Αναγνώριση προτύπων - Κριτήρια Επίδοσης Ταξινομητών

Σε πολλές εφαρμογές αναγνώρισης προτύπων αντιμετωπίζουμε ένα πρόβλημα ταξινόμησης 2 κλάσεων όπου συνήθως η μια κλάση περιέχει πρότυπα για τα οποία ισχύει μια κατάσταση ή συνθήκη X (Κλάση 1) ενώ η άλλη κλάση περιέχει πρότυπα όπου αυτή η κατάσταση ή συνθήκη δεν ισχύει (Κλάση 0). Με άλλα λόγια η ετικέτα της κλάσης είναι μια Boolean μεταβλητή $t = 0/1$ $(FALSE/TRUE)$ που αντιστοιχεί στην ύπαρξη ή στη μη-ύπαρξη της κατάστασης ή συνθήκης X.

Για παράδειγμα, σε ένα πρόβλημα αναγνώρισης ασθενών με γρίπη, τα πρότυπα $\mathbf{x}_i$ είναι διανύσματα που περιέχουν χαρακτηριστικά του ασθενούς όπως θερμοκρασία σώματος, ύπαρξη ή μη καταρροής, ύπαρξη ή μη πονόλαιμου, κλπ. Στην περίπτωση αυτή η Κλάση 1 περιέχει πρότυπα που αντιστοιχούν σε ασθενείς με γρίπη ενώ η Κλάση 0 περιέχει πρότυπα που αντιστοιχούν σε ασθενείς χωρίς γρίπη. Ο ταξινομητής δημιουργεί μια συνάρτηση διαχωρισμού $y_i = f(\mathbf{x}_i; \boldsymbol{\theta})$ η οποία παραμετροποιείται από το διάνυσμα $\boldsymbol{\theta}$ και το οποίο *μαθαίνει* κατά την εκπαίδευσή του. Η συνάρτηση $f()$ είναι ένα είδος τεστ στο οποίο υποβάλλουμε το πρότυπο $\mathbf{x}_i$:

- αν το πρότυπο βγει θετικό ($y_\iota = 1$) τότε λέμε ότι το πρότυπο ταξινομήθηκε στην Κλάση 1 (ασθενής με γρίπη)
- αν το πρότυπο βγει αρνητικό ($y_\iota = 0$) τότε λέμε ότι το πρότυπο ταξινομήθηκε στην Κλάση 0 (άτομο χωρίς γρίπη)

Ιδανικά θα έπρεπε, για κάθε πρότυπο $\mathbf{x}_i$, το τεστ να δίνει έξοδο y_ι που να ταυτίζεται με την ετικέτα t_i της κλάσης στην οποία ανήκει το πρότυπο. Δυστυχώς αυτό δεν ισχύει πάντα. Υπάρχουν 4 περιπτώσεις:

(α) Το πρότυπο βγήκε *αρνητικό*, δηλ. $y_\iota = 0$

και όντως *ανήκει στην κλάση 0*, δηλ. $t_i = 0$

Τέτοια πρότυπα καλούνται **true negatives** (πραγματικά αρνητικά)

(β) Το πρότυπο βγήκε *αρνητικό*, δηλ. $y_\iota = 0$

αλλά *ανήκει στην κλάση 1*, δηλ. $t_i = 1$

Τέτοια πρότυπα καλούνται **false negatives** (εσφαλμένα αρνητικά)

(γ) Το πρότυπο βγήκε *θετικό*, δηλ. $y_\iota = 1$

αλλά *ανήκει στην κλάση 0*, δηλ. $t_i = 0$

Τέτοια πρότυπα καλούνται **false positives** (εσφαλμένα θετικά)

(δ) Το πρότυπο βγήκε *θετικό*, δηλ. $y_\iota = 1$

και όντως *ανήκει στην κλάση 1*, δηλ. $t_i = 1$

Τέτοια πρότυπα καλούνται **true positives** (πραγματικά θετικά)

Με δεδομένο ένα σύνολο προτύπων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ αξιολογούμε την επίδοση ενός ταξινομητή χρησιμοποιώντας κάποια χαρακτηριστικά του λεγόμενου **πίνακα σύγχυσης** (**confusion matrix**) που δίνεται παρακάτω:

Πίνακας 1: Πίνακας Σύγχυσης

	Ταξινομήθηκαν στην Κλάση 0 $(\boldsymbol{y = 0})$	**Ταξινομήθηκαν στην Κλάση 1** $(\boldsymbol{y = 1})$
Negatives Ανήκουν στην Κλάση 0 $(\boldsymbol{t = 0})$	True Negatives (TN)	False Positives (FP)
Positives Ανήκουν στην Κλάση 1 $(\boldsymbol{t = 1})$	False Negatives (FN)	True Positives (TP)

Κατ' αρχήν παρατηρούμε ότι στην ιδανική περίπτωση ο πίνακας θα πρέπει να έχει μηδενικά διαγώνια στοιχεία, δηλαδή να μην υπάρχουν False Positives, ούτε False Negatives. Ένα πρώτο μέτρο εκτίμησης της επίδοσης είναι συνεπώς ο λόγος

$$\textbf{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

Προφανώς $0 \leq Accuracy \leq 1$ και επειδή θα θέλαμε ιδανικά, FN=FP=0, η επιθυμητή τιμή είναι $Accuracy = 1$. Το ελάττωμα του κριτηρίου Accuracy είναι η γενικότητά του: δεν προσφέρει πληροφορία σχετικά με την επίδοση του ταξινομητή στο πόσο καλά τα πηγαίνει στην ταξινόμηση των προτύπων της κλάσης 1. Ας επιστρέψουμε στο παράδειγμα με το τεστ γρίπης και ας υποθέσουμε ότι το σύνολο των δεδομένων μας περιέχει 10 άτομα με γρίπη και 990 άτομα χωρίς γρίπη. Ας υποθέσουμε επίσης ότι

$$\text{Από τα 10 άτομα με γρίπη} \begin{cases} 1 \text{ ταξινομείται θετικό } (TP = 1) \\ 9 \text{ ταξινομούνται αρνητικά } (FN = 9) \end{cases}$$

$$\text{Από τα 990 άτομα χωρίς γρίπη} \begin{cases} 1 \text{ ταξινομείται θετικό } (FP = 1) \\ 989 \text{ ταξινομούνται αρνητικά } (TN = 989) \end{cases}$$

Στην περίπτωση αυτή έχουμε

$$Accuracy = \frac{989 + 1}{989 + 1 + 9 + 1} = 0.99 \ (= 99\%)$$

Η τιμή του Accuracy εκ πρώτης όψεως φαίνεται εξαιρετικά υψηλή, πολύ κοντά στην ιδανική τιμή 1. Ωστόσο με μια πιο προσεκτική ματιά ανακαλύπτουμε ότι ο ταξινομητής κάνει πολύ κακή δουλειά στην αναγνώριση των ασθενών με γρίπη (πιάνει μόλις 1 στους 10!). Απλώς τυχαίνει το δείγμα των ατόμων που δεν έχουν γρίπη να είναι πολύ μεγαλύτερο (990) σε σχέση

με αυτούς που έχουν γρίπη (10) κι επειδή τα πηγαίνει πολύ καλά στην ταξινόμηση αυτών που δεν έχουν γρίπη (πιάνει 989/990) επιτυγχάνει υψηλό accuracy το οποίο δυστυχώς είναι παραπλανητικό.

Για τους παραπάνω λόγους, έχουν προταθεί δύο άλλα δημοφιλή κριτήρια που ορίζονται ως εξής:

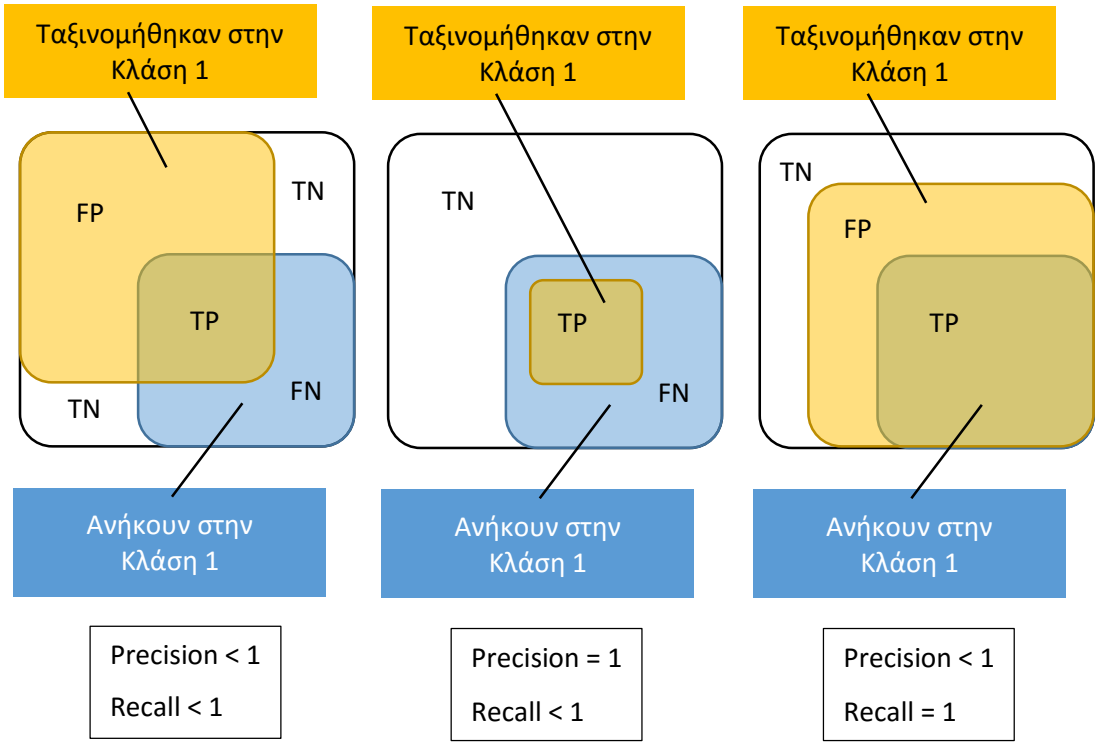
$$\textbf{Precision} = \frac{TP}{POSITIVE} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{\text{Κλάση } 1} = \frac{TP}{TP + FN}$$

Το κριτήριο *Precision* μας δείχνει το ποσοστό των προτύπων που κατηγοριοποιήθηκαν ως θετικά και ανήκουν πράγματι στην Κλάση 1. Για παράδειγμα, αν το τεστ γρίπης έχει $Precision = 0.9$ αυτό σημαίνει ότι 90 στα 100 άτομα που βγήκαν θετικά στο τεστ, έχουν όντως γρίπη.

Το κριτήριο *Recall* δείχνει το ποσοστό των προτύπων που ανήκουν στην κλάση 1 και κατηγοριοποιούνται ως θετικά. Για παράδειγμα, αν το τεστ γρίπης έχει $Recall = 0.7$ αυτό σημαίνει ότι το τεστ βγάζει θετικούς τους 70 στους 100 ασθενείς με γρίπη.

Κατ' αρχήν είναι σαφές ότι οι τιμές και των δύο κριτηρίων κυμαίνονται μεταξύ 0 και 1. Ιδανικά θα θέλαμε να μην υπάρχουν False-Negatives και False-Positives οπότε θα έχουμε $Precision = Recall = 1$. Ωστόσο, αυτό δεν είναι πάντα εφικτό. Επίσης είναι δυνατόν κάποιος ταξινομητής να έχει καλή επίδοση στο κριτήριο *Precision* αλλά όχι στο *Recall*, και αντίστροφα. Αυτό γίνεται σαφές με τη βοήθεια του παρακάτω Σχήματος όπου γίνεται γραφική παράσταση της σημασίας των κριτηρίων αυτών.



Διακρίνουμε τρεις περιπτώσεις:

(Αριστερά) Στην πιο γενική περίπτωση κάποια πρότυπα ανήκουν στις κατηγορίες TP, FN και FP, οπότε $Precision < 1$ και $Recall < 1$.

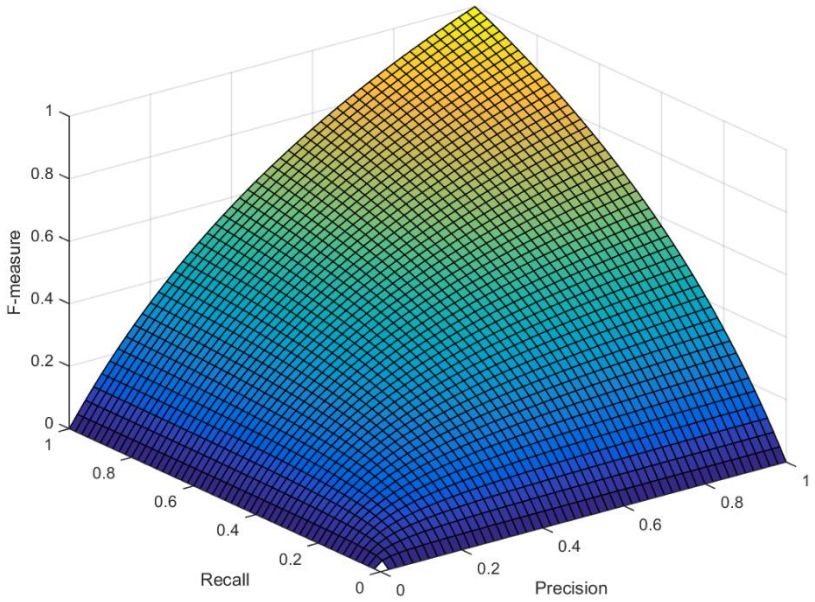
(Στο κέντρο) Στην περίπτωση που όλα τα θετικά πρότυπα ανήκουν στην κλάση 1 τότε έχουμε FP=0 και συνεπώς $Precision = 1$. Ωστόσο μπορούμε να έχουμε σημαντικό αριθμό προτύπων FN οπότε $Recall < 1$.

(Δεξιά) Στην περίπτωση που όλα τα πρότυπα που ανήκουν στην κλάση 1 ταξινομούνται θετικά τότε έχουμε FN=0 και συνεπώς $Recall = 1$. Ωστόσο μπορούμε να έχουμε σημαντικό αριθμό προτύπων FP οπότε $Precision < 1$.

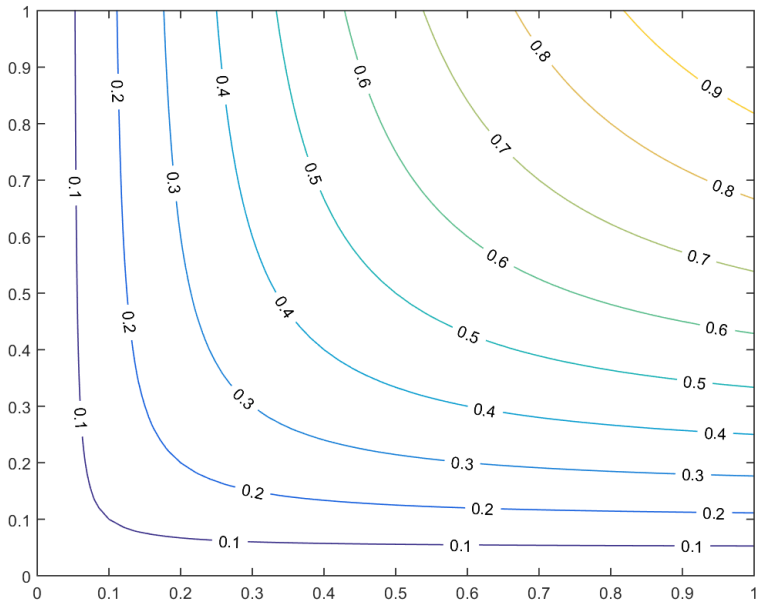
Επειδή τα κριτήρια Precision και Recall δεν αρκούν από μόνα τους για να περιγράψουν την συνολική επίδοση του ταξινομητή συνήθως συνδυάζονται στο κριτήριο **F-measure** (ή αλλιώς **F1-score**) που ορίζεται ως το πηλίκο του γεωμετρικού μέσου προς το αλγεβρικό μέσο όρο των δύο κριτηρίων

$$\text{F-measure} = \frac{Precision \cdot Recall}{(Precision + Recall)/2}$$

Είναι εύκολο να αποδειχθεί μαθηματικά ότι το F-measure κυμαίνεται επίσης μεταξύ 0 και 1 και επιτυγχάνει τη μέγιστη τιμή (1) αν και μόνο αν $Precision = Recall = 1$.



Εικόνα 1. Το F-measure ως συνάρτηση του Precision και του Recall.



Εικόνα 2. Ισοϋψείς καμπύλες του F-measure ως συνάρτηση του Precision και του Recall.

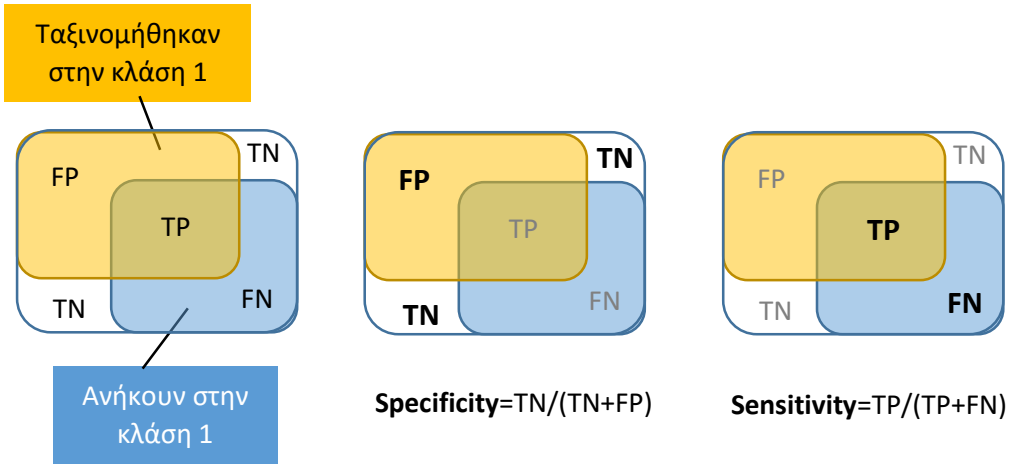
Ένα μειονέκτημα των κριτηρίων *Precision* και *Recall* είναι ότι εστιάζουν αποκλειστικά στην κλάση 1. Παρατηρήστε ότι το πλήθος των True-Negatives δεν εμπλέκεται πουθενά στον ορισμό των κριτηρίων αυτών. Αν και αυτό μπορεί να αρκεί όταν είναι πολύ σημαντική η κλάση 1 (πχ όταν γίνεται διάγνωση μιας ασθένειας) πολλές φορές η σωστή ταξινόμηση στην κλάση 0 είναι εξίσου σημαντική. Έτσι έχουν οριστεί δύο άλλα μέτρα επίδοσης τα οποία δίνουν ίση βαρύτητα και στις δύο κλάσεις:

$$\textbf{Sensitivity} \quad = \frac{TP}{\text{Κλάση } 1} = \frac{TP}{TP + FN} (= Recall = True\ Positive\ Rate)$$

$$\textbf{Specificity} = \frac{TN}{\text{Κλάση } 0} = \frac{TN}{TN + FP} (= True\ Negative\ Rate)$$

Το κριτήριο *Sensitivity* (γνωστό και ως *True-Positive Rate* ή *TPR*) είναι ουσιαστικά το ίδιο κριτήριο με το κριτήριο *Recall*. Το κριτήριο *Specificity* (γνωστό και ως *True-Negative Rate* ή *TNR*) δείχνει σε τι ποσοστό το τεστ ταξινομεί σωστά τα πρότυπα της κλάσης 0. Τα κριτήρια είναι ουσιαστικά ίδια στον μαθηματικό ορισμό τους με εξαίρεση την κλάση στην οποία εστιάζουν : το *Sensitivity* εστιάζει στην κλάση 1 ενώ το *Specificity* στην κλάση 0. Οι τιμές των κριτηρίων πάλι κυμαίνονται μεταξύ 0 και 1 και ιδανική τιμή είναι το 1.

Στο παρακάτω Σχήμα γίνεται διαγραμματική αναπαράσταση των κριτηρίων αυτών.



Όπως και στην περίπτωση *Precision*/*Recall* τα κριτήρια *Specificity*/*Sensitivity* από μόνα τους δεν περιγράφουν πλήρως την επίδοση του ταξινομητή οπότε μπορούμε να δημιουργήσουμε ένα συνδυαστικό κριτήριο με βάση αυτά. Το πιο συνηθισμένο συνδυαστικό κριτήριο βασίζεται στο γράφημα Sensitivity – Specificity που καλείται **Receiver Operating Characteristic** (**ROC**)[1]. Για να είμαστε ακριβείς, για λόγους ιστορικούς, η καμπύλη ROC είναι το γράφημα του $Sensitivity$ σε σχέση με το $(1 - Specificity)$. Το δεύτερο καλείται επίσης και **False-Positive Rate**.

