ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Χρήση τεχνικών μηχανικής μάθησης για την ασφάλεια δικτύων υπολογιστών

Δημήτρης Μπούργος Α.Μ.:2113010

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ
Υπεύθυνος
Γεώργιος Σταμούλης
Καθηγητής

Λαμία, … έτος 20/11/2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Χρήση τεχνικών μηχανικής μάθησης για την ασφάλεια δικτύων υπολογιστών

Δημήτρης Μπούργος

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΠΙΒΛΕΠΩΝ
ΣΤΑΜΟΥΛΗΣ ΓΕΩΡΓΙΟΣ
ΚΑΘΗΓΗΤΗΣ

ΣΥΝΕΠΙΒΛΕΠΩΝ
ΓΕΩΡΓΙΟΣ ΣΠΑΘΟΥΛΑΣ
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ

Λαμία, … έτος 20/11/2020

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 20.../11/2020

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται η αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Περίληψη

Στην παρούσα πτυχιακή εργασία μελετήσαμε και αξιολογήσαμε τρεις αλγορίθμους ταξινόμησης. Τον αλγόριθμο κοντινότερων γειτόνων, τη γραμμική μηχανή διανυσματικής στήριξης, και τη μη γραμμική μηχανή διανυσματικής στήριξης χρησιμοποιώντας τον πυρήνα RBF. Εφαρμόσαμε τους αλγορίθμους σε ένα σύνολο δεδομένων που αξιοποιείται στο πρόβλημα κατειλημμένης κατοικίας. Δηλαδή, στόχος των αλγορίθμων είναι να προβλέψουν, με δεδομένα εισόδου την κατανάλωση ενέργειας της κατοικίας, το αν υπάρχει κάποιο άτομο εντός του σπιτιού.

Πιο συγκεκριμένα, συγκρίναμε τους αλγορίθμους βάση ακρίβειας στα δεδομένα εισόδου, χρησιμοποιώντας την τεχνική crossvalidation, όπου το 75% του συνόλου δεδομένων χρησιμοποιείται ως σύνολο δεδομένων εκπαίδευσης, και το υπόλοιπο χρησιμοποιείται ως σύνολο δεδομένων ελέγχου.

Abstract

On the current study, we studied and assess three different classification algorithms. Namely, the nearest neighbors algorithm, the linear support vector machine, and the non linear support vector machine using RBF as a kernel. We applied the three algorithms on a dataset that is utilized for the home occupation problem. That is, having only knowledge about the energy consumption of a house, we have to guess if a person is currently occupying that house.

More specifically, we compared the three algorithms using accuracy as the assessment measure. We used a cross validation method, where 75% of the dataset is used as training dataset, and 25% of the dataset is used for testing.

Πίνακας σχημάτων

Εικόνα 1 https://i.ytimg.com/vi/IpGxLWOIZy4/maxresdefault.jpg.....12
Εικόνα 2 https://www.cleveroad.com/images/article-previews/machine-learning-features.png.....13
Εικόνα 3 https://blogs.sas.com/content/subconsciousmusings/files/2017/04/machine-learning-cheet-sheet.png.....15
Εικόνα 4 https://images.squarespace-cdn.com/content/55ff6aece4b0ad2d251b3fee/1465017787823-KXFG6O0MU5NWYF8EI6UU/?content-type=image%2Fpng.....16
Εικόνα 5 https://staesthetic.files.wordpress.com/2014/02/svm.png?w=1060 17
Εικόνα 6 https://image.slidesharecdn.com/3-lecture-graddays-170420170135/95/machine-learning-in-science-and-industry-day-3-20-638.jpg?cb=1492711393 18
Εικόνα 7 Τα συνολικά αποτελέσματα των αλγορίθμων.....21
Εικόνα 8 Αλγόριθμος κοντινότερου γείτονα για τις διάφορες τιμές της παραμέτρου k 22
Εικόνα 9 Αλγόριθμος γραμμικής μηχανής διανυσματικής στήριξης για τις διάφορες τιμές της παραμέτρου C.....23
Εικόνα 10 Αλγόριθμος μηχανής διανυσματικής στήριξης με πυρήνα rbf για τις διάφορες τιμές των παραμέτρων C και gamma 23
Εικόνα 11 Αρχικοποίηση της δομής δεδομένων για τα αποτελέσματα. 27
Εικόνα 12 Η συνάρτηση με την οποία εκτελούμε τους αλγορίθμους ταξινόμησης και κατασκευάζουμε τη δομή δεδομένων για τα αποτελέσματα του κάθε αλγορίθμου. ...27
Εικόνα 13 Οι προγραμματιστικές βιβλιοθήκες που χρησιμοποιήσαμε.....27
Εικόνα 14 Η εκτέλεση των πρώτων δύο αλγορίθμων (κοντινότερων γειτόνων και γραμμικής μηχανής διανυσμάτων στήριξης).....28
Εικόνα 15 Κεντρική επανάληψη της προσομοίωσης όπου κατασκευάζουμε τα σετ δεδομένων εκπαίδευσης και δοκιμής. 28
Εικόνα 16 Εκτέλεση του τρίτου αλγορίθμου (μηχανή διανυσματικής στήριξης με πυρήνα rbf).....29
Εικόνα 17 Με αυτόν τον κώδικα υλοποιούμε τη συγχώνευση των τριών αρχείων δεδομένων σε ένα. Ουσιαστικά κατασκευάζουμε ένα αρχείο γραμμή προς γραμμή, όπου κάθε γραμμή περιέχει την ενέργεια, τον χρόνο που έγινε η μέτρηση, και το αν βρισκόταν κάποιο άτομο στο σπίτι 30
Εικόνα 18 Με αυτήν την υλοποίηση κατασκευάζουμε όλα τα σύνολα δεδομένων που θα χρησιμοποιηθούν στους αλγορίθμους ταξινόμησης.....31

Πίνακας περιεχομένων

Περίληψη 5
Abstract 6
Πίνακας σχημάτων 7
1.Εισαγωγή 9
1.2 Πρόβλημα και σχετικές μελέτες 10
2.Smart meters 10
2.1 Προβλήματα στην ιδιωτικότητα 11
3. Μηχανική μάθηση 12
3.1 Τι είναι η μηχανική μάθηση 12
3.2 Μηχανική μάθηση με επίβλεψη 14
3.3 Μηχανική μάθηση χωρίς επίβλεψη 14
4. Αλγόριθμοι ταξινόμησης 15
4.1 Αλγόριθμος kNN 15
4.2 Μηχανές διανυσματικής στήριξης 16
4.1.1 Γραμμικές μηχανές διανυσματικής στήριξης 16
4.1.2 Μηχανή διανυσματικής στήριξης με πυρήνα 17
5 .Υλοποίηση και εκτέλεση προσομοιώσεων 18
5.1 Γλώσσα προγραμματισμού Python 18
5.2 Βιβλιοθήκη pandas και scikit-learn 19
5.3 Περιγραφή δεδομένων 19
5.4 Εκτέλεση των αλγορίθμων ταξινόμησης 20
5.5 Αποτελέσματα προσομοιώσεων 21
6. Συμπεράσματα και συζήτηση 24
6.1 Συμπεράσματα 24
6.2 Ανοιχτά ερωτήματα 24
7. Βιβλιογραφία 25
Παράρτημα: Κώδικας Python 27
Παράρτημα: Αρχείο αποτελεσμάτων 32
Παράρτημα: Σύνολο δεδομένων 36
Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 10 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 400 36
Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 20 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 500 36
Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 30 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 600 37

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 40 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 700 37

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 50 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 800 38

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 60 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 900 38

Εισαγωγή

Στην παρούσα πτυχιακή εργασία ασχοληθήκαμε με την αξιολόγηση αλγορίθμων μηχανικής μάθησης και πιο συγκεκριμένα με την αξιολόγηση τριών αλγορίθμων ταξινόμησης. Το πρόβλημα πάνω στο οποίο εκτελέσαμε τους αλγορίθμους είναι το πρόβλημα του εντοπισμού κατειλημμένης οικίας. Δηλαδή, με την καταμέτρηση της κατανάλωσης ενέργειας σε κάποιο συγκεκριμένο σπίτι σκοπός των αλγορίθμων είναι να μαντέψουν αν υπάρχει κάποιος άνθρωπος μέσα στο σπίτι ή όχι. Το μέτρο αξιολόγησης που χρησιμοποιήσαμε είναι η ακρίβεια. Δηλαδή, το ποσοστό έγκυρων προβλέψεων σε ένα σύνολο δεδομένων δοκιμής.

Για να μπορέσει κάποιος να χρησιμοποιήσει τους αλγορίθμους μηχανικής μάθησης πρέπει να έχει στην κατοχή του δύο ειδών δεδομένα. Τα δεδομένα εκπαίδευσης του αλγορίθμου και τα δεδομένα ελέγχου. Στην παρούσα πτυχιακή εργασία χρησιμοποιήσαμε τα δεδομένα που χρησιμοποιήθηκαν στη μελέτη [1]για την επίλυση του προβλήματος κατειλημμένης οικίας το οποίο όμως σέβεται την ιδιωτικότητα των ένοικων. Με αυτό το σύνολο δεδομένων μπορέσαμε με ένα μέρος να εκπαιδεύσουμε τους αλγορίθμους και με το υπόλοιπο να τους αξιολογήσουμε.

Για να υλοποιηθούν όλα τα παραπάνω χρησιμοποιήσαμε τη γλώσσα προγραμματισμού υψηλού επιπέδου Pythonκαθώς και τις βοηθητικές βιβλιοθήκες της pandasκαι scikit-learn. Η πρώτη μας βοηθάει με την επεξεργασία των δεδομένων και η δεύτερη έχει υλοποιημένους τους αλγορίθμους που αξιολογήσαμε. Οι αλγόριθμοι που αξιολογήσαμε είναι ο αλγόριθμος κοντινότερου γείτονα, οι γραμμικές μηχανές διανυσματικής στήριξης, και οι μηχανές διανυσματικής στίριξης με πυρήνα rbf, τις οποίες αναλύουμε στο κεφάλαιο Αλγόριθμοι ταξινόμησης.

Τελικά, την καλύτερη επίδοση την είχε ο αλγόριθμος κοντινότερου γείτονα. Πιο συγκεκριμένα εκτελέσαμε τους αλγορίθμους για διάφορες τιμές των παραμέτρων τους και τους αξιολογήσαμε βάση της ακρίβειας. Ο αλγόριθμος κοντινότερου γείτονα πέτυχε τη μεγαλύτερη ακρίβεια για κάθε μέγεθος κυλιόμενου παραθύρου, μετά ήρθε ο αλγόριθμος μηχανής διανυσματικής στήριξης με πυρήνα rbf και τέλος η γραμμική μηχανή διανυσματικής στήριξης.

Η πτυχιακή χωρίζεται σε έξη κύρια κεφάλαια και τρία παραρτήματα. Το πρώτο κεφάλαιο είναι η Εισαγωγή όπου αναφέρουμε γενικά την μελέτη μας καθώς και αναφέρουμε μελέτες και εργασίες σχετικές με το πρόβλημα της κατειλημμένης οικίας. Στο επόμενο κεφάλαιο Smartmeters αναλύουμε το τι είναι τα smartmeters και ποια είναι τα προβλήματα ιδιωτικότητας που μπορεί να προκύψουν. Στο κεφάλαιο Μηχανική μάθηση εξηγούμε τι είναι η μηχανική μάθηση και αναλύουμε τους αλγορίθμους που χρησιμοποιήσαμε. Στα κεφάλαια Υλοποίηση και εκτέλεση προσομοιώσεων και Αποτελέσματα προσομοιώσεων κάνουμε μία αναλυτική περιγραφή της υλοποίησης που εφαρμόσαμε και παρουσιάζουμε τα αποτελέσματα της υλοποίησης και των προσομοιώσεων. Τέλος, στο κεφάλαιο Συμπεράσματα και συζήτηση αναφέρουμε τα συμπεράσματα που προκύπτουν από τα αποτελέσματα και κάνουμε μία σχετική συζήτηση για τα αποτελέσματα και για κάποια ανοιχτά ερωτήματα.

Στα παραρτήματα μπορούμε να βρούμε τον κώδικα που υλοποιήσαμε, το αρχείο με τα αναλυτικά αποτελέσματα και κάποια παραδείγματα από τα σύνολα δεδομένων που χρησιμοποιήσαμε.

Πρόβλημα και σχετικές μελέτες

Το πρόβλημα το οποίο θα μελετήσουμε ασχολείται με την ύπαρξη ενός ατόμου εντός κάποιας οικίας. Γενικά αυτό το πρόβλημα μπορεί να βρει πολλές εφαρμογές όπως για παράδειγμα στην υλοποίηση ενός έξυπνου σπιτιού το οποίο μειώνει την κατανάλωση ενέργειας στο ελάχιστο δυνατό. Αυτό βέβαια θα πρέπει να επιτυγχάνεται χωρίς να εισβάλει στην ιδιωτική ζωή κανενός. Επομένως, θα πρέπει να έχουμε ένα σύστημα το οποίο σέβεται πλήρως την ιδιωτικότητα του κάθε ατόμου.

