Αναγνώριση Προτύπων

Αξιολόγηση μοντέλων ταξινόμησης

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών

&

Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

ARISTOTLE
UNIVERSITY OF
THESSALONIKI

Διάρθρωση διάλεξης

- Αξιολόγηση μοντέλων
- Σύγκριση μοντέλων/αλγορίθμων
- Υποεκπαίδευση/υπερεκπαίδευση/κόστος εκπαίδευσης

Αξιολόγηση Μοντέλων

- Μετρικές για την αξιολόγηση του μοντέλου
- Καθορισμός αξιόπιστων εκτιμήσεων
- Σύγκριση μοντέλων (σχετική επίδοση)

Μετρικές για την αξιολόγηση μοντέλων

- Στόχος η ικανότητα πρόβλεψης του μοντέλου (όχι η ταχύτητα κατασκευής μοντέλου ή ταξινόμησης, η κλιμάκωση κτλ)
- Confusion Matrix (Πίνακας Σύγχυσης):

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Μετρικές για την αξιολόγηση μοντέλων

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

- Η πιο συνήθης μετρική – **Ακρίβεια**:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Περιορισμοί του Accuracy

- Θεωρήστε ένα πρόβλημα 2 κλάσεων:
  - Στιγμιότυπα που ανήκουν στην Class 0 = 9990
  - Στιγμιότυπα που ανήκουν στην Class 1 = 10
- Εάν το μοντέλο προβλέψει ότι όλες οι εγγραφές πρέπει να ανήκουν στην κλάση 0, τότε:

$$\text{Accuracy} = 9990/10000 = 99.9\ \%$$

- Πολύ καλή τιμή για Accuracy, πολύ κακό μοντέλο!!!

Πίνακας κόστους (Cost Matrix)

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

C(i|j): Το κόστος λάθος ταξινόμησης μιας εγγραφής της κλάσης j ως κλάση i

Υπολογισμός του κόστους Ταξινόμησης

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

Cost vs Accuracy

Count	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

Cost	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	p	q
	Class=No	q	p

Το Accuracy είναι ανάλογο του κόστους εάν:

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

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

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

Μετρικές ευαίσθητες στο κόστος

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

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

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

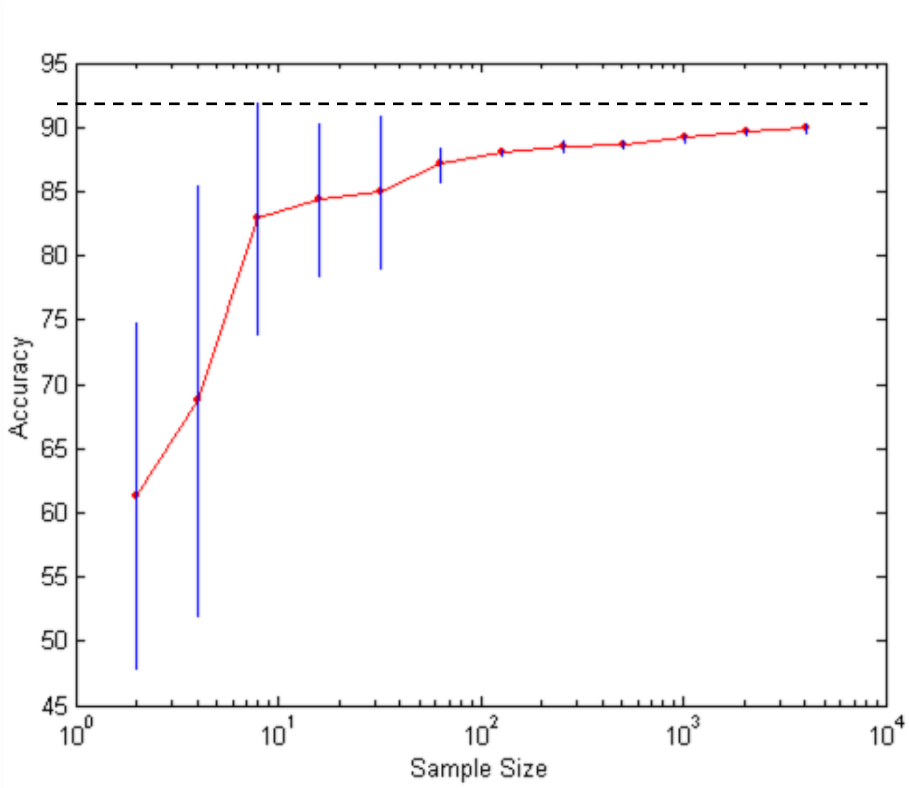
- Το Precision είναι πολωμένο προς τα: C(Yes|Yes) & C(Yes|No)
- Το Recall είναι πολωμένο προς τα: C(Yes|Yes) & C(No|Yes)
- Το F-measure είναι πολωμένο προς όλα εκτός από τα: C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

