Εφαρμογή μεθόδων μηχανικής μάθησης για διαχωρισμό σήματος και υποβάθρου σε εξωτικό σενάριο Higgs

Haka Kevin

Στην εργασία αυτήν θα προσπαθήσουμε να λύσουμε ένα πρόβλημα ταξινόμησης με τεχνικές μηχανικής μάθησης. Πιο συγκεκριμένα θα εκπαιδεύσουμε 2 μοντέλα μηχανικής μάθησης και ένα μοντέλο νευρωνικού δικτιού. Το σετ των δεδομένων είναι ένα θραύσμα από το πλήρες σετ της εργασία "*Baldi, P. et al. Searching for exotic particles in high-energy physics with deep learning. Nat. Commun. 5:4308 doi: 10.1038/ncomms5308 (2014).*" Στο σύνολο αυτό των δεδομένων θα πρέπει να διαχωρίσουμε το σήμα από τις διεργασίες υποβάθρου από τα θεωρητικά σήματα διεργασιών παραγωγής νέων σωματιδίων Higgs.

Στα δεδομένα υπάρχουν 28 χαρακτηριστικά οι 21 πρώτες (1-21) είναι χαμηλού επιπέδου χαρακτηριστικά, δηλαδή, ο διαχωρισμός των σημάτων από τις διεργασίες υποβάθρου από αυτά τα χαρακτηριστικά είναι δύσκολος, ενώ, τα υπόλοιπα 7 χαρακτηριστικά (22-28) είναι υψηλού επιπέδου, όπου έχει γίνει επεξεργασία των 21 χαρακτηριστικών χαμηλού επιπέδου βάση της γνώσης μας για τις ενδιάμεσες κατάστασης που λαμβάνουν χώρα στις δυο διεργασίες, με απώτερο σκοπό τον πιο εύκολο διαχωρισμό των σημάτων. Εμείς θα μελετήσουμε ξεχωριστά σε κάθε μοντέλο που θα κατασκευάσουμε τις δυο ομάδες χαρακτηριστικών.

Παρακάτω θα αναφέρω τα βήματα που ακολούθησα για την κατασκευή των μοντέλων και τα αποτελέσματα που έλαβα. Ο κώδικάς με τα μοντέλα βρίσκεται στο αρχείο *"HIGGS.ipynb"*.

- Αρχικά εισήγαγα τα δεδομένα από το αρχείο σε ένα DataFrame.
- Στην συνέχεια έκανα έναν ελέγχω εάν τα δεδομένα είναι πλήρη και αν ο τύπος των δεδομένων ανά στήλη είναι σωστός. Εκεί βρήκα και διόρθωσα την στήλη 17 από object σε float64 και εφόσον μιλάμε για binary classification μετέτρεψα την στήλη 0 από float64 σε bool (αν και δεν ήταν αναγκαίο).
- Έπειτα έλεγξα την συχνότητα των δυο καταστάσεων προς ταξινόμηση (Figure 1), όπου θεώρησα πως η διαφορά δεν είναι σημαντική ώστε να επέμβω.
- Προχώρησα διαχωρίζοντας τον αρχικό πίνακα στα χαρακτηριστικά χαμηλού επιπέδου, υψηλού επιπέδου και μια στήλη στόχος. Επειδή κάποιος από τα μοντέλα που θα κατασκευάσουμε μπορεί να είναι ευαίσθητο σε γεωμετρικές αποστάσεις έκανα τυποποίηση των δεδομένων ανά χαρακτηριστικό. Τέλος διαχώρισα τα

δεδομένα και από τις δυο ομάδες χαρακτηριστικών σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου (75% - 25%).

- Ως πρώτο μοντέλο διάλεξα ένα αγαπημένο μου το ExtraTreesClassifier το οποίο αποτρέπει πάρα πολύ την υπερπροσαρμογή στα δεδομένα εκμάθησης και σε γενικές γραμμές δίνει πολύ καλά αποτελέσματα. Ενώ ως δεύτερο μοντέλο διάλεξα το KNeighborsClassifier, το οποίο είναι ένα πάρα πολύ γνωστό και εβραίος χρησιμοποιούμενο.
- Σε κάθε μοντέλο κατασκεύασα ένα πλέγμα υπερπαραμέτρων (αρκετά μικρό για να τρέχει γρήγορα) όπου διερευνώ τις παραμέτρους που θα μου κάνουν το καλύτερο διαχωρισμό. Στην εκπαίδευση χρησιμοποιώ διασταυρωμένη επικύρωση με 10-τμήματα και ως κριτήριο επίδοσης χρησιμοποιώ το σκορ ROC AUC. Μετά την εκπαιδεύσει κρατάμε το καλύτερο μοντέλο από κάθε αλγόριθμο. Την ίδια διαδικασία εκτελούμε και για τις δυο ομάδες χαρακτηριστικών.
- Το τρίτο μοντέλο θα είναι ένα νευρωνικό δίκτυο. Αφότου έκανα αρκετές δοκιμές με διάφορες αρχιτεκτονικές. Κατέληξα σε ένα νευρωνικό δίκτυο με την εξής δομή:

    - Επίπεδο ισώνουμε με αριθμό νευρώνων όσες και οι μεταβλητές
    - Πυκνό επίπεδο με 64 νευρώνες και “relu” συνάρτηση ενεργοποίησης
    - Dropout επίπεδο με ποσοστό 50%
    - Πυκνό επίπεδο με 32 νευρώνες και “relu” συνάρτηση ενεργοποίησης
    - Dropout επίπεδο με ποσοστό 50%
    - Πυκνό επίπεδο με 64 νευρώνες και “relu” συνάρτηση ενεργοποίησης
    - Dropout επίπεδο με ποσοστό 50%
    - Πυκνό επίπεδο εξόδου με 1 νευρώνα

- Τα Dropout επίπεδα βοηθούν στην υπερπροσαρμογή του μοντέλου στα δεδομένα κατά την εκπαίδευση (Figure 2&3).
- Ο συνολικός αριθμός των προς εκπαίδευση παραμέτρων είναι 5665 και 4769 για το δίκτυο για τα χαρακτηριστικά χαμηλού και υψηλού επιπέδου αντίστοιχα.
- Στον νευρώνα εξόδου δοκιμαστήκαν οι συναρτήσεις ενεργοποίησης “sigmoid” και “tanh” αλλά δεν προτιμήθηκαν γιατί η απόδοση έπεφτε.
- Κατά την εκπαίδευση των δικτυών ως συνάρτηση απώλειας χρησιμοποιήθηκε το “binary crossentropy”, ως βελτιστοποιητής χρησιμοποιήθηκε ο αλγόριθμος Adam με ρυθμό εκμάθησης 0.001.
- Ο συνολικός αριθμός των εποχών που μπορεί να εκπαιδευτεί το δίκτυο ορίστηκε σε 1000. Η κάθε φουρνιά δεδομένων πριν από κάθε εκπαίδευση οριστικές σε 100.

Επιπλέων έχει οριστεί μηχανισμός πρόωρου τερματισμού της εκπαίδευσης του δικτιού που ενεργοποιείται αν η επικύρωση της συνάρτηση απώλειας δεν βελτιώνεται άλλο.

Τα τελικά αποτελέσματα των τριών μοντέλων συγκρίνονται βάση της ποσότητας ROC AUC βάση της επίδοσης τους στα δεδομένα ελέγχου:

Χαμηλού επιπέδου χαρακτηριστικά

1. ExtraTreesClassifier = 0.601
2. KNeighborsClassifier = 0.564
3. ArtificialNeuralNetwork = 0.633

Υψηλού επιπέδου χαρακτηριστικά

1. ExtraTreesClassifier = 0.701
2. KNeighborsClassifier = 0.669
3. ArtificialNeuralNetwork = 0.770

Για την ποσότητα των δεδομένων που είχαμε στην κατοχή μας προς χρήση μπορούμε να πούμε με σιγουριά πως τα δεδομένα υψηλού επιπέδου έχουν ένα προβάδισμα 0.1 για τα μοντέλα μηχανικής μάθησης και 0.14 το νευρωνικό δίκτυο, πράγμα που μας κάνει να συμπεράνουμε πως η επέμβαση των φυσικών στα δεδομένα αυξάνει την απόδοση διαχωρισμού του υποβάθρου από τα σήματα σωματιδίων Higgs, τουλάχιστον στα απλά μοντέλα που κατασκεύασα. Πιθανότατα πιο περιπλοκά μοντέλα και πιο προχωρημένες αρχιτεκτονικές στα νευρωνικά δίκτια να μην έχουν την ανάγκη από τα χαρακτηριστικά υψηλού επιπέδου στην επίδοση τους. Πέρα από αυτό, το νευρωνικό δίκτυο και στις δυο περιπτώσεις έχει ξεκάθαρο προβάδισμα στις επιδόσεις του με τα χαρακτηριστικά υψηλού επιπέδου να αυξάνουν την επίδοση αρκετά. Δεύτερο στην κατάταξη έρχεται το ExtraTreesClassifier το οποίο έχει μια μέτρια επίδοση η οποία βελτιώνεται στα υψηλού επιπέδου χαρακτηριστικά ενώ τελευταίο είναι το KNeighborsClassifier το οποίο έχει μια κακή επίδοση (οριακά καλύτερη από έναν τυχαία ταξινομητή) στα χαμηλού επιπέδου χαρακτηριστικά αλλά βελτιώνονται σχετικά στα υψηλού επιπέδου. Η παρακάτω επίδοση μας δείχνει πως το πρόβλημα είναι αρκετά σύνθετο και πως απλά μοντέλα σαν το KNeighborsClassifier δεν μπορούν να ανταπεξέλθουν σε τόσο δύσκολα πρόβλημα, από την άλλοι τα νευρωνικά δίκτυα φαίνονται αρκετά υποσχόμενα και με κατάλληλη βελτιστοποίηση στην αρχιτεκτονική του δικτιού τα αποτελέσματα μπορεί να είναι καταπληκτικά.