Όπως αναφέρεται στην μελέτη[1], το συγκεκριμένο πρόβλημα έχει μελετηθεί ευρέως στη βιβλιογραφία. για παράδειγμα στη μελέτη [2] έχει μελετηθεί το πρόβλημα χρησιμοποιώντας εξοπλισμό ο οποίος καταγράφει και αναλύει δεδομένα υψηλής συχνότητας. Επίσης άλλα παραδείγματα μπορούμε να τα βρούμε στις μελέτες [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], όπου εφαρμόζονται τακτικές όπως η εγκατάσταση ενός ή περισσοτέρων περιβαλλοντικών αισθητήρων οι οποίοι μπορεί να συμπεριλαμβάνουν αισθητήρες κίνησης πόρτας ακουστικής κάμερες επαφής ακόμα και αισθητήρες διοξείδιο του άνθρακα.

Σύμφωνα με τη μελέτη [1], το να επιλύσουμε το πρόβλημα με αισθητήρες για να εντοπίσουμε κατευθείαν την ύπαρξη ατόμου εντός της οικίας έχει πολλές προκλήσεις. όπως για παράδειγμα οι χρήστες πρέπει να τοποθετήσουν με μεγάλη προσοχή και καλιμπράρουν τους αισθητήρες κίνησης έτσι ώστε να αποφύγουν την αναγνώριση κινήσεων από κατοικίδια που πιθανώς κυκλοφορούν στο χώρο[12].

Smartmeters

Ένας έξυπνος μετρητής (smartmeter) είναι μια ηλεκτρονική συσκευή που καταγράφει την κατανάλωση ηλεκτρικής ενέργειας και μεταδίδει τις πληροφορίες στον προμηθευτή ηλεκτρικής ενέργειας για παρακολούθηση και τιμολόγηση. Οι έξυπνοι μετρητές τυπικά καταγράφουν ενέργεια ωριαία ή συχνότερα και αναφέρουν τουλάχιστον καθημερινά. Οι έξυπνοι μετρητές επιτρέπουν αμφίδρομη επικοινωνία μεταξύ του μετρητή και του κεντρικού συστήματος. Μια τέτοια προηγμένη υποδομή μέτρησης (AMI) διαφέρει από την αυτόματη ανάγνωση μετρητών (AMR) στο ότι επιτρέπει αμφίδρομη επικοινωνία μεταξύ του μετρητή και του προμηθευτή. Οι επικοινωνίες από το μετρητή στο δίκτυο μπορεί να είναι ασύρματες ή μέσω σταθερών ενσύρματων συνδέσεων, όπως του φορέα ηλεκτρικής ενέργειας (PLC). Οι επιλογές ασύρματης επικοινωνίας που χρησιμοποιούνται συνήθως περιλαμβάνουν κυτταρικές επικοινωνίες (που μπορούν να είναι ακριβές), Wi-Fi (άμεσα διαθέσιμα), ασύρματα δίκτυα adhoc μέσω Wi-Fi, ασύρματα δίκτυα mesh, ασύρματα δίκτυα χαμηλής ισχύος (LoRa), ZigBee , ασύρματο χαμηλό ρυθμό δεδομένων) και Wi-SUN (SmartUtilityNetworks).

Η χρήση έξυπνου μετρητή φαίνεται να αποτελεί έναν τρόπο αντιμετώπισης για την αύξηση των τιμών της ηλεκτρικής ενέργειας. Εκτός αυτού,ενθαρρύνει τα εμπλεκόμενα μέρη γιατί υπάρχουν πολλά οφέλη που αποδίδονται στη χρήση έξυπνου συστήματος μέτρησης.

Από την άποψη του πελάτη (τελικός χρήστης) τα κύρια οφέληπεριλαμβάνουν[14], [15]:

- Πρόσβαση σε λεπτομερή στοιχεία για τη διαχείριση της κατανάλωσης ενέργειας.
- Ακριβέστερη και έγκαιρη παράδοση τιμολόγησης.

- ❖ Δυνατότητα οφέλους λόγο ευελιξίας της ζήτησης.
- ❖ Δυνατότητα εισαγωγής λύσεων ασφάλειας του νοικοκυριού και του εξοπλισμού, μέσω της καλύτερης ποιότητας της ηλεκτρικής ενέργειας και της διαχείρισης της βλάβης.
- ❖ Άλλες, όπως ανίχνευση βλαβών οικιακών συσκευών, ανίχνευση αποβλήτων,ανίχνευση απροσδόκητης δραστηριότητας ή αδράνειας, τι θα μπορούσε να γίνει μεέξυπνοι ελεγκτές μονάδας οικίας.

Για τον προμηθευτή ενέργειας, η χρήση έξυπνουμετρητή προσφέρει, μεταξύ άλλων[14], [15]:

- ❖ Δυνατότητα εισαγωγής της προσέγγισης αντίδρασης στη ζήτηση που είναι ιδιαίτερασημαντικό στην αγορά ηλεκτρικής ενέργειας που ασχολείται με φορτία αιχμής.
- ❖ Μειωμένο κόστος μέτρησης σε σύγκριση με τη χειροκίνητη συλλογή δεδομένων.
- ❖ Ανίχνευση κακής χρήσης και απάτης.

Προβλήματα στην ιδιωτικότητα

Ένας τεχνικός λόγος ανησυχίας για την προστασία της ιδιωτικότητας είναι ότι αυτοίοι μετρητές στέλνουν λεπτομερείς πληροφορίες σχετικά με την ποσότητα ηλεκτρικής ενέργειας που χρησιμοποιείται κάθε φορά. Οι πιο συχνές αναφορές παρέχουν πιο λεπτομερείς πληροφορίες. Οι σπάνιες αναφορές ενδέχεται να έχουν ελάχιστα οφέλη για τον πάροχο, καθώς δεν επιτρέπουν τόσο καλή διαχείριση της ζήτησης για την ανταπόκριση στις μεταβαλλόμενες ανάγκες για ηλεκτρική ενέργεια. Από την άλλη πλευρά, πολύ συχνές αναφορές θα επέτρεπαν στην εταιρεία κοινής ωφελείας να συμπεράνει συμπεριφορές για τους ενοίκους ενός σπιτιού, όπως όταν τα μέλη του νοικοκυριού πιθανότατα κοιμούνται ή απουσιάζουν. Οι τρέχουσες τάσεις είναι να αυξηθεί η συχνότητα των αναφορών. Μια λύση που ωφελεί τόσο την ιδιωτικότητα του παρόχου όσο και του χρήστη θα ήταν να προσαρμόσει δυναμικά το χρονικό διάστημα.

Μια άλλη λύση αφορά την αποθήκευση ενέργειας που έχει εγκατασταθεί στο νοικοκυριό και χρησιμοποιείται για την αναμόρφωση του προφίλ κατανάλωσης ενέργειας. Στη Βρετανική Κολομβία η ηλεκτρική υπηρεσία είναι κρατική ιδιοκτησία και ως εκ τούτου πρέπει να συμμορφώνεται με τους νόμους περί απορρήτου που εμποδίζουν την πώληση δεδομένων που συλλέγονται από έξυπνους μετρητές. Πολλά μέρη του κόσμου εξυπηρετούνται από ιδιωτικές εταιρείες που είναι σε θέση να πουλήσουν τα δεδομένα τους σε τρίτους.

Στην Αυστραλία οι συλλέκτες χρεών μπορούν να χρησιμοποιήσουν τα δεδομένα για να μάθουν πότε είναι οι άνθρωποι στο σπίτι. Χρησιμοποιούμενοι ως αποδεικτικά στοιχεία σε δικαστική υπόθεση στο Ώστιν του Τέξας, οι αστυνομικές υπηρεσίες συγκέντρωσαν κρυφά δεδομένα σχετικά με τη χρήση ενέργειας έξυπνου μετρητή από χιλιάδες κατοικίες για να προσδιορίσουν ποια χρήση περισσότερης ισχύος από την "τυπική" για τον προσδιορισμό των δραστηριοτήτων καλλιέργειας μαριχουάνας.

Τα μοντέλα χρήσης δεδομένων έξυπνου μετρητή μπορούν να αποκαλύψουν πολύ περισσότερα από το πόση ενέργεια καταναλώνεται. Έρευνες έχουν δείξει ότι τα επίπεδα ισχύος δειγματοληψίας έξυπνων μετρητών σε διαστήματα δύο

δευτερολέπτων μπορούν να αναγνωρίσουν με αξιοπιστία τη χρήση διαφορετικών ηλεκτρικών συσκευών[16], [17], [18], [19].

Ο RossAnderson[20] έχει γράψει για τις ανησυχίες περί προστασίας της ιδιωτικής ζωής. Γράφει: «Δεν είναι απαραίτητο ο μετρητής μου να πει στην εταιρεία ηλεκτροπαραγωγής, πόσο μάλλον την κυβέρνηση, πόσα χρησιμοποίησα σε κάθε μισή ώρα τον περασμένο μήνα», ότι οι μετρητές μπορούν να παρέχουν «πληροφορίες στόχευσης για διαρρήκτες», ότι το λεπτομερές ιστορικό χρήσης ενέργειας μπορεί να βοηθήσει τις εταιρείες ενέργειας να πωλούν χρηστικές συμβάσεις χρηστών, και ότι μπορεί να υπάρχει ένας «πειρασμός για τους υπεύθυνους χάραξης πολιτικής να χρησιμοποιούν δεδομένα έξυπνης μέτρησης για να στοχεύσουν τις ενδεχόμενες μειώσεις ισχύος».

Μηχανική μάθηση

Τι είναι η μηχανική μάθηση

Η μηχανική μάθηση είναι γύρω μας. Στα κινητά μας τηλέφωνά, τροφοδοτώντας τα κοινωνικά δίκτυα, βοηθώντας την αστυνομία και τους γιατρούς, τους επιστήμονες και τους πολιτικούς αρχηγούς μιας χώρας.

Εικόνα 1https://i.ytimg.com/vi/IpGxLWOIZy4/maxresdefault.jpg

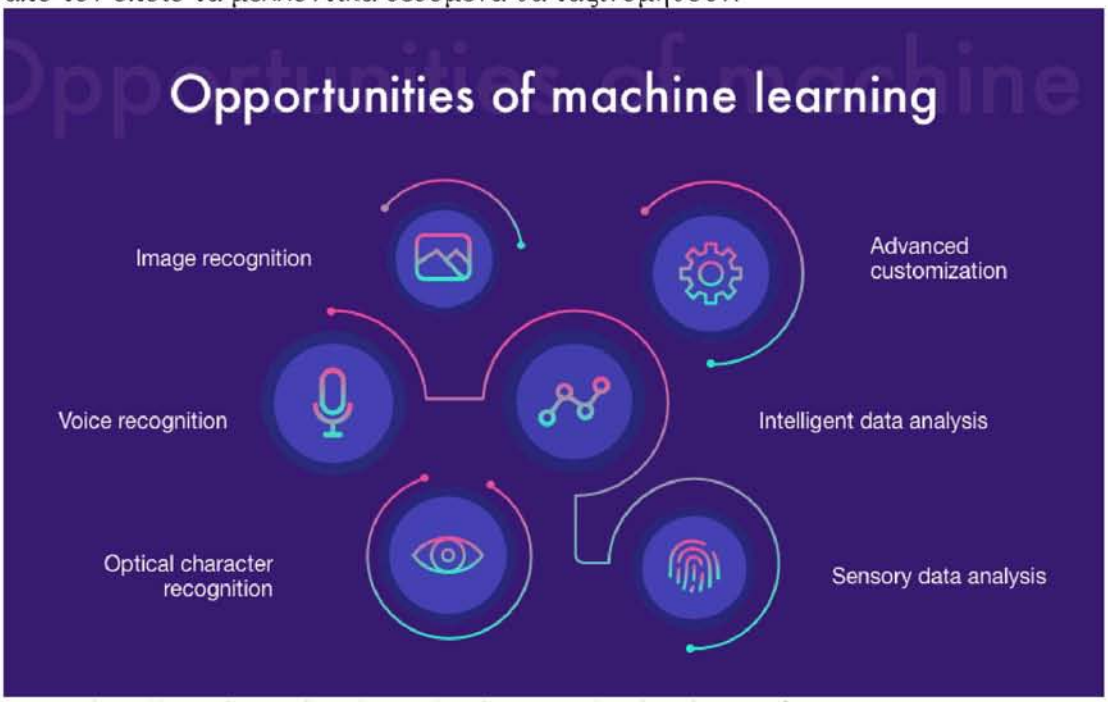
Τι είναι όμως μηχανική μάθηση?
Μια από τις ικανότητες του ανθρώπου είναι να μαθαίνει και να βελτιώνετε στα καθήκοντά του με την εμπειρία που αποκτά κατά τη διάρκεια της ζωής του.

Όταν γεννιόμαστε δεν γνωρίζουμε σχεδόν τίποτα και δεν μπορούμε να κάνουμε σχεδόν τίποτα για τον εαυτό μας, σύντομα όμως μαθαίνουμε πράγματα και γινόμαστε πιο ικανοί κάθε μέρα. Και οι υπολογιστές μπορούν να κάνουν ακριβώς το ίδιο.

