Επιστήμη των δεδομένων

Γεώργιος Αλεξάκης 58093

22 Μαρτίου 2023

Περιεχόμενα

1 Εισαγωγή **5**

2 Ταξινόμηση **7**
2.1 Εισαγωγή στους ταξινομητές 7
2.2 Naive Bayes 8
2.3 Support Vector Machine 10
2.4 K Nearest Neighbors 12

3 Παλινδρόμηση **13**

4 Ομαδοποίηση **14**

5 Μείωση διαστάσεων **15**

6 Ενισχυτική μάθηση **16**

7 Ανίχνευση ανωμαλίας **17**

8 Συμπέρασμα **18**

Κατάλογος Πινάκων

1 Δείγματα για πρόβλεψη διαβήτη 9

Κατάλογος Σχημάτων

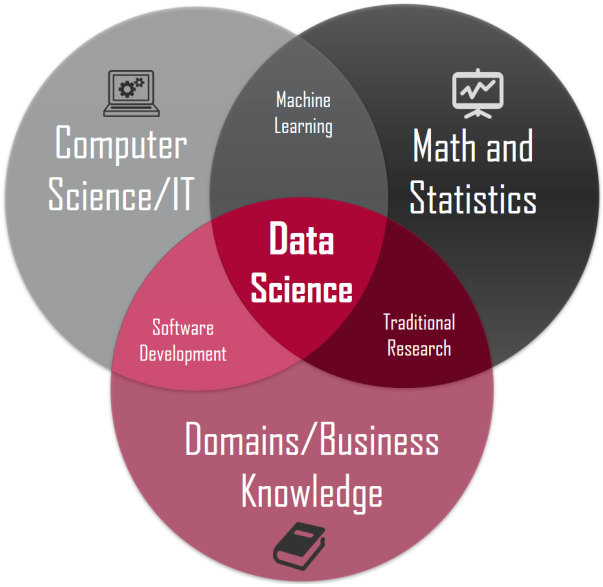
1 Γνώσεις που συνδυάζει η επιστήμη των δεδομένων 5
2 Αλγόριθμοι μηχανικής μάθησης 6
3 Τύποι ταξινομητών 7
4 Παράδειγμα SVM 10
5 Επαύξηση διάστασης με SVM 11
6 Ταξινόμηση πολλαπλών κλάσεων με SVM 11

1 Εισαγωγή

Η επιστήμη των δεδομένων (Data science) είναι μια επιστήμη που αφορά την εξαγωγή γνώσης από δεδομένα (δομημένα ή αδόμητα) και για να το πετύχει αυτό συνδυάζει γνώσεις από διάφορες άλλες επιστήμες όπως:

- Μαθηματικά
- Στατιστική
- Προγραμματισμός
- Προγνωστική ανάλυση (Predictive analysis)
- Εξόρυξη δεδομένων (Data mining)
- Τεχνητή Νοημοσύνη (Artificial intelligence)
- Μηχανική μάθηση (Machine learning)

Επιπλέον, η επιστήμη των δεδομένων συνδυάζει τα παραπάνω με τη γνώση κάποιου ειδικού τομέα (όπως η ιατρική) με σκοπό να δώσει λύση σε ένα συγκεκριμένο πρόβλημα [1]



