ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ (PATTERN RECOGNITION)

Σέργιος Θεοδωρίδης

Κωνσταντίνος Κουτρούμπας

Version 2

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ο στόχος είναι η εκτίμηση της πιθανότητας λάθους ενός συστήματος ταξινόμησης

Τεχνική μέτρησης λαθών (Error Counting Technique)

- Έστω πρόβλημα ταξινόμησης σε *M* κλάσεις.
- Έστω N_i τα σημεία της κλάσης ω_i που χρησιμοποιούνται για δοκιμή.

$$\sum_{i=1}^{M} N_i = N$$

- Έστω P_i η πιθανότητα λάθους για την κλάση ω_i
- Η συνολική πιθανότητα λάθους είναι $P = \sum_{i=1}^{M} P(\omega_i) P_i$
- Υποθέτουμε ότι ο ταξινομητής έχει σχεδιαστεί χρησιμοποιώντας ένα διαφορετικό **ανεξάρτητο** σύνολο δεδομένων.
- Υποθέτοντας ότι τα διανύσματα του συνόλου δοκιμής είναι ανεξάρτητα, η πιθανότητα να έχουμε k_i διανύσματα από την ω_i λανθασμένα ταξινομημένα είναι

$$\text{prob}\{k_i \text{ in } \omega_i \text{ wrongly classified}\} = \binom{N_i}{k_i} P_i^{k_i} (1 - P_i)^{N_i - k_i}$$

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Μια παρένθεση:

Διωνυμική κατανομή:

- Έστω ένα πείραμα που το αποτέλεσμά του *k_i* μπορεί να είναι «επιτυχία» (1) ή «αποτυχία» (0).
- Έστω *p* η πιθανότητα να έχουμε «επιτυχία» όταν διεξάγουμε το πείραμα μία φορά.
- Η πιθανότητα να έχουμε ακριβώς *k* **επιτυχίες**, αν διεξάγουμε το πείραμα *N* **φορές** είναι:
- $\Pr\{k\} = \binom{N}{k} P^k (1-P)^{1-k}$
- ***Παρατήρηση:*** για *N=1*, είναι *k=0* ή 1 και
  $\Pr\{k\} = \binom{1}{k} P^k (1-P)^{1-k} = P^k (1-P)^{1-k}$

Η ***συνάρτηση πιθανοφάνειας*** για τη ***διωνυμική κατανομή***

- Έστω ότι διεξάγουμε το πείραμα *N* φορές. Είναι $k=k_1+\ldots+k_N$.
- Η συνάρτηση πιθανοφάνειας για τη διωνυμική κατανομή είναι

$$L(p) = \prod_{i=1}^{N} \Pr\{k_i\} = \prod_{i=1}^{N} P^{k_i}(1-P)^{1-k_i} = P^{\sum_{i=1}^{N} k_i} P^{\sum_{i=1}^{N}(1-k_i)} = P^k (1-P)^{N-k}$$

Η συνάρτηση πιθανοφάνειας *L(p)* **μεγιστοποιείται** όταν *P=k/N*.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

- Αφού τα P_i’s είναι άγνωστα, εκτιμούνται μέσω της μεγιστοποίησης συνάρτησης πιθανοφάνειας της παραπάνω διωνυμικής κατανομής και προκύπτει ότι

$$\boxed{\hat{P}_i = \frac{k_i}{N_i}}$$

- Δηλ. είναι ο αριθμός των λανθασμένα ταξινομημένων διανυσμάτων προς τον αριθμό των διανυσμάτων δοκιμής της εν λόγω κλάσης.
- Συνολική πιθανότητα λάθους

$$\hat{P} = \sum_{i=1}^{M} P(\omega_i) \frac{k_i}{N_i}$$

➢ **Στατιστικές ιδιότητες**

- $$E[k_i] = N_i P_i$$
- $$\text{Ε}\tau\sigma\iota,\ E[\hat{p}] = \sum_{i=1}^{M} P(\omega_i) P_i = P$$
- $$\sigma_{k_i}^2 = N_i (1 - P_i) P_i$$
- $$\sigma_{\hat{p}}^2 = \sum_{i=1}^{M} P^2(\omega_i) \frac{P_i (1 - P_i)}{N_i}$$

Ο εκτιμητής είναι αμερόληπτος αλλά ασυμπτωτικά συνεπής. Έτσι για **μικρές τιμές του *N*, μπορεί να μην είναι αξιόπιστος.**

➢Μια εκτίμηση που έχει προκύψει από θεωρητική ανάλυση σχετικά με τον ικανό αριθμό διανυσμάτων *N* του συνόλου δοκιμής, προκειμένου η πραγματική πιθανότητα λάθους *P* να **μην ξεπερνάει** το $c\acute{P}$ (τυπικά c=1.25, $\acute{P}$ η εκτιμώμενη πιθανότητα λάθους), είναι

$$N \approx \frac{100}{P}$$

$$\text{Ε}\tau\sigma\iota,\ \gamma\iota\alpha\ P \approx 0.01,\ N \approx 10000.\ \ \Gamma\iota\alpha\ P \approx 0.03,\ N \approx 3000$$

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Εκμεταλλευόμενοι το πεπερασμένου μεγέθους σύνολο δοκιμής.