Η εκμάθηση μηχανών συγκεντρώνει τα στατιστικά στοιχεία και έχει προγραμματιστεί για να πράττει έργα. Ακριβώς όπως ο εγκέφαλός μας χρησιμοποιεί την εμπειρία για να βελτιώσει μια εργασία του, έτσι κάνουν και οι υπολογιστές. Ας υποθέσουμε ότι έχουμε έναν υπολογιστή και επιθυμούμε να μας βρει τη διαφοράς από μια εικόνα

σκύλου και μια εικόνα γάτας. Θα μπορούσατε να αρχίσετε να παρατηρείται την εικόνα και να πείτε ότι αυτό είναι ένα σκυλί και αυτή είναι μια γάτα.

Ένας υπολογιστής προγραμματισμένος να μάθει θα αναζητήσει στατιστικά πρότυπα εντός δεδομένων που θα του επιτρέψουν να αναγνωρίσει μια γάτα από ένα σκυλί. Στο μέλλον θα είναι σε θέση να καταλάβει από μόνος του ότι οι γάτες έχουν μικρότερη μύτη και τα σκυλιά έρχονται σε μεγαλύτερη ποικιλία από μεγέθη. Τότε θα είναι σε θέση να τον αντιπροσωπεύουν πληροφορίες αριθμητικά οργανωμένες στο διάστημα, αλλά ο υπολογιστής, σαν μηχάνημα, δεν είναι αυτός που προσδιορίζει αυτά τα πρότυπα αλλά ο προγραμματιστής που καθορίζει τον αλγόριθμο από τον οποίο τα μελλοντικά δεδομένα θα ταξινομηθούν.



Εικόνα 2https://www.cleveroad.com/images/article-previews/machine-learning-features.png

Ένα παράδειγμα ενός απλού αλλά πολύ αποτελεσματικού αλγορίθμου είναι να βρεθεί η βέλτιστη γραμμή χωρίζοντας τις γάτες από τους σκύλους. Όταν ο υπολογιστής βλέπει μια εικόνα που ελέγχει την πλευρά της γραμμής που πέφτει, και τότε αποφασίζει αν βρήκε γάτα ή σκύλο.

Φυσικά μπορεί να υπάρξουν και λάθη. Όσο περισσότερα δεδομένα λαμβάνει ο υπολογιστής τόσο καλύτερα τα ταιριάζει στον αλγόριθμό του και τόσο πιο ακριβής μπορεί να είναι στις προβλέψεις του.

Η μηχανική μάθηση είναι ήδη ευρέως εφαρμόσιμη. Είναι η τεχνολογία πίσω από την αναγνώριση προσώπου, κειμένου σε ομιλία, σε φίλτρα ανεπιθύμητης αλληλογραφίας στα εισερχόμενα (spam), σε ηλεκτρονικές αγορές ή σε προβολές συστάσεων, καθώς επίσης και σε απάτες με πιστωτικές κάρτες ανίχνευσης και πολλά άλλα.

Στο Πανεπιστήμιο της Οξφόρδης οι ερευνητές που ασχολούνται με τη μηχανική μάθηση συνδυάζουν τα στατιστικά δεδομένα με την επιστήμη των υπολογιστών για να αναπτύξουν αλγορίθμους που να μπορούν να λύσουν πιο σύνθετα προβλήτα αποτελεσματικότερα, χρησιμοποιώντας λιγότερη υπολογιστική δύναμη.

Από την ιατρική διάγνωση έως τα μέσα κοινωνικής δικτύωσης το δυναμικό της μηχανικής μάθησης είναι να μεταμορφώσουμε τον κόσμο μας να είναι εξωπραγματικός.

Μηχανική μάθηση με επίβλεψη

Ας δούμε με ένα παράδειγμα τι είναι η υπό επίβλεψη μάθηση
Ας υποθέσουμε ότι ένας φίλος μας μας δίνει ένα εκατομμύριο νομίσματα από τριών ειδών διαφορετικών νομισμάτων. Κάθε νόμισμα από αυτά έχει διαφορετικά βάρη. Το μοντέλο σας θα προβλέψει ποια νομίσματα είναι από τα βάρη τους.

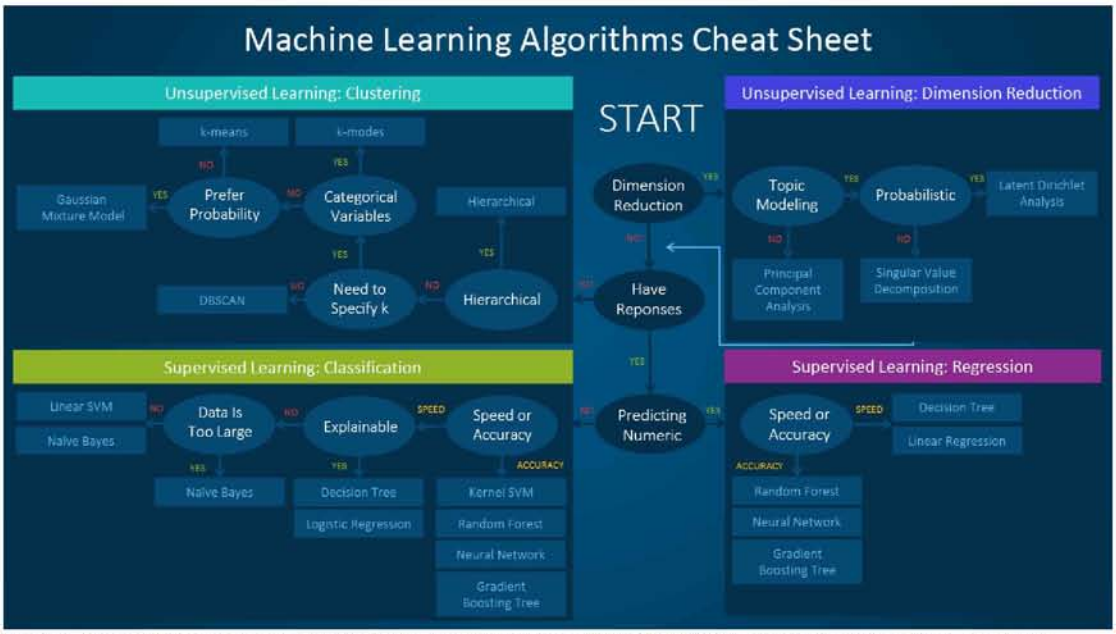
Ας υποθέσουμε ότι έχουμε τα εξής νομίσματα: ευρώ, ρουπία και σένς με τιμές δέκα, πέντε και έντεκα γραμμάρια αντίστοιχα.

Το βάρος είναι χαρακτηριστικό των κερμάτων ενώ το νόμισμα γίνεται η ετικέτα. Όταν τα δεδομένα αυτά τροφοδοτούνται στο μηχάνημα το μηχάνημα μαθαίνει ποιο χαρακτηριστικό συνδέεται που τα αναστρέφουν. Για παράδειγμα, θα μάθει ότι ένα νόμισμα είναι που είναι πέντε γραμμάρια είναι μια ρουπία.
Αν δώσουμε ένα άλλο νόμισμα στο μηχάνημα τότε με βάση το βάρος της ρουπίας θα προβλέψει το το νέο νόμισμα.

Επομένως εποπτεύει τις μαθησιακές χρήσεις ετικέτας δεδομένων για την κατάρτιση του μοντέλου. Η μηχανή υπολογίζει τα χαρακτηριστικά του αντικειμένου και συνδέει τις ετικέτες .

Μηχανική μάθηση χωρίς επίβλεψη

Για να δούμε τη διαφορά της μηχανικής μάθησης με ή χωρίς επίβλεψη ας υποθέσουμε ότι έχουμε διάφορα δεδομένα σχετικά με το άθλημα κρίκετ, όπως ένα σύνολο με τα ονόματα των παιχτών με τις αντίστοιχες βαθμολογίες του και τα wickets. Όταν τροφοδοτείτε αυτό το σύνολο δεδομένων στη μηχανή, η μηχανή αναγνωρίζει το πρότυπο της απόδοσης του παίχτη. Έτσι, γεμίζει αυτά τα δεδομένα αντίστοιχα στους άξονα x ενώ τρέχει στον άξονα y. Κατά την εξέταση των δεδομένων που θα κάνουμε είναι εμφανές ότι υπάρχουν δυο ομάδες. Στην πρώτη κλάση είναι οι παίχτες με υψηλές διαδρομές και λιγότερα wickets ενώ στην άλλη κλάση είναι οι παίχτες με τις λιγότερες διαδρομές αλλά τα περισσότερα wickets. Οπότε η διαφορά είναι ότι εδώ δεν υπάρχουν ετικέτες με δεδομένα.



Εικόνα 3https://blogs.sas.com/content/subconsciousmusings/files/2017/04/machine-learning-cheet-sheet.png

Αλγόριθμοι ταξινόμησης

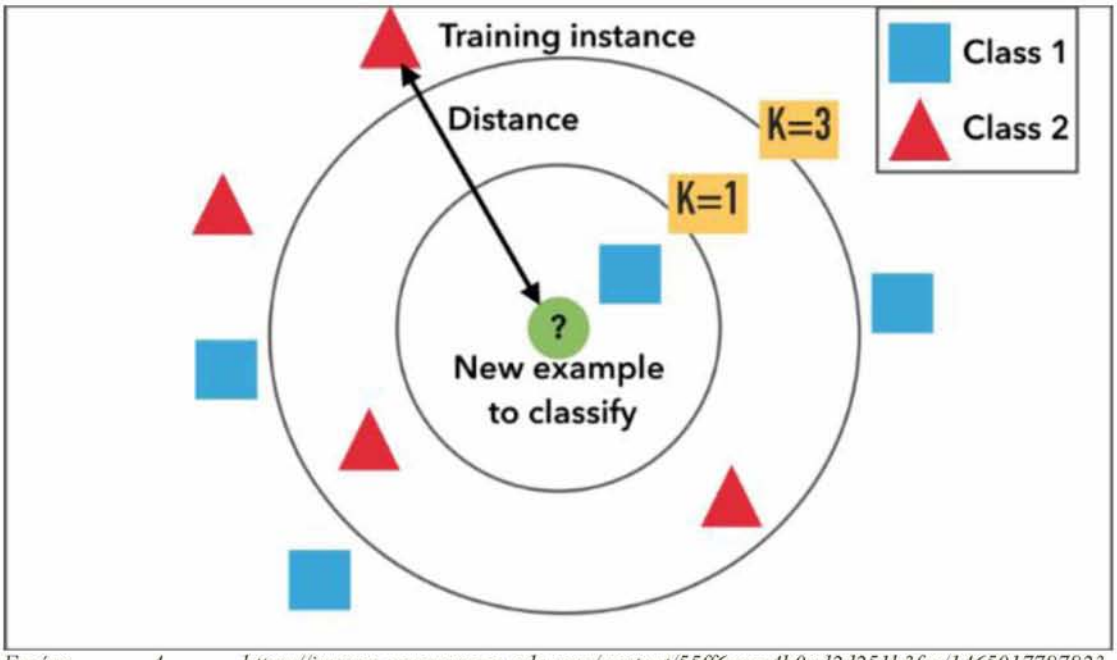
Στην παρούσα πτυχιακή εργασία, σύμφωνα με την περιγραφή που αναφέρεται στα[21], και [22], χρησιμοποιήσαμε τρεις αλγόριθμος ταξινόμησης, τον αλγόριθμο κ-κοντινότερων γειτόνων (kNN), και δύο παραλλαγές μηχανών διανυσματικής στήριξης (SVM). ο αλγόριθμος κ-κοντινότερων γειτόνων ουσιαστικά ταξινομεί ένα καινούργιο σημείο στο χώρο ανάλογα με τα ήδη ταξινομημένα σημεία που υπάρχουν. Οι μηχανές διανυσματική στήριξης ουσιαστικά διαμερίζουν τον χώρο με καμπύλες, έτσι ώστε να ξεχωρίσουν οι κλάσεις στις οποίες ανήκει το κάθε σημείο στον χώρο.

Αλγόριθμος kNN

Ο Αλγόριθμος κοντινότερων γειτόνων είναι ο πιο απλός αλγόριθμος μηχανικής μάθησης που μπορούμε να συναντήσουμε στη βιβλιογραφία. Η κατασκευή του μοντέλου του αποτελείται από την αποθήκευση των δεδομένων εκπαίδευσης. για να γίνει μία πρόβλεψη ενός νέου σημείου δεδομένων, ο αλγόριθμος βρίσκει τα κοντινότερα σημεία στο σύνολο δεδομένων εκπαίδευσης, δηλαδή τους κοντινότερους γείτονες.

Τα δεδομένα εκπαίδευσης αποτελούνται από τα διανύσματα χαρακτηριστικών καθώς και από την ετικέτα κλάσης. Για παράδειγμα, στην παρούσα πτυχιακή υπάρχουν δύο κλάσεις, είτε κάποιος βρίσκεται μέσα στο σπίτι είτε όχι. στη φάση της ταξινόμησης, όπου το κ είναι μία σταθερά που την ορίζει ο χρήστης, το σημείο δοκιμής είναι ένα διάνυσμα χαρακτηριστικών χωρίς ετικέτα, το οποίο ταξινομείται βάζοντας την ετικέτα η οποία είναι πιο συχνή ανάμεσα σε κ κοντινότερα δείγματα των δεδομένων εκπαίδευσης στο σημείο δοκιμής.

