ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

ΕΙΣΑΓΩΓΗ

Επιλογή Μοντέλου (model selection):

- το μοντέλο που έχει την απαιτούμενη πολυπλοκότητα χρησιμοποιώντας την εκτίμηση του λάθους γενίκευσης

- Αφού κατασκευαστεί μπορεί να χρησιμοποιηθεί στα δεδομένα ελέγχου για να προβλέψει σε ποιες κλάσεις ανήκουν

- Για να γίνει αυτό πρέπει να ξέρουμε τις κλάσεις των δεδομένων ελέγχου

ΕΙΣΑΓΩΓΗ

- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
- Πως να εκτιμήσουμε την απόδοση ενός μοντέλου
- Τι θα μετρήσουμε

- Μέθοδοι για την εκτίμηση της απόδοσης
- Πως μπορούνε να πάρουμε αξιόπιστες εκτιμήσεις
- Πως θα το μετρήσουμε

- Μέθοδοι για την σύγκριση μοντέλων
- Πως να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

ΕΙΣΑΓΩΓΗ

- Αφού κατασκευαστεί ένα μοντέλο, θα θέλαμε να αξιολογήσουμε/εκτιμήσουμε την ποιότητα του/την ακρίβεια της ταξινόμησης που πετυχαίνει

- Έμφαση στην ικανότητα πρόβλεψης του μοντέλου παρά στην αποδοτικότητα του (πόσο γρήγορα κατασκευάζει το μοντέλο ή ταξινομεί μια εγγραφή, κλιμάκωση κλπ.)

Μέτρα για την απόδοση της εκτίμησης

- Επικεντρώνονται στην προβλεπτική ικανότητα ενός μοντέλου
- Πίνακας (μήτρα) σύγχυσης Confusion Matrix

		πρόβλεψη PREDICTED CLASS	
		Class=Yes	Class=No
πραγματική ACTUAL CLASS	Class=Yes	f_{11} **TP**	f_{10} **FN**
	Class=No	f_{01} **FP**	f_{00} **TN**

TP (true positive) f_{11}

FN (false negative) f_{10}

FP (false positive) f_{01}

TN (true negative) f_{00}

$\mathbf{f_{ij}}$: αριθμός των εγγραφών της κλάσης i που προβλέπονται ως κλάση j

Μέτρα για την απόδοση της εκτίμησης

Πιστότητα - Accuracy

$$\text{Accuracy} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01} + f_{10}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Λόγος Λάθους

$$\text{Error rate} = \frac{f_{01} + f_{10}}{f_{11} + f_{00} + f_{01} + f_{10}}$$

		πρόβλεψη PREDICTED CLASS	
πραγματική ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	f_{11} TP	f_{10} FN
	Class=No	f_{01} FP	f_{00} TN

ErrorRate(C) = 1 – Accuracy(C)

Μέτρα για την απόδοση της εκτίμησης

Έστω ένα πρόβλημα με 2 κατηγορίες

- Παραδείγματα από την κατηγορία 0 = 9990
- Παραδείγματα από την κατηγορία 1 = 10

- Εάν ένα μοντέλο προβλέπει ότι κάθε τι ανήκει στην κατηγορία 0, η ακρίβεια είναι 9990/10000 = 99.9 %

- Η ακρίβεια σε αυτή την περίπτωση είναι παραπλανητική γιατί το μοντέλο αποτυγχάνει να ανιχνεύσει οποιοδήποτε από τα παραδείγματα που ανήκουν στην κατηγορία 1

Μέτρα για την απόδοση της εκτίμησης

Μέτρα Εκτίμησης

Πίνακας Κόστους

	PREDICTED CLASS		
	C(i|j)	Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): κόστος λανθασμένης ταξινόμησης ενός παραδείγματος της κλάσης i ως κλάση j

Μέτρα για την απόδοση της εκτίμησης

	PREDICTED CLASS	
ACTUAL CLASS	C(i|j)	**Class = +**	**Class = -**
	Class = +	**TP** F_{11}	**FN** F_{10}
	Class = -	**FP** F_{01}	**TN** F_{00}

