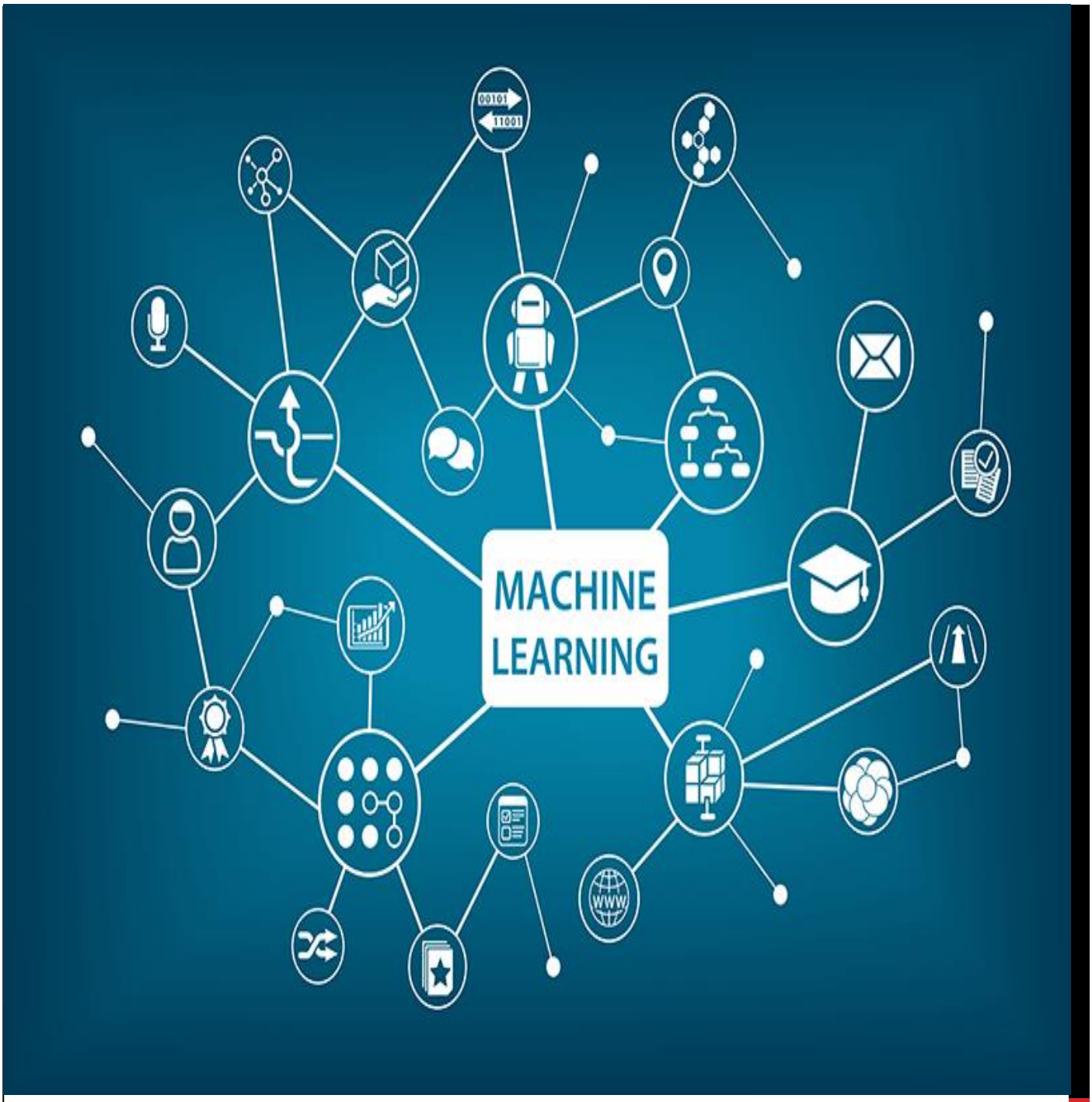


ΑΝΑΓΝΩΡΙΣΗ ΧΕΙΡΟΓΡΑΦΩΝ ΨΗΦΙΩΝ ΜΕ ΜΕΘΟΔΟΥΣ CLASSIFICATION

Στο πλαίσιο του μαθήματος περιβάλλοντα επίλυσης προβλημάτων και με υπεύθυνοκαθηγητή κ. Ηλία Χούστη χρησιμοποιήσαμε την βάση δεδομένων MNISTγια την εκπαίδευση και την σύγκριση της ταχύτητας και της ακρίβειας ορισμένων μεθόδων classification στον τομέα της μηχανικής μάθησης.