- Μέθοδος επανατοποθέτησης (Resubstitution method):

  Χρησιμοποίησε τα ίδια δεδομένα για εκπαίδευση και δοκιμή. Έχουμε υποεκτίμηση του σφάλματος. Η εκτίμηση βελτιώνεται για μεγάλες τιμές του N και μεγάλες τιμές του λόγου N/l.

- Holdout Method: Δοθέντος συνόλου N διανυσμάτων χώρισέ τα σε:

  N_1: διανύσματα εκπαίδευσης

  N_2: διανύσματα δοκιμής

  $N=N_1+N_2$

  - **Πρόβλημα:** Λιγότερα δεδομένα τόσο για εκπαίδευση, όσο και γα δοκιμή.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

- Leave-one-out Method

Τα βήματα:

- Επέλεξε ένα από τα *N* δείγματα. Εκπαίδευσε τον ταξινομητή χρησιμοποιώντας τα υπόλοιπα *N*-1 δείγματα. Έλεγξε την απόδοση του ταξινομητή χρησιμοποιώντας το επιλεγμένο δείγμα. Αν είναι λανθασμένα ταξινομημένο μέτρα ένα επιπλέον λάθος.
- Επανέλαβε το παραπάνω βήμα εξαιρώντας ένα διαφορετικό δείγμα κάθε φορά.
- Υπολόγισε την πιθανότητα λάθους παίρνοντας το μέσο όρο των καταμετρημένων λαθών.

- **Πλεονεκτήματα:**

- Χρήση όλων των διαθέσιμων δεδομένων για εκπαίδευση και δοκιμή
- Διασφάλιση ανεξαρτησίας ανάμεσα στα διανύσματα εκπαίδευσης και δοκιμής.

- **Μειονεκτήματα:**

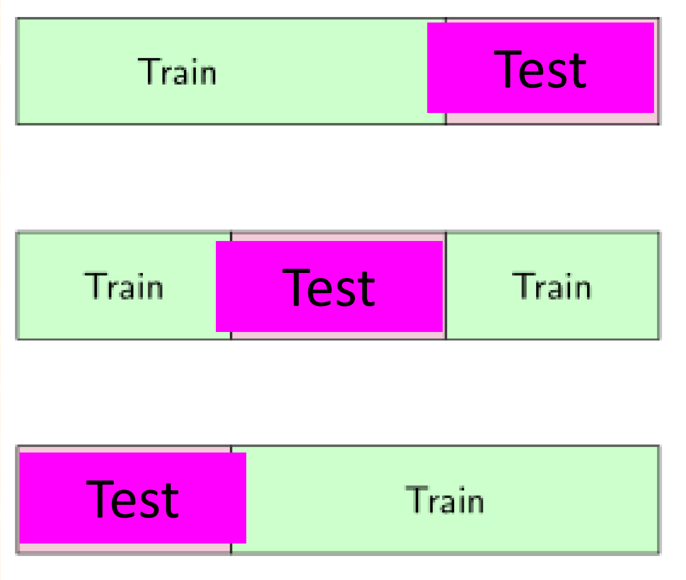
- Υψηλές υπολογιστικές απαιτήσεις.

- Παραλλαγές της μεθόδου εξαιρούν *k*>1 σημεία κάθε φορά, προκειμένου να μειώσουν την υπολογιστική πολυπλοκότητα (**k-fold cross validation**).

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Διασταυρούμενη επικύρωση (Cross-Validation):

- Πιθανόν η πιο συχνά χρησιμοποιούμενη τεχνική στην πράξη.
- Πώς δουλεύει:
  - Διαιρεί το διαθέσιμο σύνολο δεδομένων σε *K* (περίπου) ίσου μεγέθους μέρη.
  - Πραγματοποιεί εκπαίδευση *K* φορές, κάθε φορά χρησιμοποιώντας ένα μέρος για έλεγχο και τα υπόλοιπα *K*-1 μέρη για εκπαίδευση.
- Κάθε φορά τα δεδομένα που χρησιμοποιούνται για έλεγχο δεν χρησιμοποιούνται για εκπαίδευση (υψηλός βαθμός ανεξαρτησίας μεταξύ δεδομένων εκπαίδευσης και ελέγχου). Από την άλλη μεριά, τελικά, όλα τα δεδομένα χρησιμοποιούνται τόσο για εκπαίδευση, όσο και για έλεγχο.



ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Διασταυρούμενη επικύρωση (Cross-Validation):

•Μετά την ολοκλήρωση της διαδικασίας της διασταυρούμενης επικύρωσης,

- τα σφάλματα πάνω στα σύνολα ελέγχου μπορούν να συνδυαστούν ώστε να έχουμε πιο αξιόπιστη εκτίμηση του σφάλματος γενίκευσης (*K*-fold cross validation). Π.χ., μπορούμε να εκτιμήσουμε στο σφάλμα γενίκευσης ως το μέσο όρο των επιμέρους εκτιμηθέντων σφαλμάτων, ενώ η διασπορά, μας δίνει ένα μέτρο συνέπειας του αποτελέσματος.
- Μπορούμε να κρατήσουμε την εκτίμηση της διασποράς και να επανεκπαιδεύσουμε τον ταξινομητή μας χρησιμοποιώντας όλο το σύνολο δεδομένων.

•Η **ακραία περίπτωση** όπου *K=N* είναι η μέθοδος Leave one out (LOO) που αναφέρθηκε προηγουμένως.

•Στην πράξη, το *K* επιλέγεται μεταξύ 5-10.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Πίνακας Σύγχυσης (Confusion matrix)

- Πρόκειται για έναν *M*x*M* πίνακα, όπου το (*i,j*) στοιχείο του ισούται με το πλήθος των σημείων που, ενώ προέρχονται από την κλάση *i*, καταχωρούνται στην κλάση *j*.
- Δίνει πληροφορίες σχετικά με το αν κάποιες κλάσεις έχουν τη τάση να **συγχέονται** με άλλες κλάσεις.

Παράδειγμα:

- Έστω ένα πρόβλημα M=5 κλάσεων και σύνολο δεδομένων N=700 σημείων, με τα πλήθη των σημείων κάθε κλάσης να είναι N_1=100, N_2=150, N_3=200, N_4=100, N_5=150
- Ένας ταξινομητής χρησιμοποιείται για την ταξινόμησή τους σε κάποια από τις πέντε κλάσεις και τα αποτελέσματα φαίνονται στον ακόλουθο **πίνακα σύγχυσης**.

	Εκτ-ω_1	Εκτ-ω_2	Εκτ-ω_3	Εκτ-ω_4	Εκτ-ω_5	Συν.
ω_1	98	0	2	0	0	100
ω_2	0	135	0	10	5	150
ω_3	0	1	198	0	1	200
ω_4	35	0	25	40	0	100
ω_5	0	0	0	0	150	150
Συν.	133	136	225	50	156	**700**

Παρατήρηση:
Όλες οι κλάσεις είναι καλώς ταυτοποιημένες, εκτός της ω_4, η οποία **συγχέεται** με την ω_1 και την ω_3.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Συνολική ακρίβεια (Overall accuracy - OA)

- Είναι το άθροισμα των διαγωνίων στοιχείων του πίνακα σύγχυσης προς το πλήθος των σημείων του συνόλου δεδομένων, *N*.

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

- Για το προηγούμενο παράδειγμα η OA είναι $\mathrm{OA} = \frac{98+135+198+40+150}{700} = 88{,}7\%$

	Εκτ-ω_1	Εκτ-ω_2	Εκτ-ω_3	Εκτ-ω_4	Εκτ-ω_5	Συν.
ω_1	98	0	2	0	0	100
ω_2	0	135	0	10	5	150
ω_3	0	1	198	0	1	200
ω_4	35	0	25	40	0	100
ω_5	0	0	0	0	150	150
Συν.	133	136	225	50	156	**700**

Παρατήρηση:
Η ιδιαιτερότητα της ω_4 (η οποία συγχέεται με την ω_1 και την ω_3) δεν μπορεί να εκφραστεί από έναν μόνο αριθμό που προκύπτει από το συγκερασμό διαφόρων στοιχείων του πίνακα σύγχυσης.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision)

•Ανάκληση (Recall):

