Αλγόριθμοι Boosting: Adaboost

Στατιστική Μηχανική Μάθηση

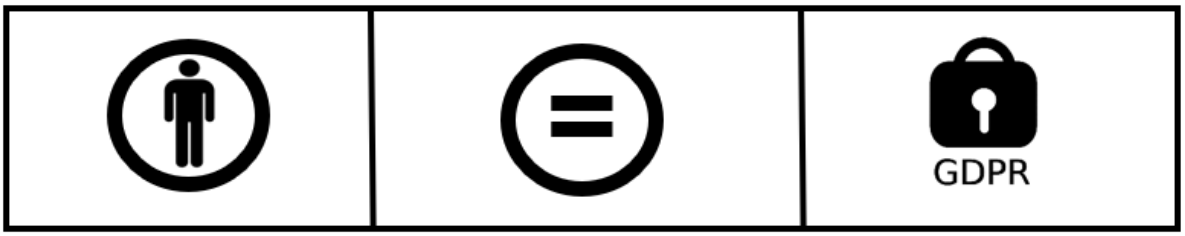
Βίννη Παναγιώτα

Α.Μ. : 1873

Εξάμηνο: 8ο

Άρτα, 2023

Πίνακας περιεχομένων



1. Εισαγωγή

Οι αλγόριθμοι Boosting αποτελούν μία δημοφιλή και ισχυρή τεχνική μηχανικής μάθησης που στοχεύει στη βελτίωση της απόδοσης αδύναμων μαθητών μέσω του συνδυασμού τους σε ένα ισχυρό συνολικό μοντέλο. Έχουν κερδίσει σημαντική προσοχή στον τομέα της μηχανικής μάθησης λόγω της ικανότητάς τους να αντιμετωπίζουν πολύπλοκα προβλήματα ταξινόμησης και παλινδρόμησης. Ένας διακεκριμένος αλγόριθμος boosting είναι ο Adaboost, ο οποίος έχει ευρέως υιοθετηθεί και έχει αποδειχθεί αποτελεσματικός σε διάφορους τομείς.

Ο Adaboost επικεντρώνεται σε προβλήματα δυαδικής ταξινόμησης. Έχει μελετηθεί εκτενώς και έχει εφαρμοστεί σε διάφορους τομείς, όπως η υπολογιστική όραση, η επεξεργασία φυσικής γλώσσας, η βιοπληροφορική και άλλα. Η ανθεκτικότητα, η απλότητα και η δυνατότητα του αλγορίθμου να αντιμετωπίζει υψηλής διάστασης δεδομένα έχουν συμβάλει στη δημοφιλία του μεταξύ των ερευνητών και των επαγγελματιών.

2. Κύριες έννοιες – Ορολογία του Adaboost

Για την καλύτερη κατανόηση του αλγορίθμου AdaBoost, είναι σημαντικό να γνωρίζουμε ορισμένες κύριες έννοιες που σχετίζονται με τον αλγόριθμο:

- **Αδύναμοι Μαθητές:** Οι αδύναμοι μαθητές, γνωστοί επίσης ως βασικοί ή συνιστώσες μαθητές, είναι απλά και σχετικά χαμηλής απόδοσης μοντέλα που έχουν ικανότητα ταξινόμησης ή παλινδρόμησης ελαφρώς καλύτερη από την τυχαία εικασία. Οι αδύναμοι μαθητές μπορεί να είναι δένδρα αποφάσεων με περιορισμένο βάθος, γραμμικά μοντέλα με απλούς κανόνες ή

οποιαδήποτε άλλα μοντέλα που επιτυγχάνουν καλύτερη απόδοση από την τύχη.

- **Βαρυτικά Δεδομένα:** Ο Adaboost αναθέτει βάρη σε κάθε παράδειγμα εκπαίδευσης, υποδηλώνοντας τη σημασία του κατά τη διάρκεια της διαδικασίας μάθησης. Αρχικά, όλα τα δεδομένα ανατίθενται ίσα βάρη. Ωστόσο, καθώς προχωρά ο αλγόριθμος, τα βάρη προσαρμόζονται με βάση την απόδοση των αδύναμων μαθητών.
- **Διαδικασία Εκπαίδευσης:** Ο αλγόριθμος εκπαιδεύει επαναληπτικά τους αδύναμους μαθητές σε νέες εκδόσεις των δεδομένων εκπαίδευσης. Σε κάθε επανάληψη, ο αδύναμος μαθητής εκπαιδεύεται για να ελαχιστοποιήσει τον βαρυτικό σφάλμα, όπου τα βάρη αντικατοπτρίζουν τη δυσκολία των δειγμάτων.
- **Συνδυασμένη Ψηφοφορία με βάρη:** Η τελική πρόβλεψη του συνόλου καθορίζεται λαμβάνοντας υπόψη τα βάρη των αδύναμων μαθητών και τις ατομικές τους προβλέψεις. Όσο μεγαλύτερο είναι το βάρος ενός αδύναμου μαθητή, τόσο μεγαλύτερη είναι η επιρροή της πρόβλεψής του στην τελική απόφαση.

