

## Αξιολόγηση Ταξινομητών

### Μέθοδοι Εκπαίδευσης Ταξινομητών

---

- **Μέθοδος επανατοποθέτησης** (resubstitution method)

Χρησιμοποίησε τα ίδια δεδομένα για εκπαίδευση και δοκιμή.

Έχουμε υποεκτίμηση του σφάλματος.

Η εκτίμηση βελτιώνεται για μεγάλες τιμές του  $N$ .

- **Μέθοδος διαχωρισμού** (holdout method)

Διοθέντος ενός συνόλου  $N$  διανυσμάτων εισόδου/εξόδου χώρισέ τα σε

$N_1$  διανύσματα εκπαίδευσης και  $N_2$  διανύσματα ελέγχου ( $N_1 + N_2 = N$ ).

**Πρόβλημα:** Λιγότερα δεδομένα τόσο για εκπαίδευση όσο και για έλεγχο.

## Μέθοδοι Εκπαίδευσης Ταξινομητών

- **Μέθοδος Leave-One-Out** (κυρίως για μικρά datasets)
  - ❖ Επίλεξε ένα από τα  $N$  δείγματα. Εκπαίδευσε τον ταξινομητή χρησιμοποιώντας τα υπόλοιπα  $N-1$  δείγματα. Έλεγχε την απόδοση του ταξινομητή στο επιλεγμένο δείγμα. Στην περίπτωση λανθασμένης ταξινόμησης αύξησε τον μετρητή λαθών κατά ένα.
  - ❖ Επανάλαβε το παραπάνω εξαιρώντας ένα διαφορετικό δείγμα κάθε φορά.
  - ❖ Υπολόγισε την πιθανότητα λάθους ως το ποσοστό των λαθών επί του συνόλου των επαναλήψεων.
  - ❖ **Πλεονεκτήματα**
    - Χρήση όλων των διαθέσιμων δεδομένων για εκπαίδευση και έλεγχο.
    - Ανεξαρτησία δεδομένων ελέγχου από αυτά της εκπαίδευσης.
  - ❖ **Μειονεκτήματα**
    - Υψηλές υπολογιστικές απαιτήσεις.

## Μέθοδοι Εκπαίδευσης Ταξινομητών

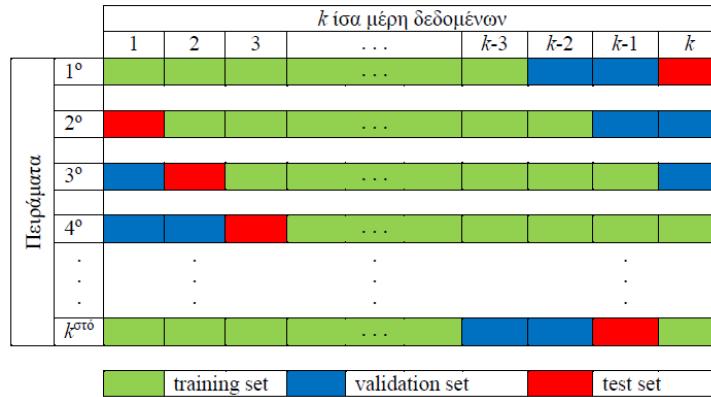
Παράδειγμα χρήσης δεδομένων τεκμηρίωσης

- **Μέθοδος  $k$ -fold cross validation ( $k$ -πλής διασταυρούμενης επικύρωσης)**
  - ❖ Χρησιμοποιείται για την ακριβέστερη αποτίμηση της γενικευτικής ικανότητας μοντέλων μηχανικής μάθησης.
  - ❖ Εκτελούμε μια τυχαία μετάθεση όλων των δεδομένων, χωρίζουμε το σύνολο των δεδομένων σε  $k$  ίσα μέρη (folds) και δημιουργούμε, για κάθε ένα από  $k$  πειράματα, τρία σύνολα δεδομένων με κυκλικό τρόπο ως εξής:
    - Το **σύνολο εκπαίδευσης** (training set) που χρησιμοποιείται για την εύρεση των παραμέτρων του ταξινομητή. Αφήνοντας ένα μέρος για το σύνολο ελέγχου, αυτό τυπικά περιλαμβάνει το 60 – 80% των υπόλοιπων  $(k-1)$  μερών (ή τα  $(k-1)$  μέρη αν δεν υπάρχουν δεδομένα τεκμηρίωσης).
    - Το **σύνολο τεκμηρίωσης** (validation set) που χρησιμοποιείται για την εύρεση βέλτιστων τιμών για τις μεταπαραμέτρους του αλγορίθμου εκπαίδευσης (μπορεί να χρησιμοποιηθεί με οποιαδήποτε μέθοδο). Αυτό περιλαμβάνει τα αδιάθετα των  $(k-1)$  μερών.
    - Το **σύνολο ελέγχου** (test set) για την αποτίμηση της γενίκευσης. Αυτό τυπικά περιλαμβάνει ένα μέρος, διαφορετικό για κάθε πείραμα.

