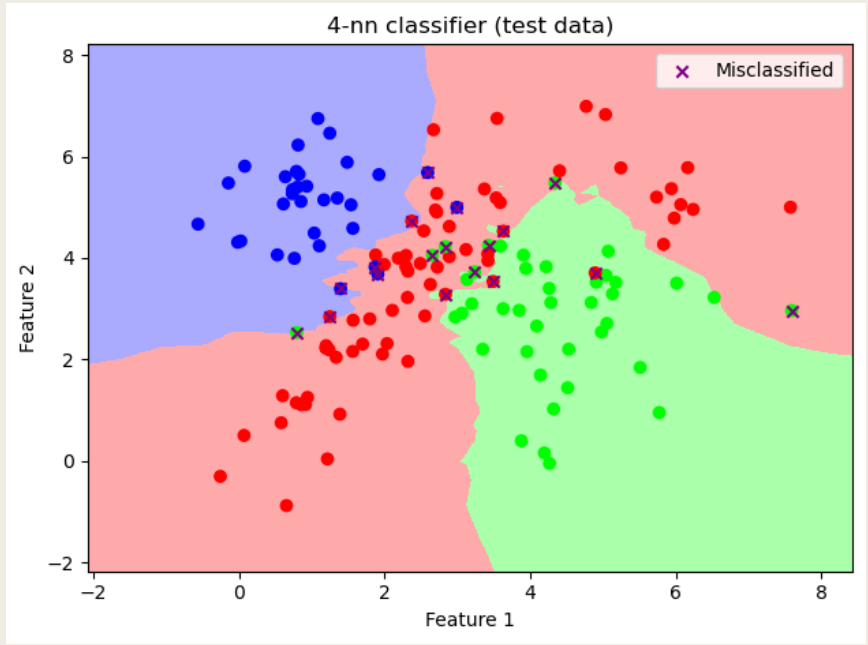
Μέρος Β
k-NN Classifier

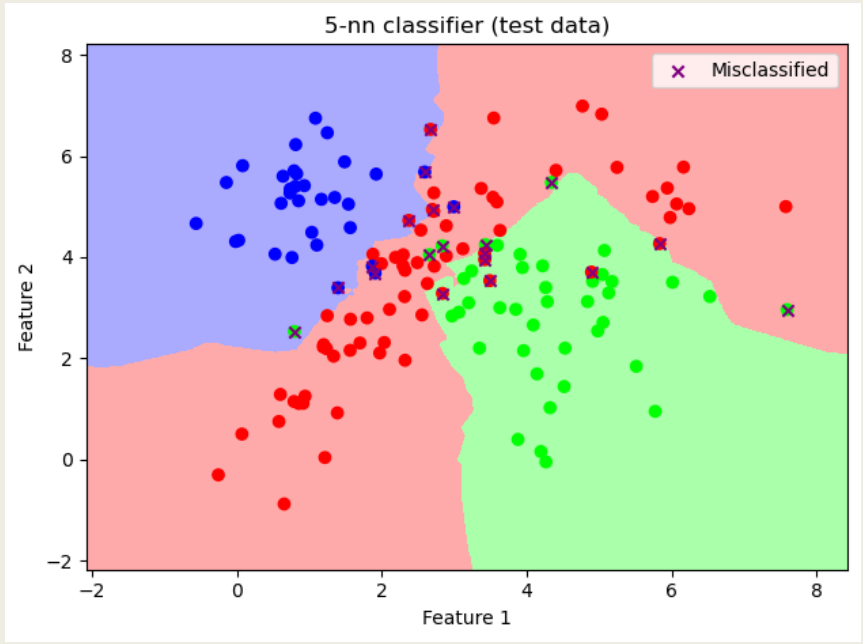
- Δοκιμάζουμε τα μοντέλα με παράμετρο k = 1,…,10 στο test set

