2η Σειρά Ασκήσεων

Ονομα: Μαυρογιώργης Δημήτρης
ΑΜ: 2016030016
ΤΗΛ 311 - Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων
Πολυτεχνειο Κρητης

June 25, 2021

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

Έστω ένα σύνολο m δεδομένων $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$ όπου $x^{(i)} \in \mathbb{R}^{n\times 1}$ είναι τα διανύσματα χαρακτηριστικών και $y^{(i)} \in \{0, 1\}$ ορίζουν την κλάση κάθε δείγματος. Στόχος μας είναι να προβλέψουμε τις τιμές $y^{(i)}$ από τις αντίστοιχες $x^{(i)}$ χρησιμοποιώντας την συνάρτηση της λογιστικής παλινδρόμησης, η οποία ορίζεται ως εξής

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

όπου $\theta = [\theta_1, \theta_1, \cdots, \theta_n]^T$ είναι οι παράμετροι του γραμμικού μοντέλου και f είναι η λογιστική συνάρτηση που ορίζεται ως:

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

Έστω $\hat{y}^{(i)} = h_\theta(x^{(i)})$ η εκτίμηση της λογιστικής συνάρτησης για το $y^{(i)}$. Σύμφωνα με τη θεωρία, στην περίπτωση της λογιστικής παλινδρόμησης, μπορούμε να υπολογίσουμε το σφάλμα με βάση τη συνάρτηση κόστους(loss function), που ονομάζεται cross-entropy, και ορίζεται ως εξής:

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

Αν αντικαταστήσουμε το $y^{(i)}$ έχουμε

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

Τότε το j-στοιχείο της κλίσης του σφάλματος υπολογίζεται ως εξής:

$$\begin{aligned}\frac{\partial J(\theta)}{\partial \theta_j} &= \frac{\partial \left(\frac{1}{m} \cdot \sum_{i=1}^{m} [-y^{(i)} \ln(h_\theta(x^{(i)})) - (1 - y^{(i)}) \cdot \ln(1 - h_\theta(x^{(i)}))]\right)}{\partial \theta_j} \\ &= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \frac{\partial \ln(h_\theta(x^{(i)}))}{\partial \theta_j} + (1 - y^{(i)}) \cdot \frac{\partial \ln(1 - h_\theta(x^{(i)}))}{\partial \theta_j} \right]\end{aligned}$$

$$= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \cdot \frac{\partial \ln(h_\theta(x^{(i)}))}{\partial \theta_j} + (1 - y^{(i)}) \cdot \frac{\partial \ln(1 - h_\theta(x^{(i)}))}{\partial \theta_j} \right]$$
$$= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \cdot \frac{\frac{\partial h_\theta(x^{(i)})}{\partial \theta_j}}{h_\theta(x^{(i)})} + (1 - y^{(i)}) \cdot \frac{\frac{\partial (1 - h_\theta(x^{(i)}))}{\partial \theta_j}}{1 - h_\theta(x^{(i)})} \right]$$

Αν παραγωγίσουμε την συνάρτηση sigmoid $f(z)$ προκύπτει

$$\begin{aligned}
\frac{\partial f(z)}{\partial z} &= \frac{\partial \left(\frac{1}{1+e^{-z}} \right)}{\partial z} \\
&= -\frac{\frac{\partial (1+e^{-z})}{\partial z}}{(1+e^{-z})^2} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}} \cdot \left(\frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} \right) \\
&= f(z) \cdot (1 - f(z))
\end{aligned}$$

Οπότε αντικαθιστώντας τη συνάρτηση h_θ στην κλίση του σφάλματος έχουμε

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \cdot \frac{\frac{\partial f(\theta^T x^{(i)})}{\partial \theta_j}}{h_\theta(x^{(i)})} + (1 - y^{(i)}) \cdot \frac{\frac{\partial (1 - f(\theta^T x^{(i)}))}{\partial \theta_j}}{1 - h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \cdot \frac{f(\theta^T x^{(i)}) \cdot (1 - f(\theta^T x^{(i)})) \cdot \frac{\partial \theta^T x^{(i)}}{\partial \theta_j}}{h_\theta(x^{(i)})} - (1 - y^{(i)}) \cdot \frac{f(\theta^T x^{(i)}) \cdot (1 - f(\theta^T x^{(i)})) \cdot \frac{\partial (\theta^T x^{(i)})}{\partial \theta_j}}{1 - h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \cdot \frac{f(\theta^T x^{(i)}) \cdot (1 - f(\theta^T x^{(i)})) \cdot x_j^{(i)}}{h_\theta(x^{(i)})} - (1 - y^{(i)}) \cdot \frac{f(\theta^T x^{(i)}) \cdot (1 - f(\theta^T x^{(i)})) \cdot x_j^{(i)}}{1 - h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left[y^{(i)} \cdot \frac{h_\theta(x^{(i)}) \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)}}{h_\theta(x^{(i)})} - (1 - y^{(i)}) \cdot \frac{h_\theta(x^{(i)}) \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)}}{1 - h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left(y^{(i)} \cdot (1 - h_\theta(x^{(i)})) \cdot x_j^{(i)} - (1 - y^{(i)}) \cdot h_\theta(x^{(i)}) \cdot x_j^{(i)} \right) \\
&= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left([y^{(i)} - y^{(i)} \cdot h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)} \cdot h_\theta(x^{(i)}) \right) \cdot x_j^{(i)} \\
&= -\frac{1}{m} \cdot \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) \cdot x_j^{(i)} \\
&= \frac{1}{m} \cdot \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x_j^{(i)}
\end{aligned}$$

Όσον αφορά την υλοποίηση θα χρησιμοποιήσουμε τη λογιστική παλινδρόμηση για να προβλέψουμε αν έανς φοιτητής γίνεται δεκτός σε ένα πανεπιστήμιο με βάση τους βαθμούς του σε δύο εξεταστικές περιόδους. Για το σκοπό αυτό υλοποιήθηκαν οι συναρτήσεις sigmoid(), στην οποία υπολογίζουμε την λογιστική συνάρτηση f(z) καθώς και τη συνάρτηση costFunction(), με την οποία υπολογίζουμε το J(ϑ) και το $\frac{\partial J(\theta)}{\partial \theta_j}$. Τέλος, υλοποιήσαμε και μία συνάρτηση predict(), με την οποία προβλέπουμε

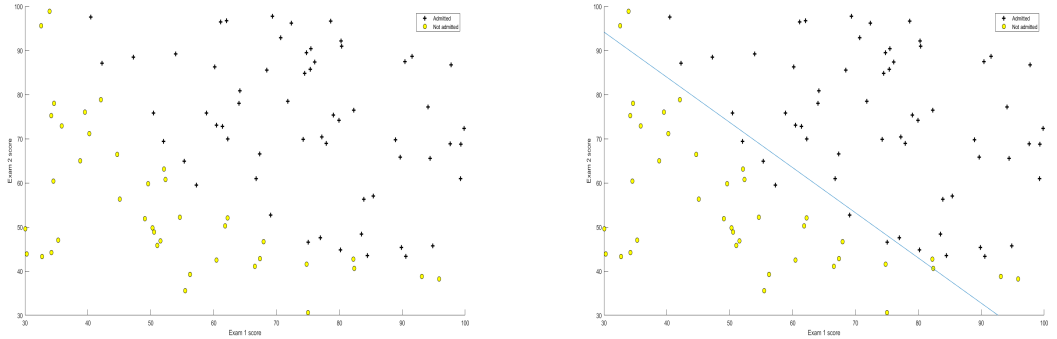
αν το label είναι 0 ή 1 χρησιμοποιώντας τις παραμέτρους ϑ του logistic regression. Πιο αναλυτικά, υπολογίζουμε τις προβλέψεις για το X, χρησιμοποιώντας ένα threshold 0.5 και ελέγχοντας αν η τιμή $f(\theta^T x)$ είναι μεγαλύτερη του theshold.

Αρχικά, ὁταν τρέχουμε τον αλγόριθμο για $\theta = 0$, το κόστος J υπολογίστηκε περίπου 0.693147, ενώ η κλίση περίπου [-0.100000, -12.009217, -11.262842]. Τα συγκεκριμένα αριθμητικά αποτελέσματα ήταν ίδια με τις αναμενόμενες τιμές που μας είχαν υποδειχθεί στην εκφώνηση.

Στη συνέχεια, με τη συνάρτηση fminunc() κα κάποιες παραμέτρους βελτίωσης υπολογίστηκε το κόστος J περίπου 0.203506 και η κλίση ήταν περίπου [-24.932758, 0.204406, 0.199616]. Παρατηρούμε δηλαδή ότι υπάρχει μείωση στο συνολικό κόστος.

Παράλληλα, βλέπουμε ότι για έναν φοιτητή που έχει γράψει 45 στο πρώτο διαγώνισμα και 85 στο δεύτερο η πιθανότητα να γίνει δεκτός από το πανεπιστήμιο είναι περίπου 0.774321.

Τέλος, όσον αφορά το συνολικό train accuracy, αυτό προκύπτει ότι είναι περίπου 89 %. Όπως φαίνεται και στις παρακάτω εικόνες φαίνεται ότι 11 από τα 100 δείγματα έχουν ταξινομηθεί λάθος, ενώ τα 89 δείγματα έχουν ταξινομηθεί σωστά.



(a) Dataset (b) Results

Άσκηση 2: Λογιστική Παλινδρόμηση με Ομαλοποίηση

Στη συγκεκριμένη άσκηση μας ζητήθηκε να εφαρμόσουμε ομαλοποιημένη λογιστική παλινδρόμηση για να προβλέψουμε αν τα μικροτσίπ από μια μονάδα κατασκευής περνούν τον έλεγχο ποιότητας. Κατά τη διάρκεια του ελέγχου ποιότητας, κάθε μικροτσίπ περνάει από διάφορες δοκιμές, για να εξασφαλιστεί ότι λειτουργεί σωστά.

Υποθέτουμε ότι υπάρχουν τα αποτελέσματα δύο διαφορετικών δοκιμών για ορισμένα μικροτσίπ. Από αυτά τα αποτελέσματα ϑα πρέπει να καθορίσουμε αν τα μικροτσίπ ϑα γίνουν αποδεκτά ή ϑα απορριφθούν. Θα χρησιμοποιήσουμε δεδομένα προηγούμενων δοκιμών για να δημιουργήσουμε ένα μοντέλο λογιστικής παλινδρόμησης.

Πιο συγκεκριμένα, απεικονίζουμε τα δεδομένα σε ένα χώρο μεγαλύτερης διάστασης όπου τα δεδομένα μπορούν να διαχωριστούν ευκολότερα. Για το σκοπό αυτό δημιουργήθηκε η συνάρτηση mapFeature()

η οποία απεικονίζει τα χαρακτηριστικά σε όλους τους όρους πολυωνύμων x_1 και x_2 μέχρι και βαθμού 6. Ειδικότερα, η συγκεκριμένη συνάρτηση υλοποιεί την παρακάτω σχέση

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

Επιπλέον, η ομαλοποιημένη συνάρτηση κόστους δίνεται από την σχέση

$$J(\theta) = \frac{1}{m} \cdot \sum_{i=1}^{m} \left(-y^{(i)} \cdot \ln(y^{(i)}) - (1 - y^{(i)}) \cdot \ln(1 - y^{(i)}) \right) + \frac{\lambda}{2m} \cdot \sum_{j=1}^{n} \theta_j^2$$

Για να βρούμε το j-στοιχείο της κλίση του σφάλματος εργαζόμαστε όπως και στην άσκηση 1 με μοναδική διαφορά ότι έχουμε να υπολογίσουμε την παρακάτω παράγωγο

$$\begin{aligned} \frac{\partial \left(\frac{\lambda}{2m} \cdot \sum_{i=1}^{n} \theta_j^2 \right)}{\partial \theta_j} &= \frac{\lambda}{2m} \cdot \sum_{i=1}^{n} \frac{\partial \left(\theta_j^2 \right)}{\partial \theta_j} \\ &= \frac{\lambda}{2m} \cdot \sum_{i=1}^{n} 2 \cdot \theta_j \\ &= \frac{\lambda}{2m} \cdot 2 \cdot \sum_{i=1}^{n} \theta_j \\ &= \frac{\lambda}{m} \cdot \sum_{i=1}^{n} \theta_j \end{aligned}$$

Συνεπώς, το j-στοιχείο της κλίσης του σφάλματος είναι

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

Αρχικά, το dataset πάνω στο οποίο εφαρμόζουμε την μέθοδο της Λογιστικής Παλινδρόμησης με Ομαλοποίηση είναι το εξής

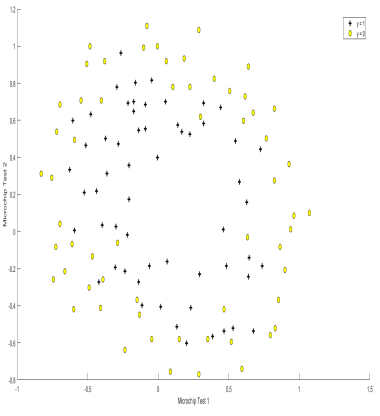
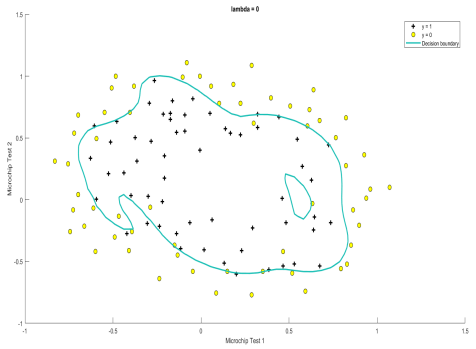


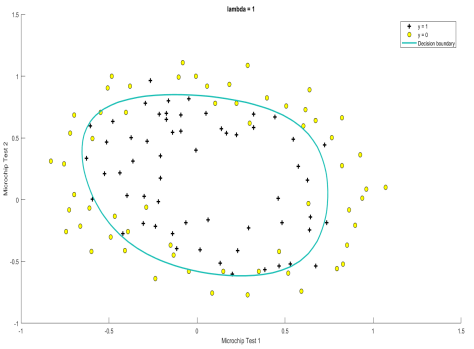
Figure 2: Dataset

Για ο σκοπό της συγκεκριμένης άσκησης δημιουργήθηκε μια συνάρτηση sigmoid() και predict() ακριβώς ίδιες με της προηγούμενης άσκησης. Επιπλέον, δημιουργήθηκε και μια συνάρτηση costFunctionReg() στην οποία υπολογίζουμε το κόστος και την κλίση του σφάλματος με βάση τους παραπάνω δύο τύπους.

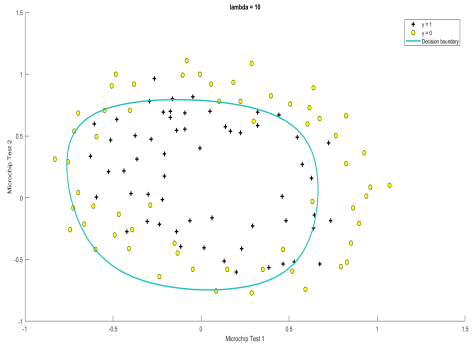
Όσον αφορά τα αποτελέσματα του αλγορίθμου, το κόστος που προκύπτει για λ=1 και ϑ αρχικοποιημένο στο 0 είναι 0.693147. Παρακάτω παρουσιάζονται τα αποτελέσματα για διαφορετικές τιμές του της παραμέτρου λ.



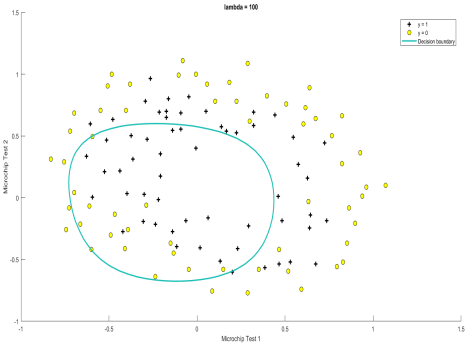
(a) λ = 0



(b) λ = 1



(c) λ = 10



(d) λ = 100

	λ=0	λ=1	λ=10	λ=100
Κόστος J	0.2226	0.5290	0.6482	0.6865
Train Accuracy	88.135593	83.050847	74.576271	61.016949

Οπως φαίνεται στο παραπάνω πινακάκι και σε συνδυασμό με τα αποτελέσματα των γραφικών παρατηρούμε ότι όσο αυξάνεται η τιμή της παραμέτρου λ, τόσο μειώνεται το Train Accuracy, ενώ το decision boundary γίνεται όλο και πιο μικρό με αποτέλεσμα να γίνονται περισσότερα σφάλματα κατα την εκπαίδευση. Τέλος, όσον αφορά το κόστος βλέπουμε ότι αυξάνεται, γεγονός που είναι λογικό με βάση και τις προηγούμενες παρατηρησεις.