Αξιόπιστες εκτιμήσεις επίδοσης

- Η επίδοση ενός μοντέλου μπορεί να εξαρτάται και από άλλες παραμέτρους πέρα από τον αλγόριθμο εκμάθησης:
  - Την κατανομή κλάσεων
  - Το κόστος λανθασμένης ταξινόμησης
  - Το μέγεθος του σετ εκπαίδευσης και του σετ ελέγχου

Καμπύλη εκμάθησης



- Η καμπύλη εκμάθησης δείχνει πώς αλλάζει το accuracy χρησιμοποιώντας διαφορετικά μεγέθη του σετ εκπαίδευσης
- Απαιτεί μια μεθοδολογία δειγματοληψίας:
  - Αριθμητική δειγματοληψία
  - Γεωμετρική δειγματοληψία
- Μικρό σετ δειγματοληψίας:
  - Πόλωση στην εκτίμηση
  - Διακύμανση στη δειγματοληψία

Μέθοδοι Εκτίμησης

- Παρακράτηση (Holdout)
  - Χρησιμοποίηση 2/3 για εκπαίδευση και 1/3 για έλεγχο
- Τυχαία υποδειγματοληψία
  - Επαναλαμβανόμενο holdout
- **Cross validation!!!**
  - Χώρισε το σετ δεδομένων σε k μη επικαλυπτόμενα τμήματα
  - k-fold: εκπαίδευσε σε k-1 τμήματα, έλεγξε με το εναπομείναν
- Στρωματοποιημένη δειγματοληψία (Stratified sampling)
  - oversampling vs undersampling
- Bootstrap
  - Δειγματοληψία με αντικατάσταση

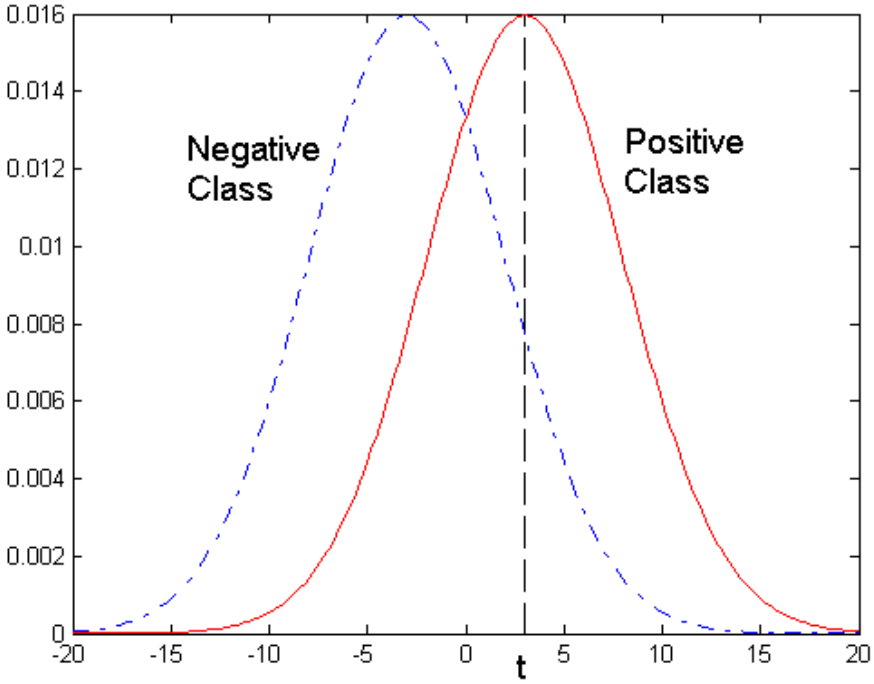
Σύγκριση μοντέλων: ROC (Receiver Operating Characteristics)

- Εμφανίστηκαν στα 1950s στο χώρο του σήματος για την ανάλυση θορυβωδών σημάτων
  - Χαρακτηρίζει το συμβιβασμό ανάμεσα σε σωστές ταξινομήσεις (true positives) και λάθος αναφορές για σωστές ταξινομήσεις (false positives)
- Οι ROC καμπύλες αναπαριστούν τα TP (στον y-άξονα) σε σχέση με τα FP (στον x-άξονα)
- Η επίδοση ενός μοντέλου ταξινόμησης αναπαρίσταται ως ένα σημείο στη ROC καμπύλη
  - Αλλάζοντας τα κατώφλια ενός αλγορίθμου, το δείγμα εκπαίδευσης ή τον πίνακα κόστους αλλάζει και η θέση του σημείου