3. Εξήγηση του αλγορίθμου Adaboost

3.1 Ιστορική Εξέλιξη του Adaboost

Ο αλγόριθμος Adaboost, συντομογραφία του Adaptive Boosting, εισήχθη από τους Yoav Freund και Robert Schapire το 1996. Το άρθρο τους "Experiments with a New Boosting Algorithm" (Πειράματα με έναν Νέο Αλγόριθμο Boosting) περιγράφει τις βασικές αρχές του Adaboost και αποδεικνύει την αποτελεσματικότητά του στη βελτίωση της απόδοσης των αδύναμων μαθητών. Ο Adaboost απέκτησε γρήγορα δημοτικότητα στην

κοινότητα της μηχανικής μάθησης και έγινε ένας από τους πλέον επιδραστικούς αλγορίθμους boosting.

3.2 Τα βήματα του Adaboost

Ο αλγόριθμος Adaboost στοχεύει στη συστηματική κατασκευή ενός συνόλου από αδύναμους μαθητές (weak learners) με την προσαρμογή των βαρών των train set βάσει της απόδοσής τους στην κατηγοριοποίηση. Ο αλγόριθμος ακολουθεί τα εξής βασικά βήματα:

1. **Αρχικοποίηση των βαρών των δειγμάτων:**
   - Ανάθεση ίσων βαρών σε όλα τα δείγματα εκπαίδευσης ($w = \frac{1}{N}$)
2. **Επανάληψη για Τ γύρους:**
   a) Εκπαίδευση ενός αδύναμου μαθητή:
      - ✓ Επιλογή ενός αδύναμου μαθητή, όπως ένα απλό δέντρο απόφασης, που έχει μια απόδοση ελαφρώς καλύτερη από την τυχαία επιλογή
      - ✓ Εκπαίδευση του αδύναμου μαθητή χρησιμοποιώντας τα δεδομένα εκπαίδευσης, λαμβάνοντας υπόψη τα βάρη των δειγμάτων που είχαν ανατεθεί στον προηγούμενο γύρο.
      - ✓ Ο στόχος του αδύναμου μαθητή είναι να ελαχιστοποιήσει τον συνολικό σφάλμα, όπου τα βάρη αντικατοπτρίζουν τη δυσκολία των δειγμάτων ($Total\ Eror = sum\ of\ weights$)

   b) Υπολογισμός του βάρους του αδύναμου μαθητή:
      - ✓ Υπολογισμός του βάρους του αδύναμου μαθητή βάσει της απόδοσής του ($amount\ of\ say\ = \frac{1}{2} * \log(\frac{1-Total\ Error}{Total\ Error})$)
      - ✓ Το βάρος υποδεικνύει τη συμβολή του αδύναμου μαθητή στο σύνολο.

   c) Ενημέρωση των βαρών των δειγμάτων:

- ✓ Αύξηση των βαρών των λανθασμένα κατηγοριοποιημένων δειγμάτων, δίνοντας τους μεγαλύτερη σημασία για τις επόμενες επαναλήψεις ($NewW = weight * e^{amount\ of\ say}$)
- ✓ Μείωση των βαρών των σωστά κατηγοριοποιημένων δειγμάτων, μειώνοντας τη σημασία τους ($NewW = weight * e^{-amount\ of\ say}$)

3. **Συνδυασμός των αδύναμων μαθητών:**
   - Συνδυασμός των προβλέψεων των αδύναμων μαθητών χρησιμοποιώντας συνδυασμένη απόφαση με βάση την ψηφοφορία με βάρη ή τον συνδυασμένο μέσο όρο για την τελική πρόβλεψη.