Άσκηση 3: Εκτίμηση Παραμέτρων (Maximum Likelihood)

Έστω n δείγματα $D = \{x_1, x_2, \cdots, x_n,\}$ παράγονται ανεξάρτητα από μια κατανομή Poisson με παράμετρο λ, όπου η PDF δίνεται από την παρακάτω σχέση

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

Αρχικά, η συνάρτηση likelihood για την κατανομή Poisson δίνεται από τη σχέση:

$$L(\lambda|x_1, x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}$$

Επειδή ο λογάριθμος είναι γνησίως μονότονη (αύξουσα) συνάρτηση, η παράμετρος λ που μεγιστοποιεί τον λογάριθμο της πιθανοφάνειας(log-likelihood), θα μεγιστοποιεί και την πιθανοφάνεια. Συνεπώς, ορίζουμε το log-likelihood ως εξής:

$$\begin{aligned}
\mathcal{L}(\lambda|x_1, x_2, \cdots, x_n) &= \ln\left(\prod_{i=1}^{n} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}\right) \\
&= \sum_{i=1}^{n} [\ln(e^{-\lambda}) + \ln(\lambda^{x_i}) - \ln(x_i!)] \\
&= \sum_{i=1}^{n} [-\lambda + x_i \cdot \ln(\lambda) - \ln(x_i!)] \\
&= \sum_{i=1}^{n} -\lambda + \sum_{i=1}^{n} x_i \cdot \ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!) \\
&= -\lambda \cdot n + \ln(\lambda) \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!)
\end{aligned}$$

Αν παραγωγίσουμε την παραπάνω log-likelihood ως προς λ και θέσουμε την παράγωγο ίση με 0 έχουμε:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\lambda|x_1, x_2, \cdots, x_n)}{\partial \lambda} = 0 \Rightarrow \\
\frac{\partial(-\lambda \cdot n + \ln(\lambda) \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!))}{\partial \lambda} = 0 \Rightarrow \\
\frac{\partial(-\lambda \cdot n)}{\partial \lambda} + \frac{\partial(\ln(\lambda) \cdot \sum_{i=1}^{n} x_i)}{\partial \lambda} - \frac{\partial(\sum_{i=1}^{n} \ln(x_i!))}{\partial \lambda} = 0 \Rightarrow \\
-n + \frac{1}{\lambda} \cdot \sum_{i=1}^{n} x_i - 0 = 0 \Rightarrow \\
\frac{1}{\lambda} \cdot \sum_{i=1}^{n} x_i = n \Rightarrow \\
\lambda = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i
\end{aligned}$$

Συνεπώς, παρατηρούμε ότι ο maximum likelihood εκτημητής είναι απλώς ο δειγματικός μέσος των n παρατηρήσεων μεσα στο δειγματικό χώρο. Αυτό το αποτέλεσμα είναι διασιθητικά σωστό καθώς η αναμενόμενη τιμή μιας Poisson τυχαίας μεταβλητής είναι ίση με την παράμετρο λ και ο δειγματικός μέσος είναι ένας σωστός εκτιμητής της αναμενόμενης τιμής.

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

Στη συγκεκριμένη άσκηση μας ζητήθηκε να υλοποιήσουμε έναν Naive ταξινομητή Bayes για την αναγνώριση των ψηφίων που υπάρχουν στη βάση δεδομένων 'digits.mat'. Στη συγκεκριμένη βάση

υπάρχουν ψηφία από το 0 ως το 9 τα οποία είναι αποθηκευμένα ως μια ασπρόμαυρη εικόνα με διαστάσεις 28×28 που μπορεί να αναπαρασταθεί ως ένα διάνυσμα 784×1. Επιπλέον, υπάρχουν δείγματα για εκπαίδευση και δείγματα δοκιμής.

Αρχικά, οπτικοποιούμε το 43ο ψηφίο της κλάσης 7. Παρακάτω φαίνεται η εικόνα του ψηφίου.

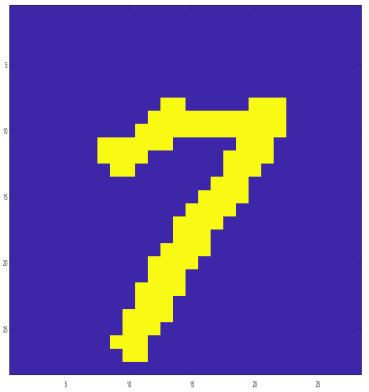


Figure 4: Digit 7

Έστω ότι στον naive Bayes ταξινομητή τα χαρακτηριστικά, τα οποία στην περίπτωσή μας, είναι τα pixel των εικόνων, είναι ανεξάρτητα και ότι ακολουθούν κατανομή Bernoulli με PDF η οποία δίνεται από τη σχέση

$$p(x) = p^x \cdot (1-p)^{1-x}$$

Η συνάρτηση likelihood για την κατανομή Bernoulli δίνεται από τη σχέση:

$$L(p|x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x) = \prod_{i=1}^{n} p^{x_i} \cdot (1-p)^{1-x_i}$$

Επειδή ο λογάριθμος είναι γνησίως μονότονη (αύξουσα) συνάρτηση, η παράμετρος λ που μεγιστοποιεί τον λογάριθμο της πιθανοφάνειας(log-likelihood), θα μεγιστοποιεί και την πιθανοφάνεια. Συνεπώς, ορίζουμε το log-likelihood ως εξής:

$$\begin{aligned}
\mathcal{L}(p|x_1, x_2, \cdots, x_n) &= \log\left(\prod_{i=1}^{n} p^{x_i} \cdot (1-p)^{1-x_i}\right) \\
&= \sum_{i=1}^{n} \log\left(p^{x_i} \cdot (1-p)^{1-x_i}\right) \\
&= \sum_{i=1}^{n} \left[\log(p^{x_i}) + \left(\log(1-p)^{1-x_i}\right)\right] \\
&= \sum_{i=1}^{n} \left(x_i \cdot \log(p) + (1-x_i) \cdot \log(1-p)\right) \\
&= \log(p) \cdot \sum_{i=1}^{n} x_i + \log(1-p) \cdot \sum_{i=1}^{n} (1-x_i)
\end{aligned}$$

Αν παραγωγίσουμε την παραπάνω log-likelihood ως προς λ και θέσουμε την παράγωγο ίση με 0 έχουμε:

$$\frac{\partial \mathcal{L}(p|x_1, x_2, \cdots, x_n)}{\partial p} = 0 \Rightarrow$$

$$\frac{\partial(\log(p) \cdot \sum_{i=1}^{n} x_i + \log(1-p) \cdot \sum_{i=1}^{n}(1-x_i))}{\partial p} = 0 \Rightarrow$$

$$\frac{\partial(\log(p) \cdot \sum_{i=1}^{n} x_i)}{\partial p} + \frac{\partial(\log(1-p) \cdot \sum_{i=1}^{n}(1-x_i))}{\partial p} = 0 \Rightarrow$$

$$\frac{1}{p} \cdot \sum_{i=1}^{n} x_i - \frac{1}{1-p} \cdot \sum_{i=1}^{n}(1-x_i) = 0 \Rightarrow$$

$$\frac{1}{1-p} \cdot \sum_{i=1}^{n}(1-x_i) = \frac{1}{p} \cdot \sum_{i=1}^{n} x_i \Rightarrow$$

$$\frac{n}{1-p} - \frac{1}{1-p} \cdot \sum_{i=1}^{n} x_i = \frac{1}{p} \cdot \sum_{i=1}^{n} x_i \Rightarrow$$

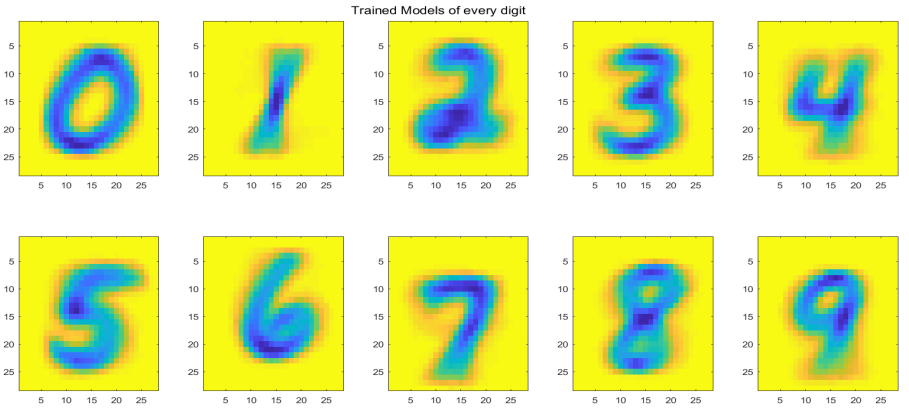
$$n \cdot p - p \cdot \sum_{i=1}^{n} x_i = (1-p) \cdot \sum_{i=1}^{n} x_i \Rightarrow$$

$$n \cdot p - p \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i - p \cdot \sum_{i=1}^{n} x_i \Rightarrow$$

$$n \cdot p = \sum_{i=1}^{n} x_i \Rightarrow$$

$$p = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Αρχικά, όσον αφορά την υλοποίηση έπρεπε να υπολογίσουμε το p^{y_i} της κάθε κλάσης με βάση τον παραπάνω τύπο που υπολογίστηκε προηγουμένως, δηλαδή αθροίζουμε όλες τις γραμμές της κάθε κλάσης και διαιρούμε με το συνολικό πλήθος εικόνων της κάθε κλάσης που είναι 500.
Στη συνέχεια, για την εκπαίδευση για κάθε κλάση και για κάθε εικόνα της κάθε κλάσης υπολογίζουμε τη log-likelihood. Ωστόσο, πριν από αυτό το βήμα επειδή κάποιες πιθανότητες έβγαιναν 0, αντικαταστήσαμε όλα τα μηδενικά του πίνακα με τις πιθανότητες με την τιμή 0.0000001, για να μπορεί να υπολογιστεί ο λογάριθος χωρίς κάποιο πρόβλημα. Επιπλέον, μόλις υπολογίσουμε και αποθηκεύσουμε όλες τις 500 log-likelihood μιας κλάσης πολλαπλασιάζουμε το ανάστροφο των εικόνων train με τον πίνακα όπου έχουμε αποθηκεύουμε τις likelihoods έτσι, ώστε να προκύψει η εκπαιδευμένη εικόνα της κάθε κλάσης. Τα αποτελέσματα φαίνονται παρακάτω



Για τη δοκιμή του Naive Bayes Classifier χρησιμοποιήθηκαν οι εικόνες δοκιμής του 'digit.mat'. Ειδικότερα, όπως και στον προηγούμενο υπολογισμό, για κάθε κλάση και για κάθε εικόνα μέσα στην κάθε κλάση υπολογίζουμε τις 10 πιθανότητες likelihood και βρίσκουμε την εικόνα με τη μέγιστη τιμή likelihood.