Θεωρήστε πάλι έναν ταξινομητή που προσπαθεί να ανιχνεύσει αν κάποιο άτομο πάσχει από γρίπη. Ο στοιχειώδης ταξινομητής είναι αυτός που αποφασίζει τελείως τυχαία με κάποια πιθανότητα p αν το πρότυπο ανήκει στην κλάση 1 και άρα με πιθανότητα $(1 - p)$ στην κλάση 0. Στην περίπτωση αυτή το ποσοστό των προτύπων που ταξινομούνται θετικά ή αρνητικά είναι ανεξάρτητο από την κλάση στην οποία ανήκουν. Συνεπώς:

$$Sensitivity = \frac{TP}{\text{Κλάση } 1} = P(y = 1 \mid t = 1) = P(y = 1) = p$$

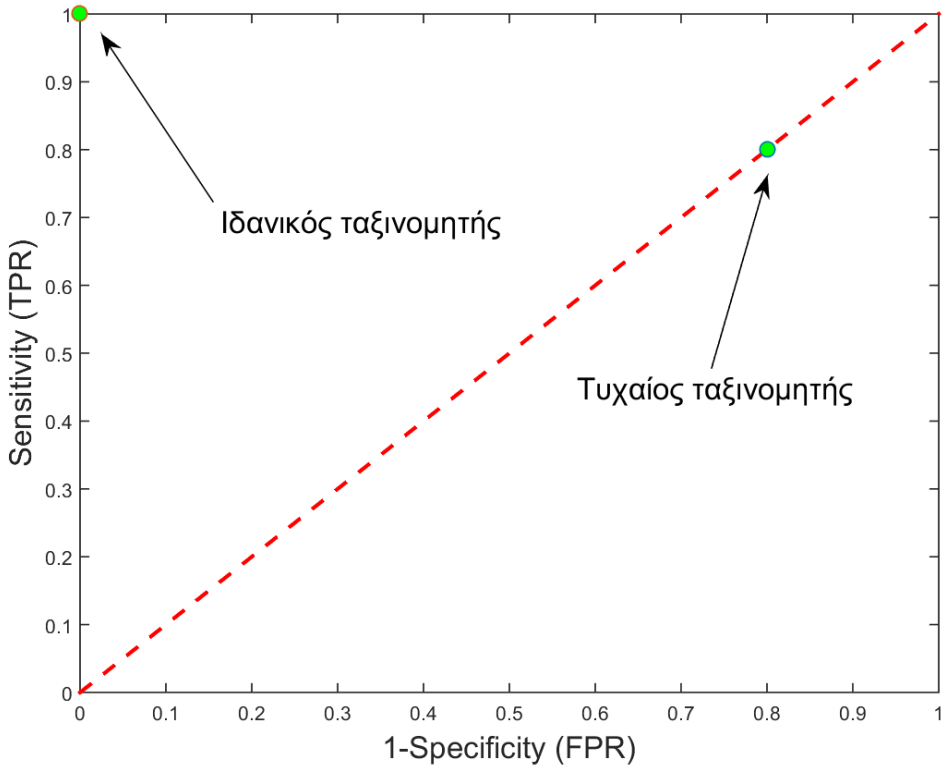
$$Specificity = \frac{TN}{\text{Κλάση } 0} = P(y = 0 \mid t = 0) = P(y = 0) = 1 - p$$

Οπότε

$$1 - Specificity = p$$

Έτσι το σημείο λειτουργίας του τυχαίου ταξινομητή βρίσκεται πάνω στην κόκκινη διαγώνιο της Εικόνα 3. Ο ιδανικός ταξινομητής βρίσκεται στο σημείο $1 - Specificity = 0$, $Sensitivity = 1$ (πάνω αριστερά στην ίδια Εικόνα). Το πάνω αριστερά τρίγωνο του Σχήματος 3 είναι η «καλή» περιοχή λειτουργίας οποιουδήποτε ταξινομητή. Όσο πιο κοντά στην πάνω αριστερή γωνία τόσο καλύτερα.

[1] Ο όρος Receiver Operating Characteristic προέρχεται από την τεχνολογία αναγνώρισης σήματος ραντάρ που επινοήθηκε κατά τον Β' Παγκόσμιο Πόλεμο.