3.3 Παράδειγμα του αλγορίθμου Adaboost

Ας θεωρήσουμε ένα πρόβλημα δυαδικής κατηγοριοποίησης όπου έχουμε ένα σύνολο δεδομένων με δύο κατηγορίες, τις θετικές (+1) και τις αρνητικές (-1). Αρχικά, όλα τα δεδομένα έχουν ίσα βάρη. Ο Adaboost ξεκινά εκπαιδεύοντας έναν αδύναμο μαθητή χρησιμοποιώντας τα δεδομένα με τα βάρη, όπου ο αδύναμος μαθητής στοχεύει στην ελαχιστοποίηση του συνολικού σφάλματος που λαμβάνει υπόψη τα βάρη των δειγμάτων. Αυτή η διαδικασία επαναλαμβάνεται για έναν καθορισμένο αριθμό επαναλήψεων, κατασκευάζοντας ένα σύνολο αδύναμων μαθητών. Τελικά, οι προβλέψεις του συνόλου συνδυάζονται χρησιμοποιώντας συνδυασμένη ψήφο με βάρη για την τελική πρόβλεψη.

4. Η συνεισφορά στην Μηχανική Μάθηση

Ο AdaBoost έχει συνεισφέρει αρκετά στον τομέα της μηχανικής μάθησης. Πιο αναλυτικά, οι κύριες συνεισφορές του AdaBoost είναι:

- **Βελτιωμένη Απόδοση:** Ο AdaBoost έχει επιδείξει εκπληκτικές βελτιώσεις απόδοσης σε σύγκριση με μεμονωμένους αδύναμους μαθητές. Συνδυάζοντας πολλούς αδύναμους μαθητές σε ένα σύνολο, ο αλγόριθμος επιτυγχάνει υψηλότερη ακρίβεια και γενίκευση, καθιστώντας το ένα πολύτιμο εργαλείο σε διάφορους τομείς.
- **Αντιμετώπιση Σύνθετων Προτύπων:** Η ικανότητα του AdaBoost να αντιμετωπίζει πολύπλοκα πρότυπα και μη γραμμικά όρια απόφασης έχει επηρεάσει την επίλυση δύσκολων προβλημάτων ταξινόμησης και παλινδρόμησης. Η ευελιξία και η προσαρμοστικότητα του συνόλου επιτρέπουν την ανίχνευση περίπλοκων σχέσεων στα δεδομένα, οδηγώντας σε βελτιωμένες δυνατότητες πρόβλεψης.
- **Ανθεκτικότητα στα δεδομένα με θόρυβο:** Η επαναληπτική διαδικασία προσαρμογής των βαρών των κακώς ταξινομημένων περιπτώσεων καθιστά το AdaBoost ανθεκτικό στα δεδομένα με θόρυβο. Εστιάζοντας σε αυτά, ο AdaBoost μπορεί να φιλτράρει αποτελεσματικά το θόρυβο και να βελτιώσει τη συνολική ανθεκτικότητα του συνόλου.
- **Επιλογή και Σημασία Χαρακτηριστικών:** Η δυνατότητα επιλογής χαρακτηριστικών του AdaBoost του επιτρέπει να αναγνωρίζει και να δίνει προτεραιότητα σε σημαντικά χαρακτηριστικά, μειώνοντας την διάσταση και επικεντρώνοντας στις πλέον σχετικές πληροφορίες. Αυτό δεν βελτιώνει μόνο την αποδοτικότητα, αλλά ενισχύει επίσης την ερμηνευσιμότητα του μοντέλου, αναδεικνύοντας τα σημαντικά χαρακτηριστικά.

5. Πρακτικές εφαρμογές

5.1 Ταξινόμηση εικόνας με τον Adaboost

Ο AdaBoost έχει αναπτύξει εκτεταμένες εφαρμογές στην ταξινόμηση εικόνας. Έχει χρησιμοποιηθεί σε διάφορα προβλήματα μηχανικής όρασης, όπως ανίχνευση προσώπου, αναγνώριση αντικειμένων και κατάτμηση εικόνας. Η ικανότητα του AdaBoost να χειριστεί πολύπλοκα μοτίβα και η ανθεκτικότητά του στο θόρυβο το καθιστούν κατάλληλο για την ταξινόμηση μιας εικόνας. Για παράδειγμα, η ανίχνευση προσώπου Viola-Jones, μία από τις πιο γνωστές εφαρμογές του AdaBoost, χρησιμοποιεί μια κατάταξη ταξινομητών AdaBoost για την ανίχνευση προσώπων σε πραγματικό χρόνο.

