ΟΝΟΜΑΤΕΠΩΝΥΜΟ : Ιωάννης-Ιάσων Γεωργακάς

ΑΜ : 2017030021

Στατιστική Μοντελοποίηση και Αναγνώριση προτύπων (ΤΗΛ 311)

Αναφορά 2ου Σετ Ασκήσεων

Θέμα 1: Λογιστική Παλινδρόμηση: Αναλυτική εύρεση κλίσης (Gradient)

Για ένα σύνολο m δεδομένων $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}),..., (x^{(m)}, y^{(m)})\}$,όπου $x^{(i)} \in R^{n\times 1}$είναι τα διανύσματα χαρακτηριστικών και $y^{(i)} \in \{0, 1\}$ορίζουν την κλάση κάθε δείγματος (label), χρησιμοποιείται η συνάρτηση της λογιστικής παλινδρόμησης προκειμένου να προβλεπτούν οι τιμές των $y^{(i)}$από τις αντίστοιχες τιμές $x^{(i)},\ i \in \{1, 2,..., m\}$.

Η συνάρτηση της λογιστικής παλινδρόμησης ορίζεται ως εξής :

$$h_{\theta}(x) = f(\theta^{T}x)$$

όπου $\theta = [\theta_1, \theta_2, \theta_3,..., \theta_n]^T$είναι οι παράμετροι του γραμμικού μοντέλου και $f()$είναι η λογιστική συνάρτηση που ορίζεται ως εξής:

$$f(z) = \frac{1}{1+e^{-z}}$$

Η λογιστική παλινδρόμηση εμπεριέχει ένα σφάλμα το οποίο μπορεί να υπολογιστεί χρησιμοποιώντας τον ακόλουθο τύπο:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(-y^{(i)} \cdot ln(\hat{y}^{(i)}) - (1 - y^{(i)}) \cdot ln(1 - \hat{y}^{(i)})) \quad (1)$$

Έστω $\hat{y}^{(i)} = h_{\theta}(x^{(i)})$η εκτίμηση της λογιστική συνάρτησης για το $y^{(i)}$τότε η εξίσωση (1) μετασχηματίζεται ως εξής:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(-y^{(i)} \cdot ln(h_{\theta}(x^{(i)})) - (1 - y^{(i)}) \cdot ln(1 - h_{\theta}(x^{(i)}))) \quad (2)$$

Για την βελτιστοποίηση του σφάλματος υπολογίζεται η κλίση(gradient) του σφάλματος J(θ) η οποία θα είναι ένα διάνυσμα ‘ισης διάστασης με το θ.

Για θ_jκαι $x_j^{(i)}$είναι η συνιστώσα των διανυσμάτων $\theta = [\theta_1, \theta_2, \theta_3,..., \theta_n]^T$και $x^{(i)} = [x_1^{(i)}, x_2^{(i)},... x_n^{(i)}]^T$ αντίστοιχα υπολογίζεται το j-στοιχείο της κλίσης του σφάλματος

$$ln(h_\theta(x^{(i)})) = ln(\frac{1}{1+e^{-\theta^T x^{(i)}}}) = -ln(1+e^{-\theta^T x^{(i)}})\ (I)$$

$$ln(1-h_\theta(x^{(i)})) = ln(1-\frac{1}{1+e^{-\theta^T x^{(i)}}}) = -ln(\frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}) \Rightarrow$$

$$ln(1-h_\theta(x^{(i)})) = ln(e^{-\theta^T x^{(i)}}) - ln(1+e^{-\theta^T x^{(i)}}) \Rightarrow$$

$$ln(1-h_\theta(x^{(i)})) = -\theta^T x^{(i)} - ln(1+e^{-\theta^T x^{(i)}})\ (II)$$

Από την εξίσωση (2) λόγω των εξισώσεων (I),(II) προκύπτει ότι:

$$(2) \Rightarrow J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(-y^{(i)}ln(h_\theta(x^{(i)})) - (1-y^{(i)})ln(1-h_\theta(x^{(i)}))) \Rightarrow$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)}ln(1+e^{-\theta^T x^{(i)}}) - (1-y^{(i)})(-\theta^T x^{(i)} - ln(1+e^{-\theta^T x^{(i)}}))) \Rightarrow$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)}ln(1+e^{-\theta^T x^{(i)}}) + \theta^T x^{(i)} + ln(1+e^{-\theta^T x^{(i)}}) - y^{(i)}\theta^T x^{(i)} - y^{(i)}\ ln(1+e^{-\theta^T x^{(i)}}))$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\ ln(1+e^{-\theta^T x^{(i)}}) - y^{(i)}\theta^T x^{(i)} + \theta^T x^{(i)}) \Rightarrow$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\ ln(1+e^{-\theta^T x^{(i)}}) - y^{(i)}\theta^T x^{(i)} + ln(e^{\theta^T \cdot x^{(i)}})) \Rightarrow$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\ ln((1+e^{-\theta^T x^{(i)}}) \cdot (e^{\theta^T \cdot x^{(i)}})) - y^{(i)}\theta^T x^{(i)}) \Rightarrow$$

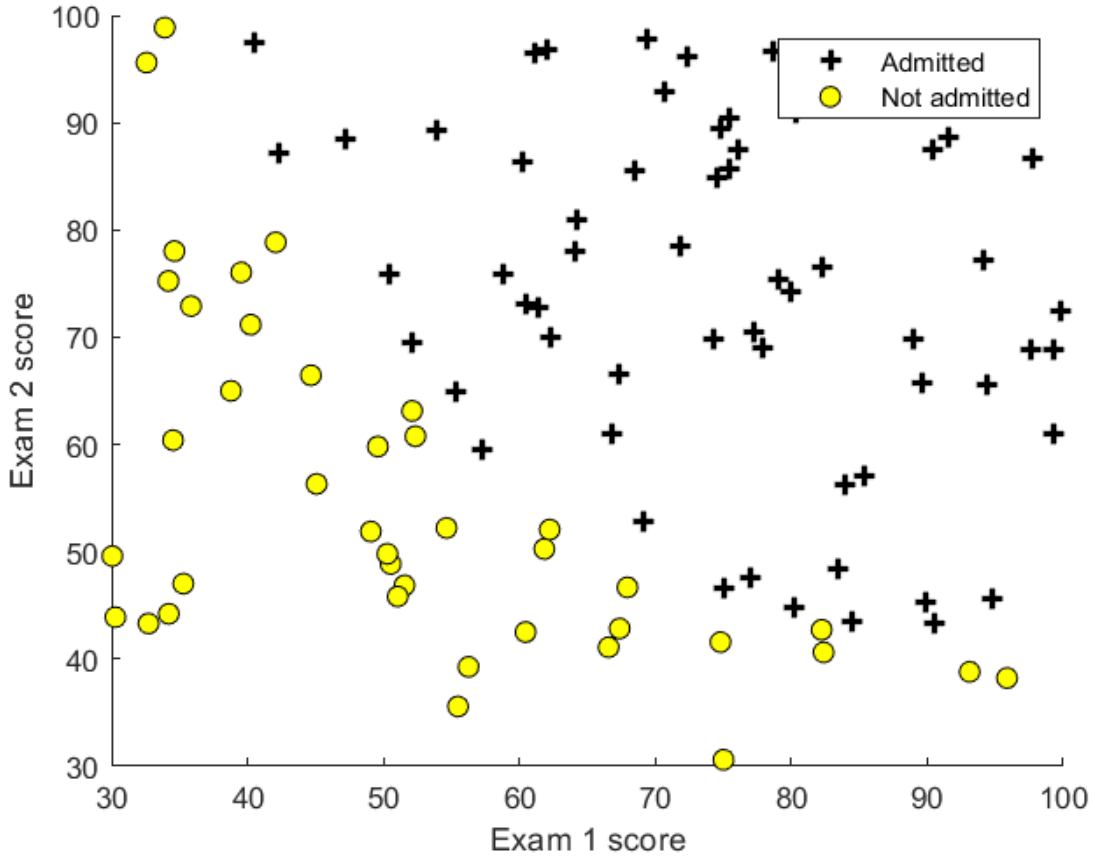
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\ ln(e^{\theta^T \cdot x^{(i)}} + 1) - y^{(i)}\theta^T x^{(i)})$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(\ \frac{x_j^{(i)} e^{\theta^T \cdot x^{(i)}}}{1+e^{\theta^T \cdot x^{(i)}}} - y^{(i)}x_j^{(i)}) = \frac{1}{m}\sum_{i=1}^{m}(\ \frac{e^{\theta^T \cdot x^{(i)}}}{1+e^{\theta^T \cdot x^{(i)}}} - y^{(i)})x_j^{(i)} \Rightarrow$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{1+e^{-\theta^T \cdot x^{(i)}}} - y^{(i)}\right)x_j^{(i)} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

Εφαρμόζοντας τη διαδικασία της λογιστική παλινδρόμηση με χρήση της MATLAB μοντελοποιήθηκε το πρόβλημα εισαγωγής ή μη ενός φοιτητή σε ένα πανεπιστήμιο με βάση τα αποτελέσματα του σε δύο εξετάσεις ώστε να προβλεφθεί αν θα γίνει δεκτός.

Αρχικά φορτώθηκαν παλαιότερες αιτήσεις φοιτητών από το αρχείο exam_scores_data1.txt της μορφής «Exam1Score, Exam2Score, [0: απόρριψη, 1: αποδοχή]» τα οποία αναπαραστάθηκαν στο δισδιάστατο χώρο με χρήση της συνάρτησης plotData.m .