Σχήμα 1: Γνώσεις που συνδυάζει η επιστήμη των δεδομένων

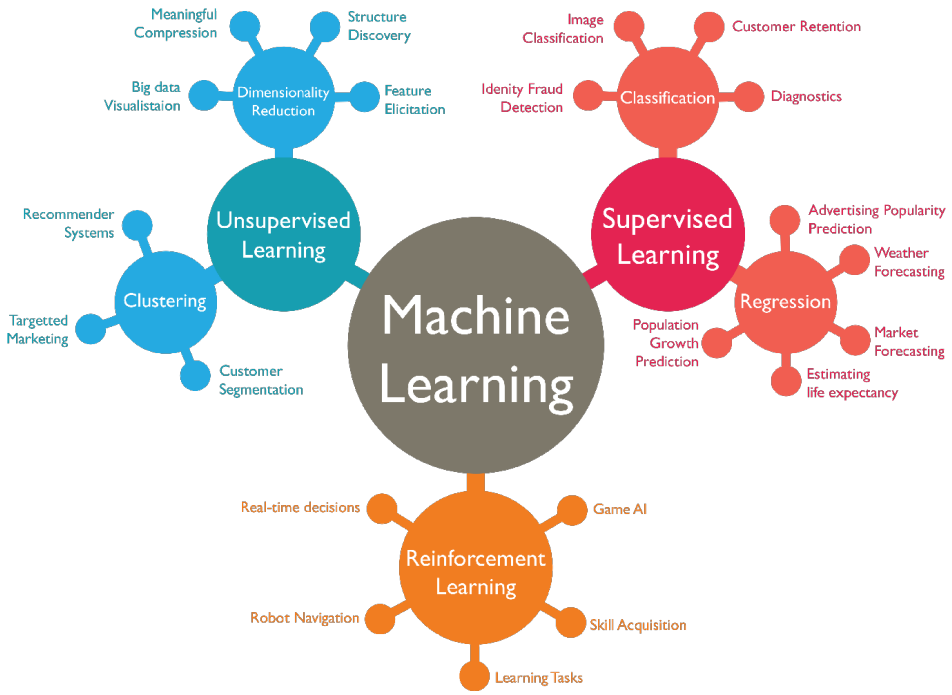
Σε αυτή την εργασία θα ασχοληθούμε κυρίως με την τεχνητή νοημοσύνη και τη μηχανική μάθηση και πιο συγκεκριμένα με τους αλγόριθμους που εφαρμόζονται στην επιστήμη των δεδομένων. Θα δοθεί αναλυτική εξήγηση για πολλούς χρήσιμους αλγόριθμους και θα υλοποιηθούν οι πιο σημαντικοί από αυτούς με τη χρήση της γλώσσας προγραμματισμού R η οποία είναι εξειδικευμένη για την επιστήμη των δεδομένων.

Οι αλγόριθμοι που θα εξετάσουμε μπορούν να χωριστούν σε πέντε ομάδες με βάση τη χρήση τους και αυτές είναι:

- Αλγόριθμοι ταξινόμησής (Classification)
- Αλγόριθμοι παλινδρόμησης (Regression)
- Αλγόριθμοι ομαδοποίησης (Clustering)

- Αλγόριθμοι μείωσης διαστάσεων (Dimentionality reduction)
- Αλγόριθμοι ενισχυτικής μάθησης (Reinforcement learning)
- Αλγόριθμοι ανίχνευσης ανωμαλίας (Anomaly detection)

Μας ενδιαφέρουν ιδιαίτερα οι πρώτες δύο ομάδες καθώς είναι αυτές που εμφανίζονται πιο συχνά. Πρόκειται για πολύ παρόμοιους αλγόριθμους οι οποίοι ανήκουν στους αλγόριθμους εποπτευόμενης μάθησης (Supervised learning) και χρησιμοποιούνται για να κάνουν προβλέψεις σύμφωνα με τα δεδομένα που τους έχουμε δώσει. Η διαφορά τους είναι ότι η ταξινόμηση προσπαθεί να προβλέψει την κλάση των καινούριων δεδομένων και να τα κατηγοριοποιήσει σύμφωνα με τις υπάρχουσες κατηγορίες, ενώ η παλινδρόμηση προσπαθεί να προβλέψει την τιμή κάποιου στοιχείου για τα καινούρια δεδομένα σύμφωνα με μια συνάρτηση που έφτιαξε από τα υπάρχοντα δεδομένα.



Σχήμα 2: Αλγόριθμοι μηχανικής μάθησης

2 Ταξινόμηση

2.1 Εισαγωγή στους ταξινομητές

Σε αυτή την ενότητα θα αναλύσουμε τους αλγορίθμους ταξινόμησης και θα δούμε τη χρήση τους. Για τους αλγορίθμους αυτούς χωρίζουμε το σύνολο των δεδομένων μας σε δυο μέρη:

- Δεδομένα εκπαίδευσης (training dataset)
- Δεδομένα επαλήθευσής (testing dataset)

Τα πρώτα τα χρησιμοποιούμε έτσι ώστε ο αλγόριθμος να βρει κάποιο μοτίβο στα δεδομένα με το οποίο θα μπορεί να κατατάσσει τα καινούρια δεδομένα που δέχεται σε κάποια από τις υπάρχουσες κλάσεις. Αυτή είναι και η διαδικασία εκπαίδευσης του μοντέλου. Αφού η εκπαίδευση τελειώσει τότε θα χρησιμοποιήσουμε τα δεδομένα επαλήθευσής για να επιβεβαιώσουμε την ορθή λειτουργία του μοντέλου. Υπάρχουν τρεις βασικοί τύποι ταξινομητών:

- Δυαδικοί (binary)
- Πολλαπλών κλάσεων (multy-class)
- Πολλαπλών ετικετών (multy-label)

Οι δυαδικοί ταξινομητές χρησιμοποιούνται όταν έχουμε μόνο δύο κλάσεις στις οποίες θέλουμε να εντάξουμε τα δεδομένα, ή όταν η απάντηση που θέλουμε να πάρουμε από το μοντέλο είναι δυαδικής φύσης. Για παράδειγμα, ένα πρόβλημα δυαδικής φύσης θα ήταν να προβλέψουμε εάν ένας ασθενής έχει ή δεν έχει μια ασθένειά σύμφωνα με τις εξετάσεις του.

Οι ταξινομητές πολλαπλών κλάσεων από την άλλη είναι ικανοί να αναγνωρίσουν περισσότερες από δύο κλάσεις και είναι πολύ χρήσιμοι για την αναγνώρισή προτύπων. Συνεχίζοντας με το προηγούμενο παράδειγμα θα θέλαμε χωρίσουμε τους ασθενείς σύμφωνα με την κατάσταση τους σε:

- υγιείς
- ήπια ασθένειά
- σοβαρή ασθένεια

΄Ετσι οι γιατροί θα μπορούν αν δράσουν ανάλογα.

Οι ταξινομητές πολλαπλών ετικετών δεν έχουν κάποια καινούρια λογική αλλά εφαρμόζουν τις λογικές των προηγούμενων προβλημάτων. Δηλαδή θα μπορούσαμε να θέλουμε να υλοποιήσουμε ένα μοντέλο το οποίο να κάνει και τις δύο προηγούμενες προβλέψεις που συζητήσαμε. Αυτά τα μοντέλα συνήθως δε χρησιμοποιούν ξεχωριστούς αλγορίθμους αλλά συνδυάζουν πολλούς ήδη γνωστούς αλγορίθμους για να φτάσουν στο αποτέλεσμα.

Σχήμα 3: Τύποι ταξινομητών

Στη συνέχεια θα αναλύσουμε τους διασημότερους αλγορίθμους και τη χρήση τους. Οι αλγόριθμοι αυτοί είναι[2, 3, 4]:

- Naive Bayes
- SVM (Support Vector Machine)
- k-NN (k Nearest Neighbors)
- Decision trees
- SGD (Stochastic Gradient Descent)
- Random Forest
- Νευρωνικά Δίκτυα Neural Networks
- LDA (Linear Discriminant Analysis)
- QDA (Quadratic Discriminant Analysis)

2.2 Naive Bayes

Σύμφωνα με το παρακάτω άρθρο [5] ο αλγόριθμος Naive Bayes είναι ένας αλγόριθμος που κάνει την υπόθεση ότι τα χαρακτηριστικά είναι υπό όρους ανεξάρτητα της δεδομένης κλάσης. Αυτή η υπόθεση στην πραγματικότητα δεν ισχύει αλλά ο αλγόριθμος πετυχαίνει πολύ υψηλή ακρίβεια και ταυτόχρονα μεγάλη υπολογιστική απόδοση. Αυτός είναι και ο λόγος που χρησιμοποιείται τόσο συχνά στην επιστήμη των δεδομένων.

Τα πλεονεκτήματα του αλγορίθμου είναι[6]:

- Η χρονική πολυπλοκότητα αυξάνεται γραμμικά με το πλήθος των δειγμάτων και των στοιχείων τους
- Χαμηλό variance (η αλλαγή των δεδομένων δεν επηρεάζει πολύ το μοντέλο)
- Μπορούμε εύκολα να προσθέσουμε και άλλα δεδομένα και το μοντέλο θα συνεχίσει την εκπαίδευση χωρίς πρόβλημα
- Δεν είναι επιρρεπής στον θόρυβο
- Δεν επηρεάζεται από έλλειψη τιμών στα δεδομένα

Τα μειονεκτήματά είναι:

- Η υπόθεση ότι τα δεδομένα είναι ανεξάρτητα τον κάνει ακατάλληλο για ορισμένα προβλήματα
- Οι συνδυασμοί στοιχείων που δεν εμφανίζονται στο σύνολο δεδομένων για κάποια πρόβλεψη θα έχει πάντα μηδενική πιθανότητα.
- Οι πιθανότητες που υπολογίζει ενδέχεται να είναι λάθος

Ο αλγόριθμος χρησιμοποιεί τον κανόνα του Bayes:

$$P(y|X) = \frac{P(y) \times P(X|y)}{P(X)}$$

Ο παραπάνω τύπος θα μας δώσει την πιθανότητα ενός δείγματος με στοιχεία X να ανήκει στην κλάση y. Το X είναι διάνυσμα με όλα τα στοιχεία του δείγματος μας. Στο παρακάτω παράδειγμα το X για

το πρώτο δείγμα θα είναι $\langle 40, 85\rangle$ και η κλάση θα είναι $y = 0$. Σκοπός μας είναι όταν πάρουμε τα στοιχεία από έναν ασθενή να μπορούμε να προβλέψουμε αν έχει διαβήτη. Θα πρέπει δηλαδή να υπολογίσουμε τις πιθανότητες $P\left(y = 0|X = \langle x_1, x_2\rangle\right)$ και $P\left(y = 1|X = \langle x_1, x_2\rangle\right)$ για τον καινούριο ασθενή με μετρήσεις $x1, x2$ και η πρόβλεψη θα είναι αυτή με τη μεγαλύτερη πιθανότητα.

Γλυκόζη	Αρτηριακή πίεση	Διαβήτης
40	85	0
40	92	0
45	63	1
45	80	0
40	73	1
45	82	0
40	85	0
30	63	1
65	65	1
45	82	0
35	73	1
45	90	0
50	68	1
40	93	0
35	80	1
50	70	1

Πίνακας 1: Δείγματα για πρόβλεψη διαβήτη

Μέχρι εδώ ήταν η γνωστή στατιστική. Αλλά όσα περισσότερα στοιχεία έχουμε ανά δείγμα τόσο πιο πολύπλοκος γίνεται ο υπολογισμός του παράγοντα $P(X|y)$. Εδώ δίνει λύση στο πρόβλημα ο αλγόριθμος Naive Bayes με την ανεξαρτησία των στοιχείων. Ανεξαρτησία σημαίνει ότι

$$P(\langle x_1, x_2, \ldots, x_n\rangle|y) = \prod_{i=1}^{n} P(x_i|y)$$

Επίσης το $P(X)$ είναι πάντα σταθερό για τα δεδομένα στοιχεία που έχουμε άρα τελικά πρέπει να υπολογίσουμε το

$$P(y) \times \prod_{i=1}^{n} P(x_i|y)$$

για κάθε y και στο τέλος να διαλέξουμε εκείνο το y που έδωσε τη μεγαλύτερη τιμή (που θα είναι και η κλάση στην οποία ανήκει το δείγμα). Βέβαια, αν τα στοιχεία δεν είναι υπό όρους ανεξάρτητα δεν ισχύει αυτή η απλοποίηση. Παρ΄όλα αυτά μπορεί να χρησιμοποιηθεί σε πολλά σύνολα δεδομένων παρόλο που κάποιες φορές δεν ισχύει και για αυτόν τον λόγο καλείται και αφελής Bayes.

Ο αλγόριθμος είναι ιδανικός για προβλήματα πολλαπλών κλάσεων λόγω του εύκολου υπολογισμού. Πιο συγκεκριμένα ο Naive Bayes χρησιμοποιείται:

- Στην ανίχνευση spam. Μπορεί διαβάζοντας τις λέξεις ενός mail να βρει την πιθανότητα να είναι spam. Κάθε λέξη που διαβάζει θα αυξάνει ή θα μειώνει την πιθανότητα με κάποιον προκαθορισμένο κανόνα μέχρι που στο τέλος θα έχουμε πάρει μια ικανοποιητική απάντηση.

- Στη συναισθηματική ανάλυση. Οι εταιρίες μπορούν να κατατάξουν τα σχόλια των καταναλωτών τους σε θετικά και αρνητικά χρησιμοποιώντας τον αλγόριθμο. Μετά μπορούν να πράξουν ανάλογα έτσι ώστε να ευνοήσουν την εταιρία.