Ένα συχνό μέτρο απόστασης για συνεχείς μεταβλητές είναι η ευκλείδεια απόσταση. Επίσης, για διακριτές μεταβλητές όπως για παράδειγμα ταξινόμηση κειμένου μπορεί να χρησιμοποιηθεί η απόσταση hamming. Ένα μειονέκτημα της απλής πλειοψηφίας για την ταξινόμηση είναι ότι, τα σημεία μιας κλάσης υπερτερούν έναντι των σημείων της άλλης κλάσης, μόνο και μόνο επειδή ο αριθμός τους είναι πολύ μεγάλος μέσα στους κ κοντινότερους γείτονες. Αυτό το μειονέκτημα μπορεί να αποφευχθεί άμα χρησιμοποιήσουμε μία παραλλαγή του αλγόριθμου με βάρη.



Εικόνα 4 https://images.squarespace-cdn.com/content/55ff6aece4b0ad2d251b3fee/1465017787823-KXFG6O0MU5NWYF8EI6UU/?content-type=image%2Fpng

Μηχανές διανυσματικής στήριξης

Στο παρόν κεφάλαιο θα δούμε την θεωρία γύρο από τις μηχανές διανυσματικής στήριξης. Επειδή θα αξιολογήσουμε δύο διαφορετικές παραλλαγές αυτού του αλγορίθμου (γραμμική μηχανή διανυσματικής στήριξης και μηχανή διανυσματικής στήριξης με πυρήνα rbf) θα διαμερίσουμε το παρόν κεφάλαιο σε δύο υποκεφάλαια ένα για κάθε είδος αλγορίθμου.

Γραμμικές μηχανές διανυσματικής στήριξης

Τα γραμμικά μοντέλα χρησιμοποιούνται ευρέως στην ταξινόμηση. Θα ξεκινήσουμε βλέποντας τη δυαδική ταξινόμηση, η οποία ισχύει και για την παρούσα πτυχιακή, καθώς η πρόβλεψη έχει να κάνει με το αν βρίσκεται κάποιο άτομο μέσα στο σπίτι ή όχι. Σε αυτήν την περίπτωση η πρόβλεψη επιτυγχάνεται βάση την παρακάτω Μαθηματικής φόρμουλας.

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_p x_p + b > 0$$

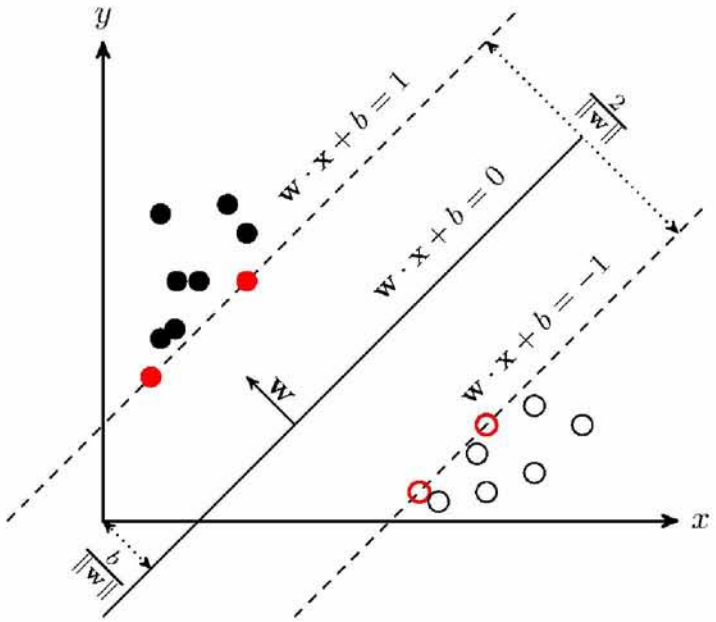
Ανάλογα με την τιμή της συνάρτησης αποφασίζουμε σε ποια κλάση ανήκει το συγκεκριμένο σημείο δοκιμής. πιο συγκεκριμένα Αν το αποτέλεσμα της συνάρτησης είναι μικρότερο του μηδενός τότε λέμε ότι το σημείο δοκιμής ανήκει στην κλάση A, διαφορετικά λέμε ότι ανήκει στην κλάση B. γενικά αυτός ο κανόνας είναι πολύ συνηθισμένος για όλα τα γραμμικά μοντέλα ταξινόμησης. σκοπός μας είναι να βρούμε τους συντελεστές w και τη σταθερά b.

Την παραπάνω συνάρτηση την συναντάμε στη βιβλιογραφία και ως εξής

$$g(x) = w^T x + w_0$$

όπου το w^Tονομάζεται διάνυσμα βαρών/συντελεστών και το w_0ονομάζεται κατώφλι. Αντίστοιχα, ανάλογα με το διάνυσμα εισόδου και την τιμή της συνάρτησης μπορούμε να κατατάξουμε το σημείο δοκιμής στις αντίστοιχες κλάσεις. Η εύρεση των

συντελεστών και του κατωφλίου πραγματοποιείται βάση κάποιο αλγόριθμο εκμάθησης. Υπάρχουν πολλοί τέτοιοι αλγόριθμοι, οι οποίοι Διαφέρουν στο πως επιλέγουν πόσο καλά προσαρμόζονται οι συντελεστές και το κατώφλι στα δεδομένα εκπαίδευσης καθώς και το αν υπάρχει κάποια κανονικοποίηση.



Εικόνα 5https://staesthetic.files.wordpress.com/2014/02/svm.png?w=1060

Στην περίπτωσή μας θέλουμε η ευθεία, το επίπεδο, ή υπερεπίπεδο να χωρίζει τις δύο ομάδες με όσο το δυνατόν ίσες αποστάσεις από κάθε σημείο των δύο ομάδων.Το υπερεπίπεδο χαρακτηρίζεται από δύο βασικά συστατικά της συνάρτησης $g(x)$. Το διάνυσμα wχαρακτηρίζει την διεύθυνση του υπερεπιπέδου και το κατώφλι w_0χαρακτηρίζει την θέση του υπερεπιπέδου στον χώρο. Επομένως, σκοπός μας είναι να βρούμε τις κατάλληλες τιμές των w και w_0 για να έχουμε τη βέλτιστη θέση του υπερεπιπέδου.

Το συγκεκριμένο πρόβλημα μπορεί να μετατραπεί σε ένα πρόβλημα βελτιστοποίησης και από εκεί και πέρα είναι θέμα μαθηματικής ανάλυσης έτσι ώστε να βρεθεί η λύση του προβλήματος. Σύμφωνα με το βιβλίο [22]το πρόβλημα που έχουμε να λύσουμε είναι το εξής:

Να υπολογιστούν τα w και w_0 του υπερεπιπέδου έτσι ώστε

$$\text{η συνάρτηση } J(w, w_0) \equiv \frac{1}{2}\|w\|^2 \text{ να έχει ελάχιστο}$$

$$\text{υπό τους περιορισμούς } y_i(w^T x_i + w_0) \geq 1, i = 1,2, \dots, N$$

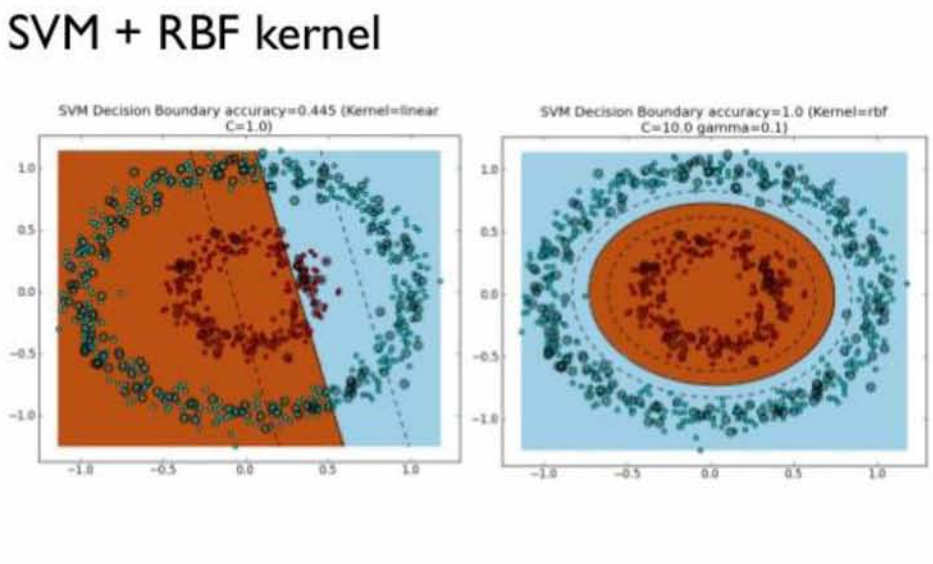
όπου το y_i αντιστοιχεί στον δείκτη της κλάσης, δηλαδή +1 για την μία και -1 για την δεύτερη κλάση.

Μηχανή διανυσματικής στήριξης με πυρήνα

Οι μη γραμμικές μηχανές διανυσματικής στήριξης ουσιαστικά ασχολούνται με δεδομένα τα οποία δεν είναι γραμμικά διαχωρίσιμα. Δηλαδή, δεν υπάρχει κάποια ευθεία γραμμή που να διαχωρίζει τις δύο κλάσεις δεδομένων. Επομένως, για να γίνουν διαχωρίσιμες οι κλάσεις γραμμικά θα πρέπει να απεικονίσουμε τα δεδομένα μας σε έναν χώρο υψηλότερης διάστασης από αυτόν που βρίσκονται αρχικά.

Σύμφωνα με ένα παράδειγμα του βιβλίου [22], μπορούμε να εκφράσουμε ένα διάνυσμα εισόδου που ανήκει στις δύο διαστάσεις, σαν ένα διάνυσμα που ανήκει στις τρεις διαστάσεις. Ένα πρόβλημα που προκύπτει είναι ότι τώρα θα έπρεπε να υπολογίσουμε εσωτερικό γινόμενο σε διανύσματα υψηλότερης διάστασης για να μπορέσουμε να διαχωρίσουμε τις δύο κλάσεις. Όμως, μπορούμε τελικά να υπολογίσουμε το εσωτερικό γινόμενο των διανυσμάτων στη μεγαλύτερη διάσταση σαν συνάρτηση των διανυσμάτων στη χαμηλότερη διάσταση, και έτσι έχουμε καλύτερους χρόνους εκτέλεσης.

Το θεώρημα του Mercer μας λέει γενικά, ότι αν έχουμε μια απεικόνιση των διανυσμάτων εισόδου σε έναν χώρο Hilbert, ο οποίος είναι ένας χώρος πλήρης γραμμικός, εξοπλισμένος με μία πράξη εσωτερικού γινομένου, τότε μπορούμε να εκφράσουμε το εσωτερικό γινόμενο των διανυσμάτων σε εκείνον τον χώρο ως μία συνάρτηση $K(x, z)$, όπου τα x, z είναι τα δύο διανύσματα στον αρχικό χώρο εισόδου. Η συγκεκριμένη συνάρτηση αποτελεί τον πυρήνα της μηχανής διανυσματικής στήριξης.



Εικόνα 6https://image.slidesharecdn.com/3-lecture-graddays-170420170135/95/machine-learning-in-science-and-industry-day-3-20-638.jpg?cb=1492711393

Υλοποίηση και εκτέλεση προσομοιώσεων

Στο παρόν κεφάλαιο θα μιλήσουμε για τα εργαλεία που χρησιμοποιήσαμε, όπως η υψηλού επιπέδου, αντικειμενοστραφείς γλώσσα προγραμματισμού Python, και οι βιβλιοθήκες pandaskαι scikit-learn. Επίσης, θα περιγράψουμε τα δεδομένα που χρησιμοποιήσαμε και θα εξηγήσουμε την υλοποίηση της μεθοδολογίας που ακολουθήσαμε.

Γλώσσα προγραμματισμού Python

HPythoneίναι μια αντικειμενοστραφείς γλώσσα προγραμματισμού σεναρίων (scripting).Επειδή είναι μια γλώσσα προγραμματισμού σεναρίων, μπορούμε πολύ γρήγορα και εύκολα να υλοποιήσουμε και να δοκιμάσουμε διάφορες μεθοδολογίες μέχρι να φτάσουμε σε ένα ικανοποιητικό αποτέλεσμα. Επίσης, εξ ορισμού δεν

υπάρχει μεταγλωττιστής για την Python, αλλά διερμηνέας, και έτσι τα συντακτικά λάθη εμφανίζονται πάντα την ώρα τις εκτέλεσης (runtime). Το πλεονέκτημα σε τέτοιου είδους γλώσσες προγραμματισμού είναι ότι δεν επιβαρυνόμαστε με τον χρόνο που απαιτείται για να πραγματοποιηθεί η μεταγλώττιση, που πολλές φορές ανάλογα και με το μέγεθος της υλοποίησης μπορεί να φτάνει ακόμα και την τάξη μεγέθους λεπτών ή ακόμα ακόμα και ωρών.

Για παράδειγμα, έστω ότι έχουμε υλοποιήσει μια εφαρμογή που θέλει ένα λεπτό για να μεταγλωττιστεί. Για κάθε μικρή αλλαγή λοιπόν που θέλουμε να κάνουμε στην προσπάθειά μας να διορθώσουμε κάποιο λογικό λάθος, χρειαζόμαστε ένα λεπτό αναμονής μέχρι να ολοκληρωθεί η μεταγλώττιση ολόκληρου του πηγαίου κώδικα από την αρχή. Ασχέτως αν η αλλαγή είναι σε πολλές γραμμές, όπως η υλοποίηση μιας νέας συνάρτησης, ή σε μία γραμμή μόνο, όπως για παράδειγμα η αλλαγή μιας λογικής έκφρασης.