Για τον υπολογισμό του accuracy για το κάθε ψηφίο και το συνολικό accuracy, έχουμε 2 counters τους οποίους αυξάνουμε μόνο όταν το index, το οποίο δείχνει παράλληλα και την κλάση i, είναι ίσο με το index του array που επιστρέφει η συνάρτηση max() της matlab. Η μοναδική διαφορά μεταξύ των counter είναι ότι τον πρώτο κάθε φορά που αλλάζουμε κλάση τον αρχικοποιούμε στο 0.

Τέλος, όσον αφορά το confusion matrix αυτός υπολογίζεται ως εξής: Συμπληρώνεται ανά γραμμή και κάθε φορά αυξάνουμε το κελί (i,j) όπου i είναι η κλάσηγραμμή που βρισκόμαστε και το j είναι το index που μας επιστρέφει η συνάρτηση max().

Confussion Matrix

0,882	0	0,002	0	0,004	0,06	0,026	0	0,024	0,002
0	0,944	0,004	0,006	0	0,026	0,008	0	0,012	0
0,016	0,024	0,75	0,074	0,014	0,004	0,018	0,026	0,066	0,008
0,002	0,02	0,008	0,832	0,01	0,046	0,01	0,026	0,002	0,026
0,002	0,002	0,01	0	0,748	0,008	0,03	0,004	0,01	0,186
0,028	0,004	0,006	0,134	0,04	0,69	0,014	0,012	0,03	0,042
0,014	0,02	0,056	0	0,014	0,066	0,824	0	0,006	0
0,002	0,042	0,018	0,006	0,03	0	0	0,79	0,02	0,092
0,012	0,032	0,024	0,098	0,022	0,036	0,002	0,006	0,698	0,07
0,006	0,016	0,006	0,016	0,102	0,014	0	0,012	0,01	0,818

Πεαρακάτω παρουσιάζεται το accuracy του ταξινομητή για κάθε ψηφίο

Digits

	0	1	2	3	4	5	6	7	8	9
Accuracy	0.882	0,944	0.75	0.832	0.748	0.69	0.824	0.79	0.698	0.818

Τέλος, όσον αφορά το συνολικό accuracy του Naive Bayesian Classifier αυτό υπολογίστηκε περίπου 0.7976 ή 79.76 %.

Άσκηση 5: Support Vector Machines (Αναλυτική βελτιστοποίηση με KKT)

Έστω δύο κλάσεις ω_1 και ω_2 με τα εξής δείγματα η κάθε μία:

$$\omega_1 : x^+ = \{[3,1]^T, [3,-1]^T, [5,1]^T, [5,-1]^T\}$$
$$\omega_2 : x^- = \{[1,0]^T, [0,1]^T, [0,-1]^T, [-1,0]^T\}$$

Θα χρησιμοποιήσουμε Support Vector Machines (SVM) για να ταξινομήσουμε τις παραπάνω δύο κλάσεις σε δύο διαφορετικές κατηγορίες.

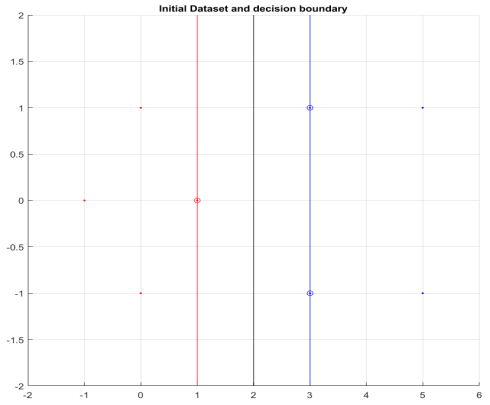


Figure 5: Linear SVM

Με βάση το παραπάνω διάγραμμα απεικόνισης το δειγμάτων, τα support vectors είναι τα εξής:

$$S_1 = [3,1]^T, \qquad S_2 = [3,-1]^T, \qquad S_3 = [1,0]^T$$

Επιπλέον, βλέπουμε ότι η βέλτιστη γραμμή διαχωρισμού των δύο κλάσεων είναι η $x = 2$, καθώς σε αυτή την περίπτωση έχουμε το μέγιστο κενό ανάμεσα στις δύο κλάσεις ω_1 και ω_2.

Για τον αναλυτικό υπολογισμό της εξίσωσης διαχωρισμού θα χρησιμοποιήσουμε πολλαπλασιαστές Lagrange και τις συνθήκες Karush-Khun-Tucker (KKT).

Αρχικά, η objective function που θέλουμε να ελαχιστοποιήσουμε είναι η εξής:

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

Οι πειορισμοί κάτω από τους οποίους θα ελαχιστοποιήσουμε την παραπάνω συνάρτηση είναι

$$w^T \cdot x^+ + b \geq 1, \forall x^+ \in \omega_1$$
$$w^T \cdot x^- + b \leq 1, \forall x^- \in \omega_2$$

Επιπλέον, έχουμε την παρακάτω σχέση

$$\mathcal{L}(w,b,\lambda) = J(w,w_0) - \sum_{i=1}^{N} \lambda_i \cdot [y_i \cdot (w^T \cdot x^+ + b) - 1]$$

Επειδή γνωρίζουμε ότι το offset της γραμμής διαχωρισμού είναι 2 και με σκοπό να περιορίσουμε το χώρο αναζήτησης, θα επιλέξουμε τιμή της παραμέτρου b ίση με -2.

$$S_1 = [3,1]^T: \quad 1 \cdot (3 \cdot w_1 + 1 \cdot w_2 - 2) \geq 1 \Rightarrow 3 \cdot w_1 + w_2 - 3 \geq 0$$
$$S_2 = [3,-1]^T: \quad 1 \cdot (3 \cdot w_1 + (-1) \cdot w_2 - 2) \geq 1 \Rightarrow 3 \cdot w_1 - w_2 - 3 \geq 0$$
$$S_3 = [1,0]^T: \quad (-1) \cdot (1 \cdot w_1 + 0 \cdot w_2 - 2) \geq 1 \Rightarrow -w_1 + 1 \geq 0$$

$$\mathcal{L}(w,b,\lambda_1,\lambda_2,\lambda_3) = \frac{1}{2} \cdot (w_1^2 + w_2^2) - \lambda_1 \cdot (3 \cdot w_1 + w_2 - 3) - \lambda_2 \cdot (3 \cdot w_1 - w_2 - 3) - \lambda_3 \cdot (-w_1 + 1)$$

Οι συνθήκες ΚΚΤ είναι οι εξής:

(1) $\Rightarrow$

$$\frac{\partial \mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3)}{\partial w_1} = w_1 - 3 \cdot \lambda_1 - 3 \cdot \lambda_2 + \lambda_3 = 0$$
$$\frac{\partial \mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3)}{\partial w_2} = w_2 - \lambda_1 + \lambda_2 = 0$$
$$\frac{\partial \mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3)}{\partial b} = -\lambda_1 - \lambda_2 + \lambda_3 = 0$$

(2) $\Rightarrow$

$$\lambda_1 \cdot (3 \cdot w_1 + w_2 - 3) = 0$$
$$\lambda_2 \cdot (3 \cdot w_1 - w_2 - 3) = 0$$
$$\lambda_3 \cdot (-w_1 + 1) = 0$$

(3) $\Rightarrow$

$$3 \cdot w_1 + w_2 - 3 \geq 0$$
$$3 \cdot w_1 - w_2 - 3 \geq 0$$
$$-w_1 + 1 \geq 0$$

(4) $\Rightarrow$

$$\lambda_i \geq 0, \qquad \forall i = 1, 2, 3$$

Διακρίνουμε τις παρακάτω 8 περιπτώσεις ανάλογα με την τιμή του κάθε λ_i

1) Αν $\lambda_1 = 0$, $\lambda_2 = 0$ και $\lambda_3 = 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 - 3 \cdot 0 - 3 \cdot 0 + 0 = 0 \Rightarrow w_1 = 0$$
$$w_2 - 0 + 0 = 0 \Rightarrow w_2 = 0$$

Αν αντικαταστήσουμε τις παραπάνω τιμές στις σχέσεις (3) προκύπτει

$$3 \cdot 0 + 0 - 3 \geq 0 \Rightarrow -3 \geq 0 \qquad \textit{Άτοπο}$$
$$3 \cdot 0 - 0 - 3 \geq 0 \Rightarrow -3 \geq 0 \qquad \textit{Άτοπο}$$
$$-0 + 1 \geq 0 \Rightarrow 1 \geq 0 \qquad \textit{Ισχύει}$$

Συνεπώς, επειδή οι παραπάνω σχέσεις δεν επαληθεύονται όλες, οι αρχική υπόθεση για τα λ δεν ισχύει.