Μέλη Ομαδας:

- Κατερίνα Κανελλοπούλου

Αεμ : 2074

- Νείλος Ψαθάς

Αεμ: 1195

- Σεβαστή Υφαντή

Αεμ: 1994

ΕΙΣΑΓΩΓΗ

Η αυτόματη αναγνώριση των ανθρώπινων χειρόγραφων ψηφίων αποτελεί έναν ιδιαίτερο τομέα του machinelearningκαθώςμπορεί να βοηθήσει στην ανάγνωση τυπωμένου ή χειρόγραφου κειμένου απ΄τον υπολογιστή και να τον μετατρέψει σε ηλεκτρονικό κείμενο.

Η αναγνώριση ψηφίων είναι δηλαδή μια μέθοδος ψηφιοποίησης (digitizing) έντυπων κειμένων έτσι ώστε να μπορούν να προσπελαστούν ηλεκτρονικά, να αποθηκευτούν σε μικρότερο χώρο, να είναι προσβάσιμα από τον παγκόσμιο ιστό και να μπορούν να χρησιμοποιηθούν περεταίρω σε διαδικασίες όπως η εξόρυξη δεδομένων (datamining).Η οπτική αναγνώριση ψηφίων αποτελεί πεδίο έρευνας για τους τομείς της αναγνώρισης προτύπων (patternrecognition), της τεχνητής νοημοσύνης (artificialintelligence) και της κατανόησης των εικόνων από τον ηλεκτρονικό υπολογιστή (computervision).

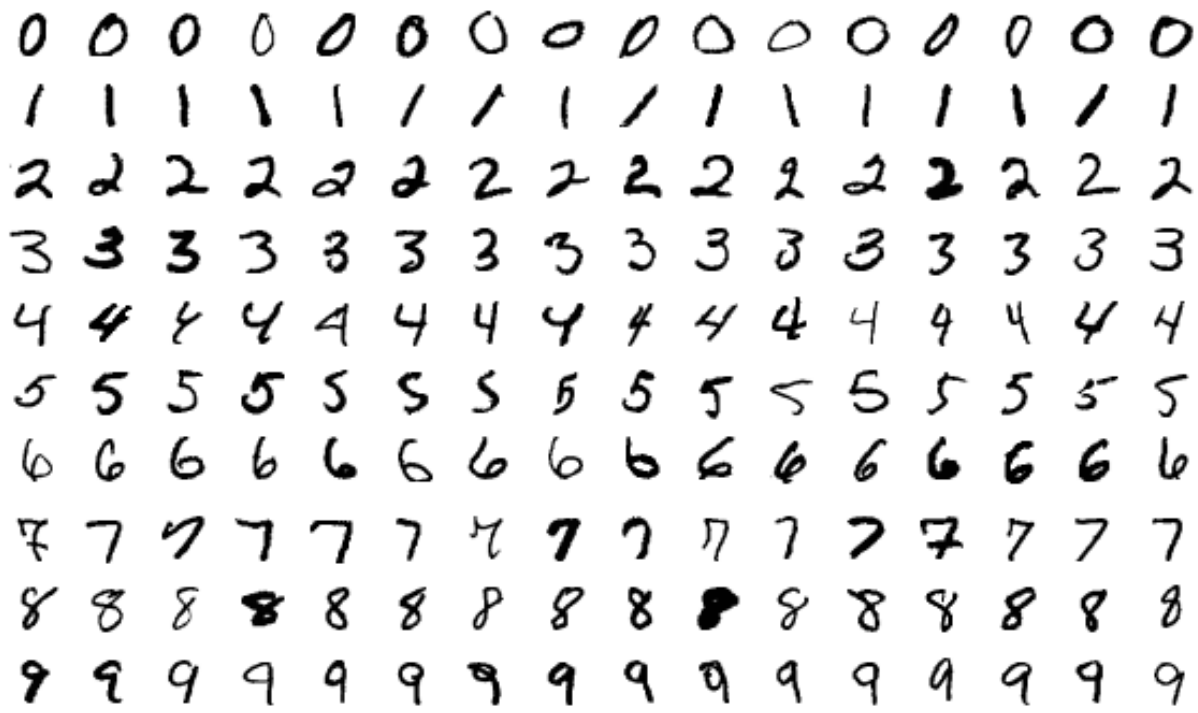
Υπάρχουν μερικές γνωστές βάσεις δεδομένων για την εξέταση της ακρίβειας και της απόδοσης των αλγορίθμων που προτείνονται από τους ερευνητές και τους επαγγελματίες της μηχανικής μάθησης.Μία από αυτές τις βάσεις δεδομένων με αναγνώριση χειρόγραφων ψηφίων είναι το MNIST.