Βιβλιοθήκη pandas και scikit-learn

Περιγραφή δεδομένων

Τα δεδομένα[23], [24], όπως αναφέρεται και στο [1], χωρίζονται σε δεδομένα που προέρχονται από δύο πειράματα Home-Aκαι Home-B. Συγκεκριμένα, εμείς χρησιμοποιήσαμε το σύνολο δεδομένων από το Home-A.

Το σύνολο δεδομένων μοιράζεται σε δύο μέρη, το αρχείο Home_A_spring_energytrace.txtto οποίο βρίσκεται στον φάκελο Home-A-spring-energy και τα αρχεία Home_A_spring_person_1.txtκαι Home_A_spring_person_2.txtπου βρίσκονται στον φάκελο Home-A-spring-occupancy. Όπως καταλαβαίνουμε από τα ονόματα των φακέλων το ένα αρχείο μας προσφέρει δεδομένα σχετικά με την κατανάλωση ενέργειας του σπιτιού, ενώ τα άλλα δύο αρχεία μας δίνουν το groundtruth δηλαδή το αν βρίσκεται κάποιο άτομο εντός του σπιτιού, και συγκεκριμένα είναι δεδομένα για δύο άτομα.

Για να υλοποιήσουμε τη μεθοδολογία μας θα πρέπει να γίνει μια αρχική προ επεξεργασία των δεδομένων έτσι ώστε να τα συγχωνεύσουμε σε ένα μεγάλο σύνολο, γιατί, για παράδειγμα, μας ενδιαφέρει αν γενικά υπάρχει κάποιος μέσα στο σπίτι, συνεπώς δεν μας ενδιαφέρει να έχουμε δύο διαφορετικά αρχεία για κάθε ένα από τα δύο άτομα.

Όπως φαίνεται στην Εικόνα 17 ξεκινάμε την υλοποίηση ανοίγοντας τρία αρχεία τα οποία περιέχουν τα δεδομένα μας, και ένα αρχείο εξόδου το οποίο θα είναι το τελικό αρχείο με τα δεδομένα συγχωνευμένα. Στην πρώτη γραμμή του αρχείου εξόδου γράφουμε τις κεφαλίδες των δεδομένων, οι οποίες είναι η μέτρηση της ενέργειας κάποια χρονική στιγμή μέσα στην ημέρα, η χρονική στιγμή που έγινε η μέτρηση, και το αν υπήρχε κάποιο άτομο εντός του σπιτιού ή όχι.

Το επόμενο βήμα είναι να προσπεράσουμε κάποιες γραμμές από τα αρχεία στα οποία εμφανίζεται το αν το σπίτι είναι κατειλημμένο. Αυτό συμβαίνει γιατί οι χρονικές στιγμές σε αυτές τις γραμμές δεν συμπίπτουν με τις χρονικές στιγμές στο αρχείο των μετρήσεων ενέργειας. Από εκεί και πέρα, κοιτάμε και τα τρία αρχεία γραμμή προς γραμμή και ελέγχουμε αν σε κάποιο από τα αρχεία των ατόμων (person_1 και person_2) η γραμμή τερματίζει με τον αριθμό 1 (δλδ. υπάρχει το συγκεκριμένο πρόσωπο μέσα στο σπίτι) και αν ναι τότε στη στήλη για το αν υπάρχει κάποιο

πρόσωπο εντός του σπιτιού γράφουμε τον αριθμό 1. Τα υπόλοιπα (χρονική στιγμή και μέτρηση) τα καταγράφουμε ως έχει. Στην (εικονα) που βρίσκετε στο παράρτημα μπορείτε να δείτε πως εμφανίζονται τα δεδομένα συνολικά.

Στη συνέχεια χρησιμοποιούμε αυτό το σύνολο δεδομένων για να κατασκευάσουμε κάποια περεταίρω σύνολα δεδομένων υπολογίζοντας την μέση τιμή και την τυπική απόκλιση των μετρήσεων ενέργειας με διάφορα μεγέθη κυλιόμενου παραθύρου. Πιο συγκεκριμένα, για τη μέση τιμή χρησιμοποιήσαμε παράθυρο μεγέθους 10, 20, 30, 40, 50, και 60, ενώ για την τυπική απόκλιση έχουμε μέγεθος 400, 500, 600, 700, 800, και 900. Τα συγκεκριμένα μεγέθη τα ομαδοποιήσαμε ανά δύο και κατασκευάσαμε έξη σύνολα δεδομένων. Δηλαδή, ένα σύνολο με μεγέθη 10, 400, ένα με 20, 500 και ου το καθεξής. Η υλοποίηση για την κατασκευή αυτών των συνόλων δεδομένων μπορεί να φανεί στην Εικόνα 18.

Σε αυτή τη φάση η βιβλιοθήκη pandastης Pythonήταν πολύ βοηθητική. Σε μία επανάληψη για κάθε ζεύγος μεγέθους παραθύρου αυτό που κάνουμε είναι να κατασκευάζουμε ένα καινούριο DataFrame, το οποίο είναι ουσιαστικά ένας πίνακας με κεφαλίδες, και να εισάγουμε δύο νέες στήλες τις οποίες τις κατασκευάζουμε με την βοηθητική συνάρτηση rolling() που υλοποιεί ένα κυλιόμενο παράθυρο. Αυτά τα δεδομένα, των οποίων παραδείγματα μπορείτε να βρείτε στο Παράρτημα: Σύνολο δεδομένων, τα χρησιμοποιούμε για να εκπαιδεύσουμε τους αλγορίθμους ταξινόμησης που θα συζητήσουμε παρακάτω.

Εκτέλεση των αλγορίθμων ταξινόμησης

Όπως είδαμε παραπάνω και στο θεωρητικό μέρος οι αλγόριθμοι ταξινόμησης που χρησιμοποιήσαμε είναι ο αλγόριθμος κοντινότερων γειτόνων, και δύο παραλλαγές των μηχανών διανυσματικής στήριξης. Όπως μπορούμε να δούμε στην Εικόνα 13 φαίνονται οι βιβλιοθήκες που χρησιμοποιήσαμε, το σημαντικό είναι ότι μπορούμε να δούμε και ποιους αλγορίθμους χρησιμοποιήσαμε. Συγκεκριμένα, από το πακέτο `sklearn.neighbors`χρησιμοποιούμε τον αλγόριθμο`KNeighborsClassifier`και από το πακέτο`sklearn.svm`χρησιμοποιούμε τους αλγορίθμους`SVC`,και `LinearSVC`.

Στη συνέχεια στην Εικόνα 12υλοποιούμε μία συνάρτηση με την οποία μπορούμε να εκτελέσουμε κάθε αλγόριθμο ταξινόμησης. Ουσιαστικά χρησιμοποιεί τον `my_classifier`για να κάνει την εκπαίδευση με την συνάρτηση`fit(x_train, y_train)` όπου τα`x_train`, και`y_train`είναι τα δεδομένα εισόδου και εξόδου / πρόβλεψης αντίστοιχα. Με την συνάρτηση `score()` γίνεται η αξιολόγηση του εκάστοτε αλγορίθμου πάνω στο σύνολο δεδομένων εκπαίδευσης και ελέγχου. Η αξιολόγηση πάνω στο σύνολο δεδομένων εκπαίδευση πραγματοποιείται για να εκλεχθεί αν υπάρχει το φαινόμενο υπέρ προσαρμογής. Τέλος, τα δεδομένα αποθηκεύονται στη δομή δεδομένων `results`.

Αφού έχουμε αρχικοποιήσει τη δομή δεδομένων των αποτελεσμάτων (Εικόνα 11), ξεκινάμε σε μία επανάληψη (Εικόνα 15) να εκτελούμε τους αλγορίθμους για κάθε σύνολο δεδομένων που έχει δημιουργηθεί με τα ζευγάρια μεγεθών κυλιόμενου παραθύρου.

Η εκτέλεση των αλγορίθμων υλοποιείται στην Εικόνα 14 και Εικόνα 16. Εδώ βλέπουμε ότι ο κάθε αλγόριθμος παραμετροποιείται περεταίρω για διάφορες τιμές

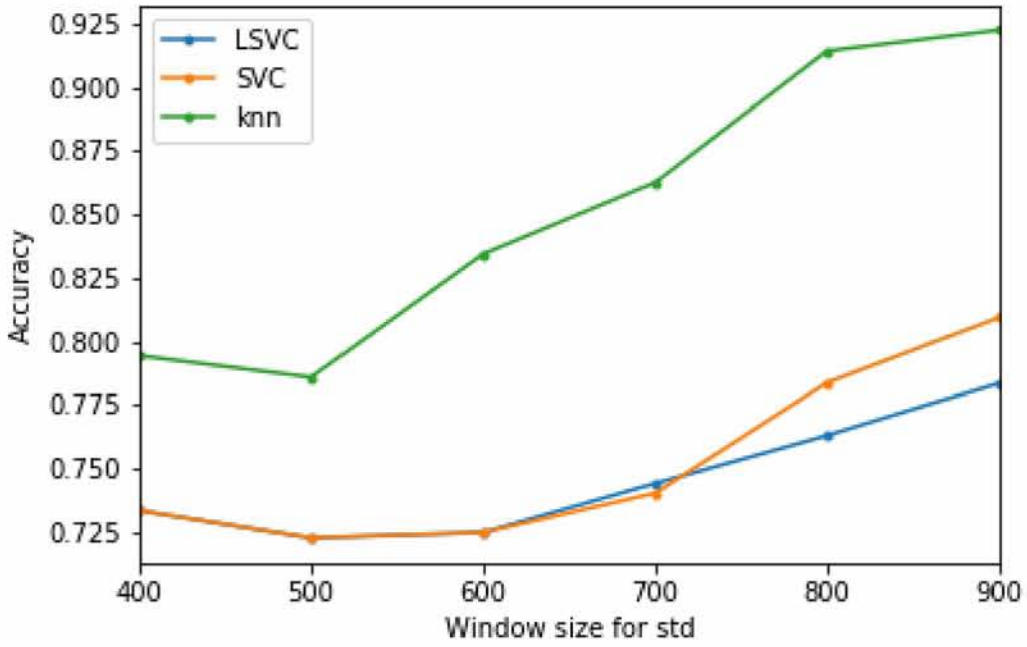
των παραμέτρων που δέχεται ο κάθε αλγόριθμος. Στον αλγόριθμο κοντινότερων γειτόνων μπορούμε να παραμετροποιήσουμε τον αριθμό τον γειτόνων που ελέγχει ο αλγόριθμος. Στη συγκεκριμένη μελέτη επιλέξαμε να δοκιμάσουμε από 1 μέχρι 9 γείτονες. Στον επόμενο αλγόριθμο (γραμμική μηχανή διανυσματικής στήριξης) επιλέγουμε τις τιμές για την παράμετρο C. Πιο συγκεκριμένα η παράμετρος έχει τις εξής τιμές $C \in \{0.1, 1, 10\}$. Τέλος, για τον αλγόριθμο μηχανής διανυσματικής στήριξης με πυρήνα rbf έχουμε τις εξής τιμές ανά ζεύγη$C \in \{0.1, 1, 10\}, \gamma \in \{0.1, 1\}$.

Τελικά, αποθηκεύουμε τα αποτελέσματα στο αρχείο`algo_results.csv` και φαίνονταιστονΠίνακας 1.

Αποτελέσματα προσομοιώσεων

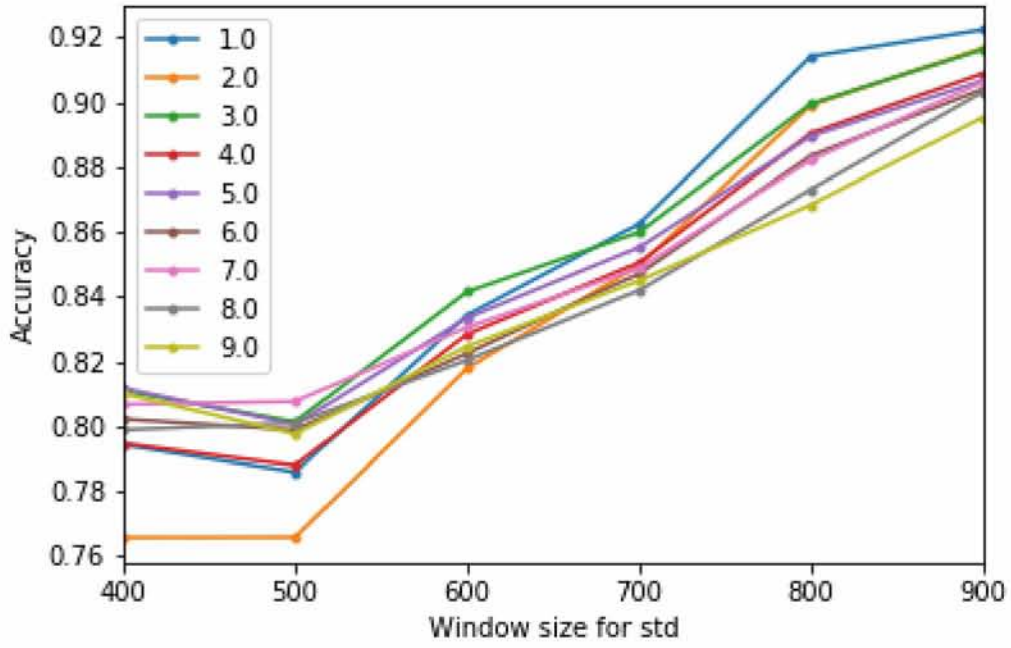
Για να αξιολογήσουμε την επίδοση των τριών αλγορίθμων που εξετάσαμε χρησιμοποιήσαμε την ακρίβεια, δηλαδή το ποσοστό επί τοις εκατό των σωστά αναγνωρίσιμων στιγμιότυπων. Στην Εικόνα 7 μπορούμε να δούμε τα συνολικά αποτελέσματα των τριών αλγορίθμων που δοκιμάσαμε στα δεδομένα που έχουμε προ αναφέρει. Ο οριζόντιος άξονας της γραφικής παράστασης συμβολίζει το μέγεθος του κυλιόμενου παραθύρου που χρησιμοποιήσαμε για τον υπολογισμό της τυπικής απόκλισης. Ο κάθετος άξονας είναι η ακρίβεια, δηλαδή από όλα τα δεδομένα πόσα μπόρεσε ο αλγόριθμος να ταξινομήσει σωστά.