## Μέθοδοι Εκπαίδευσης Ταξινομητών

- ❖ Εκπαιδεύουμε τον ταξινομητή  $k$  φορές και αποτιμούμε κάθε φορά την γενικευτική του ικανότητα στο σύνολο ελέγχου.
- ❖ Η τελική γενικευτική ικανότητα είναι ο μέσος όρος των  $k$  πειραμάτων.

**Συστηματικός τρόπος εφαρμογής  $k$ -fold cross validation**



## Πίνακας Σύγχυσης (Confusion Matrix)

- Έστω  $M$  το πλήθος των κλάσεων. Ο **πίνακας σύγχυσης** ( $\Pi\Sigma$ ) είναι ένας  $M \times M$  πίνακας  $A$  με το  $A(i, j)$  να είναι το πλήθος των διανυσμάτων εισόδου που προέρχονται από την κλάση  $\omega_i$  και ταξινομούνται στην κλάση  $\omega_j$ .
- Ο  $\Pi\Sigma$  δίνει πληροφορίες σχετικά με το αν κάποιες κλάσεις έχουν την τάση να συγχέονται με άλλες κλάσεις.

**Παράδειγμα:** Έστω ότι σε κάποιο πρόβλημα έχουμε  $M=5$ , πλήθος δεδομένων  $N=700$  και πλήθος δεδομένων ανά κλάση  $N_1=100$ ,  $N_2=150$ ,  $N_3=200$ ,  $N_4=100$  και  $N_5=150$ .

Έστω επίσης ότι ο  $\Pi\Sigma$  του ταξινομητή είναι όπως στο διπλανό σχήμα.

**Παρατήρηση:** Στο παράδειγμα αυτό, όλες οι κλάσεις είναι καλώς ταυτοποιημένες εκτός της  $\omega_4$  η οποία συγχέεται με την  $\omega_1$  και την  $\omega_3$ .

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	Εκτ- $\omega_3$	Εκτ- $\omega_4$	Εκτ- $\omega_5$	Συν.
$\omega_1$	98	0	2	0	0	100
$\omega_2$	0	135	0	10	5	150
$\omega_3$	0	1	198	0	1	200
$\omega_4$	35	0	25	40	0	100
$\omega_5$	0	0	0	0	150	150
Συν.	133	136	225	50	156	700

## Μετρικές αξιολόγησης ταξινομητών

- **Συνολική ακρίβεια** (Overall Accuracy – OA)

❖ Είναι το άθροισμα των διαγώνιων στοιχείων του ΠΣ προς το πλήθος δεδομένων N.

$$OA = \frac{\sum_{i=1}^M A(i,i)}{N}$$

❖ Στο παράδειγμα,  $OA = \frac{98+135+198+40+150}{700} = 88,7\%$

**Παρατήρηση:** Η ιδιαιτερότητα της  $\omega_4$  (η οποία συγχέεται με την  $\omega_1$  και την  $\omega_3$ ) δεν μπορεί να εκφραστεί από έναν μόνο αριθμό που προκύπτει από τον συγκερασμό διαφόρων στοιχείων του πίνακα σύγχυσης.

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	Εκτ- $\omega_3$	Εκτ- $\omega_4$	Εκτ- $\omega_5$	Συν.
$\omega_1$	98	0	2	0	0	100
$\omega_2$	0	135	0	10	5	150
$\omega_3$	0	1	198	0	1	200
$\omega_4$	35	0	25	40	0	100
$\omega_5$	0	0	0	0	150	150
Συν.	133	136	225	50	156	700

## Μετρικές αξιολόγησης ταξινομητών

- **Ανάκληση** (Recall)

❖ Είναι το ποσοστό των διανυσμάτων που προέρχονται από την κλάση  $\omega_i$  και ταξινομούνται (σωστά) στην κλάση αυτή:

$$R_i = \frac{A(i,i)}{\sum_{j=1}^N A(i,j)}$$

❖ Σχετίζεται με το ερώτημα: "Δοθέντος ενός δείγματος που προέρχεται από την κλάση  $\omega_i$ , πόσο πιθανό είναι να ταξινομηθεί σωστά από τον ταξινομητή;"