Καμπύλη ROC

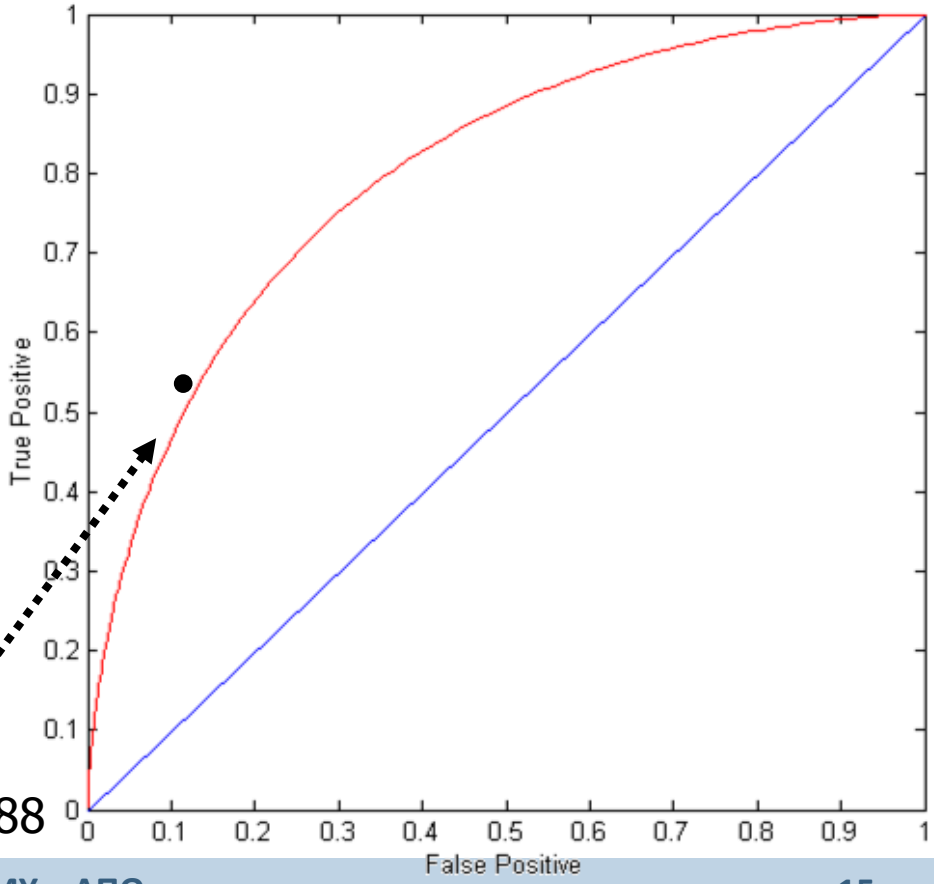
- Έστω μονοδιάστατα δεδομένα που ταξινομούνται σε 2 κλάσεις (θετική και αρνητική)

- Όσα σημεία βρίσκονται στο x > t ταξινομούνται ως θετικά



Στο κατώφλι t:

TP=0.5, FN=0.5, FP=0.12, TN=0.88

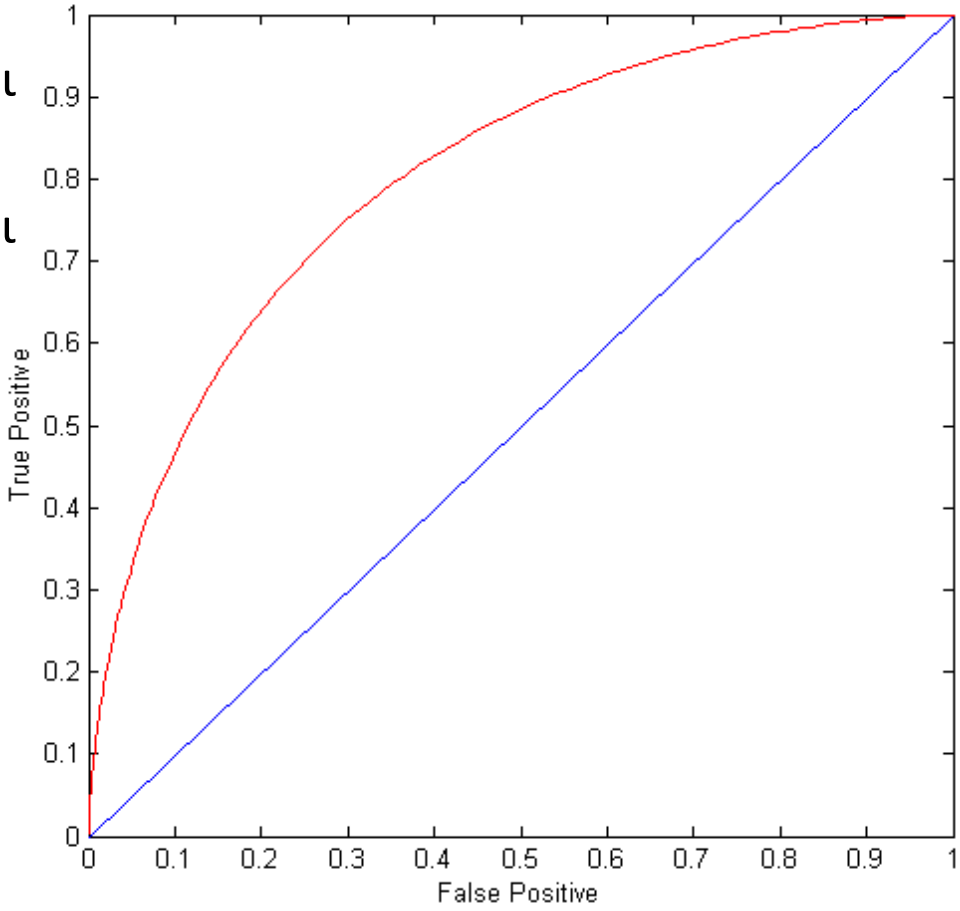


Καμπύλη ROC

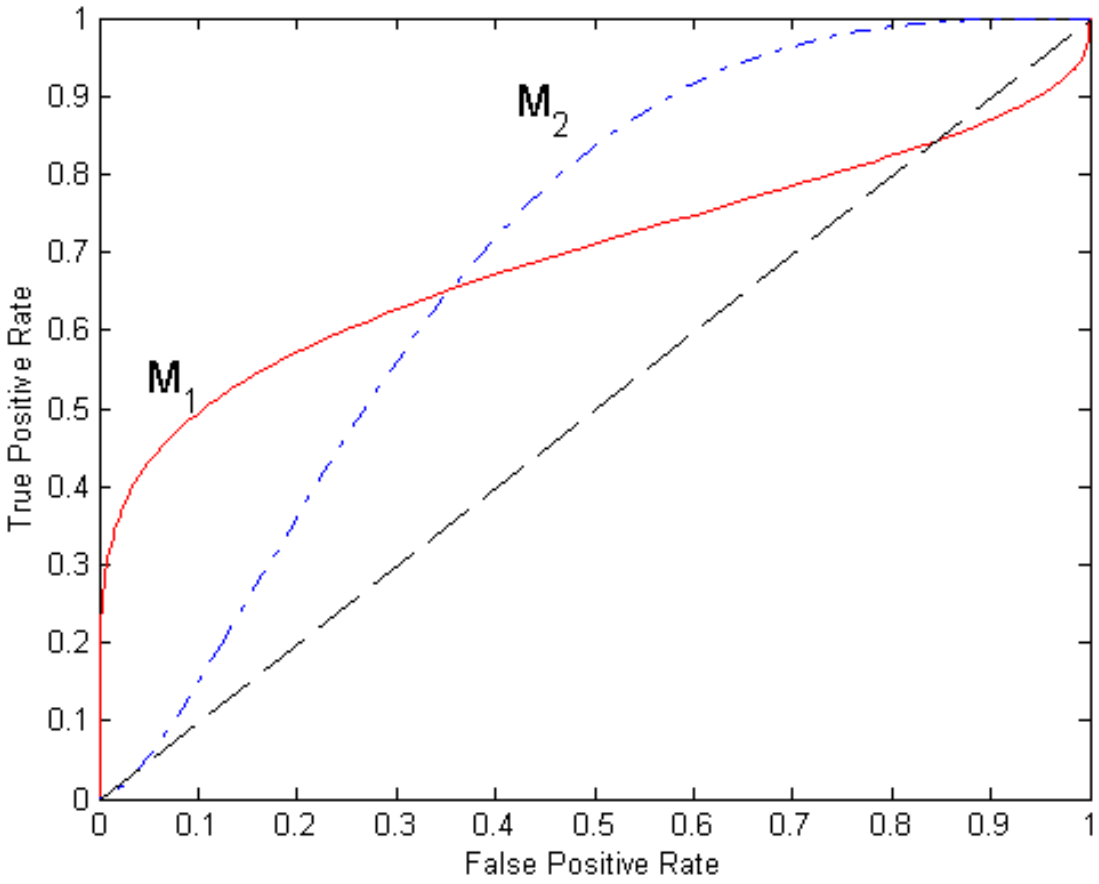
(TP,FP):