Την καλύτερη επίδοση την έχει ο αλγόριθμος κοντινότερου γείτονα. Όπως μπορούμε να δούμε και στην Εικόνα 7 αλλά και στον Πίνακας 1 ο αλγόριθμος κοντινότερου γείτονα έχει μεγαλύτερη ακρίβεια με διαφορά της τάξης των πέντε ποσοστιαίων μονάδων. Αν κοιτάξουμε μόνο το μέγεθος παραθύρου 900 μπορούμε να δούμε ότι δεύτερος στην κατάταξη έρχεται ο αλγόριθμος μηχανής διανυσματικής στήριξης με μη-γραμμικό πυρήνα, και τέλος έχουμε τη μηχανή διανυσματικής στήριξης με γραμμικό πυρήνα.

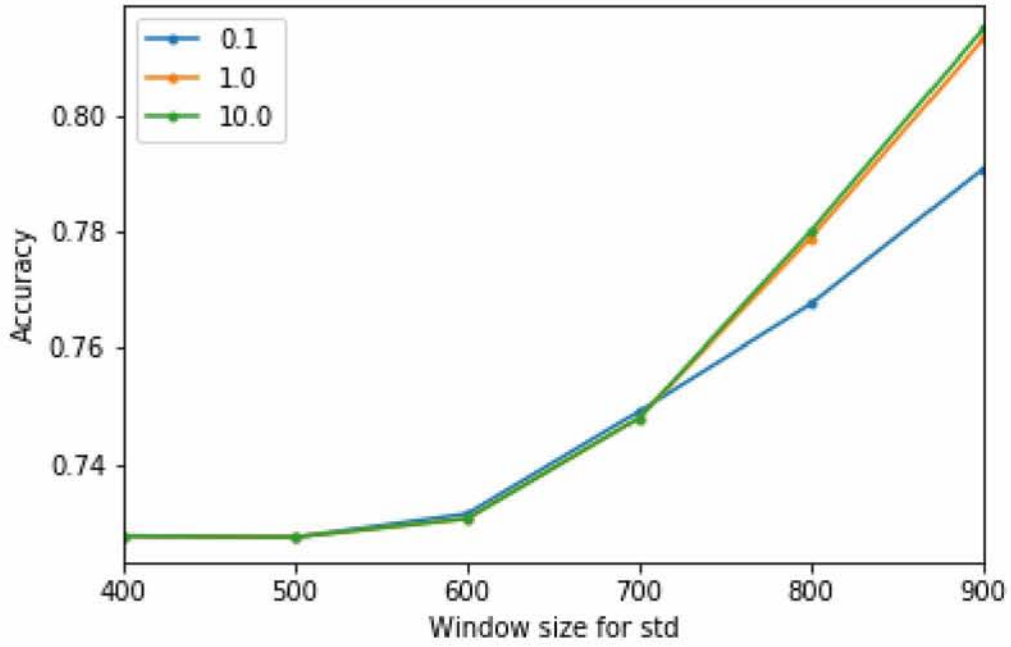


Εικόνα 7Τα συνολικά αποτελέσματα των αλγορίθμων

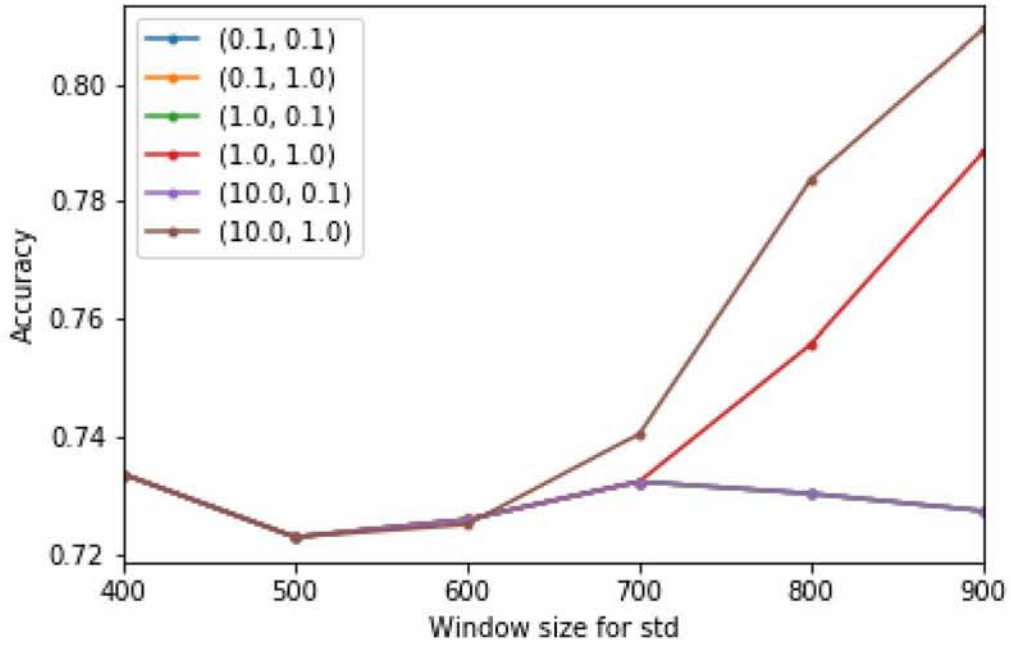
Επίσης, έχουμε δοκιμάσει τους αλγορίθμους μας για διαφορετικές τιμές των παραμέτρων τους.Στην Εικόνα 8 μπορούμε να δούμε τα αποτελέσματα για τον αλγόριθμο του κοντινότερου γείτονα. Όπως βλέπουμε η κάθε γραμμή αντιπροσωπεύει τα αποτελέσματα για μια τιμή του k. Συγκεκριμένα, έχουμε επιλέξει να εκτελέσουμε τον αλγόριθμο για τιμές από το 1 έως το 9. Όπως φαίνεται και στην εικόνα τα καλύτερα αποτελέσματα συνολικά τα πετυχαίνουμε για μέγεθος κυλιόμενου παραθύρου 900 και για $k = 1$. Αντίστοιχα, τα αποτελέσματα για τις δύο παραλλαγές των μηχανών διανυσματικής στήριξης φαίνονται στις εικόνες Εικόνα 9, Εικόνα 10.



Εικόνα 8Αλγόριθμος κοντινότερου γείτονα για τις διάφορες τιμές της παραμέτρου k



Εικόνα 9 Αλγόριθμος γραμμικής μηχανής διανυσματικής στήριξης για τις διάφορες τιμές της παραμέτρου C



Εικόνα 10Αλγόριθμος μηχανής διανυσματικής στήριξης με πυρήνα rbfγια τις διάφορες τιμές των παραμέτρων C και gamma

Στις εικόνες Εικόνα 9 και Εικόνα 10 φαίνεται ότι η καλύτερη επίδοση και των δύο αλγορίθμων επιτυγχάνεται με μέγεθος κυλιόμενου παραθύρου 900. Συγκεκριμένα, βλέπουμε ότι στη γραμμική μηχανή διανυσματικής στήριξης η καλύτερη επίδοση για το καλύτερο μέγεθος παραθύρου επιτυγχάνεται με παράμετρο $C = 10$, και η μηχανή διανυσματικής στήριξης με rbf πυρήνα πετυχαίνει καλύτερη επίδοση για τις παραμέτρους $C = 10$και $\gamma = 1$.

Συμπεράσματα και συζήτηση

Συμπεράσματα

Το πρόβλημα που είχαμε να λύσουμε ήταν να αξιολογήσουμε αλγορίθμους ταξινόμησης στο πόσο καλά μπορούν να εντοπίσουν το αν υπάρχει κάποιο άτομο εντός μιας οικίας. Πιο συγκεκριμένα, χρησιμοποιήσαμε ένα σύνολο δεδομένων το οποίο έχει χρησιμοποιηθεί και στο [1], και αξιολογήσαμε τρεις αλγορίθμους ταξινόμησης, τον αλγόριθμο κοντινότερου γείτονα, και δύο παραλλαγές των μηχανών διανυσματικής στήριξης, γραμμική και με πυρήνα rbf.

Τελικά, την καλύτερη επίδοση την είχε ο αλγόριθμος κοντινότερου γείτονα. Πιο συγκεκριμένα εκτελέσαμε τους αλγορίθμους για διάφορες τιμές των παραμέτρων τους και τους αξιολογήσαμε βάση της ακρίβειας. Ο αλγόριθμος κοντινότερου γείτονα πέτυχε τη μεγαλύτερη ακρίβεια για κάθε μέγεθος κυλιόμενου παραθύρου, μετά ήρθε ο αλγόριθμος μηχανής διανυσματικής στήριξης με πυρήνα rbf και τέλος η γραμμική μηχανή διανυσματικής στήριξης.

Ανοιχτάερωτήματα

Μέσα από την παρούσα πτυχιακή εργασία προκύπτουν διάφορα ενδιαφέροντα ανοιχτά ερωτήματα. Όπως είδαμε, η ακρίβεια των αλγορίθμων ήταν της τάξης του 90%. Δυστυχώς, με τα συγκεκριμένα δεδομένα μπορούμε να φτάσουμε μέχρι ένα συγκεκριμένο μέγεθος κυλιόμενου παραθύρου. Επομένως, ένα από τα ερωτήματα που προκύπτουν είναι τι μπορούμε να επιτύχουμε σε δεδομένα που ίσως έχουν μεγαλύτερη διάρκεια, ή ο ρυθμός δειγματοληψίας είναι μεγαλύτερος, ή και τα δύο. Ένα άλλο ερώτημα που μένει ανοιχτό είναι να γίνει αξιολόγηση σε περισσότερους αλγορίθμους ταξινόμησης, όπως και να εκτελεστεί η μηχανή διανυσματικής στήριξης με διαφορετικούς πυρήνες.

Επίσης, λόγο περιορισμού στο υπολογιστικό σύστημα που χρησιμοποιήσαμε, δεν μπορούσαμε να δοκιμάσουμε πάρα πολλές διαφορετικές τιμές των παραμέτρων για κάθε αλγόριθμο. Συνεπώς, κάποιο καλύτερο υπολογιστικό σύστημα να μας έδινε τη δυνατότητα να ερευνήσουμε τις παραμέτρους των αλγορίθμων ταξινόμησης σε πιο λεπτομερές επίπεδο.

Τέλος, θα μπορούσε να γίνει διαφορετική προεργασία στα δεδομένα, δίνοντάς τους μεγαλύτερη διάσταση, που πιθανός να επέτρεπε την ευκολότερη αναγνώριση κατειλημμένης κατοικίας. Ή να χρησιμοποιηθεί κάποιο σύνολο δεδομένων που περιέχει περισσότερες μεταβλητές εκ των προτέρων.

Βιβλιογραφία

[1] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy, "Non-Intrusive Occupancy Monitoring using Smart Meters," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings - BuildSys'13*, 2013, pp. 1–8.

[2] S. N. Patel, T. Robertson, J. A. Kientz, M. S. Reynolds, and G. D. Abowd, "At the flick of a switch: Detecting and classifying unique electrical events on the residential power line," in *9th international conference on Ubiquitous computing UbiComp07*, 2007.

[3] Y. Agarwal, B. Balaji, S. Dutta, R. K. Gupta, and T. Weng, "Duty-Cycling Buildings Aggressively : The Next Frontier in HVAC Control Categories and Subject Descriptors," in *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*, 2011.

[4] J. Chen, A. H. Kam, J. Zhang, N. Liu, and L. Shue, "Bathroom activity monitoring based on sound," in *International Conference on Pervasive Computing*, 2005, pp. 47–61.

[5] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, "Occupancy-driven energy management for smart building automation," in *BuildSys'10 - Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2010.

[6] V. L. Erickson *et al.*, "Energy efficient building environment control strategies using real-time occupancy measurements," in *BUILDSYS 2009 - Proceedings of the 1st ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, Held in Conjunction with ACM SenSys 2009*, 2009.