MNISTDATASET

Μερικές πληροφορίες:

Η βάση δεδομένων MNIST (τροποποιημένη βάση δεδομένων Εθνικού Ινστιτούτου Προτύπων και Τεχνολογίας) είναι μια μεγάλη βάση δεδομένων με χειρόγραφα ψηφία που χρησιμοποιείται συνήθως για την εκπαίδευση διαφόρων συστημάτων επεξεργασίας εικόνας. Η βάση δεδομένων χρησιμοποιείται επίσης ευρέως για την εκπαίδευση και τις δοκιμές στον τομέα της μηχανικής μάθησης. Δημιουργήθηκε με την "ανάμιξη" των δειγμάτων από τα αρχικά σύνολα δεδομένων της NIST. Οι δημιουργοί θεώρησαν ότι δεδομένου ότι το σύνολο δεδομένων κατάρτισης του NIST λήφθηκε από τους υπαλλήλους του American CensusBureau, ενώ το σύνολο δεδομένων δοκιμών λήφθηκε από Αμερικανούς μαθητές γυμνασίου, δεν ήταν κατάλληλο για πειράματα μηχανικής μάθησης. Επιπλέον, οι ασπρόμαυρες εικόνες από το NIST ομαλοποιήθηκαν ώστε να ταιριάζουν σε ένα κιβώτιο οριοθέτησης 28x28 εικονοστοιχείων και αποτελούνταν από διαβαθμίσεις του γκρι.

Η βάση δεδομένων MNIST περιέχει 60.000 εικόνες εκπαίδευσης και 10.000 εικόνες δοκιμών. Το ήμισυ του σετ εκπαίδευσης και το ήμισυ του δοκιμαστικού σετ ελήφθησαν από το σύνολο δεδομένων κατάρτισης της NIST, ενώ το άλλο μισό του σετ εκπαίδευσης και το άλλο μισό του σετ δοκιμών λήφθηκαν από το σύνολο δεδομένων δοκιμών της NIST. Έχουν υπάρξει αρκετές επιστημονικές εργασίες σχετικά με τις προσπάθειες επίτευξης του χαμηλότερου ποσοστού σφάλματος. Οι αρχικοί δημιουργοί της βάσης δεδομένων διατηρούν μια λίστα με μερικές από τις μεθόδους που δοκιμάζονται σε αυτήν. Στο πρωτότυπο έγγραφο τους, χρησιμοποιούν μια μηχανή φορέα υποστήριξης για να επιτύχουν ποσοστό σφάλματος 0,8%.Ένα εκτεταμένο σύνολο δεδομένων παρόμοιο με το MNIST που ονομάζεται EMNIST έχει εκδοθεί το 2017, το οποίο περιέχει 240.000 εκπαιδευτικές εικόνες και 40.000 εικόνες δοκιμών χειρόγραφων ψηφίων.



ΣΤΟΧΟΣ

Θα εφαρμόσουμε ένα πλήθος classification μεθόδων ώστε να βρούμε την πιο κατάλληλη μέθοδο για μια OCR (optical character recognition) εφαρμογή βασισμένη σε αυτό το dataset. Χρήση της βιβλιοθήκης sklearn για να ελέγξουμε 3 μεθόδους classification (decision tree classifier, random tree forest, logistic regression και c-support vector classification) Επιπλέον, χρήση tensorflow πακέτου για την εφαρμογή της μεθόδου softmax regression.

Η επιλογή θα γίνει βάση της ακρίβειας των προβλέψεων και του χρόνου εκτέλεσης τους. Ως μέτρο ακρίβειας θα λάβουμε υπόψη τον confusion matrix Πρόκειται για έναν πίνακα που χρησιμοποιείται για την εύρεση της ακρίβειας σε μεθόδους classification όταν οι αληθινές τιμές των test values είναι γνωστές. Δείχνει τι πιθανότητα υπάρχει να καταταγεί το κάθε στοιχείο σε κάθε κλάση. Ενώ για την χρονική μέτρηση θα γίνει χρήση του πακέτου datetime.