Αρνητική τιμή κόστους σημαίνει επιπρόσθετη «επιβράβευση» σωστής πρόβλεψης

C(M) = **TP** x C(+, +) + **FN** x C(+, -) + **FP** C(-, +) + **TN** C(-, -)

C(+, +) = C(-, -) = 0 -> όχι επιβράβευση

C(+, -) = C(-, +) = 1 -> κάθε λάθος μετρά 1

Παράδειγμα

Υπολογισμός του Κόστους της Ταξινόμησης

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Μέτρα Εκτίμησης

Ταξινόμηση που λαμβάνει υπό όψιν της το κόστος

Κατασκευή Δέντρου Ταξινόμησης

1. Επιλογή γνωρίσματος στο οποίο θα γίνει η διάσπαση
2. Στην απόφαση αν θα ψαλιδιστεί κάποιο υπο-δέντρο
3. Στον καθορισμό της κλάσης του φύλλου

Μέτρα Εκτίμησης

- Έστω p(j) τον ποσοστό των εγγραφών του κόμβου που ανήκουν στην κλάση j
- Τότε,

Leaf-label = max p(j), το ποσοστό των εγγραφών της κλάσης j που έχουν ανατεθεί στον κόμβο

Δίνουμε την κλάση i στον κόμβο που έχει το

ελάχιστο:

$$\sum_{j} p(j)C(j,i)$$

Κόστος εναντίον Ακρίβειας

Μετρήσεις	ΠΡΟΒΛΕΥΘΕΙΣΑ	
ΠΡΑΓΜ ΑΤΙΚΗ		
	Class=Yes	Class=No
Class=Yes	a	b
Class=No	c	d

Η ακρίβεια είναι ανάλογη του κόστους εάν

1. C(Yes|No)=C(No|Yes) = q
2. C(Yes|Yes)=C(No|No) = p

$N = a + b + c + d$

$\text{Accuracy} = (a + d)/N$

Κόστη	ΠΡΟΒΛΕΥΘΕΙΣΑ	
ΠΡΑΓΜ ΑΤΙΚΗ		
	Class=Yes	Class=No
Class=Yes	p	q
Class=No	q	p

$$\begin{aligned} \text{Cost} &= p\,(a + d) + q\,(b + c) \\ &= p\,(a + d) + q\,(N - a - d) \\ &= q\,N - (q - p)(a + d) \\ &= N\,[q - (q-p) \times \text{Accuracy}] \end{aligned}$$

Μέτρα Ακρίβειας

	PREDICTED CLASS	
ACTUAL CLASS		
	Class=Yes	**Class=No**
Class=Yes	**TP**	**FN**
Class=No	**FP**	**TN**

False positive rate: Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως θετικά)

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

False negative rate: Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως αρνητικά)

$$\mathrm{FNR} = \frac{FN}{TP + FN}$$

Μέτρα Ακρίβειας

Μετρήσεις	ΠΡΟΒΛΕΥΘΕΙΣΑ		
ΠΡΑΓΜ ΑΤΙΚΗ		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

$$\text{Precision (p)} = \frac{a}{a+c}$$

Πόσα από τα παραδείγματα που ο ταξινομητής έχει ταξινομήσει ως θετικά είναι πραγματικά θετικά Όσο πιο μεγάλη η ακρίβεια, τόσο μικρότερος ο αριθμός των FP

$$\text{Recall (r)} = \frac{a}{a+b}$$

Πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει Όσο πιο μεγάλη η ανάκληση, τόσο λιγότερα θετικά παραδείγματα έχουν ταξινομεί λάθος (=TPR)