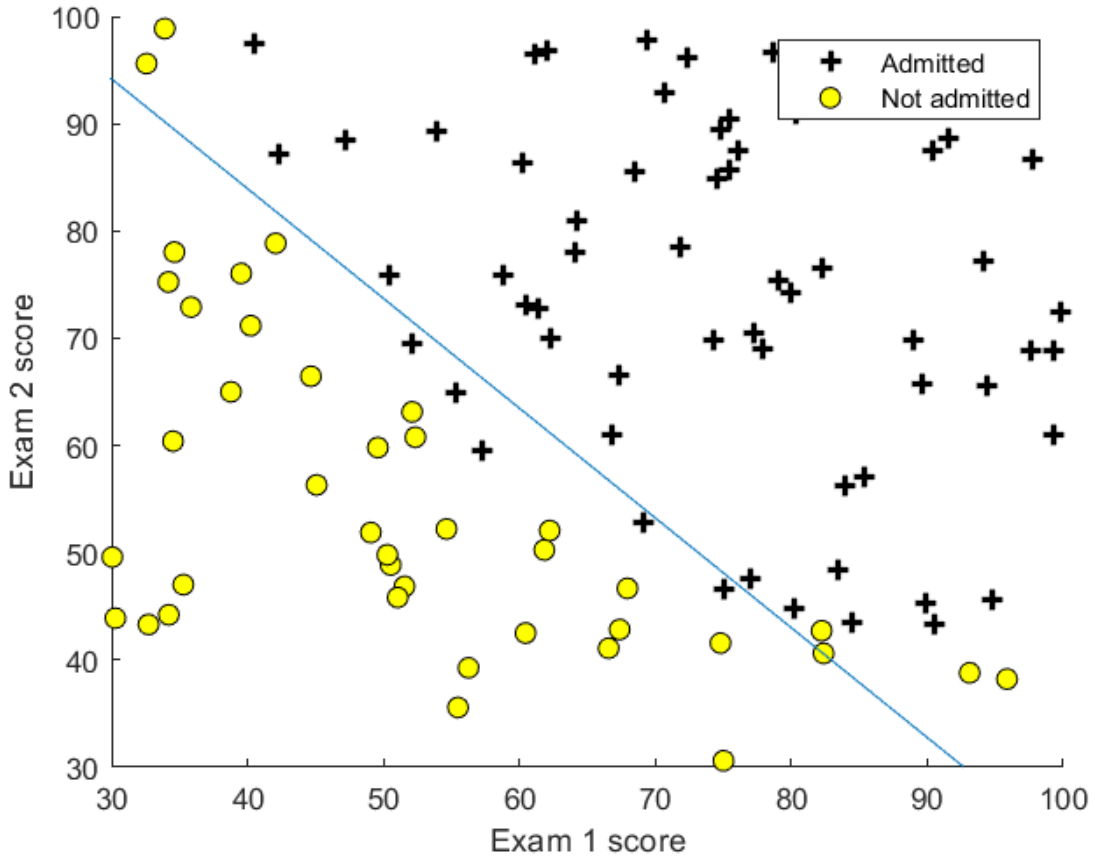
Αναπαράσταση αρχικών δειγμάτων από τις δυο εξετάσεις των φοιτητών

Στη συνέχεια υλοποιήθηκε η σιγμοειδής συνάρτηση στο αρχείο sigmoid.m , η συνάρτηση κόστους στο αρχείο costFunction.m και η συνάρτηση πρόβλεψης στο αρχείο predict.m ώστε να προβλεφθεί αν ο φοιτητής θα γίνει δεκτός με βάση διάφορες τιμές βαθμών στις δύο εξετάσεις.

Για τις αρχικές συνθήκες του θ = 0, το κόστος υπολογίζεται ίσο με τη τιμή 0.693 και η κλίση περίπου ίση με [-0.1,-12.009217,-11.262842] .

```
Plotting data with + indicating (y = 1) examples and o indicating (y = 0) examples.

Program paused. Press enter to continue.
Cost at initial theta (zeros): 0.693147
Gradient at initial theta (zeros):
 -0.100000
 -12.009217
 -11.262842

Program paused. Press enter to continue.
```

Τέλος απεικονίζεται το σύνορο απόφασης το οποίο οριοθετεί ποιες αιτήσεις φοιτητων γίνονται δεκτές με βάση τους βαθμούς του στις δυο εξετάσεις, η βελτιστοποίηση των παραμέτρων πραγματοποιήθηκε με κώδικα ο οποίος υπάρχει έτοιμος στο αρχείο My_logisticRegression.m .



Αναπαράσταση του συνόρου απόφασης των δειγμάτων των εξετάσεων των φοιτητών

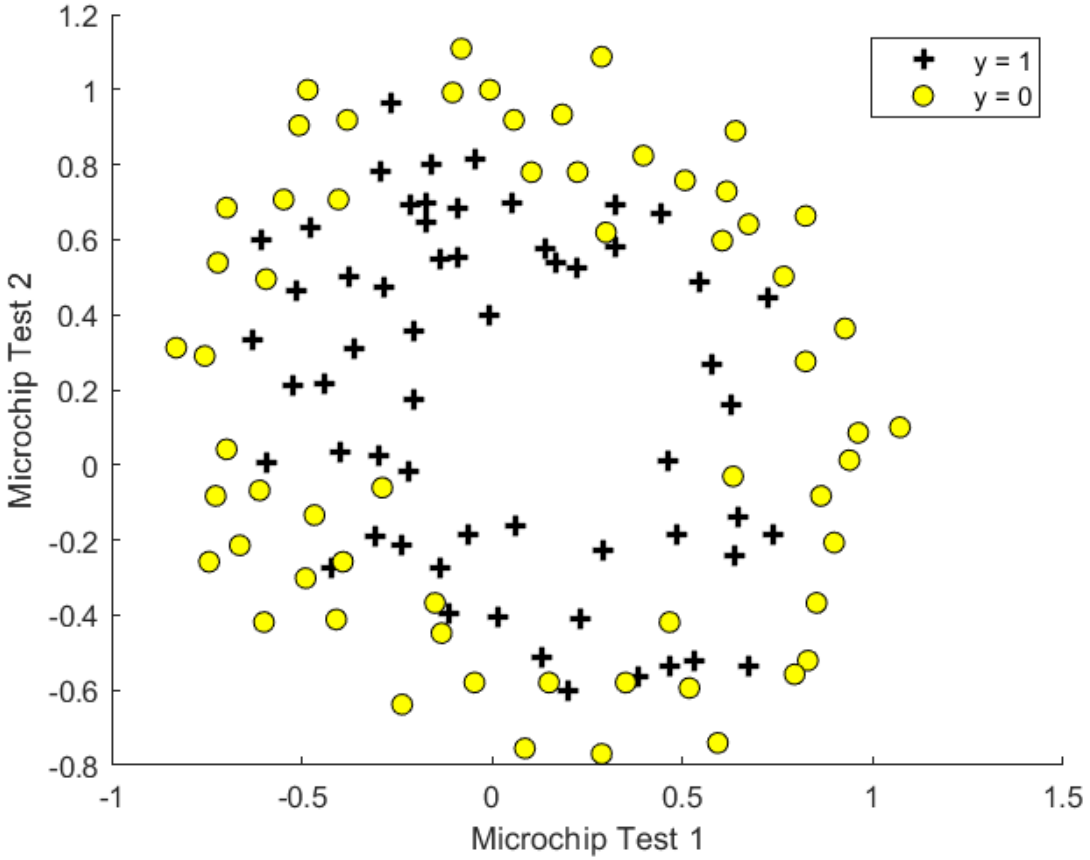
Η ακρίβεια του μοντέλου της λογιστικής παλινδρόμησης που υλοποιήθηκε υπολογίζεται στο 89% .

```
Program paused. Press enter to continue.
For a student with scores 45 and 85, we predict an admission probability of 0.776291

Train Accuracy: 89.000000

Program paused. Press enter to continue.
>>
```

Θέμα 2: Λογιστική Παλινδρόμηση με Ομαλοποίηση

Στη παρούσα άσκηση εφαρμόζεται ομαλοποιημένη λογιστική παλινδρόμηση προκειμένου να πραγματοποιηθεί η πρόβλεψη για το αν τα μικροτσίπ από μία μονάδα κατασκευής περνούν τον έλεγχο ποιότητας (QA). Χρησιμοποιώντας δεδομένα προηγούμενων δοκιμών δημιουργείται το μοντέλο της λογιστικής παλινδρόμησης το οποίο θα αποφασίζει αν τα μικροτσίπ γίνονται αποδεκτά βασιζόμενοι σε δύο διαφορετικές δοκιμές.

Αρχικά χρησιμοποιώντας τη συνάρτηση plotData.m απεικονίζονται τα δεδομένα.



Απεικόνιση των αρχικών δειγμάτων των δύο ελέγχων των μικροτσίπ

Στη συνέχεια τα δεδομένα απεικονίζονται σε χώρο μεγαλύτερης διάστασης όπου μπορούν να διαχωριστούν με μεγαλύτερη ευκολία με τη χρήση της λογιστικής παλινδρόμησης. Αυτο πραγματοποιείται υλοποιώντας τη συνάρτηση mapFeature.m η οποία απεικονίζει τα χαρακτηριστικά σε όλους τους όρους πολυωνύμων x_1και x_2μέχρι και 6ου βαθμού τα οποία ορίζονται ως εξής:

$$P(x_1, x_2) = \sum_{i=0}^{6} \sum_{j=0}^{i} x_1^{\,i-j} \cdot x_2^{\,j}$$

Με βάση την ομαλοποιημένη συνάρτηση κόστους υπολογίζεται το j-στοιχείο της κλίσης του σφάλματος.