ΜΕΘΟΔΟΙCLASSIFICATION

Αν θεωρήσουμε τα δεδομένα μας σαν ένα πίνακα ακεραίων (με προαναφερθείσα σημασία κάθε στοιχείου να είναι η φωτεινότητα του συγκεκριμένου pixel), μπορούμε να εφαρμόσουμε αρκετές classification μεθόδους ώστε να μάθουμε τη μηχανή μας να αναγνωρίζει κάποια επόμενα χειρόγραφα ψηφία βάζοντας τα στη σωστή κλάση. Το πλήθος των κλάσεων θα είναι 10 δηλαδή όσα και τα ψηφία 0,1,2,3,4,5,6,7,8,9. Χρησιμοποιήθηκαν κυρίως γνωστές από το μάθημα μέθοδοι αλλά και κάποιες γενικότερα διαδεδομένες ,υλοποιημένες σε βιβλιοθήκη της python, για να υλοποιήσουμε τον ερευνητικό μας σκοπό.

1) DECISION TREE CLASSIFIER

Μέθοδος της sklearn, ορίζουμε ως παράμετρο το maxdepth , καθορίζεται από το giniindex. Δημιουργεί ένα δέντρο όπου χρησιμοποιεί τις παρατηρήσεις ως κλαδιά διαχωρισμού για να φτάσει σε συμπέρασμα όπως η κλάση του αντικειμένου στα φύλλα του, στο maxdepth. Gini impurity είναι ο όρος που χρησιμοποιεί η μέθοδος για να δει πόσο αποτελεσματικό είναι το δέντρο που έφτιαξε. Συγκεκριμένα μετράει πόσο συχνά ένας τυχαία διαλεγμένος αριθμός θα πάρει λάθος label. Υπολογίζεται από το άθροισμα της πιθανότητας $p_i p_i$ ενός αντικειμένου με label i να επιλεχτεί επί την πιθανότητα

$$\sum_{k \neq i} p_k = 1 - p_i$$

ενός λάθους στην κατηγοριοποίηση αυτού του στοιχείου. Θέλουμε να τείνει στο 0 δηλαδή να μην υπάρχει λάθος στην κατηγοριοποίηση κανενός.

2) RANDOM FORESTS

Πρόκειται για μια μέθοδο συγχώνευσης, όπου δημιουργείται ένα συγχωνευμένο πλήθος από classification decision trees το οποίο έχει αντίστοιχο αποτέλεσμα με έκαστο tree.Ο αλγόριθμος τρέχει σύμφωνα με το Treebagging: Δοσμένου ενός training set X = x1, ..., xn με Y = y1, ..., yn αποτελέσματα, ο bagging αλγόριθμος διαλέγει επαναλαμβανόμενα ένα τυχαίο δείγμα απο το the trainingset και εφαρμόζει πανω σε αυτό ένα μοντέλο classification tree, έτσι έχω διαφορετικά fitted μοντέλα από διαφορετικό training set.Για το αποτέλεσμα κάθε πρόβλεψης ο συμψηφίζονται τα αποτελέσματα όλων των δέντρων που χρησιμοποιήθηκαν.

3) LOGISTIC REGRESSION

Συνήθως η logistic regression μέθοδος προβλέπει τη πιθανότητα μια εξαρτημένη μεταβλητή y να πάρει την τιμή 1 ή 0.Σε ένα πρόβλημα κατηγοριοποίησης με πολλές κλάσεις όπως το δικό μας, θέλουμε ένα μοντέλο που να προβλέπει τις πιθανότητες για κάθε διαφορετικό πιθανό αποτέλεσμα του εξαρτημένου y σε σχέση με τις ανεξάρτητες μεταβλητές μας. Η Multinomial logistic regression μέθοδος είναι μια λύση στο πρόβλημα της κατηγοριοποίησης που υποθέτει ότι ο γραμμικός συνδυασμός των χαρακτηριστικών που έχω παρατηρήσει (στην περίπτωσή μας η φωτεινότητα των pixel) ,και γενικότερα κάποιων παραμέτρων σχετικών με το πρόβλημα ,μπορεί να χρησιμοποιηθεί για να αποφασίσουμε την πιθανότητα το y (η εξαρτημένη μεταβλητή μας) να ανήκει σε κάθε μια κατηγορία ξεχωριστά. Για τυχαία κλάση k θα χρησιμοποιήσω την γραμμική συνάρτηση πρόβλεψης

$$f(k,i)=\beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \cdots + \beta_{M,k}x_{M,i},$$

ώστε να δω τη πιθανότητα το στοιχείο μου να ανήκει σ'αυτήν την κλάση .Με τους ίδιους συντελεστές $\beta_{m,k}$ θα υπολογίσω τις πιθανότητες για τις υπόλοιπες κλάσεις .Οι συντελεστές είναι παράγωγο των δοσμένων χαρακτηριστικών.

