ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ
Προαιρετική Άσκηση 2
2017-2018
Τοσουνίδης Βασίλειος
ΑΜ: 109 / 15

1. Εισαγωγή

Για τον σκοπό της εργασίας επιλέχθηκε ένα dataset από το Kaggle το οποίο περιλαμβάνει πληροφορίες για τα βιομηχανικά χαρακτηριστικά ορθοπαιδικών ασθενών καθώς και μία κατάταξη ανάλογα με τις τιμές τους. Τα δεδομένα είναι κατηγοριοποιημένα σε 2 αρχεία, τα οποία σχετίζονται. Στο πρώτο οι ασθενείς είναι διαχωρισμένοι σε δύο κατηγορίες “Normal” και “Abnormal” ανάλογαNormal” και “Normal” και “Abnormal” ανάλογαAbnormal” ανάλογα με τις μετρήσεις τους, ενώ στο δεύτερο σε “Normal” και “Abnormal” ανάλογαNormal”, “Normal” και “Abnormal” ανάλογαHernia” και “Normal” και “Abnormal” ανάλογαSpondylolisthesis”. Κάθε ασθενής παρουσιάζεται από έξι βιομηχανικά χαρακτηριστικά τα οποία που προέρχονται από το σχήμα και τον προσανατολισμό της λεκάνης και της οσφυϊκής σπονδυλικής στήλης.

Στο πρόβλημα που θέλουμε να λύσουμε όμως δεν χρειαζόμαστε την στήλη με την κατηγοριοποίηση αφού η συσταδοποίηση είναι κυρίως unsupervised.

2. Εργασιακό περιβάλλον

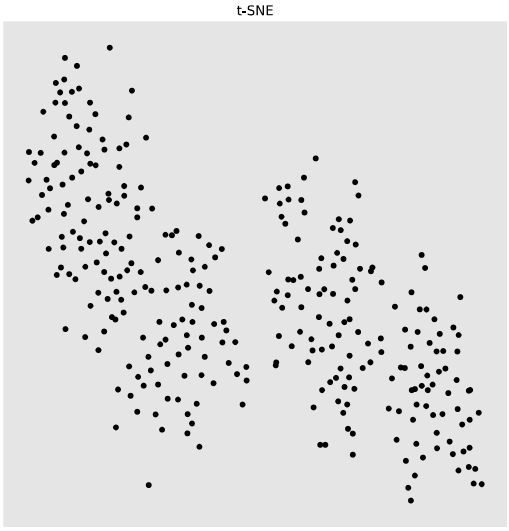
Για την ολοκλήρωση της εργασίας έγινε χρήση της γλώσσας προγραμματισμού Python Αέκδοση Β) και μερικών βιβλιοθηκών κατάλληλων για διαχείρηση και οπτικοποίηση δεδομένων αλλά και εκπαίδευση του μοντέλου. Συγκεκριμένα αυτές είναι:

- **Numpy**
- **Matplotlib:** Είναι μία βιβλιοθήκη ανοιχτού κώδικα για οπτικοποίηση δεδομένων σε plots.
- **Scikit-learn:** Είναι μια βιβλιοθήκη ανοιχτού κώδικα η οποία περιέχει πληθώρα αλγορίθμων για μηχανική μάθηση όπως random forests, k-means κλπ.
- **Pandas:** Είναι μία βιβλιοθήκη ανοιχτού κώδικα η οποία παρέχει εύκολες στη χρήση δομές δεδομένων για υψηλού επιπέδου αριθμητικούς υπολογισμούς.

3. Ανάλυση δεδομένων

Η αρχική ανάλυση έχει ήδη γίνει (προηγούμενη εργασία) οπότε δεν έγινε επανάληψη των ίδιων βημάτων. Από αυτήν βλέπουμε ότι τα δεδομένα μας χωρίζονται σε 6 στήλες (πλην της στήλης των ετικετών) και έτσι βλέπουμε ότι η συσταδοποίηση σε 6 διαστάσεις δεν είναι φυσικά εφικτή, οπότε πρέπει να εφαρμόσουμε κάποια τεχνική έτσι ώστε να μετατρέψουμε τα δεδομένα μας σε μία μορφή 2 διαστάσεων. Η τεχνική που χρησιμοποιήθηκε είναι αυτή του dimensionality reduction με τον αλγόριθμο t-SNE το οποίο στην ουσία του κάνει προβολή όλων των σημείων σε διδιάστατο χώρο. Αυτή η τεχνική χρησιμοποιείται σε περιπτώσεις που θέλουμε να κάνουμε εκπαίδευση σε δεδομένα τα οποία είναι πολυδιάστατα, όπως για παράδειγμα εικόνες, και να τα φέρουμε σε μικρότερη διάσταση ώστε να μπορέσει το μοντέλο να εκπαιδευτεί πάνω σε αυτά. Υπάρχουν πολλές τεχνικές για να επιτευχθεί αυτό το αποτέλεσμα (PCA, LDA, t-SNE), αλλά για την εργασία μετά από πειράματα επιλέχθηκε η τεχνική t-SNE η οποία έδωσε τα καλύτερα αποτελέσματα.

Αυτα φαίνονται στην παρακάτω εικόνα:



1. Dimensionality Reduction με t-SNE

4. Μοντέλο

Τα δεδομένα πλέον είναι στη σωστή μορφή και έτσι μπορούμε να τα δώσουμε στο μοντέλο για να εκπαιδευτεί. Για την συσταδοποίηση θα παρουσιαστούν 2 αλγόριθμοι οι οποίοι ενώ έχουν μερικές διαφορές βασίζονται στον ίδιο, οι οποίοι είναι ο K-Means και ο Mini Batch K-Means. Για τους δύο αυτούς αλγόριθμους χρησιμοποιήθηκε η υλοποίηση της βιβλιοθήκης Scikit-Learn. Το αποτέλεσμα φαίνεται στην παρακάτω εικόνα, μαζί με τις διαφορές των δύο αλγόριθμων: