ΜΑΘΗΣΗ ΜΗΧΑΝΗΣ, ΚΑΝΟΝΤΑΣ ΔΙΑΓΝΩΣΗ ΣΕ ΑΣΘΕΝΕΙΣ ΜΕ ΚΑΡΚΙΝΟ ΤΟΥ ΜΑΣΤΟΥ

Πετρίδης Παναγιώτης
Σαποζνίκοβ Οδυσσέας
Σιταρίδης Ιωάννης(Επιτηρητής)

ΠΕΡΙΛΗΨΗ

Έχοντας μάθει για την μάθηση μηχανής και τις εκπληκτικές τις δυνατότητες αποφασίσαμε να εφαρμόσουμε αυτά που μάθαμε σε κάτι το οποίο δεν μπορεί να θεωρηθεί ασήμαντο. Χρησιμοποιήσαμε μάθηση μηχανής προκειμένου να προβλέψουμε εάν κάποιος ασθενής με καρκίνο του μαστού έχει καλοήθη η κακοήθη όγκο. Ένα από τα κυριότερα προβλήματα που αντιμετωπίσαμε ήταν η απόδοση του μοντέλου ανάλυσης, καθώς επιλέξαμε κάτι απλό, το μοντέλο *Adaline*. Καταφέραμε να ξεπεράσουμε αυτό το πρόβλημα αλλάζοντας την συνάρτηση ενεργοποίησης του μοντέλου από μία απλή πρόσθεση στην Υπερβολική Εφαπτομένη.

1 Εισαγωγή

Το μοντέλο μας όπως αναφέραμε είναι μία παραλλαγή του μοντέλου *Adaline* το ποίο προκειμένου να κάνει τις ανάλογες προβλέψεις δέχεται κάποια στοιχεία σαν δεδομένα δίνει κάποιο συγκεκριμένο βάρος στο καθένα με βάση το πόσο σημαντικό είναι και καταφέρνει έτσι βρίσκοντας τα σωστά βάρη να κάνει κάποιες καλές προβλέψεις.Σύμφωνα με τις μετρήσεις μας καταφέρνει να κάνει προβλέψεις με 95% επιτυχία και μάλιστα στο 1 δέκατο του χρόνου που χρειάζεται το μοντέλο *Adaline*.