2) Αν $\lambda_1 = 0$, $\lambda_2 = 0$ και $\lambda_3 \neq 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 - 3 \cdot 0 - 3 \cdot 0 + \lambda_3 = 0 \Rightarrow w_1 + \lambda_3 = 0$$
$$w_2 - 0 + 0 = 0 \Rightarrow w_2 = 0$$

Επιπλέον, από τη σχέσεις (2) και εφόσον $\lambda_3 \neq 0$ προκύπτει

$$-w_1 + 1 = 0 \Rightarrow w_1 = 1$$

Αντικαθιστούμε την τιμή του w_1 στην πρώτη σχέση οπότε προκύπτει $\lambda_3 = -1$, το οποίο είναι άτοπο καθώς από την 4η συνθήκη ΚΚΤ πρέπει $\lambda_3 \geq 0$.

3) Αν $\lambda_1 = 0$, $\lambda_2 \neq 0$ και $\lambda_3 = 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 - 3 \cdot 0 - 3 \cdot \lambda_2 + 0 = 0 \Rightarrow w_1 = 3 \cdot \lambda_2$$
$$w_2 - 0 + \lambda_2 = 0 \Rightarrow w_2 = -\lambda_2$$

Επιπλέον, από τη σχέσεις (2) και εφόσον $\lambda_2 \neq 0$ προκύπτει

$$3 \cdot w_1 - w_2 - 3 = 0 \Rightarrow 3 \cdot 3 \cdot \lambda_2 - (-\lambda_2) - 3 = 0 \Rightarrow \lambda_2 = \frac{3}{10}$$

Αντικαθιστώντας την τιμή του λ_2 προκύπτουν τα εξής

$$w_1 = 3 \cdot \lambda_2 = 3 \cdot \frac{3}{10} = \frac{9}{10}$$
$$w_2 = -\lambda_2 = -\frac{3}{10}$$

Αντικαθιστώντας στις σχέσεις (3) έχουμε

$$3 \cdot \frac{9}{10} + \left(-\frac{3}{10}\right) - 3 \geq 0 \Rightarrow -\frac{6}{10} \geq 0 \qquad \textit{Άτοπο}$$
$$3 \cdot \frac{9}{10} - \left(-\frac{3}{10}\right) - 3 \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$
$$-\frac{9}{10} + 1 \geq 0 \Rightarrow \frac{1}{10} \geq 0 \qquad \textit{Ισχύει}$$

Συνεπώς, επειδή οι παραπάνω σχέσεις δεν επαληθεύονται όλες, οι αρχική υπόθεση για τα λ δεν ισχύει.

4) Αν $\lambda_1 = 0$, $\lambda_2 \neq 0$ και $\lambda_3 \neq 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 - 3 \cdot 0 - 3 \cdot \lambda_2 + \lambda_3 = 0 \Rightarrow w_1 - 3 \cdot \lambda_2 + \lambda_3 = 0$$
$$w_2 - 0 + \lambda_2 = 0 \Rightarrow \lambda_2 = -w_2$$

Επιπλέον, από τη σχέσεις (2) και εφόσον $\lambda_2 \neq 0$ και $\lambda_3 \neq 0$ προκύπτει

$$-w_1 + 1 = 0 \Rightarrow w_1 = 1$$
$$3 \cdot w_1 - w_2 - 3 = 0 \Rightarrow w_2 = 3 \cdot 1 - 3 \Rightarrow w_2 = 0$$

Αντικαθιστώντας τις τιμές των w_1 και w_2 στις παραπάνω σχέσεις έχουμε

$$w_1 - 3 \cdot \lambda_2 + \lambda_3 = 0 \Rightarrow \lambda_3 = -w_1 + 3 \cdot \lambda_2 \Rightarrow \lambda_3 = -1 + 3 \cdot 0 = -1$$
$$\lambda_2 = -w_2 = 0$$

Η τιμή $\lambda_3 = -1$ έρχεται σε αντίθεση με την 4η συνθήκη KKT όπου πρέπει το κάθε λ να είναι μεγαλύτερο ή ίσο με το 0, ενώ η τιμή $\lambda_2 = 0$ έρχεται σε αντίθεση με την αρχική υπόθεση.

5) Αν $\lambda_1 \neq 0$, $\lambda_2 = 0$ και $\lambda_3 = 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 - 3 \cdot \lambda_1 - 3 \cdot 0 + 0 = 0 \Rightarrow w_1 = 3 \cdot \lambda_1$$
$$w_2 - \lambda_1 + 0 = 0 \Rightarrow w_2 = \lambda_1$$

Επιπλέον, από τις σχέσεις (2) και επειδή $\lambda_1 \neq 0$ έχουμε

$$3 \cdot w_1 + w_2 - 3 \geq 0 \Rightarrow 3 \cdot 3 \cdot \lambda_1 + \lambda_1 - 3 \geq 0 \Rightarrow \lambda_1 = \frac{3}{10}$$

Αντικαθιστώντας την τιμή του λ_1 προκύπτουν τα εξής

$$w_1 = 3 \cdot \lambda_1 = 3 \cdot \frac{3}{10} = \frac{9}{10}$$
$$w_2 = \lambda_1 = \frac{3}{10}$$

Αντικαθιστώντας στις σχέσεις (3) έχουμε

$$3 \cdot \frac{9}{10} + \frac{3}{10} - 3 \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$
$$3 \cdot \frac{9}{10} - \frac{3}{10} - 3 \geq 0 \Rightarrow -\frac{6}{10} \geq 0 \qquad \textit{Άτοπο}$$
$$-\frac{9}{10} + 1 \geq 0 \Rightarrow \frac{1}{10} \geq 0 \qquad \textit{Ισχύει}$$

Συνεπώς, επειδή οι παραπάνω σχέσεις δεν επαληθεύονται όλες, οι αρχική υπόθεση για τα λ δεν ισχύει.

6) Αν $\lambda_1 \neq 0$, $\lambda_2 = 0$ και $\lambda_3 \neq 0$
Τότε από τη σχέσεις (2) και εφόσον $\lambda_1 \neq 0$ και $\lambda_3 \neq 0$ προκύπτει

$$-w_1 + 1 = 0 \Rightarrow w_1 = 1$$
$$3 \cdot w_1 + w_2 - 3 = 0 \Rightarrow w_2 = -3 \cdot 1 + 3 \Rightarrow w_2 = 0$$

Επιπλέον, από τις σχέσεις (1) έχουμε

$$w_2 - \lambda_1 + 0 = 0 \Rightarrow \lambda_1 = w_2 = 0$$
$$w_1 - 3 \cdot \lambda_1 - 3 \cdot 0 + \lambda_3 = 0 \Rightarrow \lambda_3 = -1 + 3 = -1$$

Η τιμή $\lambda_3 = -1$ έρχεται σε αντίθεση με την 4η συνθήκη KKT όπου πρέπει το κάθε λ να είναι μεγαλύτερο ή ίσο με το 0, ενώ η τιμή $\lambda_1 = 0$ έρχεται σε αντίθεση με την αρχική υπόθεση.
7) Αν $\lambda_1 \neq 0$, $\lambda_2 \neq 0$ και $\lambda_3 = 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 - 3 \cdot \lambda_1 - 3 \cdot \lambda_2 + 0 = 0 \Rightarrow w_1 - 3 \cdot \lambda_1 - 3 \cdot \lambda_2 = 0$$
$$w_2 - \lambda_1 + \lambda_2 = 0 - \lambda_1 - \lambda_2 + 0 = 0 \Rightarrow \lambda_1 = -\lambda_2$$