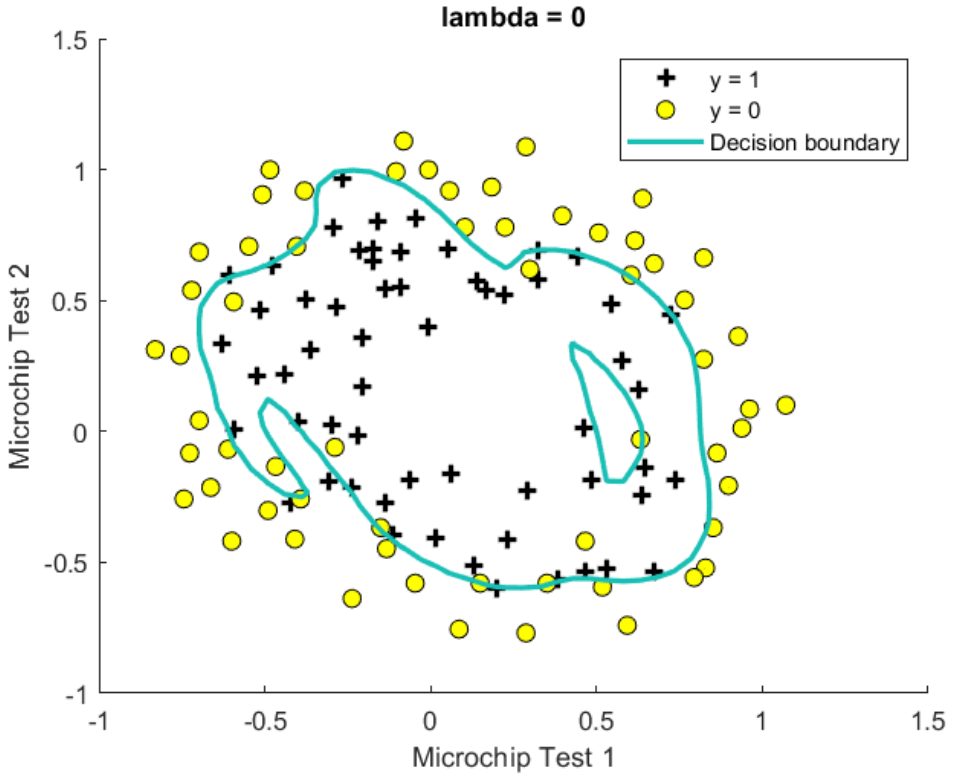
$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (-y^{(i)} ln(h_\theta(x^{(i)})) - (1 - y^{(i)}) ln(1 - h_\theta(x^{(i)}))) + \frac{\lambda}{2m} \theta_j^2$$

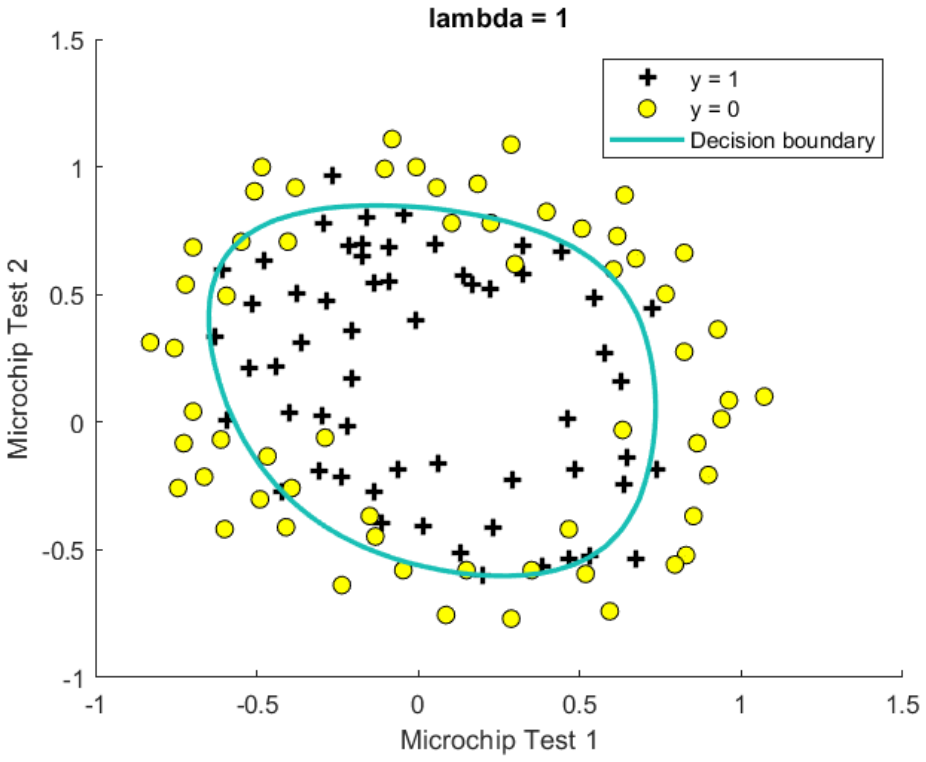
$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{2\lambda}{2m} \theta_j = \frac{\lambda}{m} \theta_j \qquad (I)$$

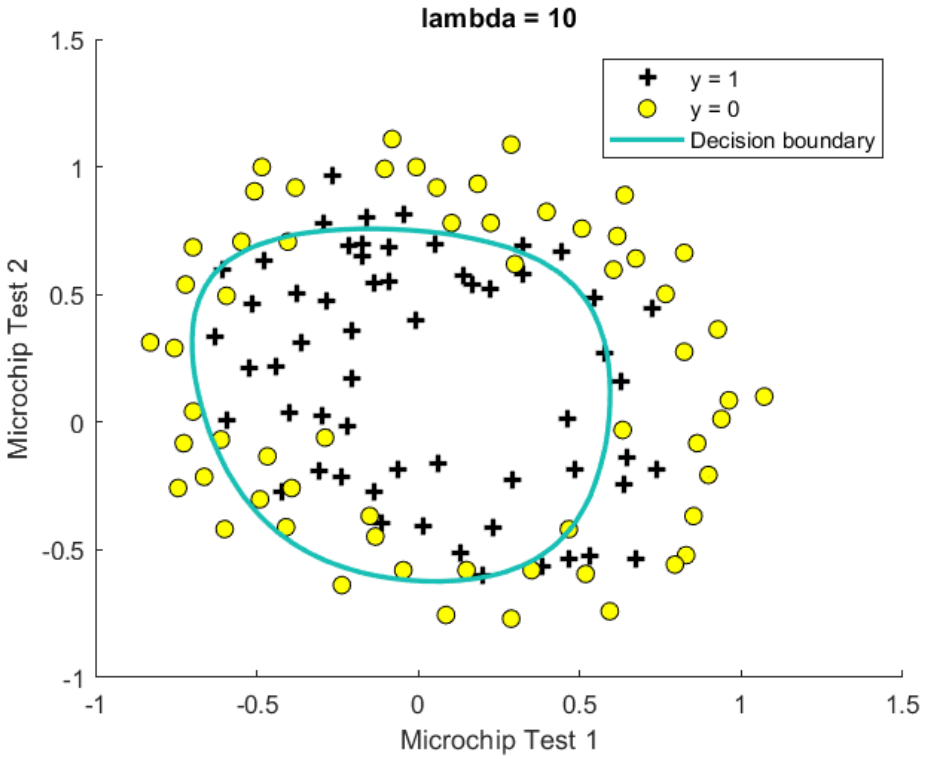
Επομένως από την σχέση (I) και το αποτέλεσμα του θέματος 1 προκύπτει ότι η κλίση του j-στοιχείου είναι ίσο με:

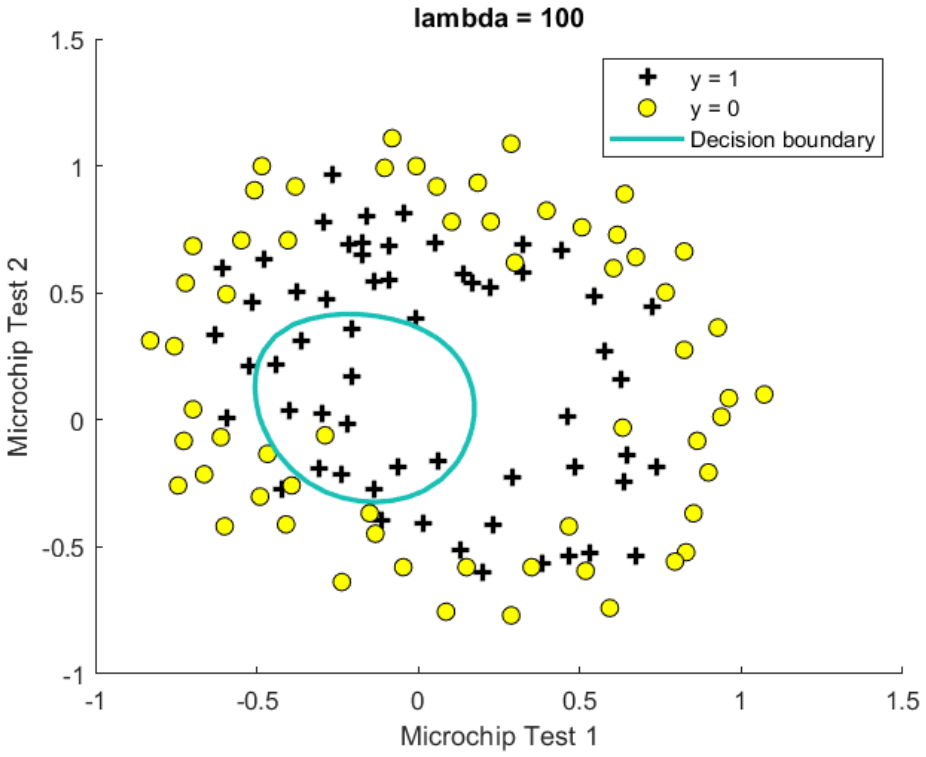
$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} \left(\frac{1}{1+e^{-\theta^T \cdot x^{(i)}}} - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} \theta_j = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j$$

Υλοποιώντας τη συνάρτηση costFunctionReg.m υπολογίζεται το κόστος της λογιστικής παλινδρόμησης.Βελτιστοποιώντας τις παραμέτρους πραγματοποιείται η εύρεση των συνόρων απόφασης για διάφορα λ = 0,1,10,100 .


lambda = 0
y = 1
y = 0
Decision boundary
Microchip Test 2
Microchip Test 1
1.5
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1
1.5


lambda = 1
y = 1
y = 0
Decision boundary
Microchip Test 2
Microchip Test 1
1.5
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1
1.5





	λ = 0.1	λ = 1	λ = 10	λ = 100
Train Accuracy	88.983051	82.20339	74.576271	60.169492

Αυξάνοντας τη παράμετρο λ παρατηρείται μείωση της ακρίβειας πρόβλεψης αν ένα μικροτσίπ θα περάσει τον έλεγχο ποιότητας του μοντέλου με βάση τα δείγματα προπόνησης.

