Απόδοση κατηγοριοποίησης

Απόδοση κατηγοριοποίησης

Ποιος είναι ο καλύτερος αλγόριθμος κατηγοριοποίησης; Για να αξιολογήσουμε την ποιότητα της κατηγοριοποίησης, που πετυχαίνει ο κάθε κατηγοριοποιητής θα πρέπει να χρησιμοποιήσουμε κάποια μέτρα ή δείκτες τα οποία μετρούν είτε την συσχέτιση ή την ομοιότητα ή ανομοιότητα του διαχωρισμού που γίνεται στα δεδομένα. Οι μέθοδοι κατηγοριοποίησης μπορούν να αξιολογηθούν με βάση τα παρακάτω μέτρα:

- **Ακρίβεια πρόβλεψης (accuracy)**. Αναφέρεται στην ικανότητα του μοντέλου να προβλέπει την κατηγορία μιας νέας περίπτωσης
- **Ταχύτητα (speed)**: Σχετίζεται με την πολυπλοκότητα της μεθόδου και το υπολογιστικό κόστος που αυτή συνεπάγεται (συμπεριλαμβανομένου την παραγωγή και τη χρήση του μοντέλου)
- **Ανθεκτικότητα (rebustness)**: Αναφέρεται στην ικανότητα των μεθόδων να πραγματοποιήσουν σωστές προβλέψεις όταν τα δεδομένα είναι ελλιπή ή είναι δεδομένα με θόρυβο.
- **Επεκτασιμότητα (scalability)**: Αναφέρεται στην ικανότητα των μεθόδων να χειριστούν πολύ μεγάλα σύνολα δεδομένων.
- **Ερμηνευσιμότητα (interpretability)**: Είναι η ικανότητα της μεθόδου να παράγει μοντέλα, τα οποία είναι κατανοητά από τον άνθρωπο.

Το σημαντικότερο μέτρο απόδοσης είναι η ακρίβεια και αυτό πρόκειται να αναλύσουμε στην παρούσα εργασία. Να επισημάνουμε όμως πως παρότι είναι το πιο σημαντικό μέτρο δε θα πρέπει να υπολογίζεται ανεξάρτητα από τα υπόλοιπα μέτρα.

Επίδοση κατηγοριοποίησης

Για την εκτίμηση της επίδοσης των αλγορίθμων κατηγοριοποίησης, μπορούν να χρησιμοποιηθούν παραδοσιακές προσεγγίσεις όπως είναι η δημιουργία ενός πίνακα σύγχυσης (confusion matrix) η οποία είναι πολύ πληροφοριακή για όλα τα είδη αλγορίθμων κατηγοριοποίησης. Η αξιολόγηση της κατηγοριοποίησης βασίζεται στον αριθμό των δειγμάτων του συνόλου ελέγχου που προβλέπονται σωστά ή όχι από τον κατηγοριοποιητή. Αυτός ο αριθμός τοποθετείται σε έναν πίνακα σύγχυσης. Οι στήλες του πίνακα σύγχυσης αντιστοιχούν στις προβλεπόμενες κλάσεις εξόδου, ενώ οι γραμμές στις πραγματικές κλάσεις.

		Προβλεπόμενη κλάση	
		Κλάση +1	Κλάση -1
Πραγματική Κλάση	Κλάση +1	TP	FN
	Κλάση -1	FP	TN

Παράδειγμα του πίνακα σύγχυσης

Σε ένα πρόβλημα δύο κλάσεων (-1,1) ο πίνακας αυτός διαμορφώνεται όπως ο Πίνακας 3.1. Κάθε κελί στον πίνακα δείχνει το πλήθος των δειγμάτων από την προβλεπόμενη κλάση i που ανήκουν ή όχι στην πραγματική κλάση j . Συμβολίζουμε ως TP (true positive) το πλήθος των αληθώς θετικών δειγμάτων, FN (false negative) το πλήθος των αληθώς αρνητικών δειγμάτων, FP (false positive) το πλήθος των ψευδώς αρνητικών δειγμάτων και TN (true negative) το πλήθος των ψευδώς θετικών δειγμάτων.

Για να εκτιμήσουμε τις επιδόσεις ενός αλγορίθμου, εισάγουμε την αναγκαία ορολογία.

- Στο κελί TP απαριθμείται το πλήθος των δειγμάτων από την κλάση +1 που σωστά με την πρόβλεψη τοποθετήθηκαν στην κλάση +1.
- Στο κελί FP απαριθμείται το πλήθος των δειγμάτων από την κλάση +1 που εσφαλμένα με την πρόβλεψη τοποθετήθηκαν στην κλάση -1.
- Στο κελί FN απαριθμείται το πλήθος των δειγμάτων από την κλάση -1 που εσφαλμένα με την πρόβλεψη τοποθετήθηκαν στην κλάση +1.
- Στο κελί TN απαριθμείται το πλήθος των δειγμάτων από την κλάση -1 που σωστά με την πρόβλεψη τοποθετήθηκαν στην κλάση -1.