Επιπλέον, από τις σχέσεις (2) και εφόσον $\lambda_1 \neq 0$ και $\lambda_2 \neq 0$ προκύπτει

$$3 \cdot w_1 + w_2 - 3 = 0$$
$$3 \cdot w_1 - w_2 - 3 = 0$$

Αθροίζοντας κατα μέλη τις παραπάνω σχέσεις έχουμε

$$2 \cdot 3 \cdot w_1 - 2 \cdot 3 = 0 \Rightarrow w_1 = 1$$

Αντικαθιστώντας την τιμή του w_1 έχουμε

$$3 \cdot w_1 - w_2 - 3 = 0 \Rightarrow w_2 = 3 \cdot 1 - 3 = 0$$

Οπότε οι από τις δύο αρχικές σχέσεις έχουμε

$$w_2 - \lambda_1 + \lambda_2 = 0 \Rightarrow 0 - \lambda_1 + \lambda_2 = 0 \Rightarrow \lambda_1 = \lambda_2$$
$$w_1 - 3 \cdot \lambda_1 - 3 \cdot \lambda_2 = 0 \Rightarrow 1 = 6 \cdot \lambda_1 \Rightarrow \lambda_1 = \frac{1}{6}\lambda_1 = -\lambda_2$$

Καταλήγουμε σε δύο σχέσεις $\lambda_1 = \lambda_2$ και $\lambda_1 = -\lambda_2$ όπου δεν μπορούν να ισχύουν και οι δύο ταυτόχρονα.
8) Αν $\lambda_1 \neq 0$, $\lambda_2 \neq 0$ και $\lambda_3 \neq 0$
Τότε από τις σχέσεις (2) και εφόσον τα λ είναι διαφορετικά το 0 προκύπτει

$$3 \cdot w_1 + w_2 - 3 = 0$$
$$3 \cdot w_1 - w_2 - 3 = 0 \Rightarrow w_2 = 3 \cdot 1 - 3 = 0$$
$$-w_1 + 1 = 0 \Rightarrow w_1 = 1$$

Από τις σχέσεις (3) ελέγχουμε αν είμαστε εντός feasible region

$$3 \cdot w_1 + w_2 - 3 \geq 0 \Rightarrow 3 \cdot 1 + 0 - 3 \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$
$$3 \cdot w_1 - w_2 - 3 \geq 0 \Rightarrow 3 \cdot 1 - 0 - 3 \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$
$$-w_1 + 1 \geq 0 \Rightarrow -1 + 1 \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$

Επιπλέον, από τις σχέσεις (1) έχουμε

$$w_2 - \lambda_1 + \lambda_2 = 0 \Rightarrow 0 - \lambda_1 + \lambda_2 = 0 \Rightarrow \lambda_1 = \lambda_2$$
$$\lambda_1 + \lambda_2 = \lambda_3 \Rightarrow \lambda_3 = 2 \cdot \lambda_1$$
$$w_1 - 3 \cdot \lambda_1 - 3 \cdot \lambda_2 + \lambda_3 = 0 \Rightarrow 1 - 3 \cdot \lambda_1 - 3 \cdot \lambda_1 + 2 \cdot \lambda_1 \Rightarrow 1 - 4 \cdot \lambda_1 = 0 \Rightarrow \lambda_1 = \frac{1}{4}$$

Οπότε καταλήγουμε στις τιμές $\lambda_1 = \lambda_2 = \frac{1}{4}$ και $\lambda_3 = 2 \cdot \lambda_1 = \frac{2}{4} = \frac{1}{2}$.

Συνεπώς, η ευθεία που προκύπτει είναι η εξής

$$w^T \cdot \vec{x} + b = 0 \Rightarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} x & y \end{bmatrix} + (-2) = 0 \Rightarrow 1 \cdot x + 0 \cdot y - 2 = 0 \Rightarrow x = 2$$

Έστω ότι οι δύο κλάσεις ω_1 και ω_2 αποτελούνται από τα εξής δείγματα

$$\omega_1 : x^+ = \{[2,2]^T, [2,-2]^T, [-2,-2]^T, [-2,2]^T\}$$
$$\omega_2 : x^- = \{[1,1]^T, [1,-1]^T, [-1,-1]^T, [-1,1]^T\}$$

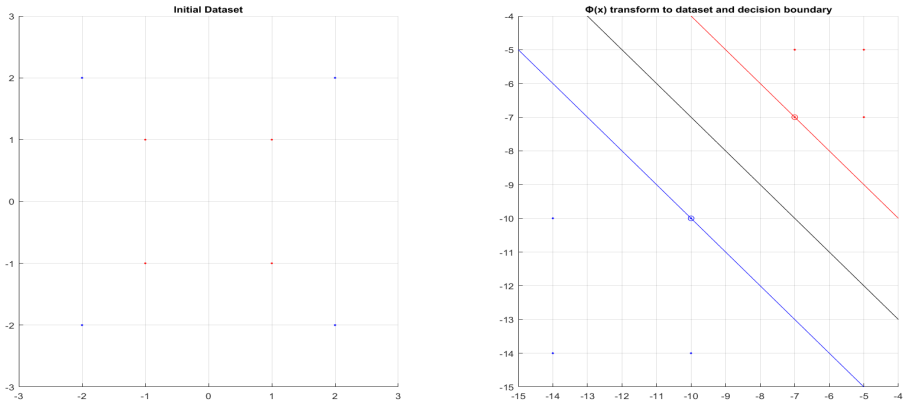


Figure 6: Not Linear SVM

Όπως φαίνεται και στο παραπάνω διάγραμμα τα αρχικά δείγματα δεν είναι γραμμικώς διαχωρίσιμα, καθώς τα δείγματα της δεύτερης κλάσης βρίσκονται ανάμεσα στα δείγματα της κλάσης 1.
Για το λόγο αυτό θα εφαρμόσουμε τον παρακάτω γραμμικό μετασχηματισμό για να γίνουν διαχωρίσιμα τα δείγματα της κάθε κλάσης.

$$\Phi(\vec{x}) = \vec{x} - \|\vec{x}\|^2 - 4$$

όπου το τετράγωνο της νόρμας των δειγμάτων της κλάσης 1 είναι ίση με $2^2 + 2^2 = 8$, ενώ της κλάσης 2 είναι $1^2 + 1^2 = 2$ Οπότε τα μετασχηματισμένα δείγματα που προκύπτουν είναι

$$\omega_1 : x^+ = \{[-10,-10]^T, [-10,-14]^T, [-14,-14]^T, [-14,-10]^T\}$$
$$\omega_2 : x^- = \{[-5,-5]^T, [-5,-7]^T, [-7,-7]^T, [-7,-5]^T\}$$

Με βάση το παραπάνω διάγραμμα απεικόνισης το δειγμάτων, τα νέα support vectors είναι τα εξής:

$$S_1 = [-10,-10]^T, \qquad S_2 = [-7,-7]^T$$

Επιπλέον, η γραμμή διαχωρισμού θα είναι η μεσοκάθετος του ευθύγραμμου τμήματος που ενώνει τα δύο support vectors. Η κλίση αυτής της μεσοκαθέτου θα είναι -1, καθώς είναι κάθετη στο ευθύγραμμο τμήμα που ενώνει τα σημεία και η εξίσωση της ευθείας που διέρχεται από τα 2 support vectors έχει κλίση 1. Όσον αφορά το offset, βλέπουμε ότι διέρχεται από το σημείο (-13,-4). Άρα, το έχουμε offset ίσο με $-4 = -1 \cdot (-13) + \beta \Rightarrow \beta = -17$. Συνεπώς, η ευθεία διαχωρισμού είναι η $y = -x - 17$.

Για τον αναλυτικό υπολογισμό της εξίσωσης διαχωρισμού θα χρησιμοποιήσουμε πολλαπλασιαστές Lagrange και τις συνθήκες Karush-Khun-Tucker (KKT). Επιπλέον, η συνάρτηση την οποία ελαχιστοποιούμε, καθώς και οι συνθήκες κάτω από τις οποίες γίνεται η ελαχιστοποίηση είναι ίδιες με την προηγούμενη περίπτωση όπου τα δείγματα ήταν γραμμικά διαχωρίσιμα.

$$S_1 = [-10, -10]^T : \quad 1 \cdot (-10 \cdot w_1 + (-10) \cdot w_2 + b) \geq 1 \Rightarrow -10 \cdot w_1 - 10w_2 + b - 1 \geq 0$$
$$S_2 = [-7, -7]^T : \quad -1 \cdot (-7 \cdot w_1 + (-7) \cdot w_2 + b) \geq 1 \Rightarrow 7 \cdot w_1 7 \cdot w_2 - b - 1 \geq 0$$

Από τις παραπάνω δυο σχέσεις βλέπουμε ότι οι ευθείες $-10 \cdot w_1 - 10w_2 + b - 1 = 0$ και $-7 \cdot w_1 - 7 \cdot w_2 - b - 1 = 0$ έχουν την ίδια κλίση αλλά διαφρετικά offsets. Επομένως, για να έχουμε 1 εξίσωση και για να μειώσουμε το χώρο αναζήτησης θα επίλέξουμε b τέτοιο ώστε να ισχύει το εξής

$$\begin{aligned} \frac{b-1}{10} &= \frac{-b-1}{-7} \\ (-7) \cdot (b-1) &= 10 \cdot (-b-1) \\ -7 \cdot b + 7 &= -10 \cdot b - 10 \\ 3 \cdot b &= -17 \\ b &= -\frac{17}{3} \end{aligned}$$