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

Τείνει να είναι πιο κοντά στο μικρότερο από τα δύο

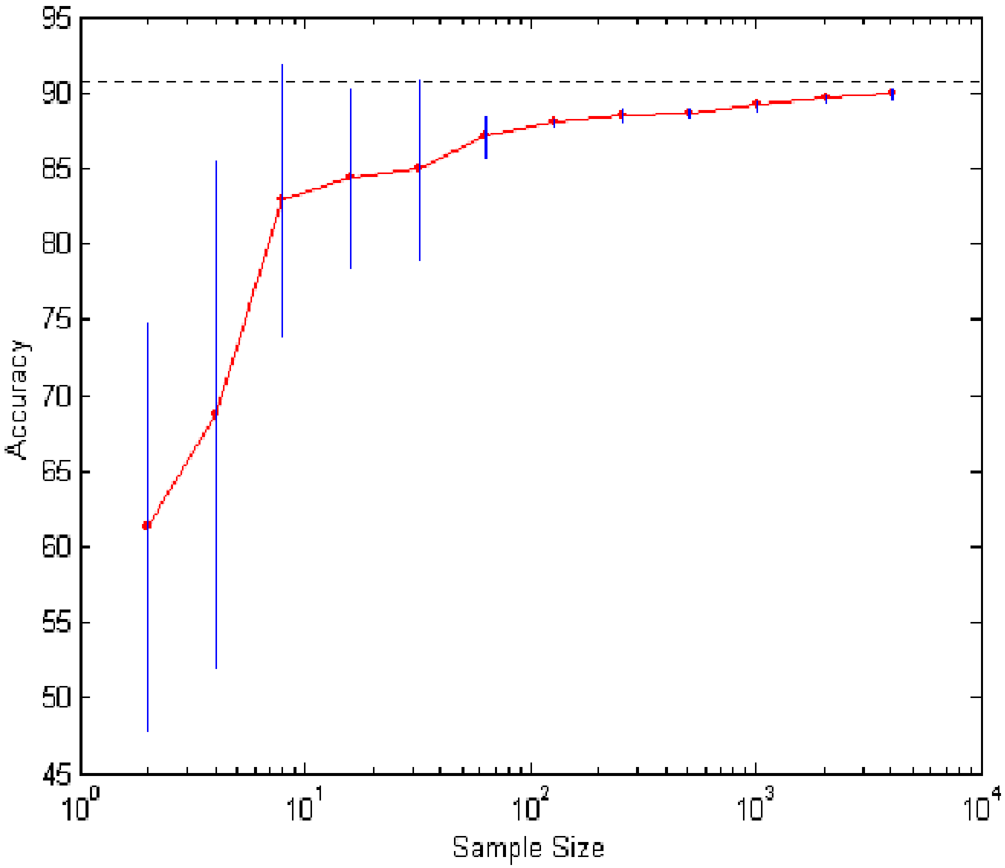
Υψηλή τιμή σημαίνει ότι και τα δύο είναι ικανοποιητικά μεγάλα

Μέτρα Ακρίβειας

- Πως μπορούμε να πάρουμε αξιόπιστες εκτιμήσεις της απόδοσης
- Η απόδοση ενός μοντέλου μπορεί να εξαρτάται από πολλούς παράγοντες εκτός του αλγορίθμου μάθησης:

1. Κατανομή των κλάσεων
2. Το κόστος της λανθασμένης ταξινόμησης
3. Το μέγεθος του συνόλου εκπαίδευσης και του συνόλου ελέγχου

Καμπύλη Μάθησης



- Η καμπύλη μάθησης δείχνει πως μεταβάλλεται η πιστότητα (accuracy) με την αύξηση του μεγέθους του δείγματος

- Επίδραση δείγματος μικρού μεγέθους:
  - Bias in the estimate
  - Variance of estimate

ROC (Receiver Operating Characteristic Curve)

- Η καμπύλη ROC δείχνει τα TPR [TruePositiveRate] (στον άξονα των y) προς τα FPR [FalsePositiveRate] (στον άξονα των x)
- Η απόδοση κάθε ταξινομητή αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