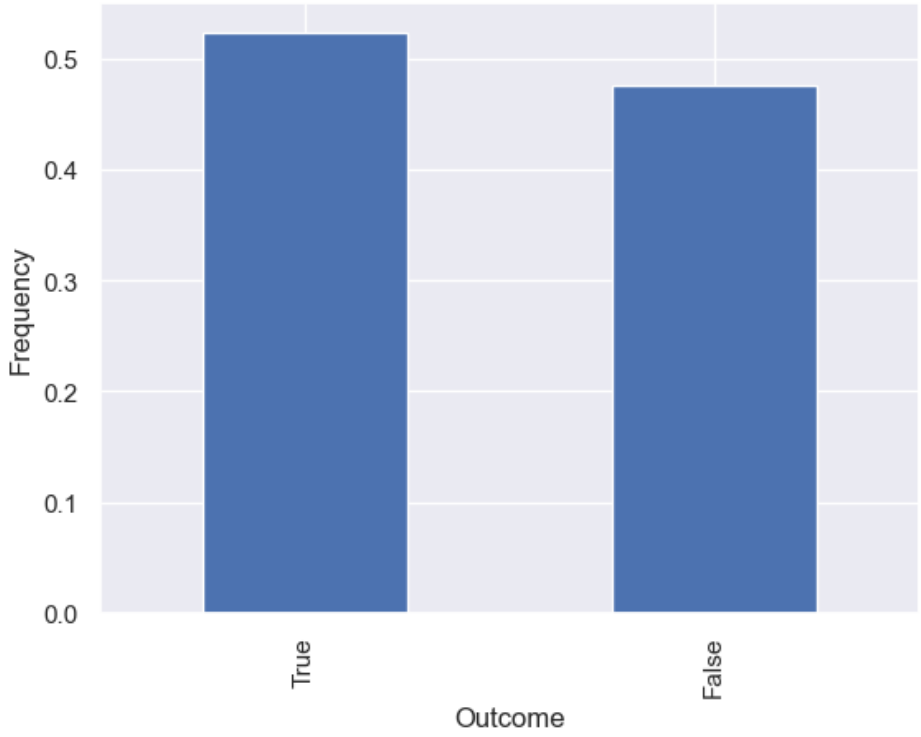

Figure 1 *Γράφημα όπου φαίνονται οι συχνότητες των δυο καταστάσεων προς ταξινόμηση.*

