ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΑΝΑΛΥΣΗΣ ΗΛΕΚΤΡΙΚΩΝ ΚΥΚΛΩΜΑΤΩΝ
ΜΑΘΗΜΑ: ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ – 9ο ΕΞΑΜΗΝΟ

Εργασία 4

(Προθεσμία: Κυριακή 24 Δεκεμβρίου 2023, 23:59)

1.

Το IRIS data set περιέχει μορφολογικές μετρήσεις για 150 φυτά iris (είδος κρίνου, αγριόκρινο). Τα δεδομένα είναι της μορφής: (μήκος σέπαλου, πλάτος σέπαλου, μήκος πετάλου, πλάτος πετάλου) σε cm. Από αυτά τα 150 φυτά, 50 είναι Iris Setosa (ω1), 50 είναι Iris Versicolour (ω2) και 50 είναι Iris Virginica (ω3).

Α. Να χωριστεί το dataset σε 80% training και 20% test δείγματα με τυχαίο τρόπο.
Β. Να εκπαιδευτεί νευρωνικό δίκτυο 2 επιπέδων για την ταξινόμηση των δειγμάτων, όπου το 1ο layer θα έχει 30 νευρώνες και συνάρτηση ενεργοποίησης θα είναι sigmoid.
Γ. Να αποτυπωθεί η εξέλιξη του test accuracy μετά από κάθε epoch εκπαίδευσης σε διάγραμμα, και να υπολογιστεί ο πίνακας σύγχυσης του τελικού μοντέλου για το test set.
Δ. Να επαναληφθεί η εκπαίδευση με συνάρτηση ενεργοποίησης την ReLu. Τί παρατηρείτε?
Ε. Πειραματιστείτε με τις παραμέτρους σχεδίασης, και προτείνετε ένα δίκτυο που επιτυγχάνει καλύτερη ακρίβεια από τα προηγούμενα. Τεκμηριώστε το αποτέλεσμα με τα απαραίτητα γραφήματα και πίνακες σύγχυσης.
ΣΤ. Στο καλύτερο δίκτυο από αυτά που εκπαιδεύσατε, υπολογίστε και αποτυπώστε τις καμπύλες precision-recall για τις τρεις κλάσεις, μεταβάλλοντας το κατώφλι της πιθανότητας κάθε κλάσης. Ποια κλάση είναι πιο εύκολα διαχωρίσιμη με βάση το AUC (Area Under Curve) που προκύπτει από τις καμπύλες?

2.

Να εκπαιδεύσετε ένα νευρωνικό δίκτυο με αρχιτεκτονική autoencoder, με 3 επίπεδα κωδικοποίησης 128,32 και 3 νευρώνων και 3 αντίστοιχα επίπεδα αποκωδικοποίησης, χρησιμοποιώντας το training set της βάσης δεδομένων MNIST Digits που χρησιμοποιήθηκε στην εργαστηριακή άσκηση 6.

Α. Ποια αρχιτεκτονική autoencoder πέτυχε το μικρότερο μέσο τετραγωνικό σφάλμα στα δεδομένα εκπαίδευσης? Αυτή που χρησιμοποιήθηκε στην εργαστηριακή άσκηση ή αυτή που εκπαιδεύσατε? Γιατί πιστεύετε ότι συνέβη αυτό?
Β. Να απεικονίσετε κατάλληλα το σύνολο των εικόνων του training set όπως κωδικοποιούνται στο latent space, με κατάλληλο χρώμα που να αντιστοιχεί στο label καθεμιάς. Κάνετε το ίδιο και για τις εικόνες του test set. Διατηρούνται τα χαρακτηριστικά της κατανομής των απεικονίσεων και στα δύο σύνολα? Σχολιάστε.

Γ. Να χρησιμοποιήσετε τον κλάδο αποκωδικοποίησης του εκπαιδευμένου autoencoder ώστε να δημιουργήσετε εικόνες ξεκινώντας από τυχαία σημεία του latent space. Εξερευνήστε το latent space δοκιμάζοντας διάφορα σημεία του χώρου και παρατηρώντας τις εικόνες που παράγονται. Αντιστοιχούν όλες οι περιοχές του latent space σε πραγματικά ψηφία? Σχολιάστε τα ευρήματά σας παραθέτοντας παραδείγματα εικόνων.

3.

Το Breast Cancer Wisconsin είναι ένα σύνολο δεδομένων που περιέχει 30 μορφολογικά χαρακτηριστικά των καρκινικών κυττάρων από 569 βιοψίες μαστού, καθώς και το είδος του όγκου για κάθε δείγμα (κακοήθης: κλάση1 ή M | καλοήθης: κλάση 0 ή B).

Α. Να χωριστεί το dataset σε 70% training και 30% test δείγματα με τυχαίο τρόπο.

