



ΗΜΜΥ · ΕΛΜΕΠΑ · 9ο εξάμηνο

Μηχανική Μάθηση και Εξόρυξη Γνώσης

Διάλεξη 4: Ταξινόμηση I
Βασικές έννοιες και αξιολόγηση

Λευτέρης Κουμάκης

Από την προηγούμενη φορά



Γραμμικά μοντέλα

$\hat{y} = \theta^T x$, κόστος MSE, κανονική εξίσωση



Gradient descent

$\theta \leftarrow \theta - \eta \nabla J(\theta)$, batch, SGD, mini-batch



Ομαλοποίηση

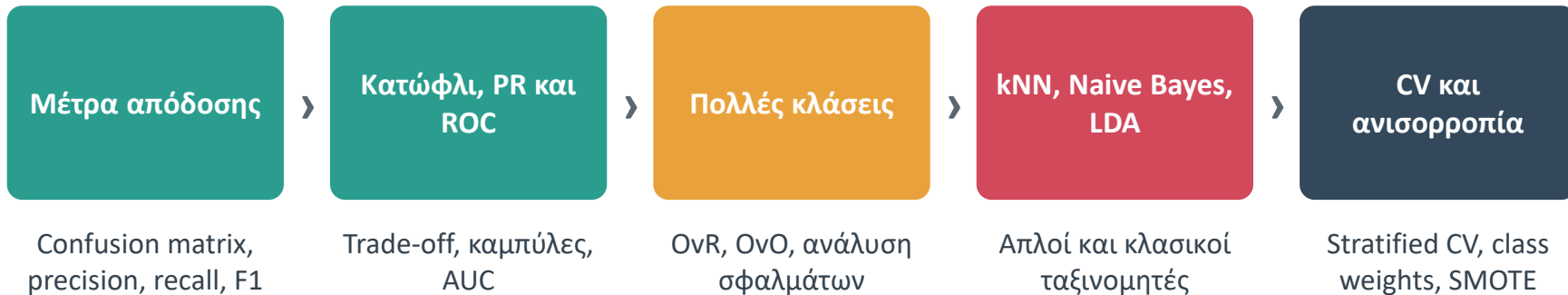
Ridge, Lasso, Elastic Net, early stopping



Λογιστική

$\hat{p} = \sigma(\theta^T x)$, κατώφλι 0,5, log loss

Ο χάρτης της σημερινής διάλεξης



Βιβλιογραφία: Géron Κεφ. 3 · ISL Κεφ. 4.4–4.5 και 5.1

To dataset MNIST

70.000

εικόνες ψηφίων

784

χαρακτηριστικά (28×28
pixels)

**60.000 /
10.000**

train / test

10

κλάσεις (0–9)

```
from sklearn.datasets import fetch_openml
mnist = fetch_openml("mnist_784", as_frame=False)
X, y = mnist.data, mnist.target
X_train, X_test = X[:60000], X[60000:]
y_train, y_test = y[:60000], y[60000:]
```

Πρώτο βήμα: δυαδικός ταξινομητής «είναι 5;»

```
y_train_5 = (y_train == "5")    # True για όλα τα 5

from sklearn.linear_model import SGDClassifier
sgd_clf = SGDClassifier(random_state=42)
sgd_clf.fit(X_train, y_train_5)
sgd_clf.predict([some_digit])    # array([ True])
```

Θετική κλάση = αυτό που ψάχνουμε (εδώ το «5»), όχι απαραίτητα κάτι «καλό».



Το ερώτημα

Πόσο καλός είναι αυτός ο ταξινομητής;

Η απάντηση είναι πιο δύσκολη απ' ό,τι φαίνεται.

ΕΝΟΤΗΤΑ 1

Μέτρα απόδοσης

Γιατί η ακρίβεια (accuracy) δεν αρκεί

Το παράδοξο της ακρίβειας

$\approx 10\%$

των εικόνων είναι «5»

$\approx 90\%$

ακρίβεια ταξινομητή που λέει πάντα
«όχι 5»

$> 93\%$

ακρίβεια SGD με 3-fold CV

Η ακρίβεια (accuracy) = σωστές προβλέψεις / όλες. Όταν μια κλάση κυριαρχεί, ακόμη και ένας άχρηστος ταξινομητής έχει υψηλή ακρίβεια. Χρειαζόμαστε μέτρα που κοιτάζουν κάθε είδος λάθους χωριστά.

Ο πίνακας σύγχυσης (confusion matrix)

		Πρόβλεψη	
		Αρνητική	Θετική
Πραγματική	αρνητική	TN Αληθώς αρνητικά	FP Ψευδώς θετικά
	θετική	FN Ψευδώς αρνητικά	TP Αληθώς θετικά

- Διαγώνιος (πράσινα): σωστές προβλέψεις
- FP: σφάλμα τύπου I, «ψευδής συναγερμός»
- FN: σφάλμα τύπου II, «χαμένη περίπτωση»
- Στο scikit-learn: `confusion_matrix(y,  $\hat{y}$ )`

Παράδειγμα: εξέταση σε 1.000 ασθενείς

	Πρόβλεψη: υγιής	Πρόβλεψη: νόσος	Σύνολο
Πραγματικά υγιής	TN = 810	FP = 90	900
Πραγματικά νόσος	FN = 20	TP = 80	100
Σύνολο	830	170	1.000

$$\text{Ακρίβεια} = (TP + TN) / \text{όλα} = (80 + 810) / 1.000 = 0,89$$

Όμως: από τις 170 θετικές διαγνώσεις, μόνο οι 80 είναι σωστές.

Precision και recall

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Από όσα είπα «θετικό», πόσα ήταν;

$$80 / 170 \approx 0,47$$

precision στο παράδειγμα

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

Από όσα ήταν θετικά, πόσα βρήκα;

$$80 / 100 = 0,80$$

recall στο παράδειγμα

Το μέτρο F1: αρμονικός μέσος

$$F1 = 2 \cdot P \cdot R / (P + R)$$

Precision	Recall	Αριθμητικός μέσος	F1
0,47	0,80	0,64	0,59
1,00	0,10	0,55	0,18
0,80	0,80	0,80	0,80

Το F1 ευνοεί ταξινομητές με ισορροπημένα precision και recall.

Περισσότερα μέτρα από τον ίδιο πίνακα

Μέτρο	Τύπος	Άλλα ονόματα	Παράδειγμα
Recall	$TP / (TP + FN)$	Ευαισθησία, TPR	0,80
Specificity	$TN / (TN + FP)$	Ειδικότητα, TNR	0,90
FPR	$FP / (FP + TN)$	1 – ειδικότητα	0,10
Precision	$TP / (TP + FP)$	Θετική προγνωστική αξία (PPV)	0,47
NPV	$TN / (TN + FN)$	Αρνητική προγνωστική αξία	0,98

Ίδιος πίνακας: $TP = 80$, $FN = 20$, $FP = 90$, $TN = 810$

Ποιο μέτρο μετρά περισσότερο;



Βίντεο για παιδιά

Υψηλό PRECISION: καλύτερα να απορριφθούν καλά βίντεο παρά να περάσει ένα ακατάλληλο



Εντοπισμός κλοπής σε κατάστημα

Υψηλό RECALL: καλύτερα ψευδείς συναγερμοί παρά χαμένοι κλέφτες



Προσυμπτωματικός έλεγχος

Υψηλό RECALL: ακολουθεί πιο ακριβής εξέταση



Φίλτρο spam

Υψηλό PRECISION: ένα σημαντικό email στο spam κοστίζει ακριβά



Ερώτηση Slido



Σε ένα dataset συναλλαγών, το 1% είναι απάτες. Ένα μοντέλο έχει ακρίβεια (accuracy) 99% στο test set. Τι συμπεραίνουμε;

A

Το μοντέλο είναι εξαιρετικό

B

Δεν ξέρουμε ακόμη αν εντοπίζει έστω και μία απάτη

Γ

Το μοντέλο υπερπροσαρμόζεται

Δ

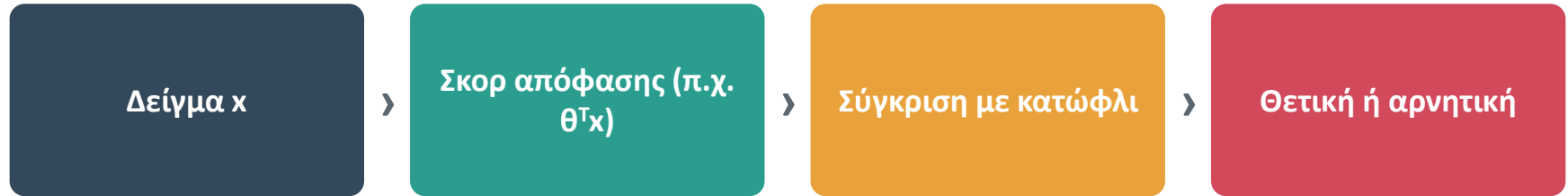
Το precision είναι σίγουρα 99%

ΕΝΟΤΗΤΑ 2

Κατώφλι, PR και ROC

Το αντιστάθμισμα precision–recall και οι καμπύλες που το αποτυπώνουν

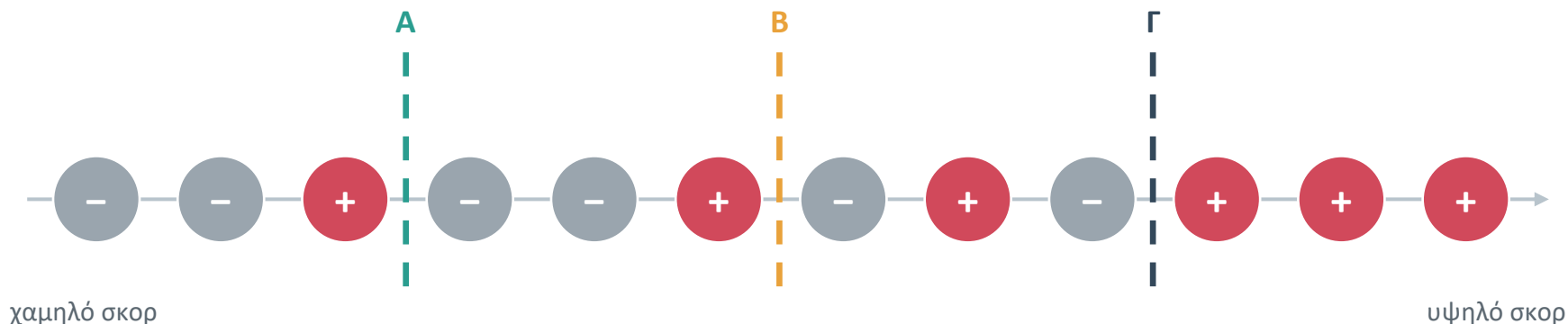
Σκορ απόφασης και κατώφλι



```
scores = sgd_clf.decision_function([some_digit])    # π.χ. array([2164.2])
threshold = 3000
y_pred = (scores > threshold)                       # array([False])
```

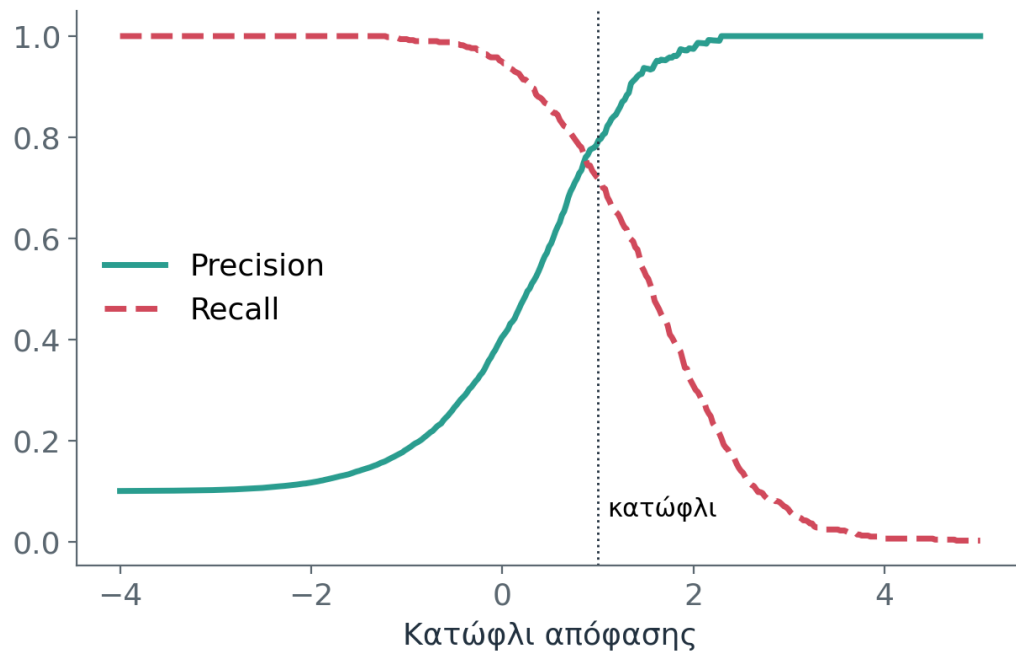
Αλλάζοντας το κατώφλι αλλάζουμε την απόφαση χωρίς επανεκπαίδευση.

Παράδειγμα: 12 δείγματα και τρία κατώφλια



Κατώφλι	Προβλέψεις «+»	TP	FP	Precision	Recall
A	9	5	4	0,56	0,83
B	6	4	2	0,67	0,67
Γ	3	3	0	1,00	0,50

Precision και recall ως συνάρτηση του κατωφλίου



- Χαμηλό κατώφλι: υψηλό recall, χαμηλό precision
- Υψηλό κατώφλι: υψηλό precision, χαμηλό recall
- Επιλογή κατωφλίου = επιχειρησιακή απόφαση

Επιλογή κατωφλίου για συγκεκριμένο στόχο

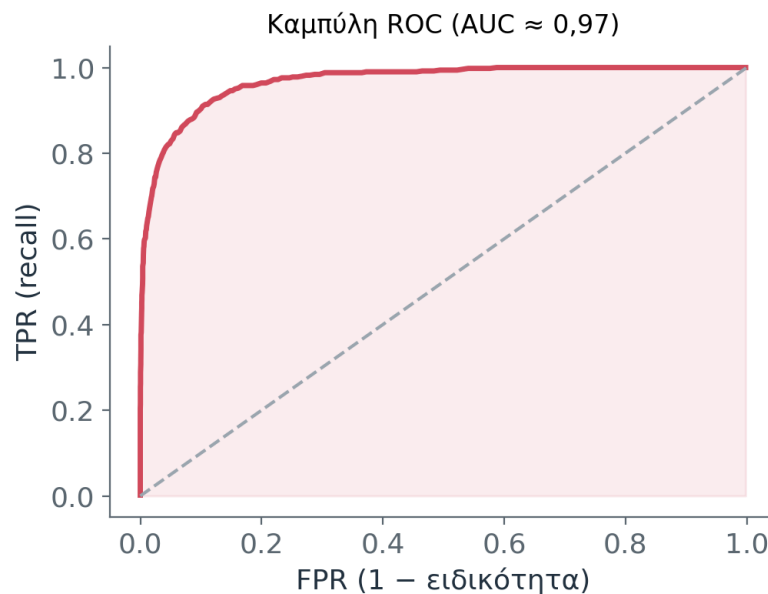
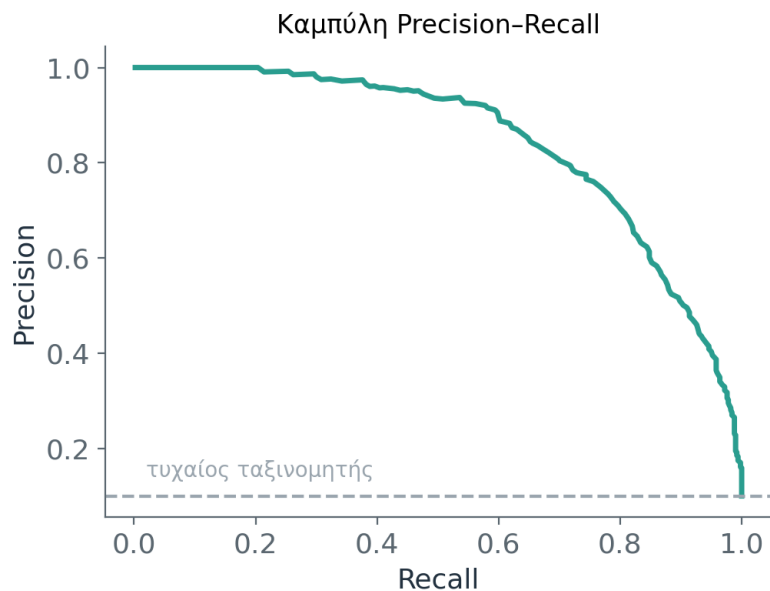
```
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import precision_recall_curve

y_scores = cross_val_predict(sgd_clf, X_train, y_train_5, cv=3,
                             method="decision_function")
precisions, recalls, thresholds = precision_recall_curve(y_train_5, y_scores)

idx = (precisions >= 0.90).argmax()          # πρώτο σημείο με  $P \geq 0,90$ 
threshold_90 = thresholds[idx]
y_pred_90 = (y_scores >= threshold_90)
```

«Θέλω 99% precision» → πάντα ρωτάμε «με πόσο recall;».

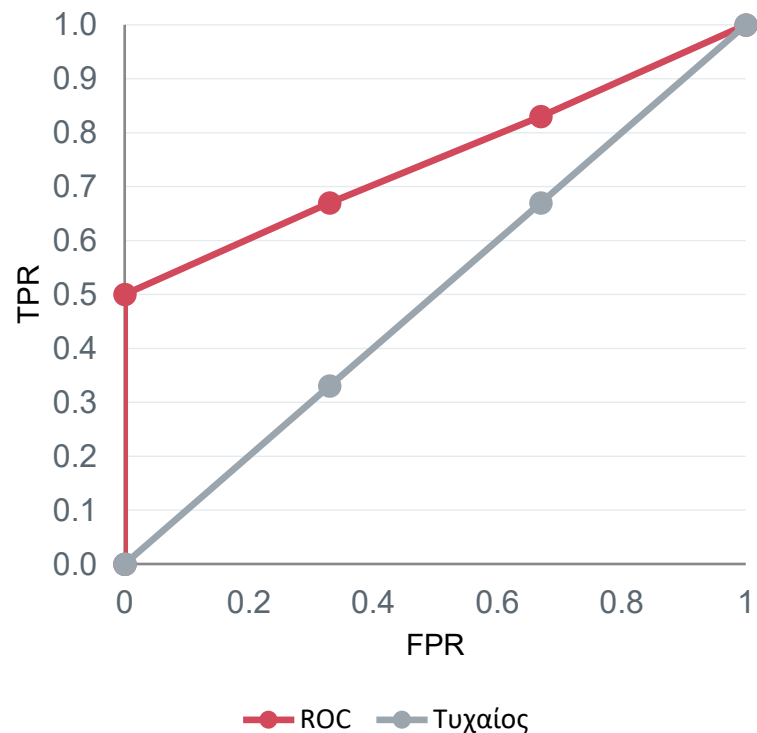
Καμπύλες Precision-Recall και ROC



Ιδανικό: PR πάνω δεξιά · ROC πάνω αριστερά

Παράδειγμα: σημεία ROC από τα 12 δείγματα

Κατώφλι	TPR = TP/6	FPR = FP/6
(όλα αρνητικά)	0	0
Γ	0,50	0,00
B	0,67	0,33
A	0,83	0,67
(όλα θετικά)	1	1



Το εμβαδόν κάτω από την καμπύλη (AUC)

0,5

τυχαίος ταξινομητής

0,7 – 0,8

αποδεκτή διάκριση (ενδεικτικά)

1,0

τέλεια κατάταξη

```
from sklearn.metrics import roc_auc_score, roc_curve
fpr, tpr, thresholds = roc_curve(y_train_5, y_scores)
roc_auc_score(y_train_5, y_scores)
```

AUC = P(σکور τυχαίου θετικού > σکور τυχαίου αρνητικού)

PR ή ROC;



Καμπύλη PR

- Θετική κλάση σπάνια (απάτη, σπάνια νόσος)
- Μας ενδιαφέρουν τα FP
- Δείχνει πιο «ειλικρινά» την απόδοση σε ανισορροπία



Καμπύλη ROC / AUC

- Κλάσεις περίπου ισορροπημένες
- Σύγκριση μοντέλων ανεξάρτητα από το κατώφλι
- Προσοχή: σε ανισορροπία φαίνεται υπερβολικά αισιόδοξη



Ερώτηση Slido



Πίνακας σύγχυσης: $TP = 40$, $FP = 10$, $FN = 60$, $TN = 890$. Ποιο είναι το precision;

A 0,40

B 0,80

Γ 0,93

Δ 0,89

ΕΝΟΤΗΤΑ 3

Πολλές κλάσεις

OnR, OnO, ανάλυση σφαλμάτων, multilabel

Ταξινόμηση πολλών κλάσεων



Ένας εναντίον όλων (OvR)

Ένας ταξινομητής ανά κλάση: «0 ή όχι», «1 ή όχι»,
...
Πρόβλεψη: η κλάση με το υψηλότερο σκορ.

10 κλάσεις $\rightarrow$ 10 ταξινομητές

Προεπιλογή για τους περισσότερους αλγορίθμους



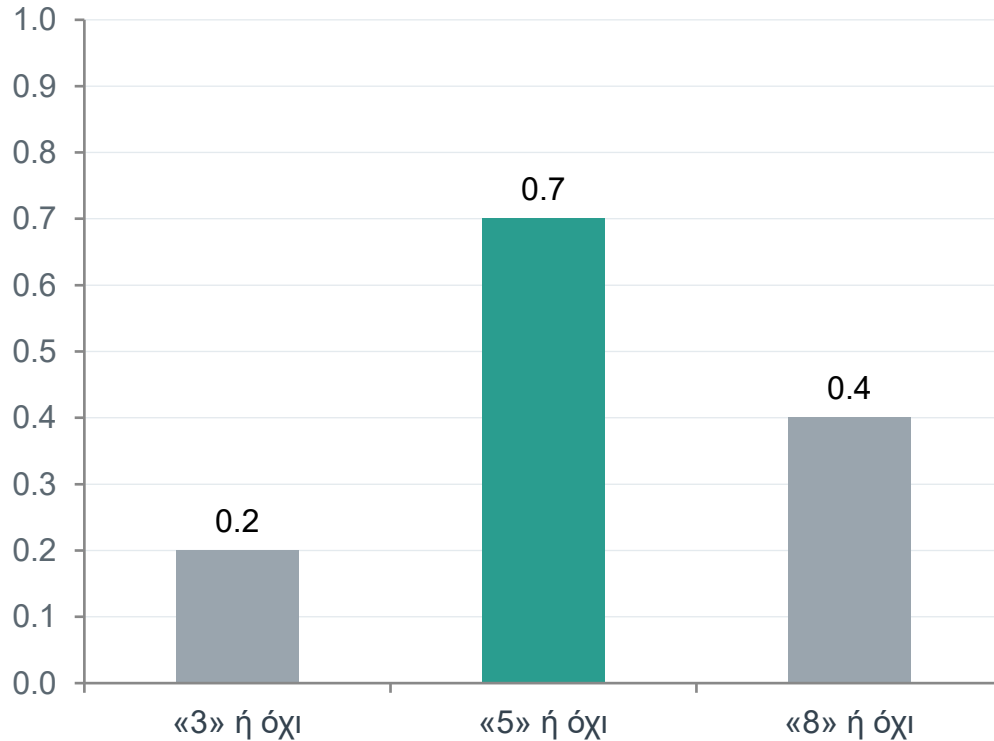
Ένας εναντίον ενός (OvO)

Ένας ταξινομητής για κάθε ζεύγος: «0 ή 1», «0 ή 2», ...
Πρόβλεψη: η κλάση με τις περισσότερες νίκες.

N κλάσεις $\rightarrow N(N-1)/2 = 45$

Κάθε ταξινομητής εκπαιδεύεται σε λιγότερα δεδομένα (χρήσιμο για SVM)

Παράδειγμα: απόφαση με OnR



Ζεύγος OnO	Νικητής
3 έναντι 5	5
3 έναντι 8	8
5 έναντι 8	5

OnR: «5» (υψηλότερο σκορ)
OnO: «5» (2 νίκες)

Ανάλυση σφαλμάτων με τον πίνακα σύγχυσης

Πραγματική κλάση	3	5	8	9
	92	1	2	5
	1	95	3	1
	4	6	80	10
	3	5	8	9
		Πρόβλεψη		

- Διαγώνιος: σωστές προβλέψεις
- Εκτός διαγωνίου: ποια ψηφία συγχέονται
- Κανονικοποίηση ανά γραμμή → ποσοστά λάθους ανά κλάση
- Ιδέες: data augmentation, νέα χαρακτηριστικά, κεντράρισμα εικόνων

ΕΝΟΤΗΤΑ 4

Απλοί και κλασικοί ταξινομητές

k-πλησιέστεροι γείτονες, Naive Bayes, LDA και QDA

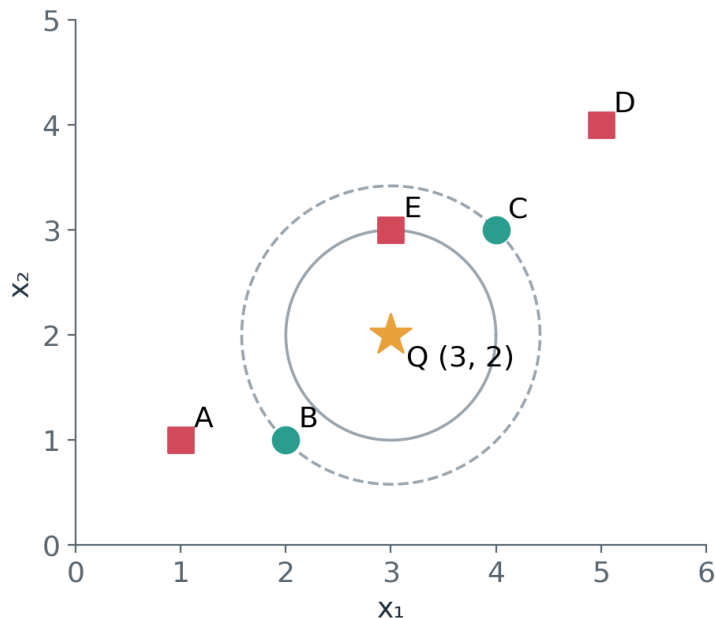
k-πλησιέστεροι γείτονες (kNN)



$$d(\mathbf{x}, \mathbf{z}) = \sqrt{[\sum_j (\mathbf{x}_j - \mathbf{z}_j)^2]} \text{ (ευκλείδεια απόσταση)}$$

Για παλινδρόμηση: μέσος όρος των τιμών των k γειτόνων

Παράδειγμα: ο kNN αλλάζει γνώμη με το k



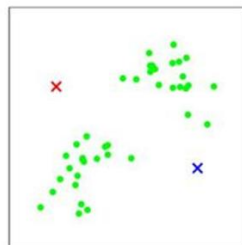
Σημείο	Κλάση	Απόσταση
E (3, 3)	κόκκινο	1,00
B (2, 1)	πράσινο	1,41
C (4, 3)	πράσινο	1,41
A (1, 1)	κόκκινο	2,24
D (5, 4)	κόκκινο	2,83

$k = 1 \rightarrow \text{κόκκινο}$ · $k = 3 \rightarrow \text{πράσινο}$ · $k = 5 \rightarrow \text{κόκκινο}$

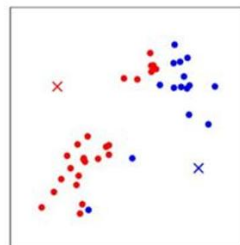
Κ-nn βήματα



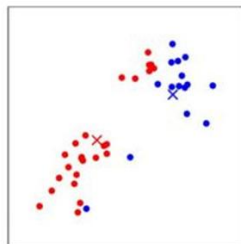
(a)



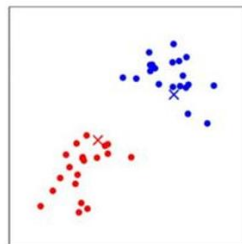
(b)



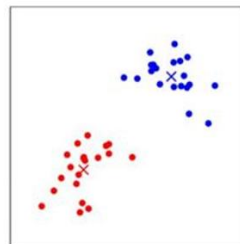
(c)



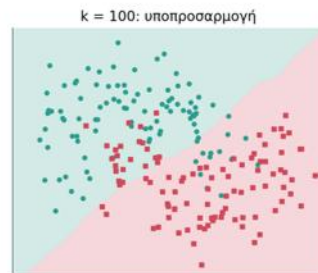
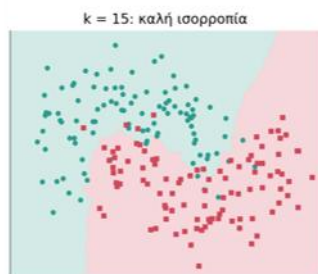
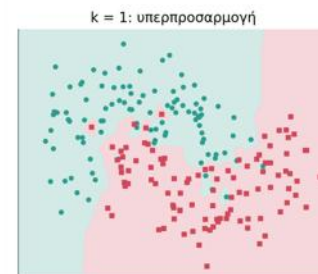
(d)



(e)



(f)



Ο kNN στην πράξη



Κλιμάκωση

Απαραίτητη: StandardScaler πριν από τον kNN



Κατάρα της διαστατικότητας

Σε πολλές διαστάσεις όλοι οι «γείτονες» απέχουν περίπου το ίδιο



Κόστος πρόβλεψης

Αποστάσεις προς όλα τα δείγματα σε κάθε πρόβλεψη (δομές KD-tree, ball tree)



Βάρη

weights="distance": οι πιο κοντινοί γείτονες μετράνε περισσότερο

Naive Bayes: ο κανόνας του Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

$$P(x|c) \approx P(x_1|c)P(x_2|c) \cdots P(x_n|c)$$

- $P(c)$: εκ των προτέρων πιθανότητα κλάσης (συχνότητα στο training set)
- $P(x_j | c)$: πόσο συχνό είναι το χαρακτηριστικό μέσα στην κλάση
- Η υπόθεση ανεξαρτησίας σπάνια ισχύει, αλλά ο NB δουλεύει εκπληκτικά καλά

Παράδειγμα: Naïve Bayes για spam

Email που περιέχει τις λέξεις «δωρεάν» και «συνάντηση»

	P(κλάση)	P(«δωρεάν» κλάση)	P(«συνάντηση» κλάση)	Γινόμενο
Spam	0,40	0,30	0,05	0,0060
Κανονικό	0,60	0,02	0,20	0,0024

$$P(\text{spam} | \text{email}) = 0,0060 / (0,0060 + 0,0024) \approx 0,71$$

Πρόβλεψη: spam

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Εξομάλυνση Laplace: το πρόβλημα του μηδενός

$$P(w|c) = \frac{\text{πλήθος}(w, c) + 1}{\text{σύνολο λέξεων στην } c + |V|}$$

	Χωρίς εξομάλυνση	Με εξομάλυνση Laplace
$P(\text{«λαχείο»} \mid \text{κανονικό})$	$0 / 1.000 = 0$	$1 / 6.000 \approx 0,00017$
Επίδραση στο γινόμενο	Μηδενίζεται	Μειώνεται, αλλά μετρούν και οι άλλες λέξεις

Υποθέσεις: 1.000 λέξεις σε κανονικά email, λεξιλόγιο $|V| = 5.000$ λέξεις

Στο scikit-learn: παράμετρος alpha (προεπιλογή 1,0)

Παραλλαγές του Naïve Bayes



GaussianNB

Συνεχή χαρακτηριστικά:
κανονική κατανομή ανά κλάση
(μ , σ^2)



MultinomialNB

Πλήθος εμφανίσεων (π.χ.
λέξεις σε κείμενο)



BernoulliNB

Διαδικά χαρακτηριστικά:
εμφανίζεται η λέξη ή όχι

LDA και QDA: γενετικοί ταξινομητές (ISL 4.4)



LDA (γραμμική)

- Κοινός πίνακας συνδιακύμανσης
- Γραμμικό όριο απόφασης
- Λιγότερες παράμετροι → χαμηλή διακύμανση
- Καλή με λίγα δεδομένα



QDA (τετραγωνική)

- Διαφορετική συνδιακύμανση ανά κλάση
- Καμπύλο (τετραγωνικό) όριο
- Περισσότερες παράμετροι → υψηλή διακύμανση
- Καλή με πολλά δεδομένα

Σύγκριση κλασικών ταξινομητών (ISL 4.5)

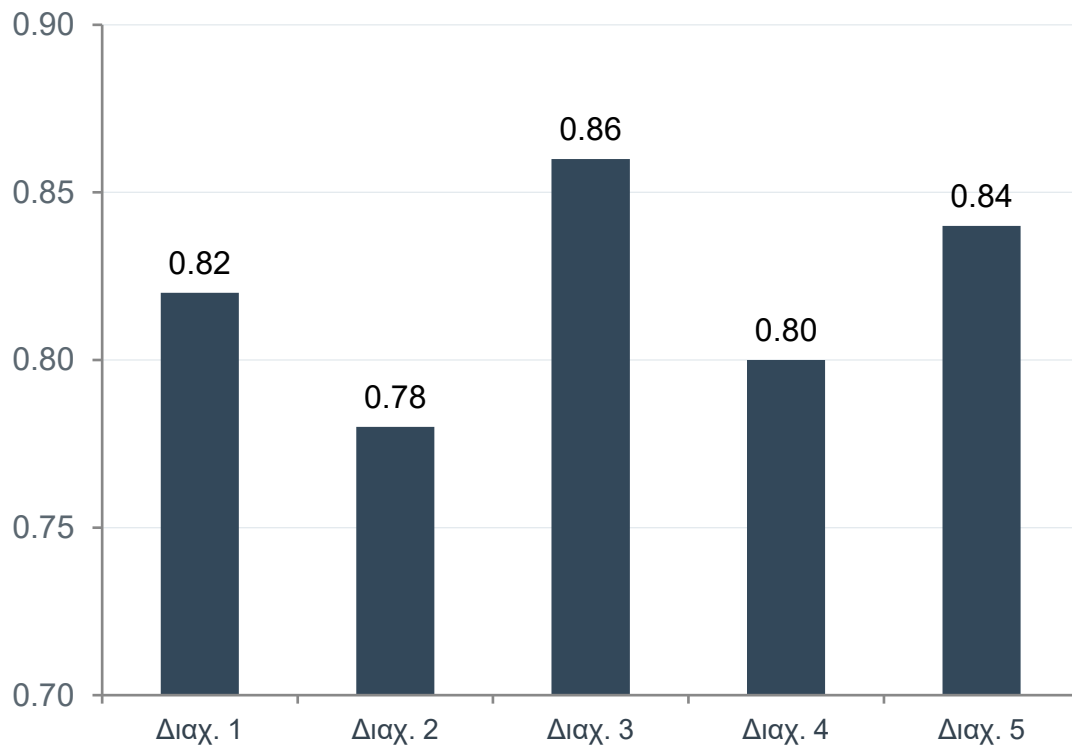
Ταξινομητής	Όριο απόφασης	Βασική υπόθεση	Πότε ταιριάζει
Λογιστική	Γραμμικό	Γραμμικά log-odds	Γραμμικό όριο, ερμηνευσιμότητα
LDA	Γραμμικό	Κανονικές, κοινή συνδιακύμανση	Λίγα δεδομένα, γραμμικό όριο
QDA	Τετραγωνικό	Κανονικές, ξεχωριστή συνδιακύμανση	Μέτρια μη γραμμικότητα
Naive Bayes	Ποικίλλει	Ανεξαρτησία χαρακτηριστικών	Κείμενο, πολλά χαρακτηριστικά
kNN	Οποιοδήποτε	Καμία (μη παραμετρικός)	Πολύπλοκο όριο, πολλά δεδομένα

ΕΝΟΤΗΤΑ 5

Επαναδειγματοληψία και ανισορροπία

Cross-validation σε βάθος (ISL 5.1) και μη ισορροπημένες κλάσεις

Το πρόβλημα ενός μόνο validation set



Δύο προβλήματα

1. Μεγάλη διακύμανση της εκτίμησης
2. Εκπαίδευση σε λιγότερα δεδομένα → απαισιόδοξη εκτίμηση

Leave-one-out cross-validation (LOOCV)

$$CV(n) = (1/n) \cdot \sum_i \sigmaφάλμα_i$$



Πλεονεκτήματα

- Σχεδόν όλα τα δεδομένα στην εκπαίδευση → χαμηλή μεροληψία
- Καμία τυχαιότητα: ίδιο αποτέλεσμα κάθε φορά



Μειονεκτήματα

- η εκπαίδευσεις: πολύ ακριβό
- Τα n μοντέλα μοιάζουν πολύ μεταξύ τους → υψηλή διακύμανση της εκτίμησης

k-fold: ο συμβιβασμός

Μέθοδος	Εκπαιδεύσεις	Μεροληψία εκτίμησης	Διακύμανση εκτίμησης
Ένα validation set	1	Υψηλή	Υψηλή
5-fold / 10-fold	5 / 10	Μέτρια	Μέτρια
LOOCV	n	Χαμηλή	Υψηλή

Πρακτική σύσταση: 5-fold ή 10-fold cross-validation.

Stratified k-fold για ταξινόμηση

1.000

δείγματα

50

θετικά (5%)

5 × 200

folds

10

θετικά σε κάθε fold

```
from sklearn.model_selection import StratifiedKFold, cross_val_score
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
cross_val_score(clf, X, y, cv=skf, scoring="f1")
```

Ο λάθος τρόπος να κάνουμε cross-validation



Λάθος

1. Επιλογή των 100 καλύτερων χαρακτηριστικών με όλα τα δεδομένα
2. Cross-validation στα επιλεγμένα

→ Αισιόδοξη εκτίμηση, ακόμη και με τυχαίες ετικέτες



Σωστό

Σε κάθε fold:

1. Επιλογή χαρακτηριστικών ΜΟΝΟ από το training fold
2. Εκπαίδευση και αξιολόγηση

→ `make_pipeline(SelectKBest(k=100), clf)`

Μη ισορροπημένες κλάσεις

0,5%

απάτες

199 : 1

νόμιμες ανά απάτη

99,5%

ακρίβεια του «ποτέ απάτη»

Συνήθη πεδία: απάτες, σπάνιες νόσοι, βλάβες μηχανημάτων, κυβερνοεπιθέσεις.

Στρατηγικές για μη ισορροπημένα δεδομένα



Σωστά μέτρα

PR-AUC, recall, F1 αντί για accuracy



Βάρη κλάσεων

`class_weight="balanced"`:
ακριβότερα τα λάθη στη
μειοψηφία



Ρύθμιση κατωφλίου

Κατώφλι με βάση το κόστος
κάθε λάθους



Undersampling

Αφαίρεση δειγμάτων της
πλειοψηφίας



Oversampling

Επανάληψη δειγμάτων της
μειοψηφίας



SMOTE

Συνθετικά δείγματα
μειοψηφίας με παρεμβολή

Παράδειγμα: `class_weight="balanced"`

$$w_c = n / (K \cdot n_c)$$

Κλάση	Δείγματα n_c	Βάρος w_c
Αρνητική	950	$1.000 / (2 \cdot 950) \approx 0,53$
Θετική	50	$1.000 / (2 \cdot 50) = 10$

$$\text{Λόγος βαρών } 10 / 0,53 = 19 = 950 / 50$$

Παράδειγμα: SMOTE με παρεμβολή

$$x_{\text{νέο}} = x + \lambda \cdot (x_{\text{γείτονας}} - x), \quad \lambda \in [0, 1]$$

	Συντεταγμένες
Δείγμα x	(2, 3)
Γείτονας	(4, 7)
Διαφορά	(2, 4)
$\lambda = 0,25 \rightarrow \text{νέο}$	$(2,5 \cdot 4)$



Προσοχή

Μόνο στο training set, μέσα σε κάθε fold.
Βιβλιοθήκη: imbalanced-learn (imblearn).

Παράδειγμα: κατώφλι με βάση το κόστος

Κόστος χαμένης απάτης (FN) = 100 € · Κόστος ψευδούς συναγερμού (FP) = 5 €

Κατώφλι	FN	FP	Κόστος FN	Κόστος FP	Σύνολο
0,5	20	90	2.000 €	450 €	2.450 €
0,3	8	250	800 €	1.250 €	2.050 €

Συμπέρασμα: το κατώφλι 0,3 είναι φθηνότερο κατά 400 €, παρά τους 160 επιπλέον ψευδείς συναγερμούς. Το κατώφλι είναι επιχειρησιακή απόφαση, όχι τεχνική λεπτομέρεια.



Ερώτηση Slido



Σχεδιάζετε προσυμπτωματικό έλεγχο για σοβαρή ασθένεια.
Όσοι βγουν θετικοί κάνουν μια δεύτερη, πιο ακριβή εξέταση.
Ποιο μέτρο πρέπει κυρίως να είναι υψηλό;

A

Precision

B

Recall (ευαισθησία)

Γ

Ειδικότητα

Δ

Ακρίβεια (accuracy)

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision Value $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

To classification_report του παραδείγματος

	precision	recall	f1-score	support
υγιής (0)	0,98	0,90	0,94	900
νόσος (1)	0,47	0,80	0,59	100
accuracy			0,89	1.000
macro avg	0,72	0,85	0,76	1.000
weighted avg	0,93	0,89	0,90	1.000

```
print(
    classification_report
    (
        y_test, y_pred)
)
```

Σε ανισορροπία: κοιτάμε το
macro avg

Οδηγός επιλογής μέτρου

Κατάσταση	Μέτρο
Ισορροπημένες κλάσεις, ίδιο κόστος λαθών	Accuracy, ROC-AUC
Ακριβά τα FP (spam, προτάσεις σε παιδιά)	Precision
Ακριβά τα FN (διάγνωση, απάτη)	Recall
Θέλουμε ισορροπία P και R	F1
Σπάνια θετική κλάση, σύγκριση μοντέλων	PR-AUC (average precision)
Γνωστό κόστος για κάθε λάθος	Αναμενόμενο κόστος

Σύνοψη

1

Η ακρίβεια παραπλανά σε μη ισορροπημένα δεδομένα

2

Confusion matrix → precision, recall, F1, ειδικότητα

3

Το κατώφλι ρυθμίζει το αντιστάθμισμα P-R

4

PR για σπάνια θετικά, ROC/AUC για ισορροπημένα

5

kNN, Naive Bayes, LDA/QDA: διαφορετικές υποθέσεις

6

Stratified k-fold, προεπεξεργασία μέσα στο fold, class weights

Για την επόμενη διάλεξη



Διάβασμα

- Géron, Κεφ. 3 (και ασκήσεις)
- ISL, Κεφ. 4.4–4.5 και 5.1
- Géron, Κεφ. 6 (για τη Διάλεξη 5)
- ISL, Κεφ. 8.1 · Κάλλιπος 4.3.1



Εξάσκηση

- Ανιχνευτής «5» στο MNIST: βρείτε το κατώφλι για precision 90%
- Συγκρίνετε SGD, kNN και Naive Bayes με 3-fold CV

Διάλεξη 5: Δένδρα Αποφάσεων



Ερωτήσεις;

koumakis@hmu.gr · Ώρες γραφείου: Δευτέρα 11:00–13:00