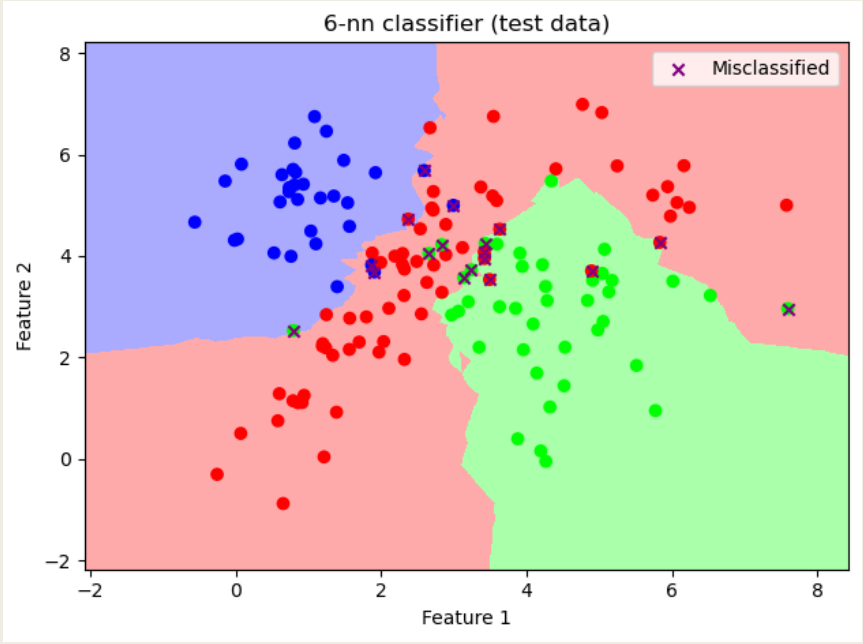


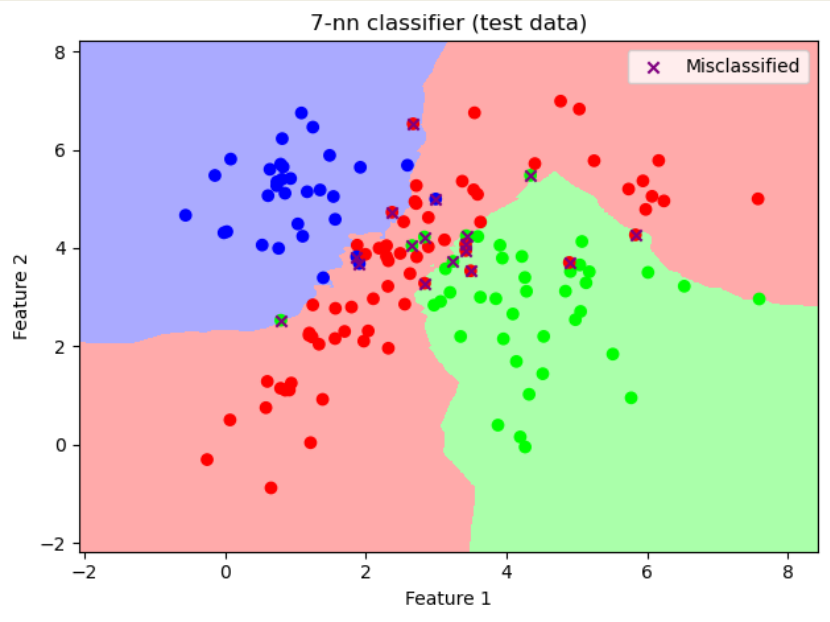


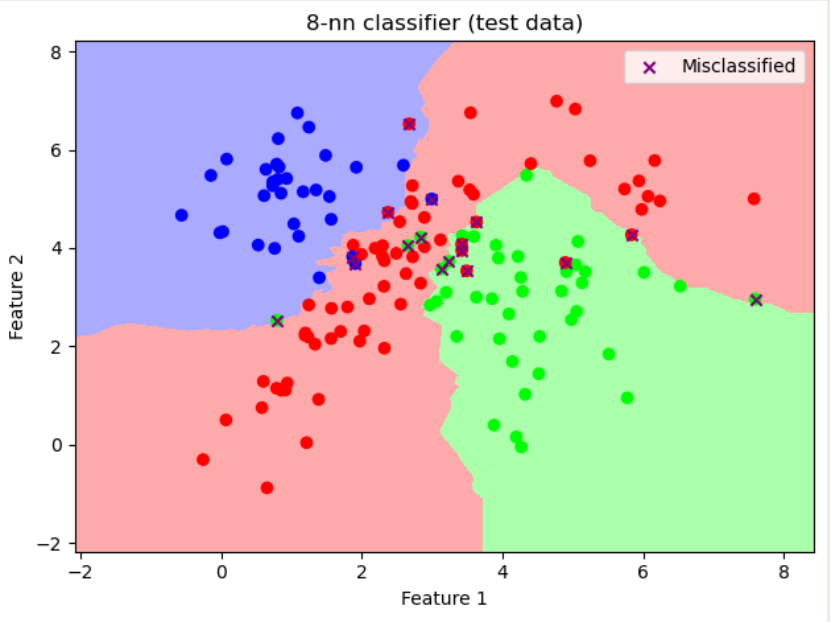

3-nn classifier (test data)
Misclassified
Feature 1
Feature 2

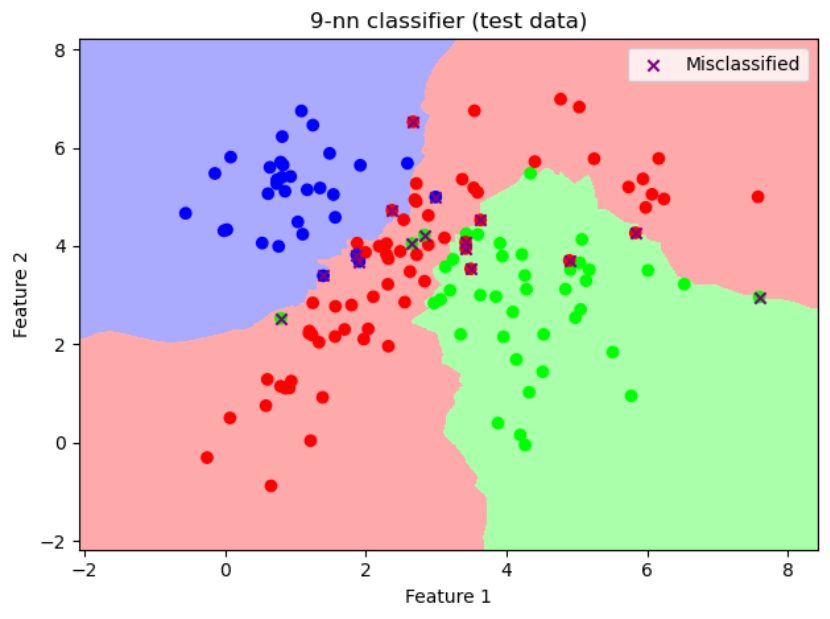

4-nn classifier (test data)
Misclassified
Feature 1
Feature 2

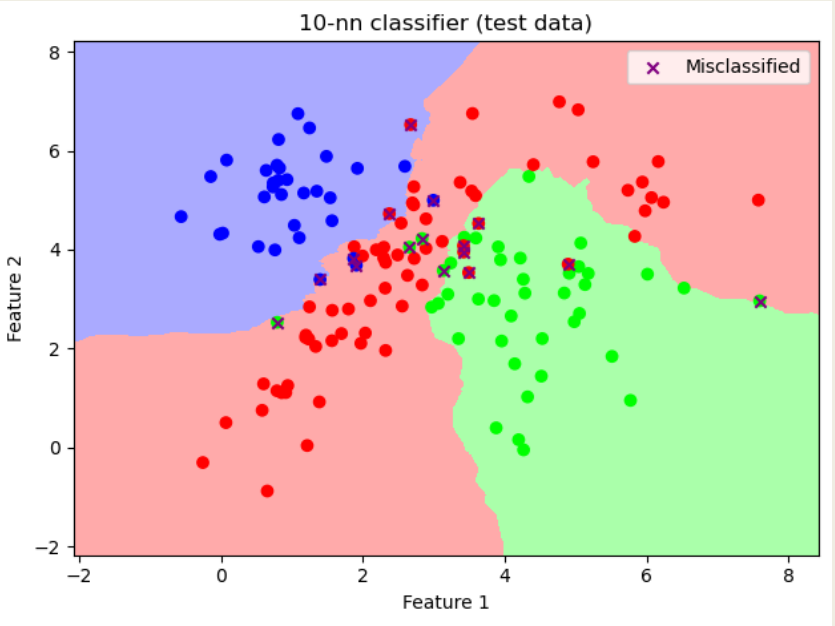

5-nn classifier (test data)
Misclassified
Feature 1
Feature 2


6-nn classifier (test data)
Misclassified
Feature 1
Feature 2


7-nn classifier (test data)
Misclassified
Feature 1
Feature 2


8-nn classifier (test data)
Misclassified
Feature 1
Feature 2


9-nn classifier (test data)
Misclassified
Feature 1
Feature 2

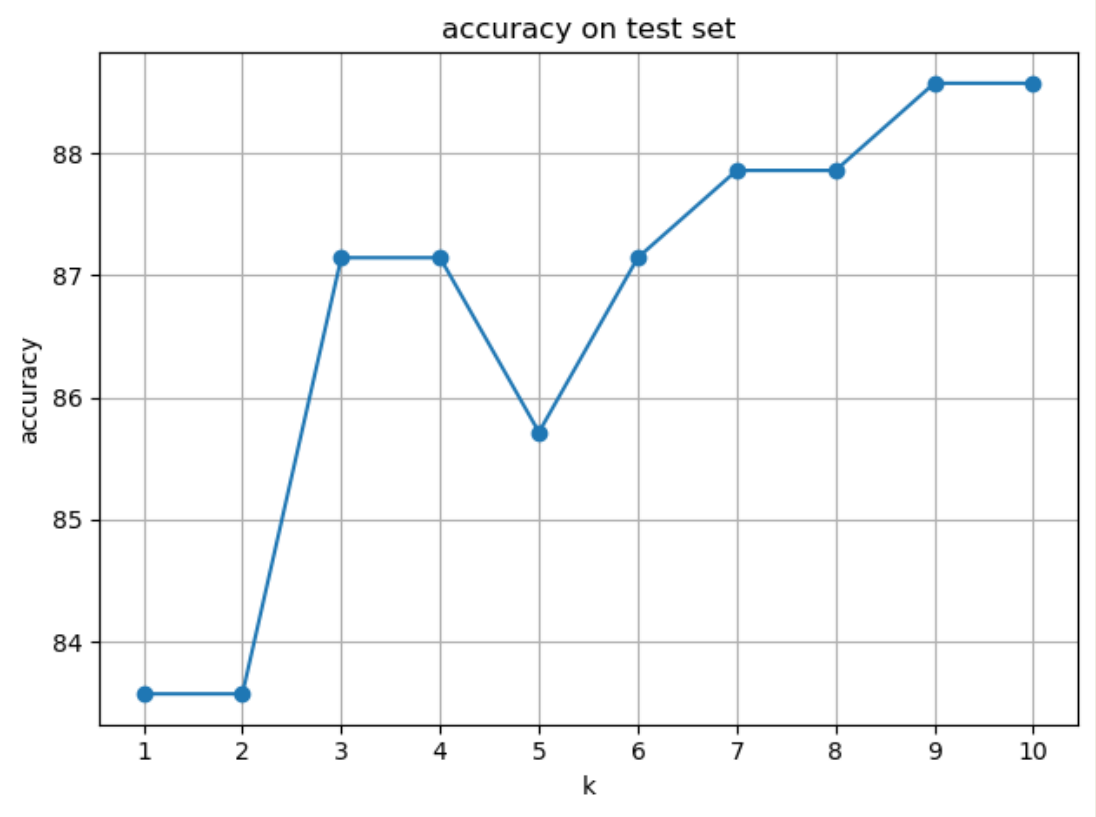

10-nn classifier (test data)
Misclassified
Feature 1
Feature 2

Σύγκριση μέρους Α και Β

Παρατίθεται το διπλανό διάγραμμα

Για k = 9 παρατηρείται η μεγαλύτερη ακρίβεια στο test set 88.57% και τα όρια των περιοχών απόφασης για τον k-NN είναι μη γραμμικά

Συγκρίνοτας με το Μέρος Α, παρατηρείται ότι ο ταξινομητής k-NN έχει διαφορετικά όρια περιοχών απόφασης και ανάλογα με την τιμή του k μπορεί να επιτευχθεί μεγαλύτερη ακρίβεια σε σχέση με τη δεύτερη περίπτωση του Bayes ταξινομητή