Επομένως, τα υπερεπίπεδα διαχωρισμού γίνονται

$$S_1 = [-10, -10]^T : \quad -10 \cdot w_1 - 10w_2 - \frac{17}{3} - 1 \geq 0 \Rightarrow -10 \cdot w_1 - 10w_2 - \frac{20}{3} \geq 0$$
$$S_2 = [-7, -7]^T : \quad 7 \cdot w_1 + 7 \cdot w_2 - \left(-\frac{17}{3}\right) - 1 \geq 0 \Rightarrow 7 \cdot w_1 + 7 \cdot w_2 + \frac{14}{3} \geq 0$$

$$\mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3) = \frac{1}{2} \cdot (w_1^2 + w_2^2) - \lambda_1 \cdot \left(-10 \cdot w_1 - 10w_2 - \frac{20}{3}\right) - \lambda_2 \cdot \left(7 \cdot w_1 + 7 \cdot w_2 + \frac{14}{3}\right)$$

Οι συνθήκες KKT είναι οι εξής:

(1) $\Rightarrow$

$$\begin{aligned} \frac{\partial \mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3)}{\partial w_1} &= w_1 + 10 \cdot \lambda_1 - 7 \cdot \lambda_2 = 0 \\ \frac{\partial \mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3)}{\partial w_2} &= w_2 + 10 \cdot \lambda_1 - 7 \cdot \lambda_2 = 0 \\ \frac{\partial \mathcal{L}(w, b, \lambda_1, \lambda_2, \lambda_3)}{\partial b} &= -\lambda_1 + \lambda_2 = 0 \end{aligned}$$

(2) $\Rightarrow$

$$\begin{aligned} \lambda_1 \cdot \left(-10 \cdot w_1 - 10w_2 - \frac{20}{3}\right) &= 0 \\ \lambda_2 \cdot \left(7 \cdot w_1 + 7 \cdot w_2 + \frac{14}{3}\right) &= 0 \end{aligned}$$

(3) $\Rightarrow$

$$\begin{aligned} -10 \cdot w_1 - 10 \cdot w_2 - \frac{20}{3} &\geq 0 \\ 7 \cdot w_1 + 7 \cdot w_2 + \frac{14}{3} &\geq 0 \end{aligned}$$

(4) $\Rightarrow$

$$\lambda_i \geq 0, \qquad \forall i = 1, 2, 3$$

Διακρίνουμε τις παρακάτω 4 περιπτώσεις ανάλογα με την τιμή του κάθε λ_i

1) Αν $\lambda_1 = 0$, $\lambda_2 = 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 + 10 \cdot 0 - 7 \cdot 0 = 0 \Rightarrow w_1 = 0$$
$$w_2 + 10 \cdot 0 - 7 \cdot 0 = 0 \Rightarrow w_2 = 0$$

Από τις σχέσεις (3) έχουμε

$$-10 \cdot 0 - 10 \cdot 0 - \frac{20}{3} \geq 0 \Rightarrow -\frac{20}{3} \geq 0 \qquad \textit{Άτοπο}$$
$$7 \cdot 0 + 7 \cdot 0 + \frac{14}{3} \geq 0 \Rightarrow \frac{14}{3} \geq 0 \qquad \textit{Ισχύει}$$

Συνεπώς, επειδή οι παραπάνω σχέσεις δεν επαληθεύονται όλες, οι αρχική υπόθεση για τα λ δεν ισχύει.

2) Αν $\lambda_1 = 0$, $\lambda_2 \neq 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 + 10 \cdot 0 - 7 \cdot \lambda_2 = 0 \Rightarrow w_1 = 7 \cdot \lambda_2$$
$$w_2 + 10 \cdot 0 - 7 \cdot \lambda_2 = 0 \Rightarrow w_2 = 7 \cdot \lambda_2$$

Επιπλέον, από τις σχέσεις (2) και εφόσον $\lambda_2 \neq 0$ προκύπτει

$$7 \cdot w_1 + 7 \cdot w_2 + \frac{14}{3} = 0 \Rightarrow 7 \cdot 7 \cdot \lambda_2 + 7 \cdot 7 \cdot \lambda_2 + \frac{14}{3} = 0 \Rightarrow \lambda_2 = -\frac{14}{294}$$

Η τιμή $\lambda_2 = -\frac{14}{294}$ έρχεται σε αντίθεση με την 4η συνθήκη KKT όπου πρέπει το κάθε λ να είναι μεγαλύτερο ή ίσο με το 0.

3) Αν $\lambda_1 \neq 0$, $\lambda_2 = 0$
Τότε από τις σχέσεις (1) έχουμε

$$w_1 + 10 \cdot \lambda_1 - 7 \cdot 0 = 0 \Rightarrow w_1 = -10 \cdot \lambda_1$$
$$w_2 + 10 \cdot \lambda_1 - 7 \cdot 0 = 0 \Rightarrow w_2 = -10 \cdot \lambda_1$$
$$\lambda_1 = \lambda_2$$

Επιπλέον, από τις σχέσεις (2) και εφόσον $\lambda_1 \neq 0$ προκύπτει

$$-10 \cdot w_1 - 10 \cdot w_2 + \frac{20}{3} = 0 \Rightarrow 10 \cdot (-10) \cdot \lambda_1 - 10 \cdot (-10) \cdot \lambda_1 + \frac{20}{3} = 0 \Rightarrow \lambda_1 = \frac{1}{30}$$

Αντικαθιστώντας στις παραπάνω σχέσεις προκύπτουν

$$w_1 = -10 \cdot \lambda_1 \Rightarrow w_1 = -10 \cdot \frac{1}{30} = -\frac{1}{3}$$
$$w_2 = -10 \cdot \lambda_1 \Rightarrow w_2 = -10 \cdot \frac{1}{30} = -\frac{1}{3}$$

Τέλος, αντικαθιστώντας στις σχέσεις (3) έχουμε

$$-10 \cdot \left(-\frac{1}{3}\right) - 10 \cdot \left(-\frac{1}{3}\right) - \frac{20}{3} \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$
$$7 \cdot \left(-\frac{1}{3}\right) + 7 \cdot \left(-\frac{1}{3}\right) + \frac{14}{3} \geq 0 \qquad \textit{Ισχύει}$$

Ωστόσο, από τις σχέσεις (1) προκύπτει ότι το $\lambda_1 = \lambda_2$ το οποίο έρχεται σε αντίθεση με την αρχική υπόθεση.
4) Αν $\lambda_1 \neq 0$, $\lambda_2 \neq 0$
Τότε από τις σχέσεις (1) έχουμε

$$\lambda_1 = \lambda_2$$
$$w_1 + 10 \cdot \lambda_1 - 7 \cdot \lambda_2 = 0 \Rightarrow w_1 = -10 \cdot \lambda_1 + 7 \cdot \lambda_1 \Rightarrow w_1 = -3 \cdot \lambda_1$$
$$w_2 + 10 \cdot \lambda_1 - 7 \cdot \lambda_2 = 0 \Rightarrow w_2 = -10 \cdot \lambda_1 + 7 \cdot \lambda_1 \Rightarrow w_2 = -3 \cdot \lambda_1$$

Επιπλέον, από τις σχέσεις (2) και εφόσον τα λ είναι διαφορετικά το 0 προκύπτει

$$-10 \cdot w_1 - 10 \cdot w_2 - \frac{20}{3} = 0 \Rightarrow -10 \cdot (-3 \cdot \lambda_1) - 10 \cdot (-3 \cdot \lambda_1) - \frac{20}{3} = 0 \Rightarrow 60 \cdot \lambda_1 = \frac{20}{3} \Rightarrow \lambda_1 = \frac{1}{9}$$

Αντικαθιστώντας την τιμή του λ_1 προκύπτει ότι

$$w_1 = -3 \cdot \lambda_1 \Rightarrow w_1 = -3 \cdot \frac{1}{9} = -\frac{1}{3}$$
$$w_2 = -3 \cdot \lambda_1 \Rightarrow w_2 = -3 \cdot \frac{1}{9} = -\frac{1}{3}$$

Από τις σχέσεις (3) ελέγχουμε αν είμαστε εντός feasible region

$$-10 \cdot \left(-\frac{1}{3}\right) - 10 \cdot \left(-\frac{1}{3}\right) - \frac{20}{3} \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$
$$7 \cdot \left(-\frac{1}{3}\right) + 7 \cdot \left(-\frac{1}{3}\right) + \frac{14}{3} \geq 0 \Rightarrow 0 \geq 0 \qquad \textit{Ισχύει}$$

Συνεπώς, η ευθεία που προκύπτει είναι η εξής

$$w^T \cdot \vec{x} + b = 0 \Rightarrow$$
$$\begin{bmatrix} -\frac{1}{3} \\ -\frac{1}{3} \end{bmatrix} \cdot \begin{bmatrix} x & y \end{bmatrix} + \left(-\frac{17}{3}\right) = 0 \Rightarrow$$
$$-\frac{1}{3} \cdot x - \frac{1}{3} \cdot y - \frac{17}{3} = 0 \Rightarrow$$
$$x + y + 17 = 0 \Rightarrow y = -x - 17$$

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

Έστω ότι έχουμε ένα σύνολο δειγμάτων $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(n)}, y^{(n)})\}$, όπου $x^{(i)} \in \mathbb{R}^{1\times 2}$ και $y^{(i)} \in \{-1, 1\}$. Σκοπός της άσκησης είναι να προβλέψουμε τις τιμές $y^{(i)}$ από τις αντίστοιχες $x^{(i)}$ χρησιμοποιώντας Support Vector Machines (SVM).