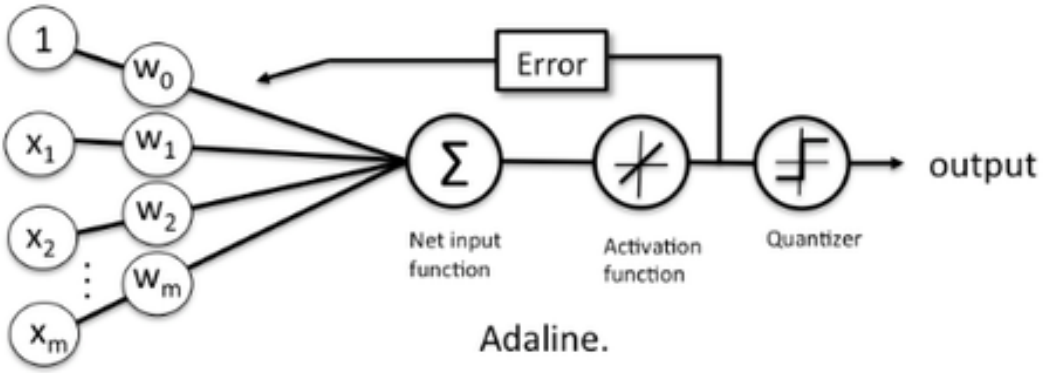
2 Πως λειτουργεί το μοντέλο

2.1 Τα Δεδομένα

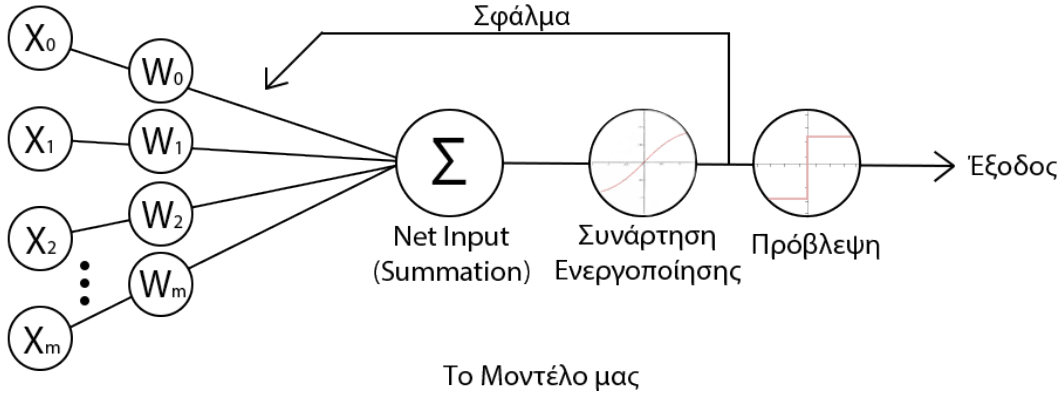
Σε αυτό το κομμάτι θα πρέπει να αναφέρουμε πως όλα τα δεδομένα που χρησιμοποιήθηκαν στην πραγματοποίηση αυτης της εργασίας συλλέχθηκαν και ανήκουν πλήρως στον Dr. WIlliam H. Wolberg, University of Wisconsin Hospitals. Madison, Wisconsin, USA .Με δωρητή Olvi Mangasarian (mangasarian@ cs.wisc.edu) και παραλήπτη by David W. Aha (aha@cs.jhu.edu). Τα δεδομένα μας δίνουν τις εξής πληροφορίες: Αριθμό ασθενή, πάχος συστάδας, ομοιομορφία του πάχους των κυττάρων, ομοιομορφία του σχήματος των κυττάρων, οριακή πρόσφυση, ενιαίο μέγεθος επιθηλιακών κυττάρων ,Bare Nuclei, το πόσο ήπια είναι η χρωματίνη, κανονικούς πυρινίσκους, μιτώσις και τέλος αν ο όγκος είναι καλοήθης η κακοήθης(Τα δεδομένα δεν είναι μεταφρασμένα να απόλυτη ακρίβεια καθώς μερικοί ιατρικοί όροι είναι δύσκολο να μεταφραστούν στα ελληνικά). Τα δεδομένα μας, αν και δεν συνηθίζετε, είναι όλα στην ίδια κλίμακα από 1-10 πράγμα που διευκολύνει την ομαδοποίηση τους. Όλα τα δεδομένα διαβάζονται από ένα αρχείο και αποθηκεύονται σε $C++vectors$.Επίσης για να μπορούμε να δούμε τα αποτελέσματα του μοντέλου σε ασθενείς που δεν έχει ξαναδεί τα χωρίσαμε σε τμήματα των 57.2% και 42.8% από τα οποία το 57.2% χρησιμοποιήσαμε για να προσαρμόσουμε το μοντέλο μας και το 42.8% ως τεστ για να δούμε τα αποτελέσματα. Αυτή είναι μια συνηθισμένη τακτική στο πεδίο μάθησης μηχανής καθώς δεν μπορούμε να κρίνουμε ένα μοντέλο μόνο πάνω σε δεδομένα που ήδη έχει δει τις απαντήσεις αλλά στην απόδοσή του σε δεδομένα που δεν έχει ξαναδεί.