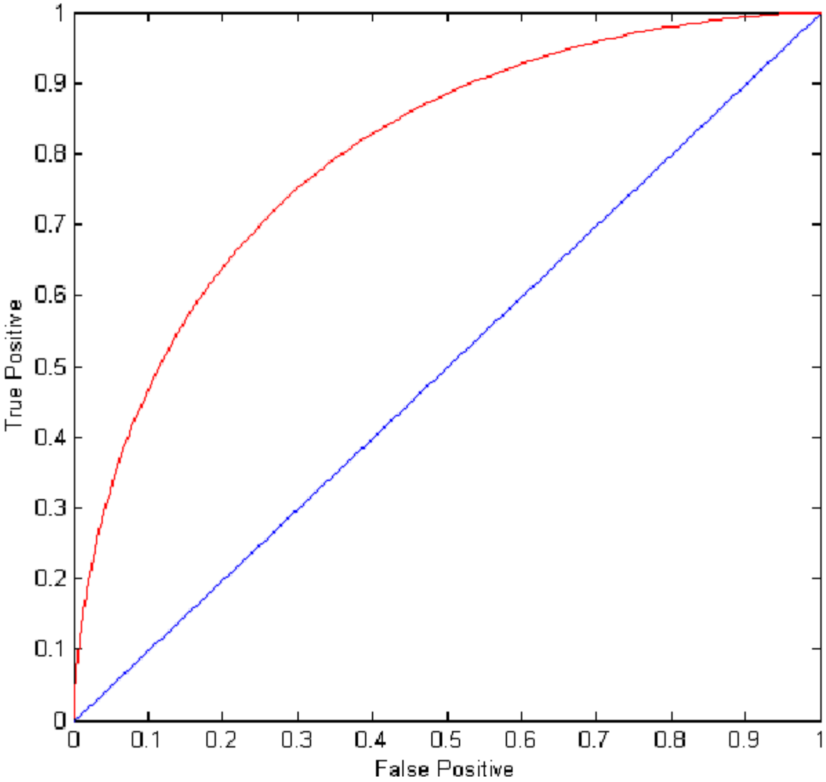
Πόσα από τα θετικά βρίσκει

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

Πόσα από τα αρνητικά βρίσκει

$$\mathrm{FPR} = \frac{FP}{TN + FP}$$

ROC (Receiver Operating Characteristic Curve)



(TP,FP): (0,0): όλα σημειώνονται ως αρνητική κλάση
(1,1): όλα σημειώνονται ως θετική κλάση
(1,0): ιδανική ταξινόμηση