- **Ακρίβεια** (Precision)

❖ Είναι το ποσοστό των διανυσμάτων που ταξινομούνται στην κλάση  $\omega_i$  και πράγματι ανήκουν στην κλάση αυτή:

$$P_i = \frac{A(i,i)}{\sum_{j=1}^N A(j,i)}$$

❖ Σχετίζεται με το ερώτημα: "Δοθέντος ενός δείγματος που ταξινομήθηκε στην κλάση  $\omega_i$ , πόσο πιθανό είναι η ταξινόμηση αυτή να είναι σωστή;"

## Μετρικές αξιολόγησης ταξινομητών

- Precision/Recall

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	Εκτ- $\omega_3$	Εκτ- $\omega_4$	Εκτ- $\omega_5$	Συν.	<b>R</b>
$\omega_1$	98	0	2	0	0	100	98/100
$\omega_2$	0	135	0	10	5	150	135/150
$\omega_3$	0	1	198	0	1	200	198/200
$\omega_4$	35	0	25	40	0	100	40/100
$\omega_5$	0	0	0	0	150	150	150/150
Συν.	133	136	225	50	156	700	
$P$	98/133	135/136	198/225	40/50	150/156		

## Μετρικές αξιολόγησης ταξινομητών

- Precision/Recall – Η περίπτωση των δύο κλάσεων

- Πολλές φορές οι παραπάνω έννοιες χρησιμοποιούνται για την περίπτωση δύο κλάσεων όπου (συνήθως) η μία (η κλάση ενδιαφέροντος) είναι πολύ μικρότερη της άλλης (πρόβλημα μη-ισοζυγισμένων κλάσεων – **class imbalance problem**).
- Εδώ εστιάζουμε στα  $R$  και  $P$  της κλάσης ενδιαφέροντος.
- Η πιθανότητα σφάλματος σε τέτοιες περιπτώσεις **δεν είναι καλό μέτρο της απόδοσης** του ταξινομητή.
- Ο πίνακας σύγχυσης τώρα γίνεται:

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	<b>R</b>
$\omega_1$	Αληθώς θετικά (true positives - <b>TP</b> )	Ψευδώς αρνητικά (false negatives - <b>FN</b> )	<b>TP/(TP+FN)</b>
$\omega_2$	Ψευδώς θετικά (false positives - <b>FP</b> )	Αληθώς αρνητικά (true negatives - <b>TN</b> )	
$P$	<b>TP/(TP+FP)</b>		

## Μετρικές αξιολόγησης ταξινομητών

- **Precision/Recall – Η περίπτωση των δύο κλάσεων**

- ❖ **Παράδειγμα:** Το πρόβλημα της ταξινόμησης ανθρώπων σε "θετικούς" (κλάση  $\omega_1$ ) και "αρνητικούς" (κλάση  $\omega_2$ ) σε μια σπάνια ασθένεια. Το **δείγμα** μας αποτελείται από  $N = 100.000$  ανθρώπους από τους οποίους οι 10 μόνο είναι θετικοί στην ασθένεια (δηλαδή ανήκουν στην  $\omega_1$ ).
- ❖ **Σενάριο 1:** Έστω ταξινομητής που ταξινομεί όλους τους ανθρώπους στην κλάση  $\omega_2$ . Ο πίνακας σύγχυσης σε αυτήν την περίπτωση θα είναι:

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	<b>R</b>
$\omega_1$	<b>TP=0</b>	<b>FN=10</b>	<b>TP/(TP+FN)=0</b>
$\omega_2$	<b>FP=0</b>	<b>TN=99990</b>	
<i>P</i>	<b>TP/(TP+FP)=0</b>		

- ❖ Η OA για έναν ταξινομητή που ταξινομεί όλους τους ανθρώπους ως "αρνητικούς" είναι 99,99% ! Ωστόσο, ο ταξινομητής αυτός δεν θα ανιχνεύσει ποτέ έναν ασθενή.
- ❖ Στην περίπτωση αυτή έχουμε:  $P=0$  και  $R=0$ .