5.2 Ανίχνευση αντικειμένων

Ο AdaBoost έχει επίσης χρησιμοποιηθεί ευρέως σε εφαρμογές ανίχνευσης αντικειμένων. Στην ανίχνευση αντικειμένων, στόχος είναι η αναγνώριση και ο εντοπισμός συγκεκριμένων αντικειμένων εντός μιας εικόνας. Μέθοδοι βασισμένες στον AdaBoost, όπως ο δημοφιλής ανιχνευτής χαρακτηριστικών Histograms of Oriented Gradients (HOG) σε συνδυασμό με ταξινομητές AdaBoost, έχουν εφαρμοστεί με επιτυχία σε ανίχνευση πεζών, ανίχνευση οχημάτων και άλλα προβλήματα ανίχνευσης αντικειμένων. Η ικανότητα του AdaBoost να μάθει διακριτικά χαρακτηριστικά και να χειριστεί μεταβαλλόμενες εμφανίσεις αντικειμένων το καθιστούν ένα χρήσιμο εργαλείο στην ανίχνευση αντικειμένων.

5.3 Φιλτράρισμα ανεπιθύμητων μηνυμάτων (spam)

Ο AdaBoost έχει χρησιμοποιηθεί αποτελεσματικά σε συστήματα φιλτραρίσματος ανεπιθύμητων μηνυμάτων (spam). Εκπαιδεύοντας ταξινομητές AdaBoost σε ένα μεγάλο αριθμό δειγμάτων ηλεκτρονικού ταχυδρομείου, περιλαμβανομένων τόσο των ανεπιθύμητων μηνυμάτων όσο και των μη ανεπιθύμητων μηνυμάτων, ο αλγόριθμος μπορεί να μάθει να διαχωρίζει τα μηνύματα ανάμεσα σε ανεπιθύμητα και νόμιμα. Η ικανότητα επιλογής χαρακτηριστικών του AdaBoost, σε συνδυασμό με την ικανότητά του να χειριστεί υψηλής διάστασης δεδομένα, το καθιστούν κατάλληλο για εφαρμογές φιλτραρίσματος ανεπιθύμητων μηνυμάτων.

5.4 Βιοπληροφορική

Στη βιοπληροφορική, ο AdaBoost έχει χρησιμοποιηθεί για διάφορες εφαρμογές, συμπεριλαμβανομένης της πρόβλεψης δομής πρωτεϊνών, της ανάλυσης έκφρασης γονιδίων και της ταξινόμησης αλληλουχιών DNA. Η ικανότητα του AdaBoost να χειριστεί υψηλής διάστασης και θορυβώδη δεδομένα, σε συνδυασμό με τις ιδιότητες επιλογής χαρακτηριστικών, το καθιστούν κατάλληλο για την ανάλυση βιολογικών συνόλων δεδομένων και την εξαγωγή σχετικών πληροφοριών.

6. Εφαρμογή (Python)

Όπως προαναφέρθηκε, ο αλγόριθμος Adaboost έχει συμβάλλει καθοριστικά στον τομέα της βιοπληροφορικής και της βιοτεχνολογίας. Έτσι, δημιουργήθηκε μια εφαρμογή, γραμμένη σε γλώσσα προγραμματισμού Python, η οποία επιδεικνύει τη χρήση του αλγορίθμου AdaBoost για εργασίες παλινδρόμησης στο σύνολο δεδομένων διαβήτη. Πιο συγκεκριμένα, αυτή η εφαρμογή φορτώνει το σύνολο δεδομένων του διαβήτη, το χωρίζει σε train sets και test sets, εκπαιδεύει έναν ταξινομητή AdaBoost στα δεδομένα εκπαίδευσης, κάνει προβλέψεις στα test data , υπολογίζει την ακρίβεια του μοντέλου και εκτυπώνει τα αποτελέσματα.

```
#Import necessary libraries
import numpy as n
from sklearn.ensemble import AdaBoostRegressor
from sklearn.datasets import load_diabetes
from sklearn.model_selection import train_test_split

#Load the diabetes dataset
diabetes = load_diabetes()

#Split the dataset into training and test sets
X_train, X_test, y_train, y_test = train_test_split(diabetes.data,
diabetes.target, test_size=0.2, random_state=42)

#Create an AdaBoost regressor
reg = AdaBoostRegressor(random_state=42)

#Fit the model
reg.fit(X_train, y_train)

#Predict the target values
y_pred = reg.predict(X_test)

#Calculate the accuracy of the model
acc = reg.score(X_test, y_test)

#Print the accuracy
print("Accuracy: {}\n".format(acc))
```