[7] V. L. Erickson and A. E. Cerpa, "Occupancy based demand response HVAC control strategy," in *BuildSys'10 - Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, 2010.

[8] X. Bian, G. D. Abowd, and J. M. Rehg, "Using sound source localization in a home environment," in *International Conference on Pervasive Computing*, 2005, pp. 19–36.

[9] A. Kamthe, L. Jiang, M. Dudys, and A. Cerpa, "Scopes: Smart cameras object position estimation system," in *European Conference on Wireless Sensor Networks*, 2009, pp. 279–295.

[10] G. R. Newsham and B. J. Birt, "Building-level occupancy data to improve ARIMA-based electricity use forecasts," in *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, 2010, pp. 13–18.

[11] S. Wang and X. Jin, "Co 2-based occupancy detection for on-line outdoor air flow control," *Indoor Built Environ.*, vol. 7, no. 3, pp. 165–181, 1998.

[12] J. Lu *et al.*, "The smart thermostat: using occupancy sensors to save energy in homes," in *Proceedings of the 8th ACM conference on embedded networked sensor systems*, 2010, pp. 211–224.

[13] E. M. Tapia, S. S. Intille, and K. Larson, "Activity recognition in the home using simple and ubiquitous sensors," in *International conference on pervasive computing*, 2004, pp. 158–175.

[14] A. Pekka *et al.*, "Definition of Smart Metering and Applications and Identification of Benefits," *Security*, 2008.

[15] J. C. P. E. Kester, M. J. G. E. I. Burgos, and J. B. Parsons, "Smart Metering Guide Energy Saving and the Customer Edition 2010," *Eur. Smart Metering Alliance*, 2011.

[16] U. Greveler, B. Justus, and D. Löhr, "Hintergrund und experimentelle Ergebnisse zum Thema Smart Meter und Datenschutz," *Fachhochschule Munst. Appl. Sci. Tech. Pap.*, 2011.

[17] L. Tien, "New smart meters for energy use put privacy at risk." 2010.

[18] S. Brinkhaus, D. Carluccio, U. Greveler, B. Justus, D. Löhr, and C. Wegener, "Smart hacking for privacy," in *Proceeding of the 28th Chaos Communication Congress (28C3)*, 2011.

[19] M. Enev, S. Gupta, T. Kohno, and S. N. Patel, "Televisions, video privacy, and powerline electromagnetic interference," in *Proceedings of the 18th ACM conference on Computer and communications security*, 2011, pp. 537–550.

[20] M. O. D. Real, "The Foundation for Information Policy Research," *UCLA Law Rev.*, no. 2010, 1701.

[21] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc., 2016.

[22] S. Theodoridis and K. Koutroumbas, *ΑναγνώρισηΠροτύπων*. Academic Press, 2009.

[23] S. Barker, S. Kalra, D. Irwin, and P. Shenoy, "Empirical characterization and modeling of electrical loads in smart homes," in *2013 International Green Computing Conference Proceedings, IGCC 2013*, 2013.

[24] S. Barker, S. Kalra, D. Irwin, and P. Shenoy, "Empirical characterization, modeling, and analysis of smart meter data," *IEEE J. Sel. Areas Commun.*, 2014.

Παράρτημα: Κώδικας Python

```
import pandas as pd
import itertools
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC, LinearSVC
```

Εικόνα 13Οι προγραμματιστικές βιβλιοθήκες που χρησιμοποιήσαμε.

```
def run_classifier(classifier_name):

    print(classifier_name, k, c_param, gamma_param)
    my_classifier.fit(x_train, y_train)


    acc_train = my_classifier.score(x_train, y_train)
    acc_test = my_classifier.score(x_test, y_test)


    results['algorithm'].append(classifier_name)
    results['k'].append(k)
    results['C'].append(c_param)
    results['gamma'].append(gamma_param)
    results['mean_win_s'].append(mean_win_s)
    results['std_win_s'].append(std_win_s)
    results['accuracy_train'].append(acc_train)
    results['accuracy_test'].append(acc_test)
```

Εικόνα 12 Η συνάρτηση με την οποία εκτελούμε τους αλγορίθμους ταξινόμησης και κατασκευάζουμε τη δομή δεδομένων για τα αποτελέσματα του κάθε αλγορίθμου.

```
results = {'algorithm': [],
           'k': [],
           'C': [],
           'gamma': [],
           'mean_win_s': [],
           'std_win_s': [],
           'accuracy_train': [],
           'accuracy_test': []}
```

Εικόνα 11 Αρχικοποίηση της δομής δεδομένων για τα αποτελέσματα.

```
for mean_win_s, std_win_s in zip(range(10, 61, 10), range(400, 901, 100)):
        df = pd.read_csv("./datasets/data_with_statistics_" + str(mean_win_s) + "_"
                         + str(std_win_s) + ".csv")

    print("Data types for each column")
    print(df.dtypes)
    print()

    df = df.dropna()  # diagrafw tis eggrafes pou exoun toulaxiston ena NaN
    # print(df)


    dataset = df.loc[:, ["energy","avg_energy","std_energy"]]


    x_train, x_test, y_train, y_test = train_test_split(
        dataset, df.occupancy.values, random_state=0
    )
```

Εικόνα 15 Κεντρική επανάληψη της προσομοίωσης όπου κατασκευάζουμε τα σετ δεδομένων εκπαίδευσης και δοκιμής.

```
### K NEAREST NEIGHBORS
    print("Running K nearest neighbors classification algorithm...")

    # arxikopoihsh twn asxetwn parametrwn
    gamma_param = None
    c_param = None


    for k in range(1, 10):
        my_classifier = KNeighborsClassifier(n_neighbors=k)
        run_classifier("knn")


    ### LINEAR SVM
    print("Running linear svc algorithm...")

    # arxikopoiw tis asxetes parametrous
    k = None
    gamma_param = None

    # edw einai san na exw to klasiko SVC() alla me parametro kernel='linear'
    for c_param in [0.1, 1, 10]:
        my_classifier = LinearSVC(C=c_param, max_iter=1000000)
        run_classifier("LSVC")
```

Εικόνα 14 Η εκτέλεση των πρώτων δύο αλγορίθμων (κοντινότερων γειτόνων και γραμμικής μηχανής διανυσμάτων στήριξης).

```
    ### SVM
    print("Running svc algorithm...")
    # arxikopoiw tis asxetes parametrous
    k = None

    for c_param, gamma_param in itertools.product([0.1, 1, 10], [0.1, 1]):
        my_classifier = SVC(kernel='rbf', C=c_param, gamma=gamma_param)
        run_classifier("SVC")

results_df = pd.DataFrame(results)
results_df.to_csv("algo_results.csv", index=False)
```

Εικόνα 16 Εκτέλεση του τρίτου αλγορίθμου (μηχανή διανυσματικής στήριξης με πυρήνα rbf).

```
ENERGY_F = "Home-A-spring-energy/Home_A_spring_energytrace.txt"
PERSON1_F = "Home-A-spring-occupancy/Home_A_spring_person_1.txt"
PERSON2_F = "Home-A-spring-occupancy/Home_A_spring_person_2.txt"
OUT_F = "final_data.csv"

with open(ENERGY_F) as energy_f, open(PERSON1_F) as p1_f, \
     open(PERSON2_F) as p2_f, open(OUT_F, "w") as out_f:

    # set headers for out file
    out_f.write("energy,time,occupancy\n")

    # get person 1 and person 2 correct line occupancy
    p1_line = p1_f.readline()
    p2_line = p2_f.readline()

    while (not p1_line.startswith("15")) and (not p2_line.startswith("15")):
        p1_line = p1_f.readline()
        p2_line = p2_f.readline()

    line = energy_f.readline()
    while line:
        line = line.strip()
        line = line.split('\t')
        # print(line)

        timestamp = line[1]
        energy = line[2]

        # print("Time: {}, energy: {}".format(timestamp, energy))

        if p1_line.endswith("1\n") or p2_line.endswith("1\n"):
            out_f.write(energy + "," + timestamp + ",1\n")
        else:
            out_f.write(energy + "," + timestamp + ",0\n")

        line = energy_f.readline()
        p1_line = p1_f.readline()
        p2_line = p2_f.readline()
```

Εικόνα 17Με αυτόν τον κώδικα υλοποιούμε τη συγχώνευση των τριών αρχείων δεδομένων σε ένα. Ουσιαστικά κατασκευάζουμε ένα αρχείο γραμμή προς γραμμή, όπου κάθε γραμμή περιέχει την ενέργεια, τον χρόνο που έγινε η μέτρηση, και το αν βρισκόταν κάποιο άτομο στο σπίτι

```
import pandas as pd

with open("final_data.csv") as data_file:
    df = pd.read_csv(data_file)


    for mean_window_size, std_window_size in zip(range(10, 61, 10),
                                                 range(400, 901, 100)):

        new_df = pd.DataFrame({"energy": df["energy"], "occupancy": df[
            "occupancy"]})

        avg_energy = new_df["energy"].rolling(mean_window_size).mean()

        new_df["avg_energy"] = avg_energy


        std_energy = new_df["energy"].rolling(std_window_size).std()


        new_df["std_energy"] = std_energy


        print(new_df.dtypes)
        print(new_df)


        new_df.to_csv("./datasets/data_with_statistics_" + str(
            mean_window_size) + "_" + str(std_window_size) + ".csv",
                      index=False)
```

Εικόνα 18 Με αυτήν την υλοποίηση κατασκευάζουμε όλα τα σύνολα δεδομένων που θα χρησιμοποιηθούν στους αλγορίθμους ταξινόμησης

Παράρτημα: Αρχείοαποτελεσμάτων

algorithm	k	C	gamma	mean_win_s	std_win_s	accuracy_train	accuracy_test
knn	1.0			10	400	1.0	0.7943731899048407
knn	2.0			10	400	0.9008275862068965	0.7658254033926355
knn	3.0			10	400	0.8987586206896552	0.8109226313611916
knn	4.0			10	400	0.8704827586206897	0.7947869259412494
knn	5.0			10	400	0.8739310344827587	0.8117501034340091
knn	6.0			10	400	0.8557241379310345	0.8022341745966074
knn	7.0			10	400	0.8551724137931035	0.8067852709971038
knn	8.0			10	400	0.8445517241379310	0.7989242863053372
knn	9.0			10	400	0.8437241379310345	0.8100951592883741
LSVC		0.1		10	400	0.7275862068965517	0.7335539925527513
LSVC		1.0		10	400	0.7275862068965517	0.7335539925527513
LSVC		10.0		10	400	0.7275862068965517	0.7335539925527513
SVC		0.1	0.1	10	400	0.7275862068965517	0.7335539925527513
SVC		0.1	1.0	10	400	0.7275862068965517	0.7335539925527513
SVC		1.0	0.1	10	400	0.7275862068965517	0.7335539925527513
SVC		1.0	1.0	10	400	0.7275862068965517	0.7335539925527513
SVC		10.0	0.1	10	400	0.7275862068965517	0.7335539925527513
SVC		10.0	1.0	10	400	0.7275862068965517	0.7335539925527513
knn	1.0			20	500	1.0	0.7859531772575251
knn	2.0			20	500	0.8949128919860627	0.7658862876254181
knn	3.0			20	500	0.8977003484320557	0.8014214046822743
knn	4.0			20	500	0.8708013937282223	0.7880434782608695
knn	5.0			20	500	0.8741463414634146	0.8005852842809364
knn	6.0			20	500	0.8596515679442509	0.7989130434782609
knn	7.0			20	500	0.8577003484320558	0.8076923076923077
knn	8.0			20	500	0.8500348432055749	0.8010033444816054
knn	9.			20	500	0.8489198606	0.7976588628