Υπάρχουν πολλές άλλες χρήσεις πέρα από αυτές. Για μια πιο αναλυτική εξήγηση των μαθηματικών του αλγορίθμου μπορείτε να διαβάσετε την εξής έρευνα [7].

2.3 Support Vector Machine

Οι μηχανές διανυσμάτων υποστήριξης (SVM) είναι ένα σύνολο μεθόδων εποπτευόμενης μάθησης μέθοδοι που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση. Ανήκουν σε μια οικογένεια γενικευμένων γραμμικών ταξινομητών. Ένα βασικό χαρακτηριστικό του είναι ότι είναι κάνει προβλέψεις που χρησιμοποιούν τη θεωρία μηχανικής μάθησης για να μεγιστοποιήσει την προγνωστική ακρίβεια ενώ ταυτόχρονα αποφεύγει την υπερβολική προσαρμογή στα δεδομένα (over-fitting)[8].

Τα πλεονεκτήματα του SVM είναι[9]:

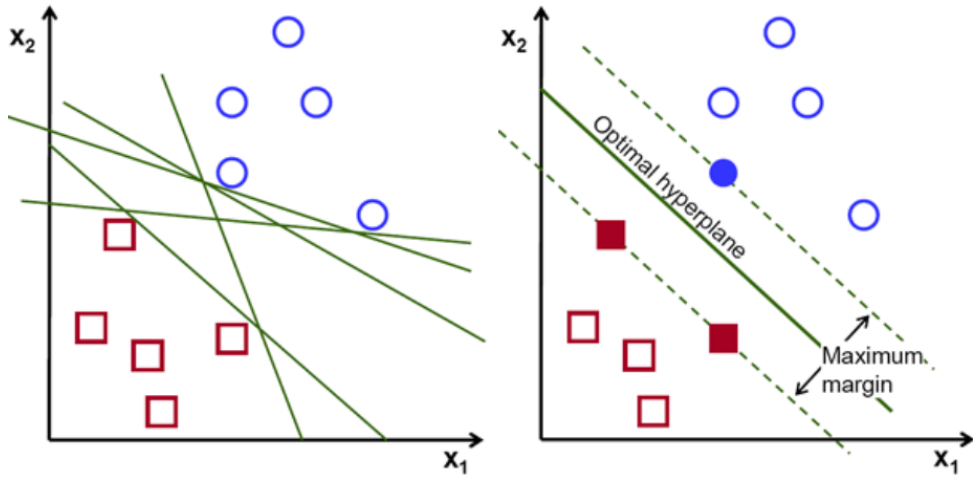
- Δουλεύει πολύ καλά όταν υπάρχει καθαρή διαφοροποίηση στα δεδομένα
- Δουλεύει καλύτερα σε μεγαλύτερες διστάσεις (πολλά στοιχεία ανά δείγμα)
- Είναι αποδοτικός όταν τα στοιχεία ανά δείγμα είναι περισσότερα από τα δείγματα
- Δεν κάνει μεγάλη χρήση μνήμης

Μειονεκτήματά:

- Δεν είναι κατάλληλος για μεγάλο σύνολο δεδομένων
- Είναι επιρρεπής στον θόρυβο
- Λόγω του τρόπου υλοποίησης του δεν μπορούμε να πάρουμε την πιθανότητα αλλά μόνο την πρόβλεψη

Ο αλγόριθμος αυτός είναι αρκετά απλός και χρησιμοποιείται κυρίως για δυαδική ταξινόμηση. Αρχικά για να τον καταλάβουμε θα πρέπει να αναπαραστήσουμε τα δεδομένα μας σαν σημεία σε έναν χώρο N διαστάσεων, όπου N τα στοιχεία που έχουμε ανά δείγμα. Το κάθε στοιχείο θα ανήκει σε μια από τις δύο κατηγορίες τις οποίες θέλουμε να αναγνωρίσουμε. Ο αλγόριθμος λοιπόν προσπαθεί να φέρει ένα υπερεπίπεδο τέτοιο ώστε να χωρίζει τα σημεία ανάλογα με την κατηγορία τους. Έτσι τα σημεία της μιας κατηγορίας θα βρίσκονται από τη μια μεριά και της δεύτερης από την άλλη.

Για να κατανοήσουμε καλύτερα τον αλγόριθμο θα κάνουμε ένα παράδειγμα σε έναν χώρο δύο διαστάσεων που μπορεί εύκολα να αντιληφθεί το ανθρώπινο μάτι.