Β. Να μετατρέψετε το 10% των training δεδομένων σε missing values με τυχαία ομοιόμορφη κατανομή, και να εκπαιδεύσετε έναν δενδρικό ταξινομητή (tree classifier) με μέγιστο βάθος 5 επίπεδα. Υπολογίστε την ακρίβεια πρόβλεψης του ταξινομητή στο test set.

Γ. Με τα ίδια training και test sets να εκπαιδεύσετε ένα random forest[1] (RF) με 100 ταξινομητές μέγιστου βάθους 3, χρησιμοποιώντας 5 features ανά δένδρο χωρίς bootstrapping των training δεδομένων. Να υπολογίσετε την ακρίβεια πρόβλεψης του RF στο test set.

Δ. Υπολογίστε το feature importance για τον δενδρικό ταξινομητή και το RF. Σχολιάστε τα μέχρι τώρα αποτελέσματα και τις διαφορές που παρατηρείτε.

Ε. Να εκπαιδεύσετε το ίδιο RF για μεταβλητό ποσοστό missing values (0% έως 80% ανά 10%) και για κάθε forest να υπολογίσετε το classification accuracy εισάγοντας στο test set μεταβλητό ποσοστό missing values (0% έως 80% ανά 10%). Απεικονίστε κατάλληλα τη συνολική συμπεριφορά της ακρίβειας του RF ταξινομητή σε σχέση με το ποσοστό missing values. Σχολιάστε την αντοχή του random forest στα missing data, τόσο στο training όσο και στο testing. Σε ποιο από τα δύο είδη missing data είναι περισσότερο ευαίσθητος?

ΣΤ. Να εκπαιδεύσετε έναν δενδρικό ταξινομητή και έναν RF ταξινομητή όπως στα προηγούμενα, για 10% missing data στο training set. Υπολογίστε τις καμπύλες precision-recall για τους δύο ταξινομητές[2] (χωρίς missing data στο test set). Ποιος ταξινομητής έχει καλύτερη συμπεριφορά και γιατί? Δεδομένου ότι οι ταξινομητές που εκπαιδεύσατε αποτελούν συστήματα πρόβλεψης της κακοήθειας όγκων, ποια χαρακτηριστικά της καμπύλης έχουν περισσότερη βαρύτητα κατά την πρακτική εφαρμογή?

Υποβάλετε τις απαντήσεις σας στην πλατφόρμα του e-class στην ενότητα εργασίες, πριν τη λήξη της προθεσμίας υποβολής. Η υποβολή σας πρέπει να αποτελείται από ***ένα μόνο συμπιεσμένο αρχείο*** *(.rar, .zip κλπ.) το οποίο θα περιέχει όλα τα απαραίτητα αρχεία για την υποβολή σας. Καταγράφετε τις απαντήσεις σας σε μία αναφορά που θα αποστείλετε σε μορφή pdf ή docx ή εναλλακτικά ενσωματωμένες σε Python notebooks ή Matlab live scripts. Οι αποδεκτές γλώσσες προγραμματισμού είναι Matlab ή*

[1] Εφόσον η βιβλιοθήκη που χρησιμοποιήσετε δεν υποστηρίζει missing data για random forests, η εκπαίδευση των random forests να γίνει χειροκίνητα ως ένα σύνολο δενδρικών ταξινομητών, με τυχαία επιλογή features και majority voting για την επιλογή της τελικής κλάσης.

[2] Για να υπολογιστεί η καμπύλη precision recall θα πρέπει να εξάγετε πιθανότητες από τον δενδρικό και τον RF ταξινομητή. Για τον RF ταξινομητή μπορείτε να χρησιμοποιήσετε το ποσοστό των προβλέψεων για την κλάση 1 από όλα τα δέντρα του ταξινομητή.

Python, όπου και θα πρέπει να συμπεριλάβετε στο συμπιεσμένο αρχείο και τα αρχεία του πηγαίου κώδικα (αρχεία .m, .py κλπ) με τα απαραίτητα σχόλια για την τεκμηρίωση εντός του κώδικα. υποβάλλοντας ***το πολύ ένα αρχείο κώδικα για κάθε άσκηση****. Σε περίπτωση που χρησιμοποιήσετε Matlab, μπορείτε να συμπεριλάβετε τον κώδικα για όλες τις ασκήσεις σε ένα αρχείο, χωρίζοντας τις απαντήσεις σε διαφορετικά cells. Επίσης, για τις ασκήσεις που περιλαμβάνουν κώδικα, αντί απάντησης στην αναφορά μπορείτε να υποβάλετε Matlab live scripts ή Python notebooks κατάλληλα δομημένα ώστε να περιλαμβάνουν και τις απαντήσεις/σχολιασμούς που ζητούνται.*

Ξάνθη, 10/12/2023