Εικόνα 3. Η καμπύλη του $Sensitivity$ ως συνάρτηση του False-Positive-Rate $= (1 - Specificity)$ *είναι γνωστή και ως Receiver Operating Characteristic (ROC). Ο ιδανικός ταξινομητής βρίσκεται πάνω αριστερά στο σημείο* $(0,1)$ *ενώ ένας ταξινομητής που επιλέγει την κλάση 1 τυχαία με πιθανότητα p βρίσκεται πάνω την κόκκινη διακεκομμένη διαγώνιο στο σημείο* (p, p).

Πολλές φορές το τεστ που κάνουμε για να ταξινομήσουμε ένα πρότυπο σε μια από τις δύο κλάσεις παράγει συνεχείς τιμές x. Για παράδειγμα, θεωρήστε την περίπτωση όπου χρησιμοποιούμε το ύψος ενός ανθρώπου ως χαρακτηριστικό για να ταξινομήσουμε ένα άτομο στην κατηγορία «άνδρας» ή «γυναίκα». Τότε χρησιμοποιούμε κάποιο κατώφλι θ για να αποφασιστεί η κλάση στην οποία θα ταξινομηθεί το πρότυπο:

- Αν $x_i < \theta$ τότε το πρότυπο ταξινομείται στην κλάση «γυναίκα» (πχ Κλάση 0)
- Αν $x_i > \theta$ τότε το πρότυπο ταξινομείται στην κλάση «άνδρας» (πχ Κλάση 1)

Έστω για παράδειγμα, ότι η κατανομή του ύψους των προτύπων της κλάσης 0 είναι Γκαουσιανή με μέση τιμή $\mu_0 = 1.6$ και διασπορά $\sigma_0 = 0.2$, ενώ η κατανομή των προτύπων της κλάσης 1 είναι Γκαουσιανή με μέση τιμή $\mu_1 = 1.8$ και διασπορά $\sigma_1 = 0.2$. Με άλλα λόγια

$$p(x|\text{Κλάση } 0) = f_0(x) = N(1.6, 0.2)$$

$$p(x|\text{Κλάση } 1) = f_1(x) = N(1.8, 0.2)$$

όπου $N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)$.

Έχουμε:

$$TP = p(x > \theta, \text{Κλάση } 1) = p(x > \theta|\text{Κλάση } 1) \cdot p(\text{Κλάση } 1)$$

$$= P_1 \int_{\theta}^{\infty} f_1(x)dx$$

Παρομοίως:

$$FP = p(x > \theta | \text{Κλάση } 0) \cdot p(\text{Κλάση } 0) = P_0 \int_{\theta}^{\infty} f_0(x)dx$$

$$TN = p(x < \theta | \text{Κλάση } 0) \cdot p(\text{Κλάση } 0) = P_0 \int_{-\infty}^{\theta} f_0(x)dx$$

$$FN = p(x < \theta | \text{Κλάση } 1) \cdot p(\text{Κλάση } 1) = P_1 \int_{-\infty}^{\theta} f_1(x)dx$$

Με βάση τα παραπάνω έχουμε:

$$Sensitivity = TPR = \frac{TP}{TP + FN} = \frac{\int_{\theta}^{\infty} f_1(x)dx}{\int_{\theta}^{\infty} f_1(x)dx + \int_{-\infty}^{\theta} f_1(x)dx} = \int_{\theta}^{\infty} f_1(x)dx$$