Ακρίβεια και ρυθμός σφάλματος κατηγοριοποίησης

Οι δύο πιο γνωστοί δείκτες για την αξιολόγηση της κατηγοριοποίησης είναι η ακρίβεια ή ορθότητα (accuracy, AC ή success rate) ή αλλιώς πιστότητα, και το ποσοστό σφάλματος (error rate, ER). Η ακρίβεια είναι το πλήθος των ορθών προβλέψεων προς το σύνολο των δειγμάτων ελέγχου, και ορίζεται ως εξής :

$$AC = \frac{TP+TN}{TP+TN+FP+FN}$$

Ισοδύναμα, η συνολική απόδοση του κατηγοριοποιητή μπορεί να εκφραστεί και με το ποσοστό σφάλματος, που είναι το πλήθος εσφαλμένων προβλέψεων προς το σύνολο εγγραφών, και ορίζεται ως εξής :

$$ER = \frac{FP+FN}{TP+TN+FP+FN}$$

Ορισμένα πρόσθετα μέτρα για τις επιδόσεις ενός κατηγοριοποιητή είναι τα ακόλουθα:

$$SENSITIVITY = \frac{TP}{TP+FN}$$

$$SPECIFICITY = \frac{TN}{TN+FP}$$

$$PRECISION = \frac{TP}{TP+FP}$$

$$NEGATIVE\ PREDICTIVE\ VALUE = \frac{TN}{TN+FN}$$

$$FALSE\ NEGATIVE\ RATE = \frac{FN}{TP+FN}$$

Καμπύλες ROC

Ένα εξίσου ισχυρό μέτρο για την αξιολόγηση της ανά κλάση ακρίβειας ενός αλγορίθμου είναι οι καμπύλες ROC, οι οποίες δείχνουν τη σχέση μεταξύ των αναληθών θετικών και των αληθώς θετικών. Οι καμπύλες ROC σχεδιάζονται σε έναν δυσδιάστατο επίπεδο χώρο.

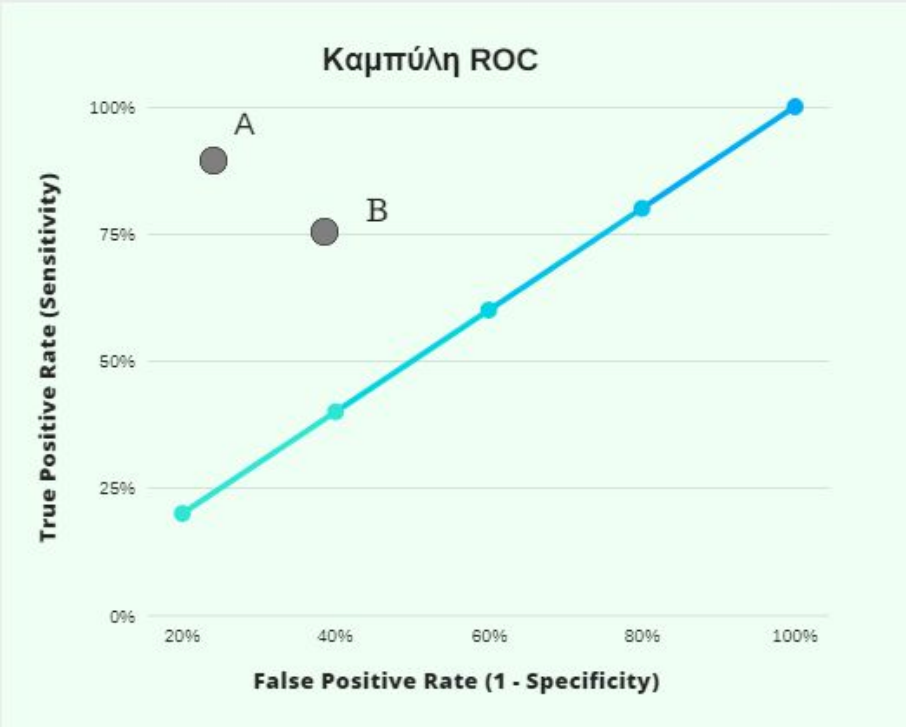
Ο οριζόντιος άξονας εκφράζει το μέγεθος 1-specificity και ονομάζεται False Positive Rate.

$$FPR = 1 - SPECIFICITY = \frac{FP}{TN+FP}$$

Ενώ ο κατακόρυφος άξονας εκφράζει το μέγεθος sensitivity το οποίο ονομάζεται και True Positive Rate.

$$TPR = SENSITIVITY = \frac{TP}{TP+FN}$$

Ουσιαστικά, ο οριζόντιος άξονας εκφράζει το ποσοστό των αρνητικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν λάθος, και ο κατακόρυφος άξονας εκφράζει το ποσοστό των θετικών παρατηρήσεων, οι οποίες κατηγοριοποιήθηκαν σωστά.



Σημεία στον χώρο καμπύλων ROC

Στο Σχήμα 3.3 απεικονίζεται ο δυσδιάστατος χώρος των καμπύλων ROC. Η διαγώνια είναι ένας κατηγοριοποιητής που προβλέπει τυχαία την κλάση. Οι κατηγοριοποιητές που βρίσκονται κάτω από τη διαγώνια γραμμή είναι χειρότεροι από την τυχαία πρόβλεψη. Οι κατηγοριοποιητές που βρίσκονται πάνω από τη διαγώνια γραμμή είναι καλύτεροι από την τυχαία πρόβλεψη.