- (0,0): δηλώνει ότι όλα ταξινομούνται στην αρνητική κλάση
- (1,1): δηλώνει ότι όλα ταξινομούνται στη θετική κλάση
- (0,1): ιδανικό

- Κύρια διαγώνιος:
  - Τυχαία ταξινόμηση
- Κάτω από την κύρια διαγώνιο:
  - Οι ταξινομήσεις είναι ανάποδα από την πραγματική κλάση



Χρήση της καμπύλης ROC για σύγκριση μοντέλων



- Κανένα δεν είναι καλύτερο από το άλλο
- Το M_1 είναι καλύτερο για μικρές τιμές FPR
- Το M_2 είναι καλύτερο για μεγάλες τιμές FPR
- Κάτω από την καμπύλη
  - Ιδανικά:
    - Area = 1
  - Τυχαία ταξινόμηση:
    - Area = 0.5

Κατασκευή ROC καμπύλης

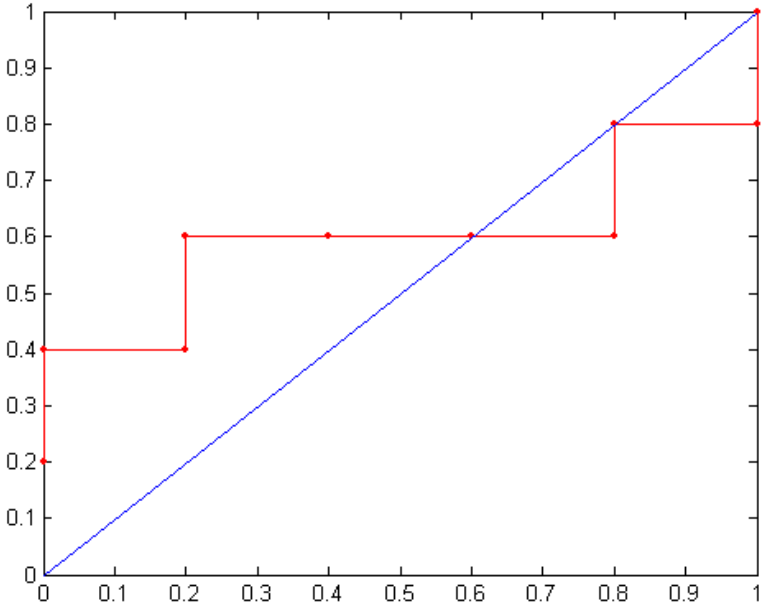
Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Χρήση ενός ταξινομητή που παράγει posterior πιθανότητα για κάθε στιγμιότυπο P(+|A)
- Ταξινομήστε σε φθίνουσα σειρά τις P(+|A)
- Ξεκινήστε από κάτω
- Εφαρμόστε ένα κατώφλι για κάθε μοναδική τιμή για το P(+|A), ταξινομώντας όλα τα από πάνω στιγμιότυπα ως θετικά.
- Επιλέξτε το επόμενο σημείο και θεωρήστε όλα τα από πάνω του θετικά, και όλα τα από κάτω του αρνητικά
- Υπολογίστε τον αριθμό των TP, FP, TN, FN σε κάθε κατώφλι:
  - TP rate, TPR = TP/(TP+FN)
  - FP rate, FPR = FP/(FP + TN)

Κατασκευή ROC καμπύλης

	Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=		0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
	TP	5	4	4	3	3	3	3	2	2	1	0
	FP	5	5	4	4	3	2	1	1	0	0	0
	TN	0	0	1	1	2	3	4	4	5	5	5
	FN	0	1	1	2	2	2	2	3	3	4	5
→	TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
→	FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

Καμπύλη:



Έλεγχος σημαντικότητας

- Δεδομένων 2 μοντέλων:
  - Μοντέλο M_1: accuracy = 85%, έλεγχος σε 30 στιγμιότυπα
  - Μοντέλο M_2: accuracy = 75%, έλεγχος σε 5000 στιγμιότυπα

- Είναι το M_1 καλύτερο από το M_2;
  - Πόση εμπιστοσύνη έχουμε στο accuracy των M_1 και M_2;
  - Μπορεί η μετρική επίδοσης (accuracy score) να ερμηνευτεί ως τυχαία διακύμανση στο σετ ελέγχου;