7. Συμπεράσματα και μελλοντικές Κατευθύνσεις

7.1 Σύνοψη των βασικών ευρημάτων

Ο αλγόριθμος AdaBoost πρόκειται για έναν ισχυρό αλγόριθμο boosting που κατασκευάζει ένα σύνολο από αδύναμους μαθητές για να βελτιώσει την απόδοση της ταξινόμησης και της παλινδρόμησης. Τα κύρια ευρήματά του σύμφωνα με ό,τι αναφέρθηκε προηγουμένως είναι:

- Ο AdaBoost είναι ένας αποτελεσματικός αλγόριθμος boosting που αξιοποιεί τα πλεονεκτήματα πολλαπλών αδύναμων μαθητών για να βελτιώσει την απόδοση πρόβλεψης.
- Ο αλγόριθμος προσαρμόζει επαναληπτικά τα βάρη των δειγμάτων και συνδυάζει τις προβλέψεις των αδύναμων μαθητών χρησιμοποιώντας είτε κατά κύριο λόγο την ψηφοφορία με βάρη είτε τον μέσο όρο.
- Ο AdaBoost μπορεί να επεκταθεί για να αντιμετωπίσει προβλήματα πολυκατηγορικής ταξινόμησης, παλινδρόμησης και επιλογής χαρακτηριστικών.
- Η κατανόηση των θεωρητικών βάσεων του AdaBoost, όπως ο συμβιβασμός ανάμεσα σε διακύμανση και απόκλιση, αλλά και η θεωρία της στατιστικής μάθησης, παρέχει εισηγήσεις για τη συμπεριφορά του και τη γενίκευσή του.
- Ο AdaBoost έχει εφαρμοστεί με επιτυχία σε διάφορες πραγματικές εφαρμογές, όπως η ταξινόμηση εικόνας, η ανίχνευση αντικειμένων, το φιλτράρισμα ανεπιθύμητων μηνυμάτων, η βιοπληροφορική και η ανίχνευση ανωμαλιών.

7.2 Μελλοντικές Κατευθύνσεις Έρευνας

Παρά την επιτυχία του AdaBoost, υπάρχουν ακόμα πεδία για περαιτέρω έρευνα και εξερεύνηση σε αλγορίθμους boosting. Ορισμένες πιθανές μελλοντικές κατευθύνσεις έρευνας περιλαμβάνουν:

- Την ανάπτυξη νέων αδύναμων μαθητών που είναι προσαρμοσμένοι σε συγκεκριμένους τομείς ή χαρακτηριστικά δεδομένων για να ενισχύσουν την απόδοση του AdaBoost.
- Την εξερεύνηση στρατηγικών συνόλου πέραν της ψηφοφορίας με βάρη ή του μέσου όρου, όπως οι μέθοδοι συνδυασμού με βάση την κλίση ή προσεγγιστικές μεθόδους μάθησης.
- Την έρευνα τρόπων βελτίωσης της ανθεκτικότητας του AdaBoost έναντι εκτροπών και θορύβου στα δεδομένα.
- Την επέκταση του AdaBoost για την αντιμετώπιση ελλιπών δεδομένων.
- Τη μελέτη της ερμηνευσιμότητας των μοντέλων AdaBoost και την ανάπτυξη τεχνικών για την εξαγωγή σημαντικών πληροφοριών από τις συνθέτες προβλέψεις.

Βιβλιογραφία

1. Wikipedia, Adaboost, https://en.wikipedia.org/wiki/AdaBoost
2. A survey on object detection in optical remote sensing images, ScienceDirect, https://www.sciencedirect.com/science/article/abs/pii/S0924271616300144?fr=RR-2&ref=pdf_download&rr=7ca7b5a42ef3fd5e
3. Impact of Cross-Validation on Machine Learning Models for Early Detection of Intrauterine Fetal Demise, MDPI, https://www.mdpi.com/2075-4418/13/10/1692
4. Homogeneous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data, NCIB, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9955103/
5. Importance analysis of psychosociological variables in frailty syndrome in heart failure patients using machine learning approach, Scientific Reports, https://www.nature.com/articles/s41598-023-35037-3
6. Step-by-Step Guide to Implement Machine Learning VI – AdaBoost, Code Project, https://www.codeproject.com/Articles/4114375/Step-by-Step-Guide-to-Implement-Machine-Learning