Θέμα 3: Εκτίμηση Παραμέτρων (Maximum Likelihood)

Έστω n δείγματα $D = \{x_1,..., x_n\}$παράγονται ανεξάρτητα από μία κατανομή poisson με παράμετρο λ.

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}\ x = 0, 1, 2,.... , \lambda > 0$$

$$\lambda_{ML} = argmax_\lambda L(\lambda)$$

$$p(D|\lambda) = p(x_1, x_2,.. , x_n|\lambda) = \prod_{i=1}^{N} p(x_i|\lambda) = \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

$$L(\lambda) = ln(\ p(D|\lambda)) = ln \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \sum_{i=1}^{N} ln \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \sum_{i=1}^{N} (ln\ (\lambda^{x_i} e^{-\lambda}) - ln(x_i!)) \Rightarrow$$

$$L(\lambda) = \sum_{i=1}^{N} (ln\ (\lambda^{x_i}) + ln(e^{-\lambda}) - ln(x_i!)) = \sum_{i=1}^{N} (x_i ln\ (\lambda\ ) - \lambda\, ln(e\ ) - ln(x_i!))$$

$$\frac{dL(\lambda)}{d\lambda} = 0 \Leftrightarrow \sum_{i=1}^{N} (x_i \frac{ln\ (\lambda\ )}{d\lambda} - \frac{\lambda}{d\lambda} ln(e\ ) - \frac{ln(x_i!)}{d\lambda}) = 0 \Leftrightarrow \sum_{i=1}^{N} (x_i \frac{1}{\lambda} - 1 - 0) = 0 \Leftrightarrow$$

$$\frac{x_1+x_2+...+x_n}{\lambda} - N = 0 \Leftrightarrow \lambda = \frac{x_1+x_2+...+x_n}{N} \Leftrightarrow \lambda = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Επομένως η παράμετρος λ η οποία μεγιστοποιεί τον εκτιμητή πιθανοφάνειας είναι ίση με:

$$\lambda = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Θέμα 4: Εκτίμηση Παραμέτρων και Ταξινόμηση (ML - Naïve Bayes Classifier)

Στη παρούσα άσκηση υλοποιείται ένας naive-bayes ταξινομητής για αναγνώριση ψηφίων.
Τα δεδομένα της άσκησης φορτώνονται από το αρχείο digits.mat.π και υπάρχουν 10 κλάσεις που αντιστοιχούν στα ψηφία από το 0 έως το 9. Κάθε πίνακας χαρακτηριστικών είναι μια ασπρόμαυρη εικόνα με διαστάσεις 28 × 28 ο οποίος μπορεί να αναπαρασταθεί σαν ένα διάνυσμα 784 × 1 με στοιχεία τις τιμές 0 ή 1. Στον naive bayes ταξινομητή γίνεται η υπόθεση ότι τα χαρακτηριστικά είναι εικονοστοιχεία(pixel) των εικόνων τα οποία είναι ανεξάρτητα και κάθε χαρακτηριστικό(pixel) κάθε κλάσης μοντελοποιείται σύμφωνα με την κατανομή bernoulli.

Έστω x^{y_i} η τυχαία μεταβλητή η οποία παριστάνει το i-οστο χαρακτηριστικό του ψηφίου y και p^{y_i} η αντίστοιχη παράμετρος της κατανομής bernoulli του i-οστου χαρακτηριστικού του ψηφίου y.

Υπολογίζεται ότι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου p^{y_i} δοσμένων n δειγμάτων $x^{y_i} = \{x_1^{y_i}, x_2^{y_i}, \ldots, x_n^{y_i}\}$είναι :

$$p(x^{y_i}|p^{y_i}) = \prod_{i=1}^{n} p(x_1^{y_i}, x_2^{y_i}, \ldots, x_n^{y_i}|p^{y_i}) = \prod_{i=1}^{n} (p^{y_i})^s \cdot (1 - p^{y_i})^{n-s}$$

$$\widehat{y} = argmax\, L(p^{y_i})$$

$$L(p^{y_i}) = ln(p(x^{y_i}|p^{y_i})) = ln((p^{y_i})^s \cdot (1 - p^{y_i})^{n-s}) = ln(p^{y_i})^s + ln\,(1 - p^{y_i})^{n-s} \Rightarrow$$

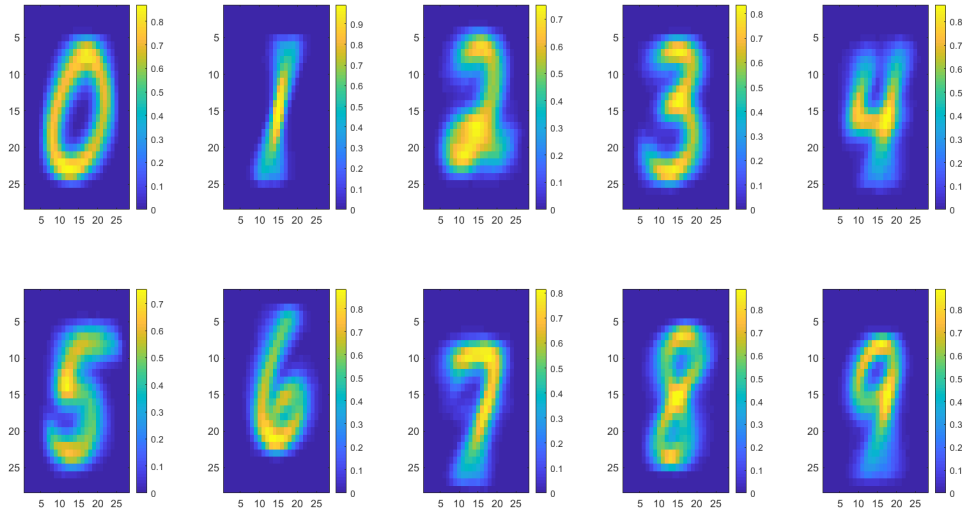
$$L(p^{y_i}) = s \cdot ln(p^{y_i}) + (n - s) \cdot ln\,(1 - p^{y_i})$$

$$\frac{\partial L(p^{y_i})}{\partial p^{y_i}} = \frac{s}{p^{y_i}} - \frac{n-s}{1-p^{y_i}} = 0 \Leftrightarrow s \cdot (1 - p^{y_i}) = n \cdot p^{y_i} - s \cdot p^{y_i} \Leftrightarrow s = n \cdot p^{y_i} \Rightarrow$$

$$p^{y_i} = \frac{s}{n}$$

Συνεπώς το p^{y_i}το οποίο μεγιστοποιεί τη πιθανοφάνεια είναι ο αριθμός των δειγμάτων που λαμβάνουν τη τιμή 1 προς τον συνολικό αριθμό δειγμάτων.

Χρησιμοποιώντας τον εκτιμητή μέγιστης πιθανοφάνειας εκπαιδεύονται τα μοντέλα για κάθε ψηφίο με βάση τα δοσμένα δείγματα εκπαίδευσης. Η παράμετρος $p^{\hat{y}_i}$ αντιστοιχεί στο i-στο εικονοστοιχείο του ψηφίου y και αναπαριστά την φωτεινότητα κάθε εικονοστοιχείου.



Οπτικοποίηση των εκπαιδευμένων μοντέλων για κάθε ψηφίο

Χρησιμοποιώντας τα εκπαιδευμένα μοντέλα, ταξινομούνται τα ψηφία στο σύνολο δοκιμής (test set), εκτίμηση (y hat) σε ποια κλάση ανήκει το συγκεκριμένο ψηφίο πραγματοποιείται σύμφωνα με τον εξής τύπο:

$$\hat{y} = argmax_{Digitk\ prob} p(x|Digitk\ prob\ )p(Digitk\ prob) \Rightarrow$$

$$\hat{y} = argmax_{Digitk\ prob} \frac{1}{10} p(x_1, x_2, .., x_n|Digitk\ prob\ )$$

Οι εκ των προτέρων πιθανότητες(a-priori) των ψηφίων είναι ίσες με 1/10 καθώς κάθε ψηφίο έχει τον ίδιο αριθμό δειγμάτων δοκιμής. Η πιθανοφάνεια $p(x_i|Digitk\ prob\ )$εκφράζει την πιθανότητα το i-οστο εικονοστοιχείο του ψηφίου k (Digitk prob) να λαμβάνει τη τιμή του pixel της εικόνας του training set.