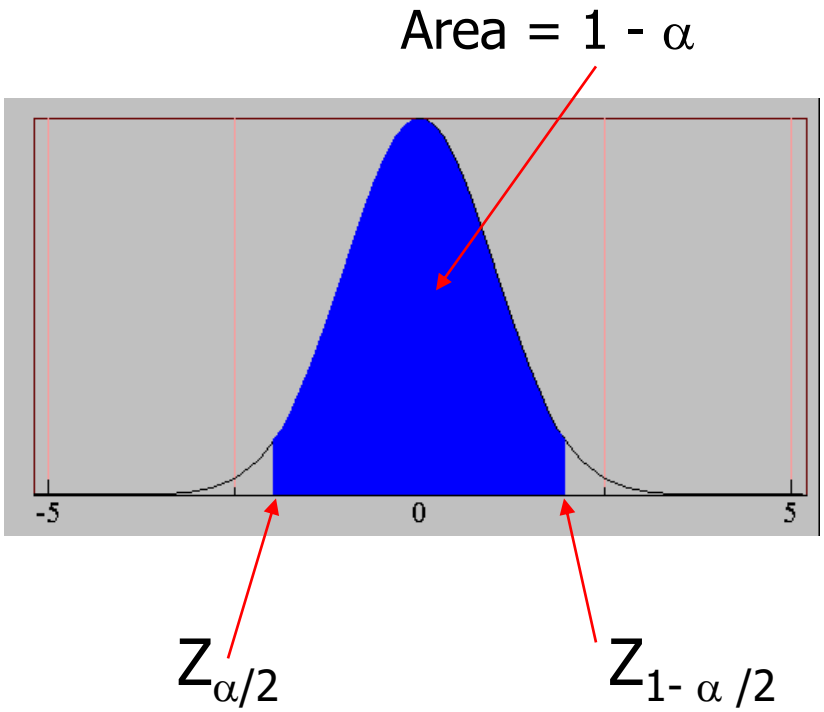
Περιθώριο εμπιστοσύνης για το Accuracy

- Η πρόβλεψη μπορεί να θεωρηθεί ως μια δοκιμή Bernoulli
  - Μια δοκιμή Bernoulli έχει 2 πιθανά αποτελέσματα
  - Πιθανά αποτελέσματα για την πρόβλεψη της ταξινόμησης: σωστή ή λάθος
  - Η συλλογή των δοκιμών Bernoulli έχει πολυωνυμική κατανομή:
    - $x \sim Bin(N, p)$, όπου x: αριθμός σωστός προβλέψεων
    - Π.χ.: Στρίψτε ένα κέρμα 50 φορές, πόσες φορές θα φέρει κορώνα?
      Αναμενόμενος αριθμός κορώνα = $N \times p = 50 \times 0.5 = 25$
- Άρα η κατανομή του x είναι κανονική με μέση τιμή $N \times p$ και τυπική απόκλιση $N \times p \times (1-p)$.
- Δεδομένου του x (# σωστών προβλέψεων) ή αντίστοιχα, acc=x/N, και N (# στιγμιοτύπων ελέγχου) μπορούμε να προβλέψουμε το p (την πραγματική ακρίβεια του μοντέλου);

Περιθώριο εμπιστοσύνης για το Accuracy

- Για μεγάλα σετ ελέγχου (N > 30),
  - acc έχει κανονική κατανομή με μέση τιμή p και διακύμανση p(1-p)/N, δηλ.

$$P(Z_{\alpha/2} < \frac{acc - p}{\sqrt{p(1-p)/N}} < Z_{1-\alpha/2}) = 1-\alpha$$



- Το περιθώριο εμπιστοσύνης για το p είναι:

$$p = \frac{2 \times N \times acc + Z^2_{\alpha/2} \pm \sqrt{Z^2_{\alpha/2} + 4 \times N \times acc - 4 \times N \times acc^2}}{2(N + Z^2_{\alpha/2})}$$

Περιθώριο εμπιστοσύνης για το Accuracy

- Θεωρήστε ένα μοντέλο με accuracy 80% πάνω σε ένα τεστ ελέγχου 100 στιγμιοτύπων
  - N=100, acc = 0.8
  - Έστω 1-α = 0.95 (95% confidence)
  - Από τον πίνακα πιθανοτήτων, $Z_{\alpha/2}$=1.96

1-α	$Z_{\alpha/2}$
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

N	50	100	500	1000	5000
p(lower)	0.670	0.711	0.763	0.774	0.789
p(upper)	0.888	0.866	0.833	0.824	0.811