Μέρος Γ
Support Vector Machine classifier

- **Linear SVM**

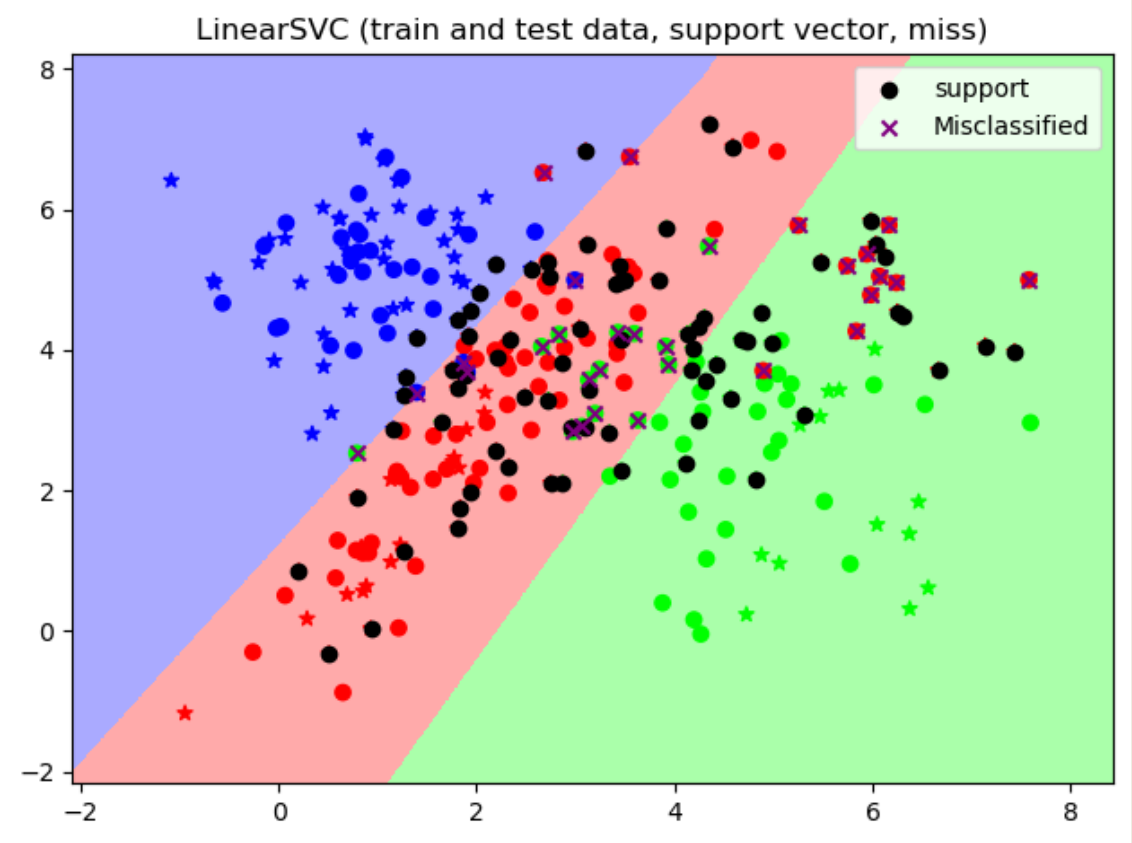
Με την τεχνική cross-validation (10-fold) εκπαιδεύουμε τον ταξινομητή για διάφορες τιμές της παραμέτρου C και για την καλύτερη τιμή της παραθέτουμε τα αποτελέσματα στο test set

C = 1

Accuracy test: 78.57%

Accuracy train: 80.71%

75 support vector



Μέρος Γ
Support Vector Machine classifier

- RBF kernel SVM

Με την τεχνική cross-validation (10 fold) εκπαιδεύουμε τον ταξινομητή για διάφορες τιμές της παραμέτρου C και gamma, και για τις καλύτερες τιμές παραθέτουμε τα αποτελέσματα στο test set