2.2 Κάνοντας μια πρόβλεψη

Πώς ακριβώς όμως μπορούμε να κάνουμε προβλέψεις; Ας παρατηρήσουμε την παρακάτω εικόνα:



Το μοντέλο, *Adaline*



Όπως μπορούμε να δούμε από την εικόνα η διαδικασία είναι πολύ απλή και όχι μόνο αυτό αλλά προσπαθεί και να μιμηθεί τους νευρώνες που έχουμε στο μυαλό μας. Τα δεδομένα είναι στα αριστερά και το καθένα από αυτά πολλαπλασιάζετε με ένα ξεχωριστό βάρος το οποίο θα μπορούσαμε να πούμε πως καθορίζει το πόσο σημαντικό είναι αυτό το στοιχείο. Για παράδειγμα το μοντέλο μπορεί να φανταστεί πώς το μέγεθος του όγκου είναι πολύ σημαντικό ενώ η ομοιομορφία του σχήματος των κυττάρων όχι τόσο Στην συνέχεια προσθέτει όλα τα γινόμενα των δεδομένων με τα αντίστοιχα βάρη και την τιμή που παίρνει την βάζει σε μία συνάρτηση(Συνάρτηση Ενεργοποίησης) η οποία στο μοντέλο*Adaline*είναι $y = x$ ενώ στο δικό μας μοντέλο έχουμε την υπερβολική εφαπτομένη. Επίσης προκειμένου να πάρουμε μία δυαδική τιμή (ναι η όχι) θεωρούμε ότι αν το $\hat{y}$ είναι μεγαλύτερο του μηδενός τότε είναι 1 αλλιώς -1. Παρακάτω ακολουθούν οι μαθηματικοί τύποι
με $x_0 = 1$ (πάντα).

$$\hat{y} = tanh\left(\sum_{i=0}^{m} w_i \cdot x_i\right)$$

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

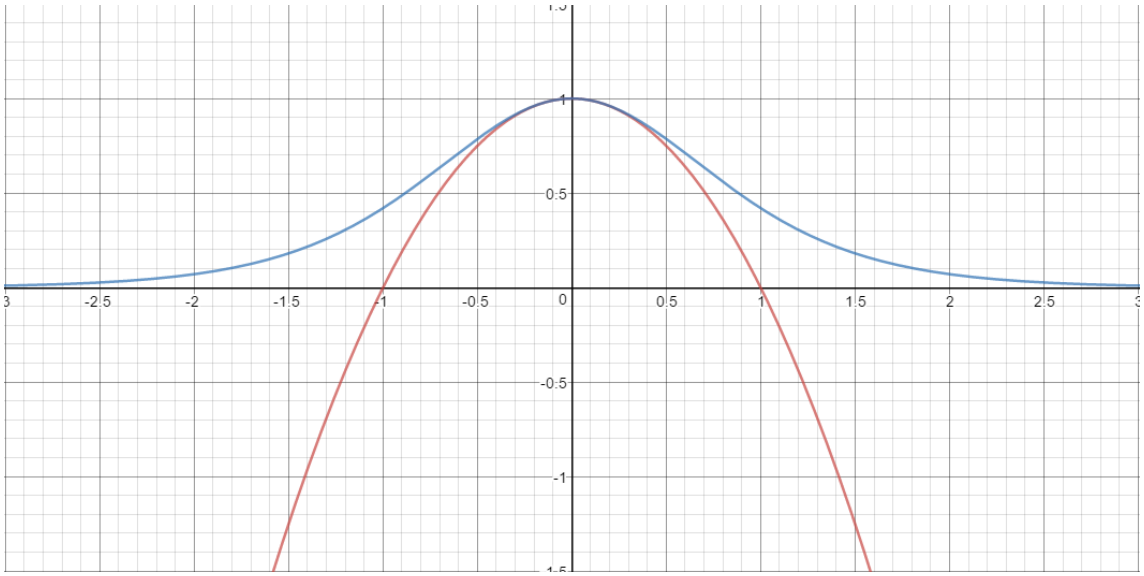
2.3 Βρίσκοντας τα σωστά βάρη

Πώς όμως μπορούμε να βρούμε τα σωστά βάρη; Πρώτα θα πούμε τον τρόπο πούμε για το μοντέλο *Adaline* και στην συνέχεια για το δικό μας μοντέλο που είναι μία παραλλαγή του. Ο μαθηματικός