Η Λαγκρανζιανή του δυικού προβλήματος είναι η παρακάτω

$$\begin{aligned}\mathcal{L}(\lambda) &= \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda^{(i)} \cdot \lambda^{(j)} \cdot y^{(i)} \cdot y^{(j)} \cdot x^{(i)} \cdot (x^{(j)})^T \\ &= \begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}^T \cdot \begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{bmatrix} - \frac{1}{2} \cdot \begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{bmatrix}^T \cdot H \cdot \begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{bmatrix}\end{aligned}$$

όπου $H = y^{(i)} \cdot y^{(j)} \cdot x^{(i)} \cdot (x^{(j)})^T$ και επιπλέον μεγιστοποιούμε την παραπάνω Λαγκρανζιανή υπό τους παρακάτω περιορισμούς

$$\begin{aligned}\lambda^{(i)} \geq 0, \qquad & i = 1, 2, \cdots, n \\ \sum_{i=1}^{n} \lambda^{(i)} \cdot y^{(i)} &= 0\end{aligned}$$

Ωστόσο, η συνάρτηση quadprog() που θα χρησιμοποιήσουμε ελαχιστοποιεί την παρακάτω σχέση

$$f^T \cdot x + \frac{1}{2} \cdot x^T \cdot H \cdot x \quad \textit{έτσι ώστε} \quad \begin{cases} A \cdot x \leq b, \\ A_{eq} \cdot x = b_{eq}, \\ l_b \leq x \leq u_b \end{cases}$$

Αν πολλαπλασιάσουμε τη Λαγκρανζιανή με -1 τότε προκύπτει

$$\mathcal{L}(\lambda) = \begin{bmatrix}-1\\ \vdots \\ -1\end{bmatrix}^T \cdot \begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{bmatrix} + \frac{1}{2} \cdot \begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{bmatrix}^T \cdot H \cdot \begin{bmatrix}\lambda_1\\ \vdots \\ \lambda_n\end{bmatrix}$$

Έτσι προκύπτουν οι πίνακες f και H που θα χρησιμοποιηθούν ως ορίσματα στη συνάρτηση quadprog. Όσον αφορά τα υπόλοιπα ορίσματα για το vectors A και b χρησιμοποιούμε ένα κενό vector, καθώς ο συγκεκριμένος περιορισμός δεν μας χρειάζεται. Επιπλέον, ο πίνακας A_{eq} θα είναι ο y^T, ενώ το b_{eq} το θέλουμε να είναι ίσο με 0, για να ικανοποιείται ο δεύτερος περιορισμός. Τέλος, για τον πρώτο περιορισμό θέλουμε το vector l_b να είναι ένα vector αρχικοποιημένο με μηδενικά, ενώ το vector u_b το αρχικοποικούμε με την τιμή Inf (άπειρο), εφόσον θέλουμε το κάθε $\lambda^{(i)} \geq 0$.

Έπειτα, από τον υπολογισμό των λ, βρίσκουμε το w, w_0 και το width of margin με βάση τους παρακάτω τύπους

$$\begin{aligned} w &= \sum_{i=1}^{n} \lambda^{(i)} \cdot y_{(i)} \cdot x_{(i)} \\ w_0 &= \frac{1}{m} \cdot \sum_{i=1}^{m} (y_i - x_i \cdot w) \\ d &= -\frac{w_0}{max\{\|w\|\}} \end{aligned}$$

Τα αποτελέσματα που προκύπτουν από την εκτέλεση του αλγορίθμου είναι τα εξής

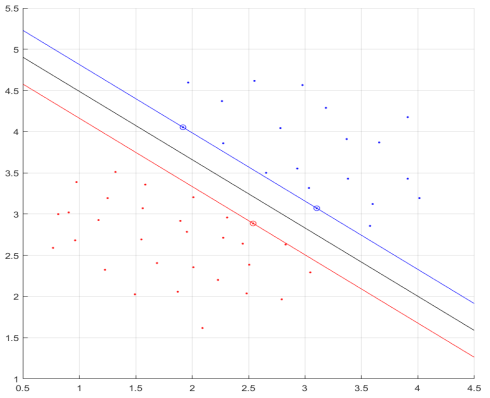
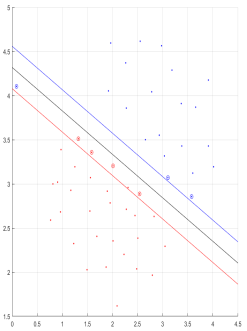


Figure 7: SVM using only 50 out of 51 samples

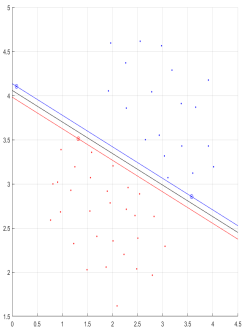
Για το δεύτερο μέρος της άσκησης θα χρησιμοποιήσουμε και τα 51 δείγματα του αρχείου και θα γίνει χρήση κάποιων slack variables, τα οποία χρησιμοποιούνται όταν υπάρχει θόρυβος στα δεδομένα έτσι ώστε να γίνονται ανεκτά κάποια δείγματα στη λάθος πλευρά της επιφάνειας απόφασης. Η Λαγκρανζιανή που θα χρησιμοποιήσουμε είναι ίδια με του προηγούμενου ερωτήματος. Ωστόσο, προστίθεται ακόμα η εξής συνθήκη

$$0 \leq \lambda^{(i)} \leq C, \qquad i = 1, 2, \cdots, n$$

Το μοναδικό πράγμα που αλλάζει από άποψη κώδικα είναι ότι αντί για την αρχικοποίηση σε Inf του vecotr u_b το αρχικοποιούμε με C σε κάθε θέση του. Οπότε τα αποτελέσματα που προκύπτουν είναι τα εξής:



(a) C = 10



(b) C = 100

Αυτό που παρατηρούμε είναι ότι όσο η τιμή της παραμέτρου C πλησιάζει στο 0 έχουμε μέγιστη ανοχή και κατ΄ επέκταση ενδέχεται κάποιο δείγμα που ανήκει πχ στην κλάση 1 να κατηγοριοποιηθεί στην κλάση 2, ενώ όσο αυξάνεται η τιμή του C έχουμε μηδενική ανοχή.
Στο τρίτο και τελευταίο μέρος της άσκησης, θα χρησιμοποιήσουμε κάποια δείγματα τα οποία δεν είναι γραμμικώς διαχωρίσιμα με την ίδια Λαγκρανζιανή των προηγούμενων ερωτημάτων. Ακριβώς επειδή τα δείγματα δεν είναι γραμμικώς διαχωρίσιμα θα χρησιμοποιήσουμε ως τρίτο χαρακτηριστικό το τετράγωνο του μέτρου κάθε δείγματος. Γι'αυτό το λόγο θα προσθέσουμε μία τρίτη στήλη στον πίνακα X με αυτό το χαρακτηριστικό. Τα βήματα του αλγορίθμου για τον υπολογισμό των λ, w, w_0

και width of margin είναι ίδια με του προηγούμενου ερωτήματος. Τα αποτελέσματα που προέκυψαν είναι τα παρακάτω

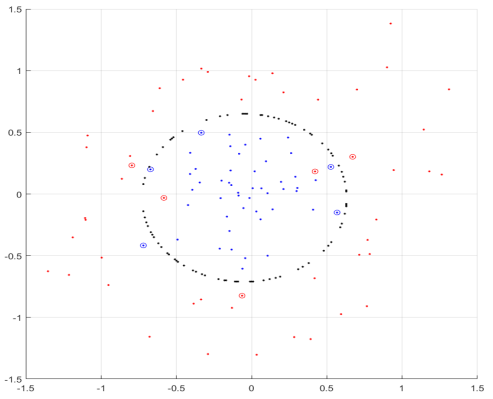


Figure 9: SVM - Not Linearly Separatable samples

Από το παραπάνω διάγραμμα παρατηρούμε ότι προκύπτει ένα αρκετά ικανοποιητικό decision boundary. Ωστόσο, υπάρχουν κάποια ελάχιστα δείγματα της κόκκινης κλάσης (2 δείγματα) εντός της περιοχής απόφασης της μπλε κλάσης, ενώ συμβαίνει και το αντίθετο, δηλάδή υπάχει ένα μπλε δείγμα στην περιοχή απόφασης της κόκκινης κλάσης.