	0					271777	762542
LSVC		0.1		20	500	0.7273867595818815	0.7228260869565217
LSVC		1.0		20	500	0.7273867595818815	0.7228260869565217
LSVC		10.0		20	500	0.7273867595818815	0.7228260869565217
SVC		0.1	0.1	20	500	0.7273867595818815	0.7228260869565217
SVC		0.1	1.0	20	500	0.7273867595818815	0.7228260869565217
SVC		1.0	0.1	20	500	0.7273867595818815	0.7228260869565217
SVC		1.0	1.0	20	500	0.7273867595818815	0.7228260869565217
SVC		10.0	0.1	20	500	0.7273867595818815	0.7228260869565217
SVC		10.0	1.0	20	500	0.7273867595818815	0.7228260869565217
knn	1.0			30	600	1.0	0.8343895226024504
knn	2.0			30	600	0.9145070422535211	0.8179129700042248
knn	3.0			30	600	0.9146478873239436	0.8415716096324461
knn	4.0			30	600	0.8974647887323943	0.828474862695395
knn	5.0			30	600	0.8942253521126761	0.8335445711871567
knn	6.0			30	600	0.8795774647887324	0.8225602027883396
knn	7.0			30	600	0.8787323943661972	0.830587241233629
knn	8.0			30	600	0.8680281690140845	0.8204478242501057
knn	9.0			30	600	0.8646478873239437	0.8246725813265737
LSVC		0.1		30	600	0.7314084507042253	0.7249683143219265
LSVC		1.0		30	600	0.7305633802816901	0.7249683143219265
LSVC		10.0		30	600	0.7305633802816901	0.7249683143219265
SVC		0.1	0.1	30	600	0.7325352112676057	0.7258132657372202
SVC		0.1	1.0	30	600	0.7325352112676057	0.7258132657372202
SVC		1.0	0.1	30	600	0.7325352112676057	0.7258132657372202
SVC		1.0	1.0	30	600	0.7325352112676057	0.7258132657372202
SVC		10.0	0.1	30	600	0.7325352112676057	0.7258132657372202
SVC		10.0	1.0	30	600	0.7319718309859155	0.7249683143219265
knn	1.0			40	700	1.0	0.8625106746370623

knn	2.0			40	700	0.9204270462633451	0.8497011101622545
knn	3.0			40	700	0.9232740213523132	0.8599487617421008
knn	4.0			40	700	0.9003558718 86121	0.8505550811272417
knn	5.0			40	700	0.9010676156 58363	0.8552519214346712
knn	6.0			40	700	0.8896797153024911	0.8471391972672929
knn	7.0			40	700	0.8875444839857651	0.8488471391972673
knn	8.0			40	700	0.8787188612099645	0.8420153714773698
knn	9.0			40	700	0.8727402135231317	0.8450042698548249
LSVC		0.1		40	700	0.7490391459074733	0.7425277540563621
LSVC		1.0		40	700	0.7479003558718861	0.7450896669513236
LSVC		10.0		40	700	0.7479003558718861	0.7442356959863364
SVC		0.1	0.1	40	700	0.7407829181494662	0.7322801024765158
SVC		0.1	1.0	40	700	0.7407829181494662	0.7322801024765158
SVC		1.0	0.1	40	700	0.7407829181494662	0.7322801024765158
SVC		1.0	1.0	40	700	0.7407829181494662	0.7322801024765158
SVC		10.0	0.1	40	700	0.7407829181494662	0.7322801024765158
SVC		10.0	1.0	40	700	0.7471886120996442	0.740392826643894
knn	1.0			50	800	1.0	0.9141130772550712
knn	2.0			50	800	0.9533812949640288	0.8990073370738023
knn	3.0			50	800	0.9522302158273381	0.8994389296504101
knn	4.0			50	800	0.9378417266 18705	0.8903754855416487
knn	5.0			50	800	0.9302158273381295	0.8895123003884333
knn	6.0			50	800	0.9217266187050359	0.8834700043159258
knn	7.0			50	800	0.9138129496402878	0.8821752265861027
knn	8.0			50	800	0.9096402877697841	0.8731117824773413
knn	9.0			50	800	0.9054676258992805	0.8683642641346568
LSVC		0.1		50	800	0.7674820143884892	0.7505394907207596
LSVC		1.0		50	800	0.7788489208633094	0.7630556754423824
LSVC		10		50	800	0.78	0.7630556754

		.0					423824
SVC		0.1	0.1	50	800	0.7520863309352518	0.7302546396201985
SVC		0.1	1.0	50	800	0.7520863309352518	0.7302546396201985
SVC		1.0	0.1	50	800	0.7520863309352518	0.7302546396201985
SVC		1.0	1.0	50	800	0.7735251798561151	0.7557186016400518
SVC		10.0	0.1	50	800	0.7520863309352518	0.7302546396201985
SVC		10.0	1.0	50	800	0.7922302158273381	0.7837721191195511
knn	1.0			60	900	1.0	0.9223385689354275
knn	2.0			60	900	0.9675636363636364	0.9166666666666666
knn	3.0			60	900	0.9666909090909090	0.9162303664921466
knn	4.0			60	900	0.9572363636363637	0.9088132635253054
knn	5.0			60	900	0.9534545454545454	0.9066317626527051
knn	6.0			60	900	0.9479272727272727	0.9040139616055847
knn	7.0			60	900	0.9413818181818182	0.9061954624781849
knn	8.0			60	900	0.9368727272727273	0.9031413612565445
knn	9.0			60	900	0.9310545454545455	0.8952879581151832
LSVC		0.1		60	900	0.7906909090909091	0.7539267015706806
LSVC		1.0		60	900	0.8129454545454545	0.7805410122164049
LSVC		10.0		60	900	0.8146909090909091	0.7835951134380453
SVC		0.1	0.1	60	900	0.7639272727272727	0.7273123909249564
SVC		0.1	1.0	60	900	0.7639272727272727	0.7273123909249564
SVC		1.0	0.1	60	900	0.7639272727272727	0.7273123909249564
SVC		1.0	1.0	60	900	0.8132363636363636	0.7883944153577661
SVC		10.0	0.1	60	900	0.7639272727272727	0.7273123909249564
SVC		10.0	1.0	60	900	0.8315636363636364	0.8093368237347295

Πίνακας 1 Αποτελέσματα εκτέλεσης τριών αλγορίθμων ταξινόμησης

Παράρτημα: Σύνολο δεδομένων

Στο παρόν παράρτημα παρουσιάζονται τα δεδομένα με όλους τους συνδυασμούς μεγέθους κυλιόμενου παραθύρου που έχουμε παράγει. Λόγο χωρικού περιορισμού θα παρουσιάσουμε μόνο μέρος του κάθε αρχείου.

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 10 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 400

energy	occupancy	avg_energy	std_energy
0.7897333329999999	1	1.011968333400001	0.39873107099411
0.83145	1	0.9742866667000009	0.39830100914741695
0.836683333	1	0.9346083333000008	0.39796661569696684
0.842283333	1	0.8980199999000007	0.39767866744745733
0.809316667	1	0.8567283333000008	0.3973659487924234
0.954633333	1	0.8326599999000008	0.39719169302094887
1.0429	1	0.8153333332000008	0.397143076791542
0.720166667	1	0.7864066666000007	0.3966632317657811
0.5767	1	0.7920899999000006	0.39639484481466775
0.5748833329999999	1	0.7978749999000008	0.39545270352722955
1.7255666669999998	1	0.8914583333000008	0.39791144525622757
2.1751	1	1.0258233333000009	0.4039378034018035
2.206783333	1	1.162833333300001	0.4099910720970858
2.2022	1	1.298825000000001	0.41595853280902045
1.120516667	1	1.329945000000001	0.4157986382356616
0.695583333	1	1.304040000000001	0.41523415886272486
0.5679833329999999	1	1.256548333300001	0.4147357265796459
0.581516667	1	1.2426833333000011	0.4139623198206425
0.5730333329999999	1	1.2423166666000012	0.4131967563486581
0.58325	1	1.2431533333000009	0.41240697600567877

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 20 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 500

energy	occupancy	avg_energy	std_energy
1.341666667	1	0.49808833339999836	0.4500119639925646
1.221933333	1	0.5327925000499982	0.4500081656635248
1.167366667	1	0.5680316667499983	0.4499762368827842
1.15355	1	0.6026850000999984	0.4499537973894389
1.11505	1	0.6354775000999984	0.4498659337191125
1.131483333	1	0.6690233333999984	0.4497970406543182
1.211566667	1	0.7065533333999985	0.44984336013515785
1.20985	1	0.7442125000499984	0.4497862212168019
1.215366667	1	0.7824000000499984	0.45003016612230873
1.235566667	1	0.8216083333999984	0.44990780602200103
1.220233333	1	0.8598608333999984	0.44995478860606253
1.219116667	1	0.8981375000999984	0.4501020852170156
1.198	1	0.9353866667499984	0.4501151870081692
1.18005	1	0.9781691667499984	0.45013422769025774

1.18445	1	1.0223141667499984	0.4501348122838145
1.1863	1	1.0665633333999984	0.4501648922256385
1.1781	1	1.1103525000499985	0.4501758753055763
1.186583333	1	1.1513116666999985	0.4500364404871166
1.186933333	1	1.1892824999999987	0.44990301857406545
1.17415	1	1.1958658333499987	0.44974310176573323

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 30 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 600

energy	occupancy	avg_energy	std_energy
0.7032666670000001	0	1.3255511111999991	0.5305609111782589
0.70165	0	1.268867777866666	0.5301946501931484
0.6978	0	1.2113972222999994	0.5298922760347944
0.750933333	0	1.156376111166666	0.5296009783956546
0.7737166670000001	0	1.101526111166666	0.5293030419947145
0.733066667	0	1.045326111166666	0.5290142827441995
0.7299	0	0.9875494444999994	0.5287123963626101
0.851883333	0	0.9317405555999994	0.5283160600885995
0.8411333329999999	0	0.8757227777999994	0.5282415903571778
0.8402166670000001	0	0.8490438889333327	0.5279158513565155
0.83265	0	0.8333700000333327	0.527715349682148
0.8328	0	0.819642222266666	0.5275775976087623
0.82965	0	0.8263105555999994	0.5273805412336683
0.8290000000000001	0	0.8327466666999995	0.5272064141106241
0.828016667	0	0.8380327778333327	0.5270152625001159
0.8269833329999999	0	0.8433222222666662	0.5268398848929128
0.8288	0	0.8480450000333328	0.5266607798369514
0.829483333	0	0.8472038888999994	0.5263755234784582
0.834916667	0	0.8466022222333328	0.526092653395416
0.626583333	0	0.8389472222333328	0.5258735242489638

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 40 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 700

energy	occupancy	avg_energy	std_energy
1.05625	0	0.7560812499999989	0.5485213819863969
1.0442833329999999	0	0.7725829166499989	0.5482137485761758
1.017433333	0	0.7864891666499989	0.5479465174946568
0.940866667	0	0.8004341666499988	0.5476801555646694
0.714166667	0	0.8071633333249988	0.5474308949766868
0.276333333	0	0.7993958333249986	0.5475965878543017
0.274766667	0	0.7938266666749987	0.5477529534047852
0.274316667	0	0.7882970833499987	0.5478458255826302
0.35283333299999997	0	0.7847266666749987	0.548126969523737
0.42795	0	0.7807987499999987	0.5481180759566677
0.43498333299999997	0	0.7792954166499987	0.5481986907237941

0.45756666700000004	0	0.7783770833249988	0.5483029084674671
0.41616666700000005	0	0.7750712499999988	0.5484041790675395
0.41715	0	0.7726049999999988	0.5485207233081625
0.392316667	0	0.7694833333499987	0.5486511153645901
0.407416667	0	0.7666670833499987	0.5487744406324072
0.405	0	0.7639004166749988	0.5488959908161145
0.4311	0	0.7617441666749987	0.5489024089412092
0.40856666700000005	0	0.7588525000249986	0.5489341402115354
0.40516666700000004	0	0.7559508333749986	0.5489649266601244

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 50 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 800

energy	occupancy	avg_energy	std_energy
0.436516667	0	0.7361173333399996	0.5316381915271998
0.414516667	0	0.7203023333399997	0.5315537646679148
0.39756666700000004	0	0.7041013333399997	0.5315325154964355
0.40828333299999997	0	0.6877536666599997	0.5315230308150993
0.412816667	0	0.6718843333399996	0.5315088163827921
0.4088	0	0.6562693333399997	0.5314979086276179
1.0835333329999999	0	0.6540823333399998	0.5313459348840446
1.36015	0	0.6576143333399997	0.5313931178661743
1.2623166670000001	0	0.6592126666799998	0.5315521178341781
1.18565	0	0.6573070000199998	0.5314482880847089
1.249666667	0	0.6566193333599998	0.5314987421119278
1.269333333	0	0.6561926666799996	0.5316173368743704
1.239183333	0	0.6553999999999996	0.5316611602169475
1.2338	0	0.6546536666599997	0.5317177352030431
1.222366667	0	0.6531173333399997	0.5317512481982553
1.231883333	0	0.6519419999999997	0.5318060347902852
1.2159166670000001	0	0.6642246666799997	0.5318437871263916
1.211166667	0	0.6769693333599998	0.5317983144892611
1.2142	0	0.6897613333599998	0.5317582001975034
1.220033333	0	0.7027960000199998	0.5317213720969973

Δεδομένα με μέσο όρο κυλιόμενου παραθύρου μεγέθους 60 και τυπικής απόκλισης κυλιόμενου παραθύρου μεγέθους 900

energy	occupancy	avg_energy	std_energy
0.498616667	0	0.7909180555833294	0.5182718253280851
0.40785	0	0.7905516666999961	0.5182010352193822
0.406816667	0	0.7904850000333296	0.5181732900720413
0.493883333	0	0.7919527777999961	0.5180955929048999
0.40901666700000006	0	0.7915861111333294	0.5180860663433112
0.40578333299999997	0	0.7918472222333295	0.5180788694792173
0.48631666700000004	0	0.7929191666833295	0.5179968794582002
0.40095	0	0.7930422222333295	0.5179357899433487
0.39606666700000004	0	0.7927663888999962	0.5180780826942312

0.48945	0	0.7942613888999961	0.5179651262325046
0.40835	0	0.7934138888999961	0.5180048220801999
0.506266667	0	0.7954216666833294	0.5180109126314878
0.584816667	0	0.797412777799996	0.5179275339465367
0.49981666700000005	0	0.7990166666999959	0.5179111694538706
0.5016166670000001	0	0.8006575000333293	0.5178809157258889
0.584766667	0	0.8035452778166625	0.5178115551769669
0.5004	0	0.801629166699996	0.5177896147930456
0.50055	0	0.7972797222499959	0.5176922647391147
0.583283333	0	0.7939066666833291	0.5175467715593178
0.5136666670000001	0	0.7889972222499959	0.517438615432556