4)SUPPORT VECTOR CLASSIFICATION

Πρόκειται για μια μέθοδο με πρωταρχική χρήση την δυαδική κατηγοριοποίηση αλλά μπορούμε να τη χρησιμοποιήσουμε για πολλές κλάσεις μέσω της svm.SVC μεθόδου της βιβλιοθήκης sklearn της γλώσσας python. Η svm μέθοδος στις δυο διαστάσεις βρίσκει ένα διάνυσμα το οποίο έχει την μεγαλύτερη δυνατή απόσταση από τα δυο πιθανά σύνολα μου. Στις πολλές διαστάσεις όπως το πρόβλημα μας ,δημιουργεί ένα υπερεπίπεδο (hyperplane). Η δημιουργία του υπερεπιπέδου μπορεί να γίνει με δύο τρόπους:

A) oneversus the rest (υπολογίζει ποια ανήκουν στην πρώτη κλάση σε σχέση με την πληροφορία όλων, έπειτα από τα απομείναντα στοιχεία υπολογίζει ποια ανήκουν στη δεύτερη κλάση σε σχέση με όλα κ.ο.κ.)

B) onevsone (υπολογίζει έναυ περεπίπεδο για κάθε κλάση ξεχωριστά ,δηλ. πως ξεχωρίζει η κάθε κλάση από τις υπόλοιπες, και έπειτα συνδυάζει την πληροφορία όλων των υπερεπίπεδων που βρήκε)

Στην svm χρησιμοποιείται και το kernel trick δηλ. τα δεδομένα μου αποκτούν μια επιπλέον διάσταση στην υπολογιστική τους αναπαράσταση για ευκολότερο διαχωρισμό.

5) SOFTMAX REGRESSION WITH TENSORFLOW

Όπως υποδηλώνει το όνομα, στην softmax regression (SMR), αντικαθιστούμε τη sigmoid logistic συνάρτηση από τη λεγόμενη συνάρτηση softmax φ:

$$P(y = j|z^{(i)} = \Phi_{sotmax}(z^{(i)}) = \frac{e^{z^{(i)}}}{\sum_{j=0}^{k} e^{z^{(i)}}}$$

Υπάρχουν μόνο δέκα πιθανά πράγματα που μπορεί να είναι μια δεδομένη εικόνα. Θέλουμε να μπορούμε να δούμε μια εικόνα και να δώσουμε τις πιθανότητες να είναι κάθε ψηφίο. Αυτή είναι μια κλασική περίπτωση όπου το softmax regression είναι ένα φυσικό, απλό μοντέλο. Γενικά, μας δίνει μια λίστα τιμών μεταξύ 0 και 1 που προσθέτουν μέχρι 1. Έχει δύο βήματα: πρώτα προσθέτουμε τα στοιχεία της εισροής μας σε ορισμένες κατηγορίες, και στη συνέχεια μετατρέπουμε τα στοιχεία αυτά σε πιθανότητες. Για να συγκεντρώσουμε τα στοιχεία που αποδεικνύουν ότι μια δεδομένη εικόνα είναι σε μια συγκεκριμένη κλάση, κάνουμε ένα σταθμισμένο άθροισμα των εντάσεων των pixels. Το βάρος είναι αρνητικό εάν το pixel που έχει υψηλή ένταση είναι διάφορο από την εικόνα που βρίσκεται σε αυτή την κατηγορία και θετικό αν υπάρχουν αποδεικτικά στοιχεία ότι ταιριάζει.