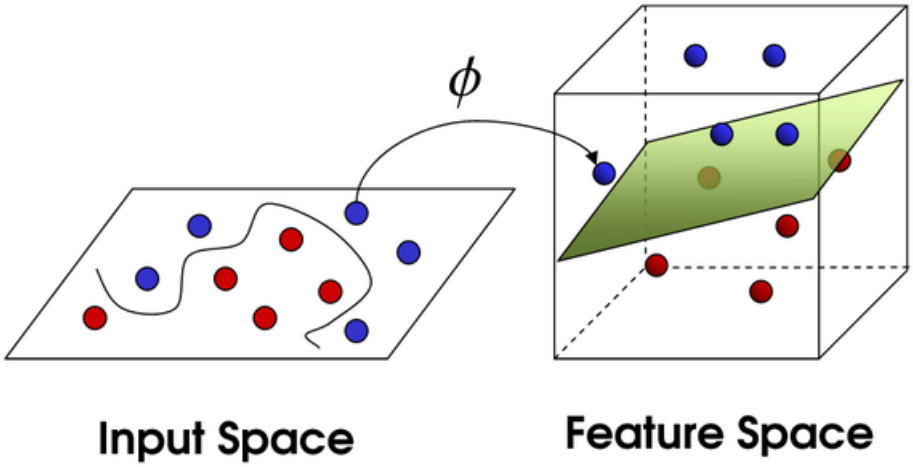


Σχήμα 4: Παράδειγμα SVM

Στον χώρο δύο διαστάσεων το υπερεπίπεδο είναι απλώς μια ευθεία. Εμείς λοιπόν θέλουμε να βρούμε μια ευθεία που να χωρίζει τα τετράγωνα με τους κύκλους. Όπως μπορούμε να δούμε από το πρώτο διάγραμμα υπάρχουν πολλές ευθείες που μπορεί να χωρίζουν τα σημεία μας αλλά εμείς

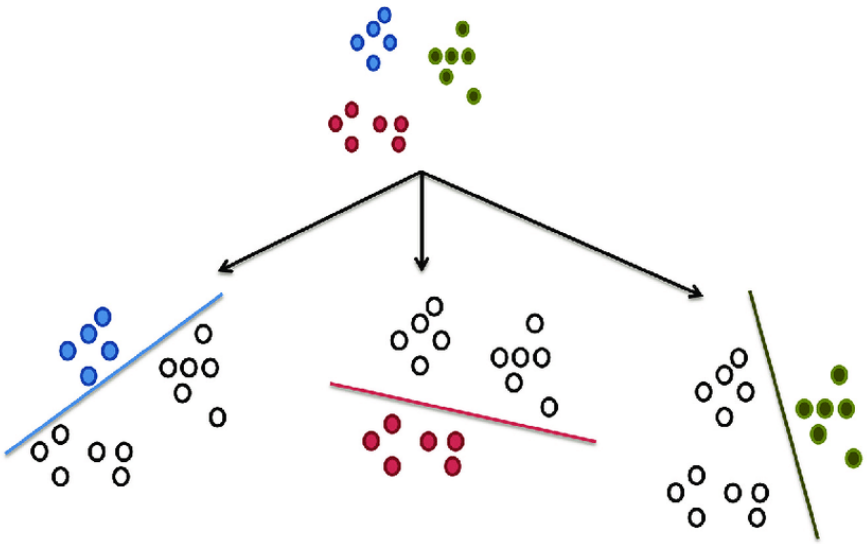
θέλουμε να διαλέξουμε την ευθεία που να έχει τη μέγιστη απόσταση και από τις δύο μεριές. Επομένως, όταν δώσουμε ένα καινούριο δείγμα στον αλγόριθμο και θέλουμε να το κατατάξει σε μια από τις δύο ομάδες πρέπει απλά να το τοποθετήσει σε αυτόν τον χώρο και μετά να δει από ποια μεριά της ευθείας βρίσκεται. Ανάλογα με τη μεριά αυτή θα αποφασίσουμε και την κλάση του αντικειμένου.