Υπολογίζεται η ακρίβεια της ταξινόμησης (classification accuracy) ως ο λόγος των ψηφίων που έχουν ταξινομηθεί σωστά ως προς το σύνολο των δειγμάτων ελέγχου ο οποίος είναι ίσος με 79.76% .

```
>> classifier_accuracy

classifier_accuracy =

    0.7976
```

Τέλος υπολογίζεται ο πίνακας confusion matrix στον οποίο κάθε στοιχείο (i,j) παριστάνει πόσο συχνά η εικόνα του ψηφίου i ταξινομείται ως ψηφίο j.

```
confusion_matrix =

   441     0     1     0     2    30    13     0    12     1
     0   472     2     3     0    13     4     0     6     0
     8    12   375    37     7     2     9    13    33     4
     1    10     4   416     5    23     5    13    10    13
     1     1     5     0   374     4    15     2     5    93
    14     2     3    67    20   345     7     6    15    21
     7    10    28     0     7    33   412     0     3     0
     1    21     9     3    15     0     0   395    10    46
     6    16    12    49    11    18     1     3   349    35
     3     8     3     8    51     7     0     6     5   409
```

Confusion Matrix του Testing Data

Θέμα 5: Support Vector Machines (Αναλυτική βελτιστοποίηση με ΚΚΤ)

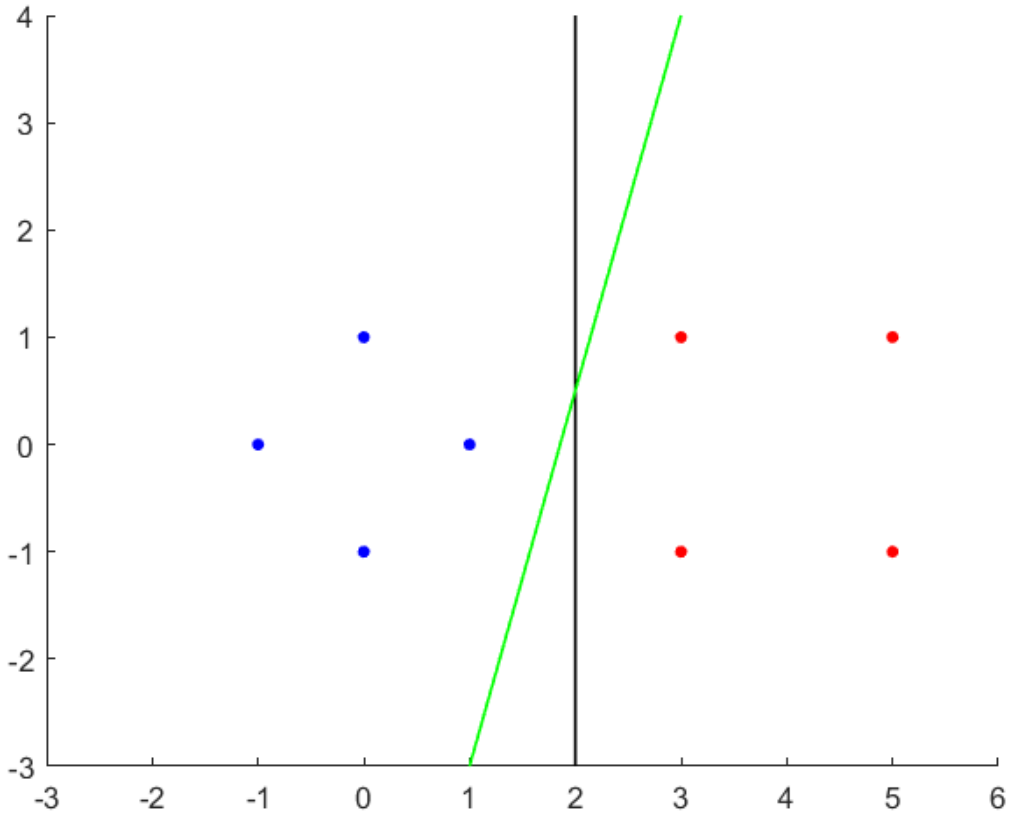
Στη παρούσα άσκηση χρησιμοποιείται Support Vector Machines(SVM) σε ένα πρόβλημα ταξινόμησης σε δυο κατηγορίες ω_1και ω_2. Τα δείγματα στο χώρο σχεδιάστηκαν με τη χρήση του λογισμικού Matlab.

Μέρος Α:

Τα ακόλουθα δείγματα εκπαίδευσης $x = [x_1, x_2]$για τις δύο κλάσεις:

$$\omega_1 : x^+ = \{[3,1]^T, [3,-1]^T, [5,1]^T, [5,-1]^T\}$$

$$\omega_2 : x^- = \{[1,0]^T, [0,1]^T, [0,-1]^T, [-1,0]^T\}$$



Απεικόνιση γραμμικών υπερεπιπέδων διαχωρισμού των δύο κατηγοριών με βάση τον αλγόριθμο SVM

Σχεδιάζοντας στο χώρο τα παραπάνω δείγματα διαισθητικά προτείνεται ως το βέλτιστο γραμμικό υπερεπίπεδο διαχωρισμού με βάση τον αλγόριθμο SVM η πράσινη γραμμή και η μαύρη γραμμή (g(x) =2) διότι αφήνει το μεγαλύτερο περιθώριο(margin) ανάμεσα στις δύο κλάσεις δειγμάτων το οποίο καθορίζει αν ο γραμμικός ταξινομητής θα είναι ικανοποιητικός και σε δεδομένα εκτός των δειγμάτων εκπαίδευσης.

Χρησιμοποιώντας πολλαπλασιαστές Lagrange και τις συνθήκες Karush-Khun-Tucker (KKT) βρίσκεται αναλυτικά η βέλτιστη γραμμή διαχωρισμού.

$$g(x) = w^T \cdot \bar{x} + w_0 = 2$$

Η απόσταση του υπερεπιπέδου διαχωρισμού από το υπερεπίπεδο g(x) = 2 ορίζεται ώς εξής:

$$d = \frac{|w_0|}{||w||_2}$$

Η κανονικοποιημένη απόσταση ενός δειγματος x από το υπερεπίπεδο g(x) = 2 ορίζεται ως εξής:

$$z = \frac{|g(x)|}{||w||_2} = \frac{|g(x)|}{\sqrt{w_1^2 + w_2^2}}$$

Η κανονικοποίηση πραγματοποιείται προκειμένου η τιμή του g(x) για τα κοντινότερα στο υπερεπίπεδο δείγματα(support vectors) x^+ για την κλάση ω_1και x^- για την κλαση ω_2 να γίνεται

$$g(x^+) = < w \cdot x^+ > \ + w_0 = 1 \rightarrow \omega_1 = 1$$
$$g(x^-) = < w \cdot x^- > \ + w_0 = -1 \rightarrow \omega_2 = -1$$

Επομένως το εύρος(margin) του περιθωρίου ορίζεται ως εξής:

$$\gamma = \frac{1}{||w||_2} + \frac{1}{||w||_2} = \frac{2}{||w||_2}$$

Ο στόχος του αλγορίθμου SVM είναι να μεγιστοποιηθεί το εύρος του margin $max(\gamma) = max(\frac{2}{||w||_2})$ ή ισοδύναμα ελαχιστοποιείται η objective function δηλαδή:

$J(w, w_0) = \frac{1}{2}||w||_2^2 = \frac{1}{2}(w_1^2 + w_2^2)$

Με περιορισμούς: $w^T \cdot x^+ + w_0 \geq 1, \ \forall x^+ \in \omega_1$

$w^T \cdot x^- + w_0 \leq 1, \ \forall x^- \in \omega_2$

s.t. $f_i(w_1, w_0) = g_i(w^T \cdot x_i + w_0) - 1 \geq 0$

όπου

$g_i = 1$, για ω_1

$g_i = -1$, για ω_2

Ο κάθε ένας από τους παραπάνω περιορισμούς ορίζει ένα ξεχωριστό εύρος τιμών (περιοχών) μέσα στον χώρο $\mathbb{R}^2$. Η τομή αυτών των περιοχών εκφράζει την τελική περιοχή μέσα στην οποία η γραμμή διαχωρισμού θα πρέπει να βρίσκεται (feasible region). Ορίζεται η βοηθητική συνάρτηση Lagrange η οποία θα ελαχιστοποιηθεί ως εξής:

$$L(w, w_0, \lambda) = J(w, w_0) - \sum_{i=1}^{N} \lambda_i f_i(w, w_0) = \frac{1}{2}||w||_2^2 - \sum_{i=1}^{N} \lambda_i [g_i(w^T \cdot x_i + w_0) - 1] = 0 \ (1)$$

Η παράμετρος w που ελαχιστοποιεί την παραπάνω σχέση θα πρέπει απαραίτητα να ικανοποιεί τις παρακάτω συνθήκες KKT :

$$\frac{\partial L((w,w_0,\lambda)}{dw} = 0 \Leftrightarrow w - \sum_{i=1}^{N} \lambda_i g_i x_i = 0 \Leftrightarrow w = \sum_{i=1}^{N} \lambda_i g_i x_i \quad (2)$$

$$\frac{\partial L((w,w_0,\lambda)}{dw_0} = \sum_{i=1}^{N} \lambda_i g_i = 0 \quad (3)$$

$$\lambda_i \geq 0,\ i = 1,2,..,N$$

$$\lambda_i[g_i(w^{\mathrm{T}} \cdot x_i + w_0) - 1] = 0,\ i = 1,2,..,N \quad (4)$$

Αντικαθιστώντας τα παραπάνω προκύπτει ότι:

$$max\{\sum_{i=1}^{N} \lambda_\iota - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\lambda_i\lambda_j g_i g_j x_i^T x_j)\}$$

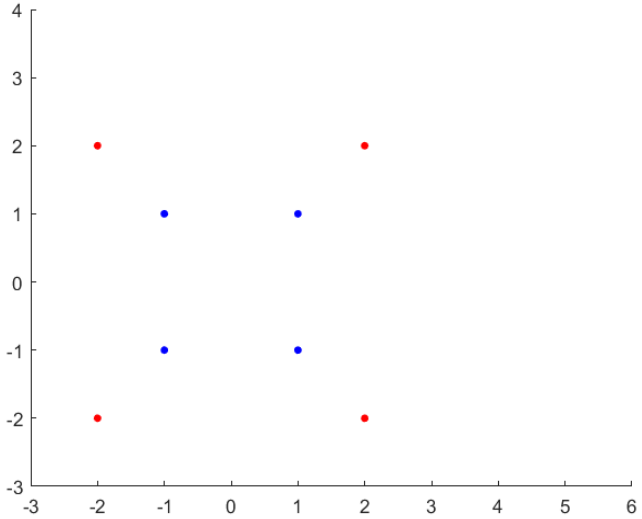
$$\text{s.t} \sum_i \lambda_i g_i = 0$$

$$\lambda_i \geq 0,\ i = 1,2,..,N$$

Συνεπώς τα πιθανά support vectors είναι τα $w = [1\ 0\ 2]^T$ ή $w = [1\ \ -1\ \ -3]^T$, τα οποία θα ληφθούν από τον αλγόριθμο SVM και με βάση τους παραπάνω περιορισμούς ΚΚΤ θα επιλεχθεί το βέλτιστο.