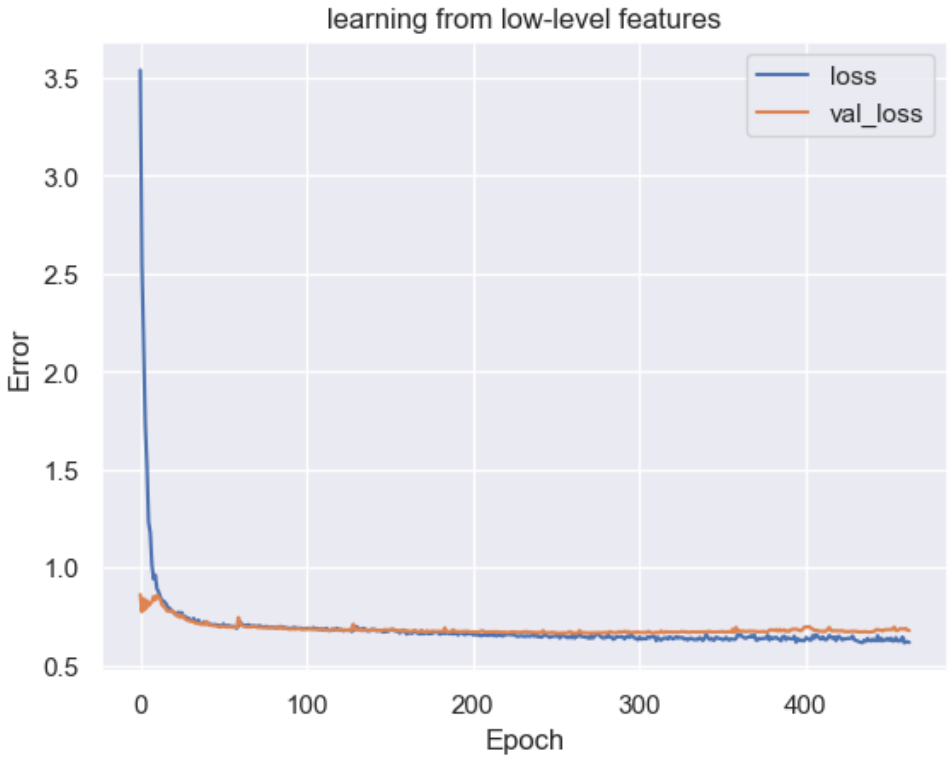

Figure 2 *Μεταβολή της συνάρτηση απώλειας (binary crossentropy) κατά την εκπαίδευση του δικτιού από τα δεδομένα χαμηλού επιπέδου. Με μπλε το σφάλμα από τα δεδομένα εκμάθησης και πορτοκαλή από την επικύρωση.*

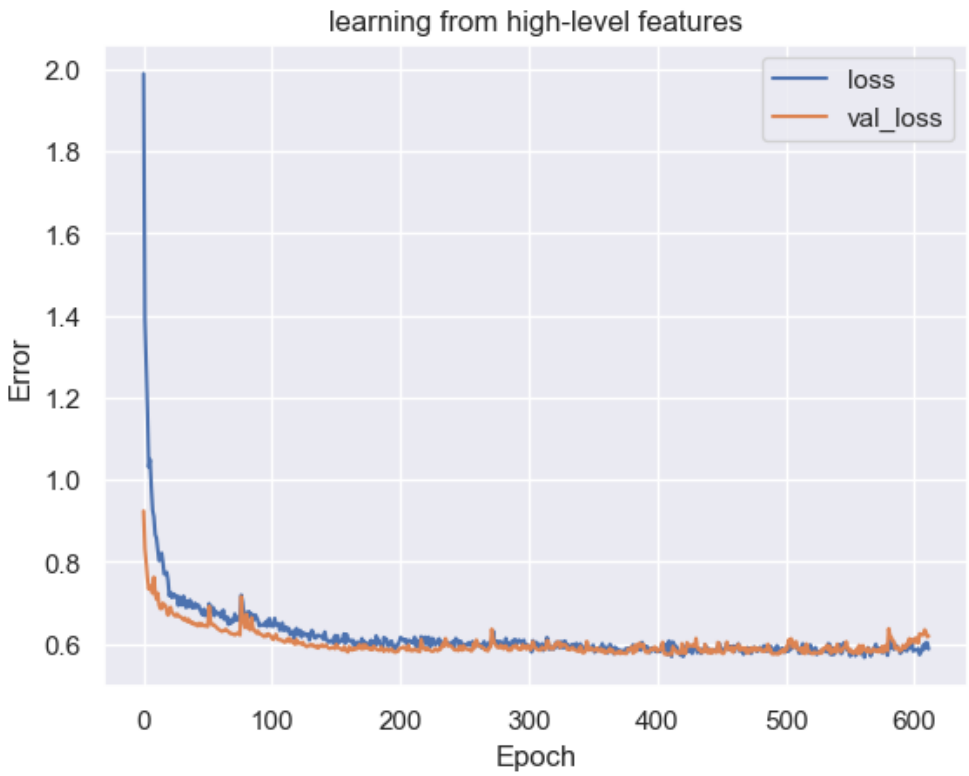

Figure 3 *Μεταβολή της συνάρτηση απώλειας (binary crossentropy) κατά την εκπαίδευση του δικτιού από τα δεδομένα υψηλού επιπέδου. Με μπλε το σφάλμα από τα δεδομένα εκμάθησης και πορτοκαλή από την επικύρωση.*