# Θεωρία Αποφάσεων

## Task Force 1

Απαλλακτική Εργασία του φοιτητή

Τσιαμήτρου Κωνσταντίνου – ΑΜ 235913 – 9ο έτος

# Εισαγωγή

Το αντικείμενο της παρούσας εργασίας είναι η εξερεύνηση ενός συνόλου δεδομένων στον κλάδο της υγειονομικής περίθαλψης. Το σύνολο δεδομένων που αξιοποιήθηκε, χρησιμοποιείται για την πρόβλεψη της Στεφανιαίας Νόσου (ΣΝ), με βάση κάποιους παράγοντες κινδύνου όπως ο διαβήτης και το κάπνισμα.

Η ανεξάρτητη μεταβλητή (η μεταβλητή την οποία θα προσπαθήσουμε να προβλέψουμε) είναι ο 10ετής κίνδυνος ΣΝ, και είναι μια δυαδική μεταβλητή, καθώς ένας ασθενής είτε εμφανίζει την νόσο είτε όχι.

Για την πρόβλεψη αυτή, αναπτύχθηκε ένας δυαδικός ταξινομητής με τη βοήθεια ενός Νευρωνικού Δικτύου με 13 εισόδους, 1 κρυφό επίπεδο (διαφόρων μεγεθών) και μια έξοδο.

Το σύνολο δεδομένων είναι εξαιρετικά ανισορροπημένο. Για αυτόν τον λόγο αξιοποιήθηκαν κάποιες τεχνικές προ-επεξεργασίας του συνόλου δεδομένων (αναλύονται στη συνέχεια).

# Baseline Νευρωνικό Δίκτυο

Πριν εφαρμοστούν οι είσοδοι στο Νευρωνικό Δίκτυο, κρίθηκε απαραίτητη η προεπεξεργασία του συνόλου δεδομένων, για τον εξής λόγο:

Δοκιμάστηκαν οι παρακάτω τοπολογίες για το ανεπεξέργαστο σύνολο δεδομένων:

(με μπλε είναι το Mean accuracy και το πορτοκαλί είναι το standard deviation για 10-fold Cross-Validation)

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |

Παρατηρώ ότι παρόλη την αύξηση στον αριθμό των Νευρώνων στο κρυφό επίπεδο, το mean accuracy παραμένει σταθερά γύρω στο 50%, όσο είναι (περίπου) και το ποσοστό των θετικών δειγμάτων προς τα αρνητικά.

Αν ο αριθμός των Νευρώνων του κρυφού επιπέδου αυξηθεί αρκετά (κοιτάξτε το τελευταίο διάγραμμα με τους 50 Νευρώνες) παρατηρώ μια αύξηση στο accuracy του δικτύου. Αυτό μάλλον οφείλεται στο ότι το δίκτυο απλώς αποστηθίζει τα διανύσματα εισόδου και άρα η γενικευτική του ικανότητα δεν θα είναι καλή.

Προκειμένου να αντιμετωπιστεί το παραπάνω φαινόμενο, αξιοποιήθηκαν οι παρακάτω τεχνικές προ-επεξεργασίας των δεδομένων.

## Προ-επεξεργασία του συνόλου δεδομένων

Εφαρμόστηκαν τρείς τεχνικές προ-επεξεργασίας δεδομένων.

1. Standardization

Η τεχνική αυτή προσθέτει κάποιες επιθυμητές ιδιότητες στο dataset. Αυτές οι ιδιότητες είναι:

1. Αφαίρεση outliers

Αφότου έχουμε κάνει scale τα δεδομένα μας ώστε να έχουν μέση τιμή μηδέν και τυπική απόκλιση ένα, οι outliers (δηλαδή ακραίες τιμές) είναι φανερές. Οι outliers αποπροσανατολίζουν το δίκτυο κατά την εκπαίδευσή του, και συνεπώς, η αφαίρεσή τους προσδοκούμε ότι θα έχει θετικά αποτελέσματα στο accuracy. Οι outliers εντοπίζονται, συγκρίνοντας την τιμή τους με την μέση τιμή του κάθε χαρακτηριστικού. Αν κάποιο δείγμα έχει σε κάποιο πεδίο τιμή μεγαλύτερη από δύο φορές την μέση τιμή, αφαιρείται από το dataset , προτού διοχετευτεί στην είσοδο του Νευρωνικού Δικτύου προς εκπαίδευση. (Αυτό το βήμα εφαρμόστηκε μόνο για τα χαρακτηριστικά age, trestbps, chol, thalach, oldpeak)