Μέρος Β:

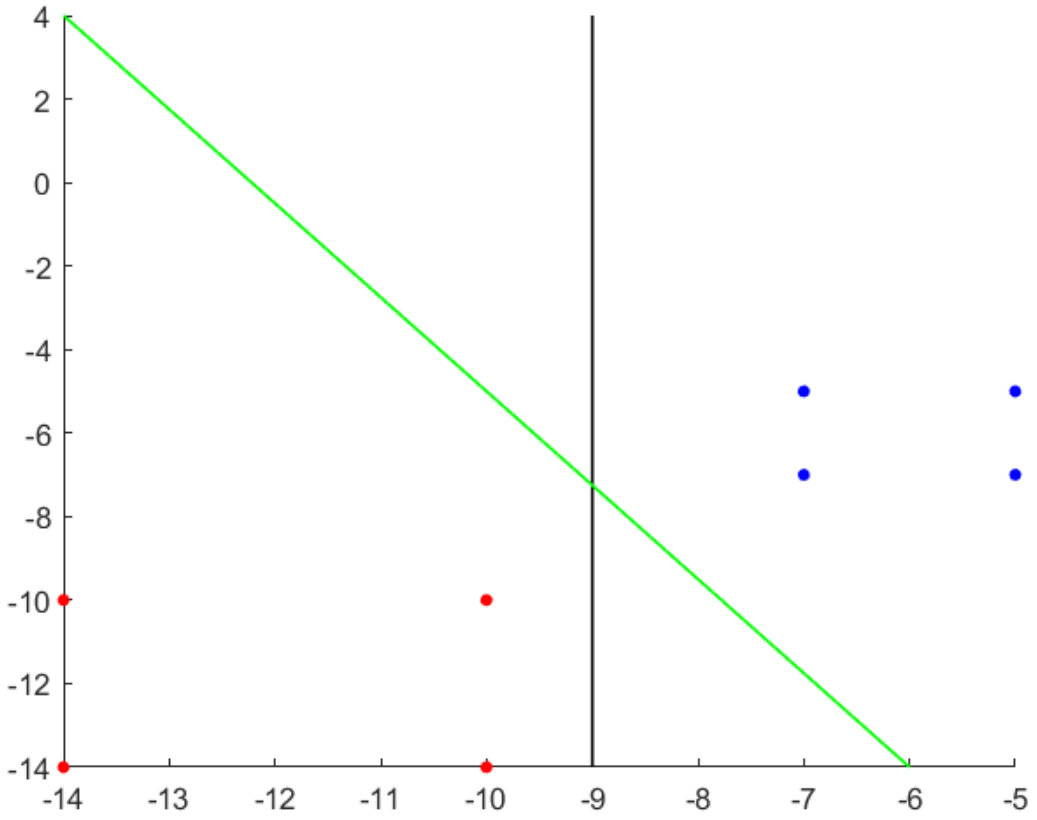
Σχεδιάζοντας τα δείγματα στο χώρο, παρατηρείται ότι τα δείγματα δεν είναι γραμμικώς διαχωρίσιμα με αποτέλεσμα να μην μπορεί να οριστεί κάποιο κατάλληλο support vectors με χρήση του αλγορίθμου SVM .



Απεικόνιση γραμμικών υπερεπιπέδων διαχωρισμού των δύο κατηγοριών με βάση τον αλγόριθμο SVM

Συνεπώς πραγματοποιείται ένας μετασχηματισμός των δειγμάτων προκειμένου να είναι τα δείγματα γραμμικά διαχωρισιμα . Ο μετασχηματισμός είναι της μορφής :

$$\Phi(x) = x - ||x||_2^2 - 4, \text{ όπου } ||x|_2^2 = x_1^2 + x_2^2$$



Απεικόνιση γραμμικών υπερεπιπέδων διαχωρισμού των δύο κατηγοριών με βάση τον αλγόριθμο SVM

Με την εφαρμογή του παραπάνω μετασχηματισμού διαπιστώνεται ότι δύο πιθανά γραμμικά υπερεπίπεδα τα οποία θα πραγματοποιούσαν το βέλτιστο διαχωρισμό των κλάσεων είναι η πράσινη και η μαύρη γραμμή καθώς αφήνουν το μεγαλύτερο περιθώριο(margin) ανάμεσα στις δύο κλάσεις δειγμάτων το οποίο καθορίζει αν ο γραμμικός ταξινομητής θα είναι ικανοποιητικός και σε δεδομένα εκτός των δειγμάτων εκπαίδευσης.

Ομοίως με τη διαδικασία που ακολουθήθηκε στο μέρος Α, προκύπτει το νέο πρόβλημα βελτιστοποίησης ώς εξής:

$$max\{\sum_{i=1}^{N}\lambda_{\iota} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\lambda_i\lambda_j g_i g_j \Phi(x_i)^T\Phi(x_j))\}$$

$$\text{s.t}\sum_i \lambda_i g_i = 0$$

$$\lambda_i \geq 0,\ i = 1,2,..,N$$

Συνεπώς τα πιθανά support vectors είναι τα $w = [1\ 0\ -9]^T$ ή $w = [1\ -7\ -2.33]^T$, τα οποία θα ληφθούν από τον αλγόριθμο SVM και με βάση τους παραπάνω περιορισμούς ΚΚΤ θα επιλεχθεί το βέλτιστο.

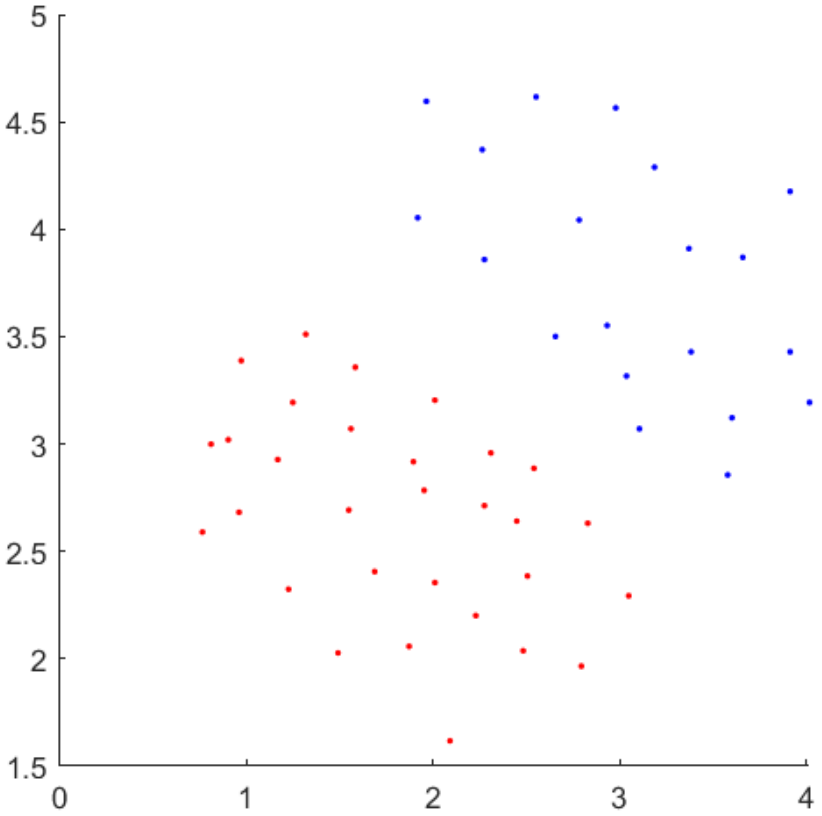
Θέμα 6: Support Vector Machines (Εφαρμογή σε τεχνητό σύνολο δεδομένων)

Στη παρούσα άσκηση δίνεται ένα σύνολο n παραδειγμάτων $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}),.., (x^{(n)}, y^{(n)})$όπου $x^{(i)} \in \Re^{1\times 2}$και $y^{(i)} \in \{-1, 1\}$. Όλα τα δεδομένα βρίσκονται σε έναν πίνακα Χ όπου οι γραμμές είναι τα δείγματα και οι στήλες τα χαρακτηριστικά . Ο σκοπός της άσκησης είναι να γίνει η πρόβλεψη των τιμών $y^{(i)}$ από τις αντίστοιχες τιμες των $x^{(i)}$, $i \in \{1, 2,.., n\}$, χρησιμοποιώντας Support Vector Machines (SVM).

$$\hat{y} = \begin{cases} 1, \text{ εάν } \boldsymbol{w}\cdot\boldsymbol{x}^T + w_0 > 0 \\ -1, \text{ εάν } \boldsymbol{w}\cdot\boldsymbol{x}^T + w_0 < 0 \end{cases}$$

Οι παράμετροι του SVM προσδιορίζονται λύνοντας ένα πρόβλημα βελτιστοποίησης με περιορισμούς.