## Μετρικές αξιολόγησης ταξινομητών

- **Precision/Recall – Η περίπτωση των δύο κλάσεων**

- ❖ **Παράδειγμα:** Το πρόβλημα της ταξινόμησης ανθρώπων σε "θετικούς" (κλάση  $\omega_1$ ) και "αρνητικούς" (κλάση  $\omega_2$ ) σε μια σπάνια ασθένεια. Το **δείγμα** μας αποτελείται από  $N = 100.000$  ανθρώπους από τους οποίους οι 10 μόνο είναι θετικοί στην ασθένεια (δηλαδή ανήκουν στην  $\omega_1$ ).
- ❖ **Σενάριο 2:** Έστω ταξινομητής που ταξινομεί όλους τους ανθρώπους στην  $\omega_2$  εκτός έναν της  $\omega_1$  που τον ταξινομεί σωστά. Ο πίνακας σύγχυσης τώρα θα είναι:

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	<b>R</b>
$\omega_1$	<b>TP=1</b>	<b>FN=9</b>	<b>TP/(TP+FN)=1/10</b>
$\omega_2$	<b>FP=0</b>	<b>TN=99990</b>	
<i>P</i>	<b>TP/(TP+FP)=1</b>		

- ❖ Η OA για τον ταξινομητή αυτό θα είναι  $99.991/100.000 = 99,99\%$  !
- ❖ Στην περίπτωση αυτή:  $P=1$  και  $R=1/10$ .

## Μετρικές αξιολόγησης ταξινομητών

- **Precision/Recall – Η περίπτωση των δύο κλάσεων**

❖ **Παράδειγμα:** Το πρόβλημα της ταξινόμησης ανθρώπων σε "θετικούς" (κλάση  $\omega_1$ ) και "αρνητικούς" (κλάση  $\omega_2$ ) σε μια σπάνια ασθένεια. Το **δείγμα** μας αποτελείται από  $N = 100.000$  ανθρώπους από τους οποίους οι 10 μόνο είναι θετικοί στην ασθένεια (δηλαδή ανήκουν στην  $\omega_1$ ).

❖ **Σενάριο 3:** Έστω ταξινομητής που ταξινομεί όλους τους ανθρώπους στην  $\omega_1$ . Ο πίνακας σύγχυσης σε αυτήν την περίπτωση θα είναι:

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	<b>R</b>
$\omega_1$	<b>TP=10</b>	<b>FN=0</b>	<b>TP/(TP+FN)=1</b>
$\omega_2$	<b>FP=99990</b>	<b>TN=0</b>	
<b>P</b>	<b>TP/(TP+FP)=1/10000</b>		

❖ Η ΟΑ για έναν ταξινομητή που ταξινομεί όλους τους ανθρώπους ως "θετικούς" είναι  $10/100.000 = 0,01\%$  !

❖ Στην περίπτωση αυτή:  $P=0.0001$  και  $R=1$ .

## Μετρικές αξιολόγησης ταξινομητών

- **Precision/Recall – Η περίπτωση των δύο κλάσεων**

❖ Συνήθως οι P και R δεν παίρνουν ταυτόχρονα μεγάλες τιμές. Το σε ποιά από τις δύο θα δώσουμε μεγαλύτερη βαρύτητα εξαρτάται από το εκάστοτε πρόβλημα.

❖ Αν π.χ. στο παράδειγμα θέλουμε να ανιχνεύσουμε όλους τους ασθενείς που πράγματι έχουν την ασθένεια (έστω και αν κάποιους τους κατατάξουμε εσφαλμένα ως ασθενείς – FP), θέλουμε μεγάλη τιμή για το R.

❖ Αν όμως οι επιπλέον εξετάσεις που απαιτούνται για την ανιχνευση της νόσου είναι πολύ ακριβές, θα θέλαμε επιπλέον και μια σχετικά μεγάλη τιμή για το P.

❖ Συνδυασμός των P και R.

❖ **F1-score ή F1-measure:**

- Ο αρμονικός μέσος των P & R
- $$\text{F1-score} = \frac{2}{\frac{1}{R} + \frac{1}{P}} = \frac{2 * R * P}{R + P}$$

	Εκτ- $\omega_1$	Εκτ- $\omega_2$	<b>R</b>
$\omega_1$	Αληθώς θετικά (true positives - <b>TP</b> )	Ψευδώς αρνητικά (false negatives - <b>FN</b> )	<b>TP/(TP+FN)</b>
$\omega_2$	Ψευδώς θετικά (false positives - <b>FP</b> )	Αληθώς αρνητικά (true negatives - <b>TN</b> )	
<b>P</b>	<b>TP/(TP+FP)</b>		