ΜΕΤΡΗΣΕΙΣ

Αποτελέσματα του συνολικού πειράματος είναι τα εξής:

1) Απαιτούμενος χρόνος για:

α. την εφαρμογή του decision tree t = 11.593750 sec

β. την εφαρμογή των random forests t = 3.437500 sec

γ. την εφαρμογή του logistic regression t = 62.171875 sec

δ. την εφαρμογή του support vector t = None

ε. την εφαρμογή του softmax regression t = 1.265625 sec

2) Το ποσοστό ακρίβειας για:

α. την εφαρμογή του decision tree AccuracyScore: 88.3%

β. Την εφαρμογή των random forests Accuracy Score: 94.6%

γ. Την εφαρμογή του logistic regression Accuracy Score: 91.7%

δ. Την εφαρμογή του support vector Accuracy Score: None

ε. Την εφαρμογή του softmax regression Accuracy Score: 91.9%

ΠΑΡΑΤΗΡΗΣΕΙΣ

1) Παρατηρούμε πως πιο γρήγορος αλγόριθμος για το classification 60.000 εικόνων με ψηφία είναι το softmax regression. Το decision tree επίσης παράγει αποτελέσματα σε αμελητέο χρόνο. Όσον αφορά το logistic regression , απ' τις μετρήσεις φαίνεται πως απαιτεί αρκετά περισσότερο χρόνο επειδή χρησιμοποιεί όλα τα δεδομένα για να παράγει τα αποτελέσματά του. Αυτό τον καθιστά ακατάλληλο για μεγάλα datasets.

2) Το μεγαλύτερο πρόβλημα που σχετίζεται με τα decision trees είναι ότι αποτελούν αρκετά μεροληπτικά μοντέλα. Μπορούμε να δημιουργήσουμε ένα decision tree model στο trainin gset μας το οποίο θα ξεπεράσει όλους τους άλλους αλγόριθμους, αλλά θα αποδειχθεί ότι είναι ένας κακός προγνωστικός παράγοντας στο test set μας. Αυτό προκαλείται λόγο του overfitting. Overfitting συμβαίνει όταν το δέντρο σχεδιάζεται έτσι ώστε να ταιριάζει απόλυτα σε όλα τα δείγματα στο trainset. Έτσι καταλήγει σε κλάδους με αυστηρούς κανόνες αραιών δεδομένων. Αυτό επηρεάζει την ακρίβεια κατά την πρόβλεψη δειγμάτων που δεν αποτελούν μέρος του trainset.
Το πρόβλημα του overfitting υπερνικάται σε μεγάλο βαθμό με τη χρήση random forests, τα οποία δεν είναι παρά μια πολύ έξυπνη επέκταση των decision trees.

3) Σημασία ακρίβειας
Στις μεθόδους randomforests και logistic regressionόπου βλέπουμε το ποσοστό ακρίβειας να ανεβαίνει μπορούμε να δούμε από τον confusion matrix ότι η πιθανότητα να έχω μπερδέψει ένα ψηφίο με κάποια από τα υπόλοιπα 9 εξαλείφεται έτσι σε περίπτωση λάθους μπορώ πιο εύκολα να μελετήσω τα δεδομένα μου ώστε να βρω μέσω του συλλογισμού μου ποιο είναι το σωστό.

4) Η εφαρμογή του C-SUPPORT VECTOR CLASSIFICATION δεν ήταν δυνατή γιατί το μοντέλο δεν είναι πολύ αποτελεσματικό με μεγάλο αριθμό παρατηρήσεων. Επιπλέον μπορεί να μην χρησιμοποιήθηκε ο κατάλληλος πυρήνας. Με τους μη γραμμικούς πυρήνες, τα SVM μπορούν να είναι πολύ δαπανηρά για να εκπαιδεύσουν τεράστια δεδομένα.(Δεν κατάφερε να τρέξει σε διάρκεια 1 ώρας επομένως, θεωρήθηκε ακατάλληλο μοντέλο για τα συγκεκριμένα δεδομένα).

Βιβλιογραφία:

- MazdakFatahi. [MNIST handwritten digits Description and using.]
- Weiran Wang. [ANN for Handwritten Digits Recognition.]
- Wikipedia:
- https://en.wikipedia.org/wiki/Random_forest
- https://en.wikipedia.org/wiki/Random_forest
- https://www.kdnuggets.com/2016/07/softmax-regression-related-logistic-regression.html
- https://www.youtube.com/watch?v=1NxnPkZM9bc&t=373s
- https://www.youtube.com/watch?v=N1vOgolbjSc
- https://www.edvancer.in/logistic-regression-vs-decision-trees-vs-svm-part2/
- https://pdfs.semanticscholar.org/8a3d/c02a337f6a07d0f6da254992defa007a8322.pdf