Σύγκριση της επίδοσης 2 μοντέλων

- Δεδομένων 2 μοντέλων, M_1 και M_2, ποιο είναι καλύτερο;
  - Το M_1 ελέγχθηκε με το D_1 (μέγεθος=n_1), με ρυθμό σφάλματος* = e_1
  - Το M_2 ελέγχθηκε με το D_2 (μέγεθος=n_2), με ρυθμό σφάλματος = e_2
  - Έστω ότι D_1 και D_2 ανεξάρτητα
  - Εάν τα n_1 και n_2 είναι ικανά μεγάλα, τότε:

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

*Ρυθμός σφάλματος: 1-accuracy

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

Σύγκριση της επίδοσης 2 μοντέλων

- Έλεγχος εάν η διαφορά είναι στατιστικά σημαντική:

$$d = |e_1 - e_2|$$

  - $d \sim N(d_t, \sigma_t)$, όπου d_t η πραγματική διαφορά
  - Αφού D_1 και D_2 ανεξάρτητα, οι τυπικές τους αποκλίσεις αθροίζονται: $\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$

$$= \frac{e_1(1-e_1)}{n_1} + \frac{e_2(1-e_2)}{n_2}$$

  - Στο επίπεδο εμπιστοσύνης (1-α),

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

Παράδειγμα

- Δεδομένα: M_1: $n_1 = 30$, $e_1 = 0.15$
  M_2: $n_2 = 5000$, $e_2 = 0.25$
- $d = |e_2 - e_1| = 0.1$ (2-sided test)

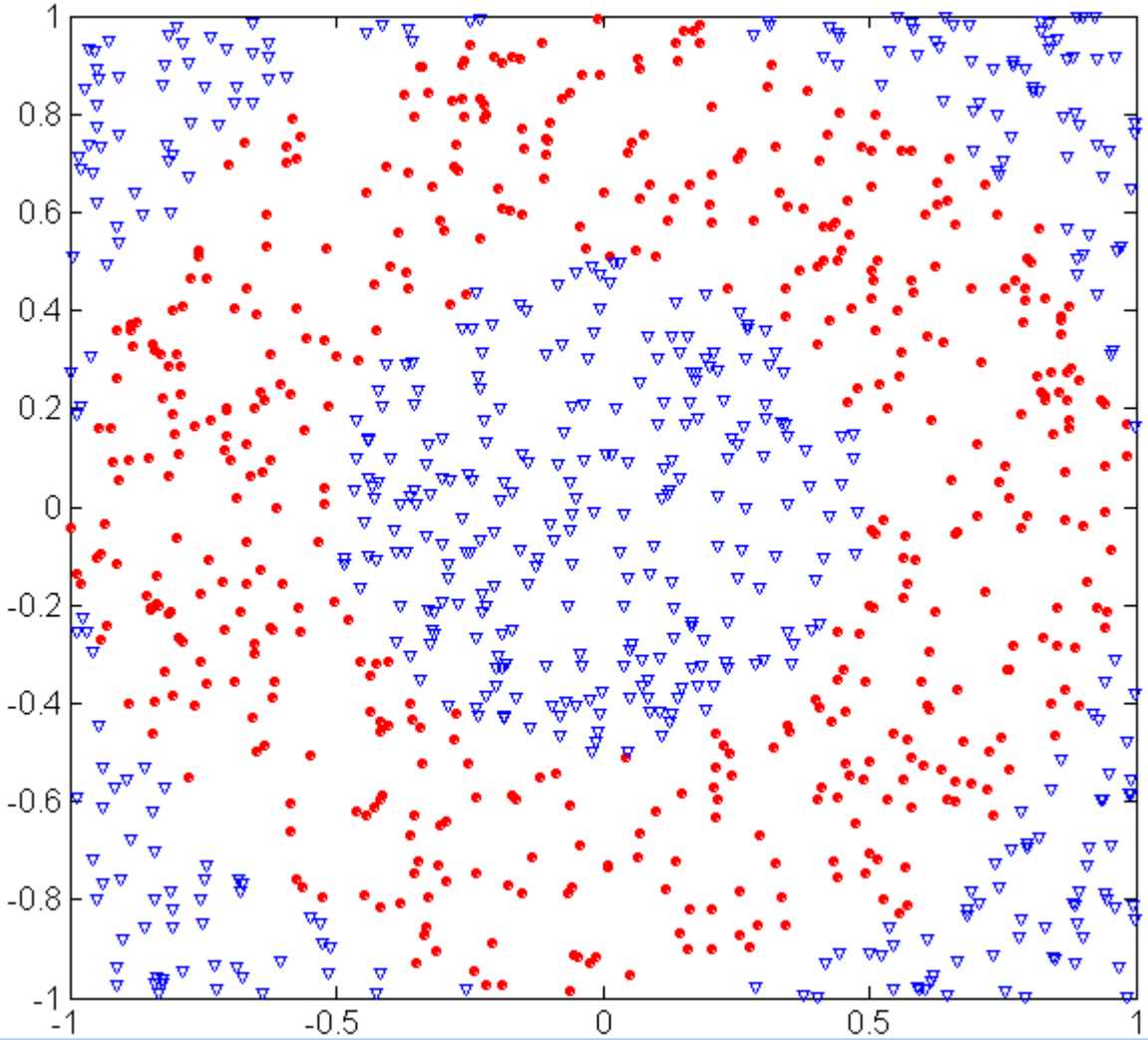
$$\hat{\sigma}_t = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- Στο 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 μπορεί να μην είναι στατιστικά σημαντική