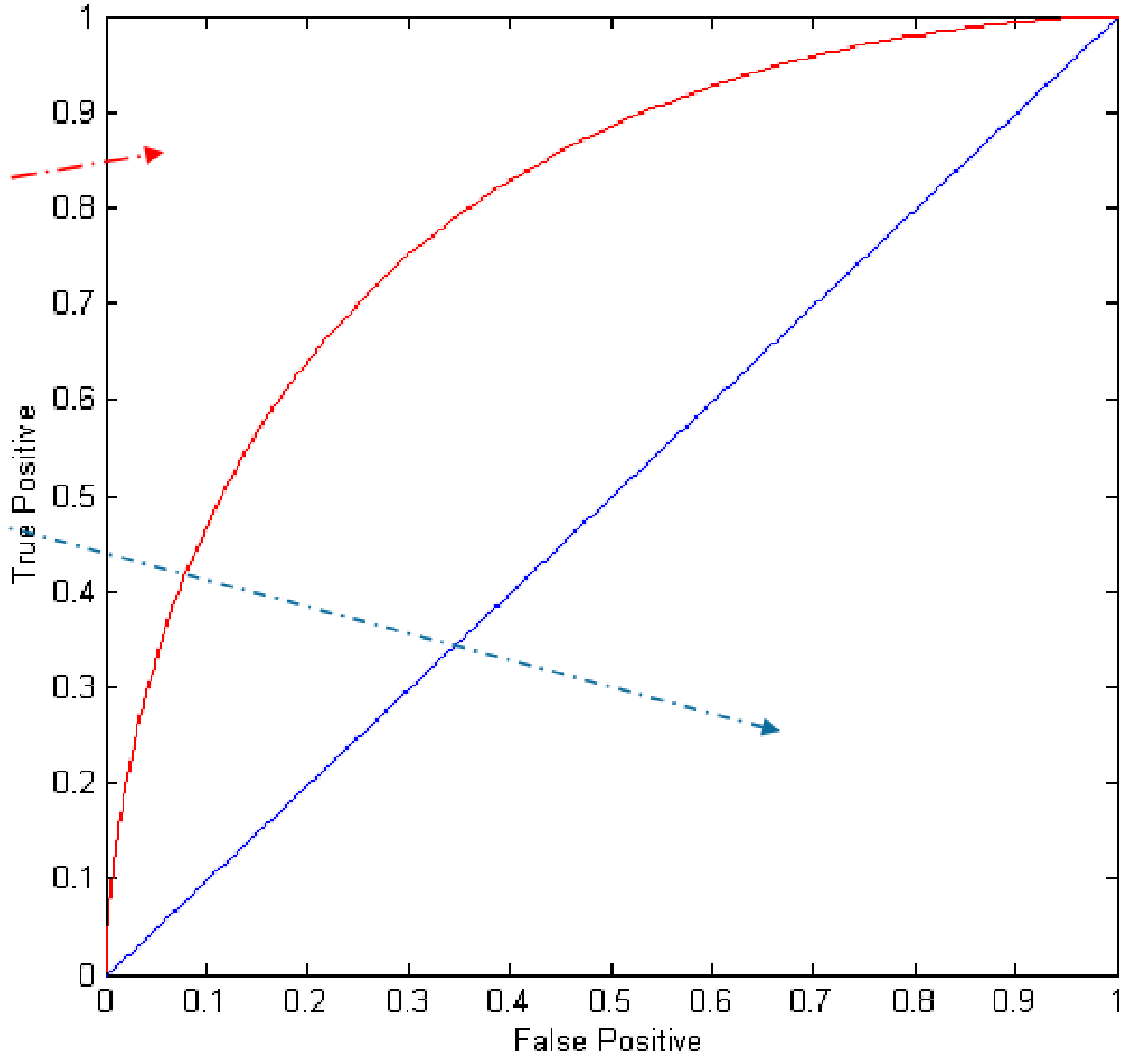
Διαγώνιος: Τυχαία εκτίμηση κλάσης

$$\mathrm{TPR} = \frac{TP}{TP + FN} \qquad \mathrm{FPR} = \frac{FP}{TN + FP}$$

ROC (Receiver Operating Characteristic Curve)

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

Κάτω από τη διαγώνιο Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



Αποτίμηση Μοντέλου

- Δοσμένων δύο μοντέλων, έστω M_1 και M_2, ποιο είναι καλύτερο;

M_1 ελέγχεται στο D_1 (size=n_1), error rate = e_1

M_2 ελέγχεται στο D_2 (size=n_2), error rate = e_2

Έστω D_1 και D_2 είναι ανεξάρτητα

Αν τα n_1 και n_2 είναι αρκετά μεγάλα, τότε:

$$e_1 \sim N(\mu_1, \sigma_1)$$
$$e_2 \sim N(\mu_2, \sigma_2)$$

$$\hat{\sigma}_i = \frac{e_i(1-e_i)}{n_i}$$

Αποτίμηση Μοντέλου

Για να ελέγξουμε αν η απόδοση διαφέρει στατιστικά σημαντικά: d=e1-e2

d ~ $N(d_t,\sigma_t)$ όπου d_t είναι η πραγματική διαφορά

Αφού τα D1 και D2 είναι ανεξάρτητα, η διασπορά τους προστίθεται

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2 = \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

επίπεδο εμπιστοσύνης (1-α)

$$d_t = d \pm Z_{\alpha/2}\hat{\sigma}_t$$

Παράδειγμα

Έστω: M1: n1 = 30, e1 = 0.15
M2: n2 = 5000, e2 = 0.25

d = |e2 – e1| = 0.1

$$\hat{\sigma}_d = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

Σε επίπεδο εμπιστοσύνης 95%, $Z_{\alpha/2}$=1.96

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> το διάστημα περιέχει το 0 => η διαφορά δεν είναι στατιστικά σημαντική

Αξιολογήσεις

Μετρικές Αξιολόγησης Μεθόδων Κατηγοριοποίησης						
	Χρόνος Εκπαίδευσης	Χρόνος Εκτέλεσης	Ανοχή στο Θόρυβο	Χρήση Προϋπάρχου σας Γνώσης	Ακρίβεια	Κατανοησιμ ότητα
Δέντρα Απόφασης	Μικρός	Μικρός	Μικρή	Όχι	Μέτρια	Καλή
Νευρωνικά Δίκτυα	Μεγάλος	Μικρός	Καλή	Όχι	Καλή	Μικρή
Bayesian Μέθοδοι	Μεγάλος	Μικρός	Καλή	Ναι	Καλή	Καλή