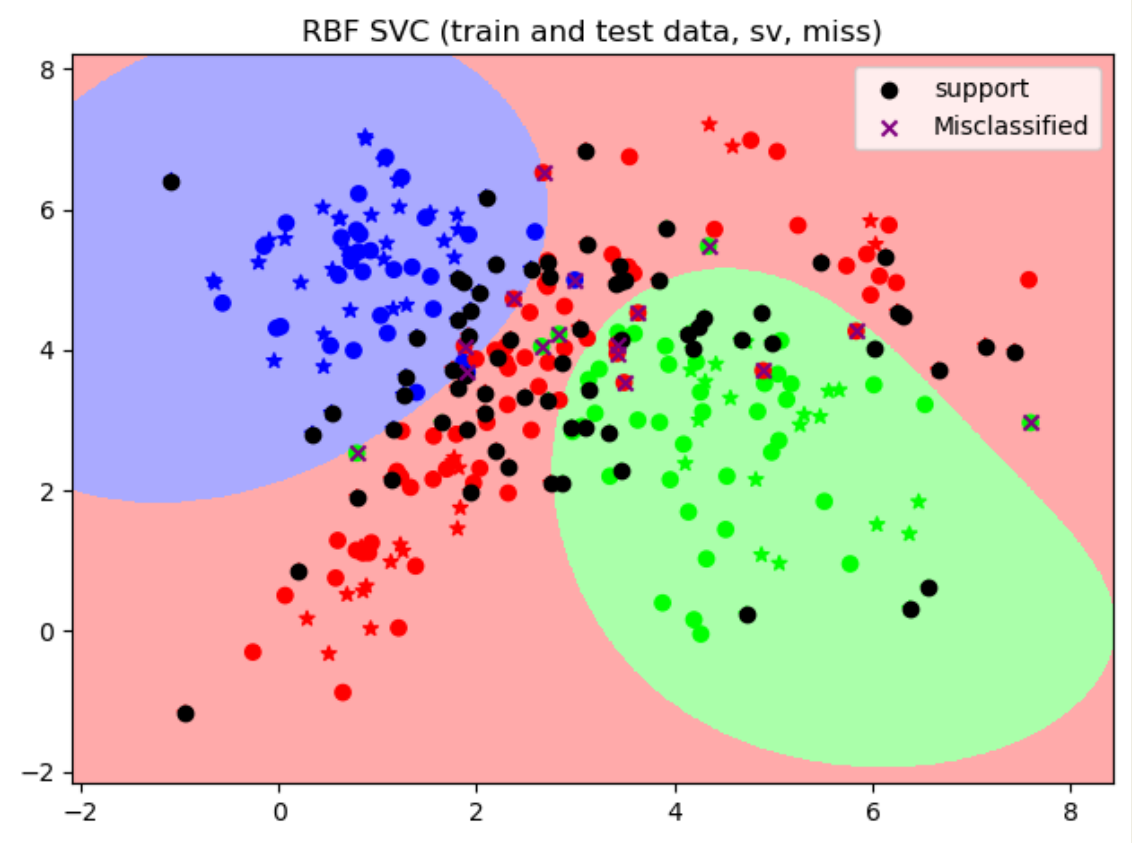
C = 1

gamma = 0.1

Accuracy test: 88.57%

Accuarcy train: 92.85%

72 support vector



Σύγκριση ταξινομητών

- Με RBF SVM και k-NN επιτυγχάνεται ίδια ακρίβεια στο test set, ωστόσο στο SVM συναντάται μεγαλύτερη πολυπλοκότητα στο μοντέλο καθώς και στην εύρεση των βέλτισων υπερ-παραμέτρων, διότι έχουμε 2 πλέον παραμέτρους σε αντίθεση με τον k-NN που χρειάζεται tuning μονο η παράμετρος k.
- Ο Bayes classifier δεν επιτυγχάνει βέλτιστο αποτελέσμα λόγω των υποθέσεων για τις κανονικές κατονομές και του πίνακα Σ

Μέρος Δ

- Data: σε training και test (80%-20%), ουσιαστικά τα test data αποτελούν τα validation data, διότι δεν γνωρίζουμε το test set.
- Κανονικοποίση MinMaxScaler στο [-1,1]
- PCA (0.98) για μείωση των features

Δοκιμές ταξινομητών για διάφορες παραμέτρους και καλύτερο accuracy στο test set:

- SVM linear, polynomial, rbf (88%)
- Nearest Centroid (80.5%)
- Gaussian NB (80.5%)
- MLP classifier (82.7%)
- HistGradientBoostingClassifier (85.5%)
- XGBoost (85.2%, eta = 0.5)

Βέλτιστο Μοντέλο

- Grid Search + Cross validation για την εύρεση των υπερ-παραμέτρων με νευρωνικά δίκτυα SVM
- Εστιάσαμε για την εύρεση του καλύτερου SVM classifier για το δοθέν dataset
- Το καλύτερο μοντέλο που βρέθηκε ήταν για **polynomial kernel SVM** και παραμέτρους:

{class_weight = 'balanced', C = 1, coef0 = 0.2, degree = 2, gamma = 0.6}

Με accuracy στο validation set 88%