$$1 - Specificity = FPR = \frac{FP}{TN + FP} = \int_{\theta}^{\infty} f_0(x)dx$$

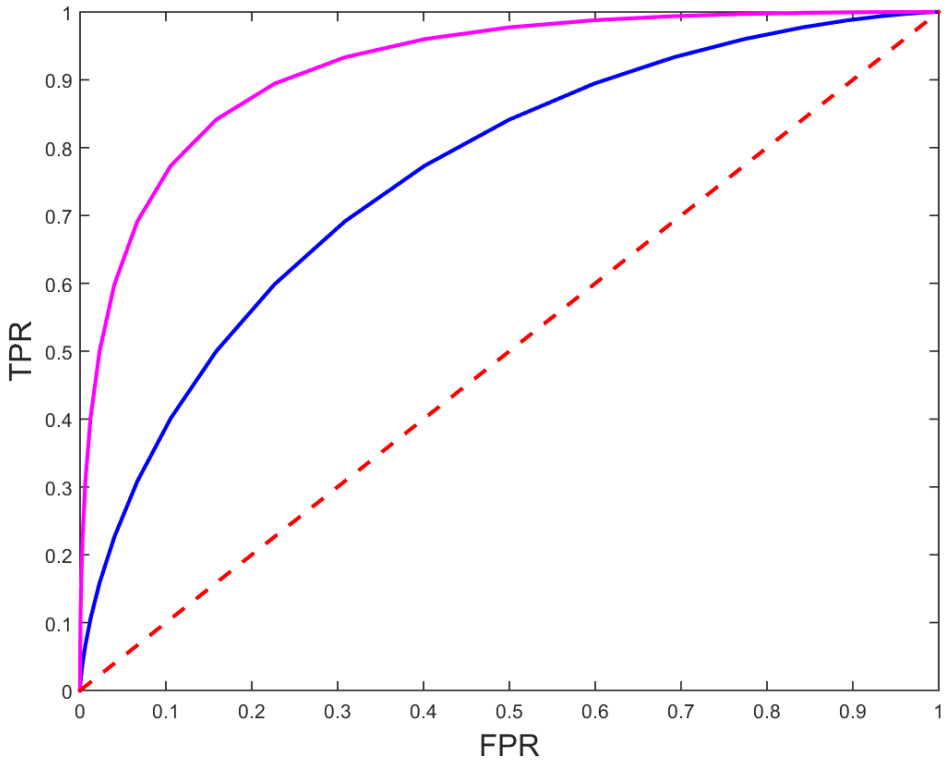
Στο παρακάτω Σχήμα δίνεται οι καμπύλη ROC (καμπύλη TRP-FPR) για όλες τις τιμές του θ από $-\infty$ έως ∞ (μπλε γραμμή). Ας ονομάσουμε αυτόν τον ταξινομητή Α.

Για λόγους σύγκρισης δίνεται και η καμπύλη ROC για έναν άλλο υποθετικό ταξινομητή Β όπου χρησιμοποιείται κάποιο άλλο χαρακτηριστικό x' για το οποίο οι κατανομές των δύο κλάσεων είναι λίγο πιο απομακρυσμένες μεταξύ τους, συγκεκριμένα

$$f_0(x') = N(1.5, 0.2)$$

$$f_1(x') = N(1.9, 0.2)$$

Βλέπουμε ότι η μωβ γραμμή που αντιστοιχεί στον ταξινομητή Β προσεγγίζει περισσότερο το βέλτιστο σημείο (0,1) καθώς ο ταξινομητής αυτός διαχωρίζει καλύτερα τα πρότυπα των δύο κλάσεων. Θυμίζουμε ότι στον ταξινομητή Α οι μέσες τιμές των κλάσεων είναι 1.6, 1.8 ενώ οι μέσες τιμές για τον ταξινομητή Β είναι 1.5, 1.9. Ο ταξινομητής Β είναι καλύτερος από τον Α. Μέτρο αξιολόγησης των ταξινομητών είναι η επιφάνεια κάτω από τις καμπύλες ROC. Όσο πιο μεγάλη είναι τόσο καλύτερος ο ταξινομητής. Άνω όριο του εμβαδού είναι το 1.



Εικόνα 4. Μπλε γραμμή: Καμπύλη ROC για τον ταξινομητή Α με $f_0(x) = N(1.6, 0.2)$, $f_1(x) = N(1.8, 0.2)$. Μωβ γραμμή: Καμπύλη ROC για τον ταξινομητή Β με $f_0(x) = N(1.5, 0.2)$, $f_1(x) = N(1.9, 0.2)$. Προφανώς ο ταξινομητής Β είναι καλύτερος από τον Α καθώς οι δύο κατανομές απέχουν περισσότερο μεταξύ τους. Αυτό αποτυπώνεται στις καμπύλες ROC: η μωβ καμπύλη πλησιάζει πιο κοντά στο βέλτιστο σημείο $(0,1)$ πάνω αριστερά. Ένα μέτρο ποιότητας είναι το εμβαδόν κάτω από κάθε ROC το οποίο καλείται Area Under the Curve (AUC). Όσο το AUC είναι πιο κοντά στο 1 τόσο καλύτερος είναι ο ταξινομητής.