Λύσαμε το πρόβλημα όπου είχαμε πάνω από μία ευθεία αλλά υπάρχει ένα μεγαλύτερο πρόβλημα. Αυτό είναι η περίπτωση που δεν μπορούμε να χωρίσουμε τα στοιχεία μας με μια ευθεία. Εδώ έρχεται το λεγόμενο ¨kernel trick¨ για να λύσει το πρόβλημα. Το kernel είναι μια συνάρτηση σύμφωνα με την οποία επαυξάνουμε τη διάσταση του χώρου μας έτσι ώστε με την καινούρια διάσταση που θα δημιουργηθεί τα σημεία μας να είναι διαχωρίσιμα από πλέον ένα επίπεδο. Όταν γυρίσουμε πίσω στην αρχική διάσταση θα φαίνεται σαν τα σημεία μας να είναι διαχωρισμένα με μια καμπύλη.



Σχήμα 5: Επαύξηση διάστασης με SVM

Ένα τελευταίο πρόβλημα που θα μπορούσαμε να λύσουμε θα ήταν να τροποποιήσουμε τον αλγόριθμο έτσι ώστε να δουλεύει για πάνω από δύο κλάσεις. Ένας τρόπος να το κάνουμε αυτό θα ήταν με τη μέθοδο ONE versus ALL. Αυτό που κάνουμε είναι να φέρουμε μια ευθεία για κάθε κλάση που έχουμε η οποία θα χωρίζει τον χώρο στα στοιχεία που ανήκουν σε αυτήν την κλάση και σε όλα τα υπόλοιπα. Έτσι συνδυάζοντάς την πληροφορία και από τις τρεις ευθείες μπορούμε να πάρουμε τη σωστή απάντηση.



Σχήμα 6: Ταξινόμηση πολλαπλών κλάσεων με SVM

Βέβεα κάνοντας αυτή τη μέθοδο θα εμφανιστουν σημεία στον χώρο τα οποία μπορεί να ανοίκουν σε πάνω από μια κλάση και και άλλα που δεν θα ανήκουν σε καμία. Θα πρέπει λοιπόν να λάβουμε υπόψη μας όχι μόνο από ποια μεριά της κάθε ευθείας βρισκόμαστε αλλά και την απόσταση από αυτές για να μπορέσουμε να καταλήξουμε σε ένα σωστό συμπέρασμα.

Μερικές χρήσεις του αλγορίθμου είναι[10]:

- Ανίχνευση προσώπου
- Κατηγοριοποίηση κειμένου
- Ταξινόμηση εικόνων
- Βιοπληροφορική (πχ. ανίχνευση καρκίνου)
- Αναγνώριση γραφικού χαρακτήρα
- Γενικευμένος προγνωστικος έλεγχος (έλεγχος χαοτικών συστημάτων)

2.4 K Nearest Neighbors

3 Παλινδρόμηση

4 Ομαδοποίηση

5 Μείωση διαστάσεων

6 Ενισχυτική μάθηση

7 Ανίχνευση ανωμαλίας

8 Συμπέρασμα

Αναφορές

[1] Wikipedia. https://en.wikipedia.org/wiki/Data_science. Accessed: 2023-03-13.

[2] Rohit Garg. 7 types of classification algorithms. https://analyticsindiamag.com/7-types-classification-algorithms/. Accessed: 2023-03-15.

[3] Machine learning classification - 8 algorithms for data science aspirants. https://data-flair.training/blogs/machine-learning-classification-algorithms/. Accessed: 2023-03-15.

[4] Ekin Keserer. 8 types of machine learning classification algorithms. https://www.akkio.com/post/5-types-of-machine-learning-classification-algorithms. Accessed: 2023-03-15.

[5] Webb, Geoffrey I and Keogh, Eamonn and Miikkulainen, Risto. Naïve Bayes. *Encyclopedia of machine learning*, 15:713–714, 2010.

[6] Pavan Vadapalli. Naive Bayes explained: Function, Advantages and disadvantages applications in 2023. https://www.upgrad.com/blog/naive-bayes-explained/. Accessed: 2023-03-18.

[7] Rish, Irina and others. An empirical study of the naive Bayes classifier. 3(22):41–46, 2001.

[8] Jakkula, Vikramaditya. Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5):3, 2006.

[9] Dhiraj K. Top 4 advantages and disadvantages of Support Vector Machine or SVM. https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107. Accessed: 2023-03-18.

[10] Real-Life Applications of SVM (Support Vector Machines). https://data-flair.training/blogs/applications-of-svm/. Accessed: 2023-03-22.