1. Down-sampling

Η αφαίρεση των παραπάνω δειγμάτων από το σύνολο δεδομένων μπορεί να θεωρηθεί ως down-sampling, (υποδειγματοληψία) διότι, με τον τρόπο που περιεγράφηκε παραπάνω, αναγνωρίζονται και αφαιρούνται από το dataset συγκεκριμένα δείγματα που ανήκουν σε κάποιες «προβληματικές» κλάσεις δειγμάτων.

## Αρχιτεκτονική Νευρωνικού Δικτύου

Η αρχιτεκτονική που έχει επιλεγεί για το Νευρωνικό Δίκτυο είναι πολύ απλή. Έχει 13 εισόδους (μια για κάθε χαρακτηριστικό), 1 κρυφό επίπεδο (διαφόρων μεγεθών) και μια έξοδο (μια έξοδος αρκεί ακόμα και για One-hot-encoding αφού αντιμετωπίζουμε ένα binary classification πρόβλημα).

Το κρυφό επίπεδο έχει ως συνάρτηση ενεργοποίησης την leaky ReLU (προκειμένου να αντιμετωπιστεί το φαινόμενο των «νεκρών» Νευρώνων) και το επίπεδο εξόδου έχει ως συνάρτηση ενεργοποίησης την sigmoid. Η τελευταία, προσδίδει μια πιθανότητα στο διάστημα [0,1] στο κάθε δείγμα. Η πιθανότητα αυτή αντιπροσωπεύει το likelihood να ανήκει το δείγμα στην μία ή στην άλλη κλάση (οι δύο αυτές πιθανότητες είναι συμπληρωματικές). Τέλος, με τη χρήση της binary cross-entropy loss function, του optimizer “adam” και της μετρικής απόδοσης accuracy έγινε compile το δίκτυο.

Έγινε προσπάθεια, να ελαχιστοποιηθούν οι Νευρώνες του κρυφού επιπέδου, και ταυτόχρονα στόχος ήταν να αυξηθεί το accuracy όσο γίνεται (σε min-max sense).

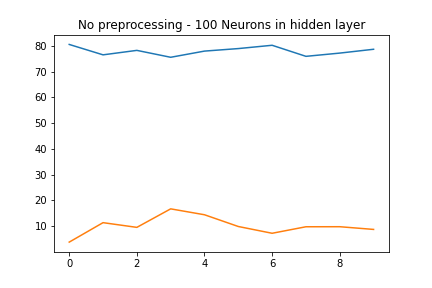
## Αποτελέσματα Νευρωνικού Δικτύου με προ-επεξεργασμένο σύνολο δεδομένων

Παραθέτω τα αντίστοιχα cases με το προηγούμενο εδάφιο για λόγους σύγκρισης:

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

Παρατηρώ ότι για 20 και 50 Νευρώνες στο κρυφό επίπεδο, το δίκτυο έχει ικανοποιητική απόδοση (χωρίς απώλεια της γενικευτικής του ικανότητας).

Τέλος, να επισημάνω ότι αντίστοιχη απόδοση πετυχαίνει το Δίκτυο με 100 Νευρώνες στο κρυφό επίπεδο για ανεπεξέργαστο σύνολο δεδομένων, οπότε μπορώ να θεωρήσω ότι ο αρχικός στόχος για μείωση των διαστάσεων του κρυφού επιπέδου έχει επιτευχθεί.



## Πηγές

1. [www.google.com](http://www.google.com)
2. <https://machinelearningmastery.com/binary-classification-tutorial-with-the-keras-deep-learning-library/>
3. Γνώσεις Ψηφιακής Επεξεργασίας Σημάτων