τύπος είναι ο εξής:

$$w_i = w_i + \eta \cdot (y - \hat{y}) \cdot x_i$$

Στην πραγματικότητα μπορούμε να φανταστούμε ότι το παραπάνω γινόμενο σαν να πολλαπλασιάζετε με την παράγωγό της συνάρτησης $y = x$ είναι πάντα 1. Επιπρόσθετα θα παρατηρήσουμε ότι υπάρχει και μια μεταβλητή η. Αυτή η μεταβλητή ορίζει το πόσο μεγάλα βήματα θέλουμε να κάνουμε προς τις κατάλληλες τιμές. Αν είναι πολύ μικρή τότε το μοντέλο μας θα αργήσει να φτάσει στα σωστά βάρη ενώ αν είναι πολύ μεγάλη θα τα προσπεράσει. Άλλα στην δική μας περίπτωση η συνάρτηση ενεργοποίησης είναι $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ πράγμα που σημαίνει ότι πρώτα θα πρέπει να βρούμε την παράγωγό της υπερβολικής εφαπτομένης, η οποία είναι $1 - tanh^2(x)$. Υπάρχει όμως κάτι ακόμα που μπορούμε να κάνουμε για να βελτιώσουμε την απόδοση του μοντέλου και αυτό είναι αντί να χρειάζεται να υπολογίσουμε την υπερβολική εφαπτομένη κάθε φορά να προσπαθήσουμε να την προσεγγίσουμε με μία άλλη συνάρτηση και αυτή είναι η $y = 1 - x^2$. Όχι μόνο αυτό αλλά με αυτή την συνάρτηση μπορούμε επίσης να φτάσουμε πιο γρήγορα στο επιθυμητό αποτέλεσμα καθώς έχει πιο ακραία αλλαγή κλήσης όσο πιο κοντά είναι στο 1, -1. Ακολουθεί ένα γράφημα για να μπορέσουμε να καταλάβουμε πόσο κοντά είναι στην παράγωγο της υπερβολικής εφαπτομένης:



Με μπλε είναι η παράγωγός της υπερβολικής εφαπτομένης ενώ με κόκκινο η $y = 1 - x^2$(το γράφημα είναι από τον ισότοπο *Demsos.com*)

Έτσι με αυτές τις αλλαγές έχουμε:

$$w_i = w_i + \eta \cdot (y - \hat{y}) \cdot (1 - \hat{y}^2) \cdot x_i$$

3 Αποτελέσματα

Με το μοντέλο αυτό καταφέραμε να έχουμε αποτελέσματα με έως και 95% ακρίβεια στα δεδομένα για τεστ ενώ 98% στα δεδομένα για την προσαρμογή του μοντέλου. Και μάλιστα σε λιγότερα από 500 περάσματα των δεδομένων προσαρμογής του μοντέλου. Σε σύγκριση με το απλό μοντέλο *Adaline* το οποίο χρειάστηκε πάνω από 5000 (μέσο όρο) περάσματα για να φτάσει παρόμοια αποτελέσματα.Πρακτικά αυτο συμαίνει οτι πλέον ασθενείς οι οποίοι δεν έχουν πρόσβαση σε κάποιο γιατρό μπορούν να μάθουν αν ο όγκος που έχουν είναι καλοήθης η κακοήθης μέσα σε κλάσματα του δευτερολέπτου με πολύ υψηλή ακρίβεια.Παρά τα θεαματικά αυτά αποτελέσματα το μοντέλο αυτό δεν είναι τέλειο ακόμα γίνονται κάποια λάθη λόγο της μικρής ποσότητας δεδομένων που μας παρέχονται. Έτσι ως τρόποι βελτίωσής του θα μπορούσαν να είναι περισσότερα δεδομένα.

Πηγές

[1] *The University of Sydney* ($http://www.cs.usyd.edu.au/\ comp4302/ann3-6s.pdf$)

[2] *Desmos Online Graphing Calculator* ($https://www.desmos.com/$)

[3] *Wikipedia Hyperbolic function* ($https://en.wikipedia.org/wiki/Hyperbolic_function$)

[4] *Wikipedia ADALINE* ($https://en.wikipedia.org/wiki/ADALINE$)