Μέρος 1ο: Γραμμικά διαχωρίσιμα δείγματα:



Τα αρχικά παραδείγματα του αρχείου twofeature1.txt

Η λαγκρανζιανή του δυικού προβλήματος είναι :

$$\overline{L}(\lambda) = \sum_{i=1}^{n} \lambda^{(i)} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda^{(i)}\lambda^{(j)}y^{(i)}y^{(j)}K(x^{(i)}, x^{(j)}) \text{με } K(x^{(i)}, x^{(j)}) = x^{(i)} \cdot (x^{(j)})^{T}$$

Η οποία μεγιστοποιείται κάτω από τους εξής περιορισμούς:

$$\lambda^{(i)} \geq 0, \ i = 1,\ldots,n$$

$$\sum_{i=1}^{n} \lambda^{(i)}y^{(i)} = 0$$

Συμπληρώνοντας τον απαραίτητο κώδικα matlab στο αρχείο svm_exercise1.m ελαχιστοποιείται κατάλληλα η συνάρτηση ώστε να υπολογιστούν οι πολλαπλασιαστές Lagrange $\lambda^{(i)}, \ i = 1,\ldots,n$.

Συγκεκριμένα λαμβάνοντας υπόψη τους παραπάνω περιορισμούς υπολογίζονται τα κατάλληλα ορίσματα της συνάρτησης quadprog(H,f,A,b,Aeq,Beq) όπου οι παράμετροι της συνάρτησης ορίζονται ως εξής:

- Ο πίνακας H είναι ίσος με $\mathrm{H} = y^{(i)}y^{(j)}K(x^{(i)}, x^{(j)})$
- Ο πίνακας f είναι ίσος με τους συντελεστές του όρου $\sum_{i=1}^{n} \lambda^{(i)}$αλλά με αντίθετο πρόσημο
- Ο πίνακας A είναι ίσος με το πίνακα -eye(n)
- Ο πίνακας b είναι ίσος με το πίνακα zeros(n,1)
- Ο πίνακας Aeq είναι ίσος με το πίνακα y'
- Ο πίνακας Beq είναι ίσος με το μηδέν(0)

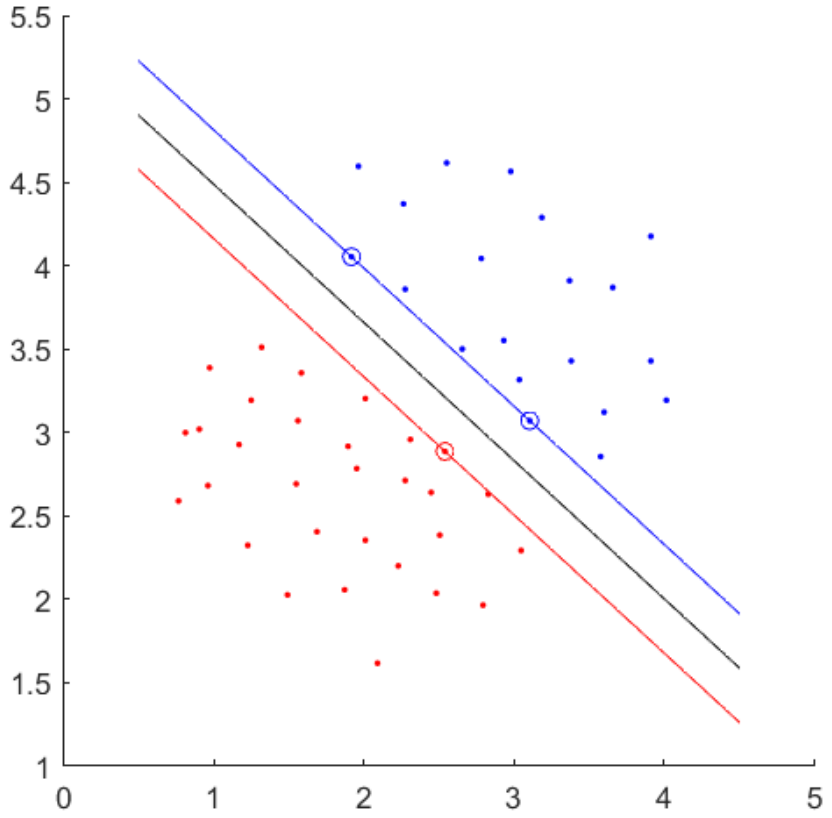
Στη συνέχεια υπολογίζονται τα βάρη w, το bias w_0 καθώς και το πλάτος της λωρίδας μεταξύ των support vectors σύμφωνα με τους ακόλουθους τύπους:

$$w = \sum_{i=1}^{N_s} \lambda_i y_i x_i$$

$$w_0 = -\frac{(max_{y_i=-1}(x_i^T \cdot w) + min_{y_i=1}(x_i^T \cdot w))}{2}$$

$$width = \frac{2}{||w||}$$

Τέλος σχεδιάζονται η γραμμή διαχωρισμού των δύο κλάσεων και οι παράλληλες γραμμές σε αυτή οι οποίες περνούν από τα support vectors.Για το γραφικό σχεδιασμό των γραμμών χρησιμοποιήθηκαν μόνο τα 50 από τα 51 παραδείγματα του αρχείου twofeature1.txt .



Οι γραμμές διαχωρισμού των δύο κλάσεων και οι παράλληλες τους.

Μέρος 2ο: Μεταβλητές Περιθωρίου (slack variables):

Στο δεύτερο μέρος της παρούσας άσκησης προστίθεται θόρυβος στα δεδομένα ώστε να προσεγγίσει τη πραγματικότητα το πρόβλημα της βελτιστοποίησης. Για να διαχειριστούν δεδομένα με την ύπαρξη του θορύβου χρησιμοποιούνται μεταβλητές περιθωρίου ώστε να γίνονται ανεκτά ορισμένα παραδείγματα τα οποία είναι στη λάθος πλευρά της επιφάνειας απόφασης. Η παράμετρος η οποία καθορίζει το επίπεδο ανοχής συνηθως συμβολίζεται με C και λαμβάνει τιμές στο διάστημα $(0, +\ infinity)$,με 0^{+}να αντιστοιχεί σε μέγιστη ανοχή και το +άπειρο σε μηδενική ανοχή.

Η Λαγκρανζιανή του δυικού προβλήματος είναι ίδια με αυτή του πρώτου μέρους αλλά την μεγιστοποιούμε κάτω από τους παρακάτω περιορισμούς:

$$0 \leq \lambda^{(i)} \leq C,\ \ i = 1,..,n$$

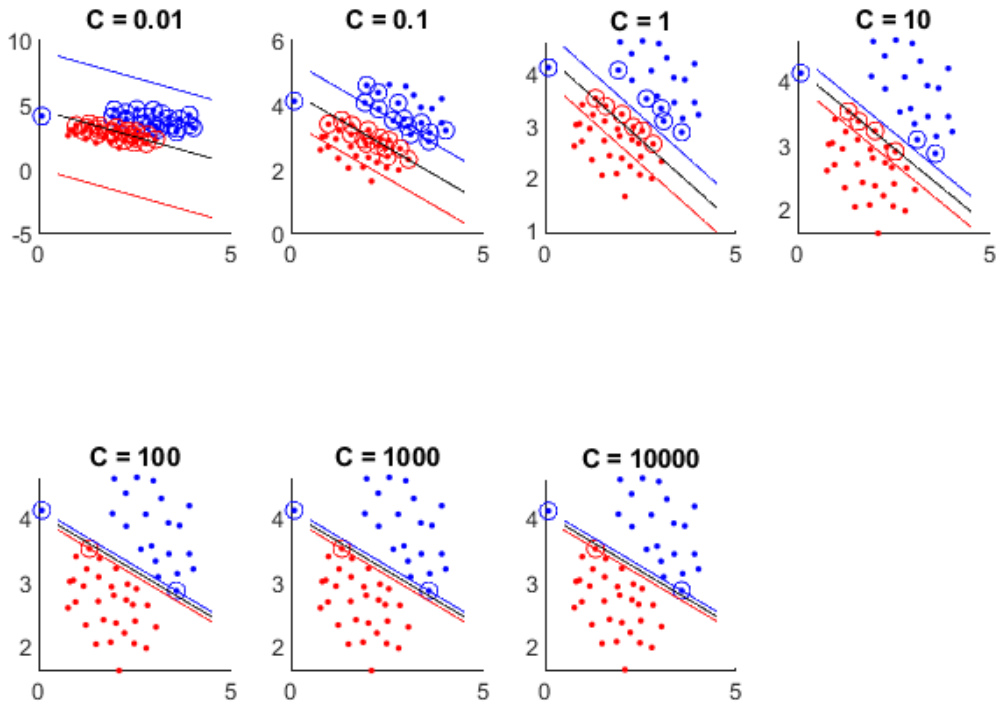
Συμπληρώνοντας τον απαραίτητο κώδικα matlab στο αρχείο svm_exercise2.m ελαχιστοποιείται κατάλληλα η συνάρτηση ώστε να υπολογιστούν οι πολλαπλασιαστές Lagrange $\lambda^{(i)},\ i = 1,\ldots,n$. Συγκεκριμένα λαμβάνοντας υπόψη τους παραπάνω περιορισμούς υπολογίζονται τα κατάλληλα

ορίσματα της συνάρτησης quadprog(H,f,A,b,Aeq,Beq,lb,ub) όπου οι παράμετροι της συνάρτησης ορίζονται όμοια με το πρώτο μέρος και οι δύο επιπλέον παράμετροι lb,ub ορίζονται ως εξής:

- Το lb είναι ίσο με μηδέν το οποίο είναι το κάτω όριο του περιορισμού
- Το ub είναι ίσο με C το οποίο είναι το άνω όριο του περιορισμού

Κατόπιν υπολογίζονται τα βάρη w, το bias w_0 καθώς και το πλάτος της λωρίδας μεταξύ των support vectors όμοια με το πρώτο μέρος.