- Το ποσοστό των διανυσμάτων που προέρχονται από την κλάση *i* και ταξινομούνται (σωστά) στην κλάση αυτή: $R_i = \frac{A(i,i)}{\sum_{j=1}^{N} A(i,j)}$
- **Σχετίζεται** με το ακόλουθο **ερώτημα:** «Δοθέντος ενός δείγματος που **προέρχεται** από την **κλάση** *i*, θα ταξινομηθεί αυτό σωστά από τον ταξινομητή;»

•Ακρίβεια (Precision):

- Το ποσοστό των διανυσμάτων που ταξινομούνται στην κλάση *i* και πράγματι ανήκουν στην κλάση αυτή: $P_i = \frac{A(i,i)}{\sum_{j=1}^{N} A(j,i)}$
- **Σχετίζεται** με το ακόλουθο **ερώτημα:** «Δοθέντος ενός δείγματος που **ταξινομήθηκε** στην **κλάση** *i*, ποσό πιθανό είναι η ταξινόμηση αυτή να είναι σωστή;»

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision)

Παράδειγμα: Για το προηγούμενο παράδειγμα η τιμές R και P για κάθε κλάση φαίνονται παρακάτω

	Εκτ-ω_1	Εκτ-ω_2	Εκτ-ω_3	Εκτ-ω_4	Εκτ-ω_5	Συν.	***R***
ω_1	98	0	2	0	0	100	98/100
ω_2	0	135	0	10	5	150	135/150
ω_3	0	1	198	0	1	200	198/200
ω_4	35	0	25	40	0	100	40/100
ω_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		

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision) – Η περίπτωση των δύο κλάσεων

•Τις περισσότερες φορές οι παραπάνω έννοιες χρησιμοποιούνται για την περίπτωση δύο κλάσεων όπου (συνήθως) η μία (κλάση ενδιαφέροντος) είναι πολύ μικρότερη της άλλης (πρόβλημα μη σταθμισμένων 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** γίνεται:

	Εκτ-ω_1	Εκτ-ω_2	***R***
ω_1	Αληθώς θετικά (true positives -**TP**)	Ψευδώς αρνητικά (false negatives - **FN**)	**TP/(TP+FN)**
ω_2	Ψευδώς θετικά (false positives - **FP**)	Αληθώς αρνητικά (true negatives - **TN**)	
P	**TP/(TP+FP)**		

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision) – Η περίπτωση των δύο κλάσεων

Παράδειγμα: Το πρόβλημα της ταξινόμησης ανθρώπων σε «θετικούς» (κλάση ω_1) και «αρνητικούς» (κλάση ω_2) σε μια σπάνια ασθένεια. Το **δείγμα** μας αποτελείται από N=100000 ανθρώπους, από τους οποίους οι 10 μόνο είναι θετικοί στην ασθένεια (ανήκουν στην ω_1).

Σενάριο 1: Έστω ο **ταξινομητής** που **ταξινομεί *όλους* τους ανθρώπους στην κλάση ω_2**.

- Ο πίνακας σύγχυσης γίνεται:

	Εκτ-ω_1	Εκτ-ω_2	***R***
ω_1	**TP**=0	**FN**=10	**TP/(TP+FN)**=0
ω_2	**FP**=0	**TN**=99990	
P	**TP/(TP+FP)**=0		

- Η ΟΑ για έναν ταξινομητή που ταξινομεί όλους τους ανθρώπους ως «αρνητικούς» είναι 99990/100000=99,99%.!! Ωστόσο ο ταξινομητής αυτός δεν θα ανιχνεύσει ποτέ έναν ασθενή.
- Είναι P=0 και R=0.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision) – Η περίπτωση των δύο κλάσεων

Παράδειγμα: Το πρόβλημα της ταξινόμησης ανθρώπων σε «θετικούς» (κλάση ω_1) και «αρνητικούς» (κλάση ω_2) σε μια σπάνια ασθένεια. Το **δείγμα** μας αποτελείται από N=100000 ανθρώπους, από τους οποίους οι 10 μόνο είναι θετικοί στην ασθένεια (ανήκουν στην ω_1).

Σενάριο 2: Έστω ο **ταξινομητής** που **ταξινομεί *όλους* τους ανθρώπους στην κλάση ω_2 εκτός από 1 της ω_1 που τον ταξινομεί σωστά**.

• Ο πίνακας σύγχυσης τώρα γίνεται:

	Εκτ-ω_1	Εκτ-ω_2	***R***
ω_1	**TP**=1	**FN**=9	**TP**/(**TP**+**FN**)=1/9
ω_2	**FP**=0	**TN**=99990	
P	**TP**/(**TP**+**FP**)=1		

•Η ΟΑ για έναν ταξινομητή που ταξινομεί όλους τους ανθρώπους ως «αρνητικούς» είναι 99991/100000=99,99%.!!

•Είναι P=1 και R=1/9.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision) – Η περίπτωση των δύο κλάσεων

Παράδειγμα: Το πρόβλημα της ταξινόμησης ανθρώπων σε «θετικούς» (κλάση ω_1) και «αρνητικούς» (κλάση ω_2) σε μια σπάνια ασθένεια. Το **δείγμα** μας αποτελείται από N=100000 ανθρώπους, από τους οποίους οι 10 μόνο είναι θετικοί στην ασθένεια (ανήκουν στην ω_1).

Σενάριο 3: Έστω ο **ταξινομητής** που **ταξινομεί *όλους* τους ανθρώπους στην κλάση ω_1**.

• Ο πίνακας σύγχυσης τώρα γίνεται:

	Εκτ-ω_1	Εκτ-ω_2	***R***
ω_1	**TP**=10	**FN**=0	**TP/(TP+FN)**=1
ω_2	**FP**=99990	**TN**=0	
P	**TP**/(**TP**+**FP**)=1/10000		

•Η OA για έναν ταξινομητή που ταξινομεί όλους τους ανθρώπους ως «αρνητικούς» είναι 10/100000=0.01%.!!

•Είναι P=0.0001 και R=1.

ΑΞΙΟΛΟΓΗΣΗ ΤΑΞΙΝΟΜΗΤΗ

Ανάκληση (Recall) – Ακρίβεια (Precision) – F1 score - Η περίπτωση των δύο κλάσεων

Παρατηρήσεις:

- Συνήθως οι *P* και *R* είναι δεν παίρνουν ταυτόχρονα μεγάλες τιμές. Το σε ποια από τις δύο θα δώσουμε μεγαλύτερη βαρύτητα εξαρτάται από το εκάστοτε πρόβλημα.
- Αν π.χ., στο παράδειγμά μας θέλουμε να ανιχνεύσουμε όλους τους ασθενείς που πράγματι έχουν την ασθένεια (έστω και αν κάποιους τους κατατάξουμε κατά λάθος ως ασθενείς – αυξημένο **FP**), θέλουμε μεγάλη τιμή για το *R*.
- Αν όμως οι επιπλέον εξετάσεις που απαιτούνται για την ανίχνευση της νόσου είναι πολύ ακριβές, θα θέλαμε επιπλέον και μια σχετικά μεγάλη τιμή για το *P*.

Συνδυασμός των *R* και *P*:

F1-score: Ο **αρμονικός μέσος** των *R* και *P*, $F1-score = \frac{2}{\frac{1}{R}+\frac{1}{P}} = \frac{2*R*P}{R+P}$

	Εκτ-ω_1	Εκτ-ω_2	***R***
ω_1	Αληθώς θετικά (true positives -**TP**)	Ψευδώς αρνητικά (false negatives - **FN**)	**TP/(TP+FN)**
ω_2	Ψευδώς θετικά (false positives - **FP**)	Αληθώς αρνητικά (true negatives - **TN**)	
P	**TP/(TP+FP)**		

Διασταυρούμενη επικύρωση (cross validation) για επιλογή τιμών παραμέτρων:

•Έστω ένας ταξινομητής που περιλαμβάνει σύνολο παραμέτρων που πρέπει να καθοριστούν από το χρήστη.

• Η διασταυρούμενη επικύρωση μπορεί να χρησιμοποιηθεί για τον βέλτιστο προσδιορισμό των τιμών αυτών των παραμέτρων, ως εξής:

- Διαιρεί το διαθέσιμο σύνολο **εκπαίδευσης** σε *K* (περίπου) ίσου μεγέθους μέρη.
- Για κάθε συνδυασμό τιμών παραμέτρων πραγματοποιεί *K* εκπαιδεύσεις, χρησιμοποιώντας κάθε φορά ένα διαφορετικό μέρος για **επικύρωση** και τα υπόλοιπα *K*-1 μέρη για εκπαίδευση.
- Εκτιμά το σφάλμα γενίκευσης με βάση τα αποτελέσματα των παραπάνω *K* εκπαιδεύσεων, για κάθε συνδυασμό τιμών των παραμέτρων.
- Επιλέγει το συνδυασμό τιμών που δίνει το μικρότερο σφάλμα γενίκευσης.
- Για το συνδυασμό αυτών των τιμών εκπαιδεύει τον ταξινομητή χρησιμοποιώντας **όλο** το σύνολο εκπαίδευσης και ελέγχει χρησιμοποιώντας το σύνολο ελέγχου.