Χρησιμοποιώντας όλα τα 51 παραδείγματα του αρχείου twofeature1.txt σχεδιάζονται η γραμμή διαχωρισμού των δύο κλάσεων και οι παράλληλες γραμμές σε αυτή οι οποίες περνούν από τα support vectors για διαφορετικά επίπεδα ανοχής με C = 0.01, 0.1 ,1, 10 ,100, 1000 ,10000 .



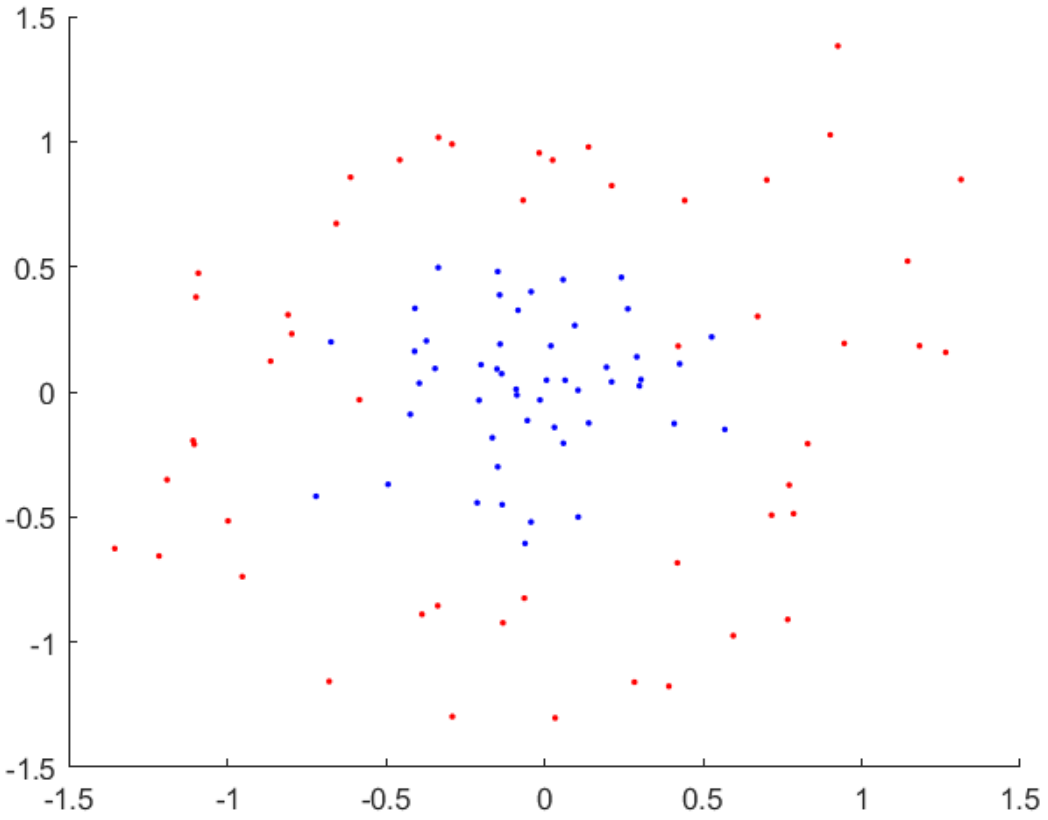
Οι γραμμές διαχωρισμού των δύο κλάσεων και οι παράλληλες τους για διαφορετικά επίπεδα ανοχής C = 0.01, 0.1, 1, 10, 100, 1000, 10000

Παρατηρώντας τις γραφικές παραστάσεις των γραμμών διαχωρισμού για διαφορετικά επίπεδα ανοχής C γίνεται αντιληπτό ότι όσο πλησιάζει η ανοχή στο θόρυβο τη μέγιστη τιμή της(0^+) τόσο

γίνεται πολυπλοκότερο να διαχωριστούν τα παραδείγματα των δύο κλάσεων μεταξύ τους. Αντιθέτως όσο μειώνεται η ανοχή των παραδειγμάτων στο θόρυβο τόσο καλύτερος διαχωρισμός επιτυγχάνεται μεταξύ των δύο κλάσεων.

Μέρος 3ο: Μη γραμμικά διαχωρίσιμα παραδείγματα:

Στο τρίτο μέρος της άσκησης τα δεδομένα φορτώνονται από το αρχείο twofeature2.txt τα οποία δεν είναι γραμμικά διαχωρίσιμα και χρησιμοποιείται η Λαγκρανζιανή του δεύτερου μέρους. Για κάθε δυάδα $x = (x_1, x_2)$ κάθε παραδείγματος προστίθεται ένα τρίτο χαρακτηριστικό το οποίο είναι το μέτρο του x στο τετράγωνο δηλαδή $x = (x_1, x_2) \rightarrow (x_1, x_2, ||x||_2^2)$ όπου $||x||_2^2 = x_1^2 + x_2^2$ προκειμένου τα παραδείγματα να μετασχηματιστούν σε γραμμικώς ανεξάρτητα.

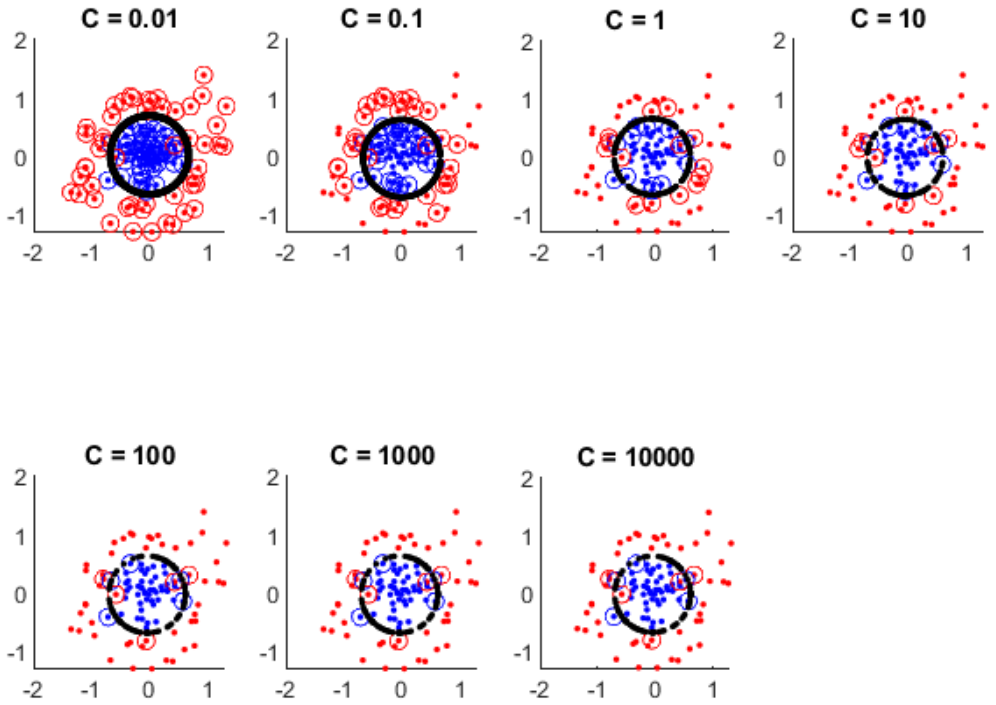


Τα αρχικά παραδείγματα του αρχείου twofeature2.txt

Συμπληρώνοντας τον απαραίτητο κώδικα matlab στο αρχείο svm_exercise3.m ελαχιστοποιείται κατάλληλα η συνάρτηση ώστε να υπολογιστούν οι πολλαπλασιαστές Lagrange $\lambda^{(i)}$, $i = 1,\ldots,n$. Συγκεκριμένα λαμβάνοντας υπόψη τους παραπάνω περιορισμούς υπολογίζονται τα κατάλληλα

ορίσματα της συνάρτησης quadprog(H,f,A,b,Aeq,Beq,lb,ub) όπου οι παράμετροι της συνάρτησης ορίζονται όμοια με το δεύτερο μέρος.

Στη συνέχεια υπολογίζονται τα βάρη w και το bias όμοια με το δεύτερο μέρος. Τέλος χρησιμοποιώντας όλα τα 51 παραδείγματα του αρχείου twofeature2.txt σχεδιάζονται η γραμμή διαχωρισμού των δύο κλάσεων και οι παράλληλες γραμμές σε αυτή οι οποίες περνούν από τα support vectors για διαφορετικά επίπεδα ανοχής με C = 0.01, 0.1 ,1, 10 ,100, 1000 ,10000 .



Τα υπερεπίπεδα διαχωρισμού των κλάσεων για διαφορετικά επίπεδα ανοχής του θορύβου C = 0.01, 0.1, 1, 10, 100, 1000, 10000

Παρατηρώντας τα υπερεπίπεδα διαχωρισμού των κλάσεων διαπιστώνεται ότι μειώνοντας την ανοχή του θορύβου στα παραδείγματα επιτυγχάνεται καλύτερος διαχωρισμός μεταξύ των κλάσεων.