Υπερεκπαίδευση (Παράδειγμα)



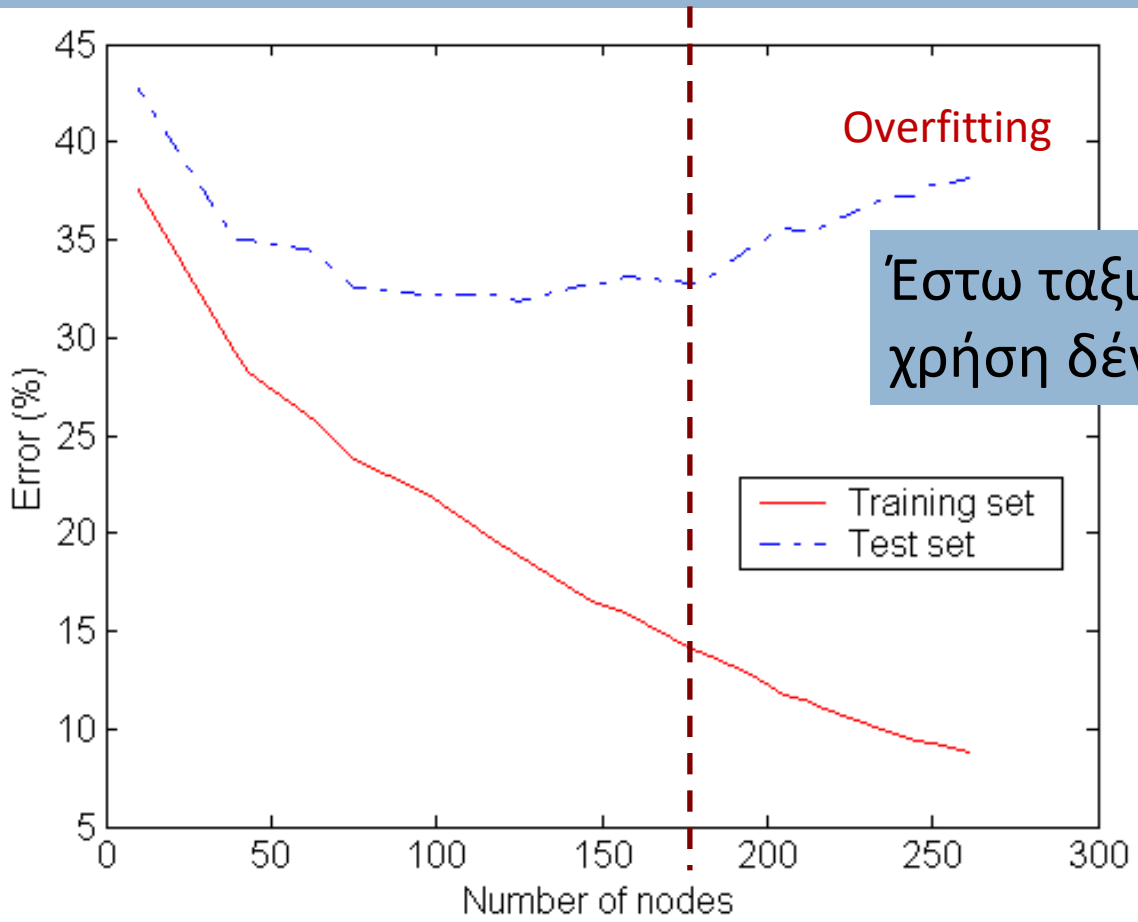
- 500 κυκλικά και 500 τριγωνικά σημεία.
- Για τα κυκλικά σημεία:

  $0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$
- Για τα τριγωνικά σημεία:

  $\text{sqrt}(x_1^2+x_2^2) > 0.5$ ή

  $\text{sqrt}(x_1^2+x_2^2) < 1$

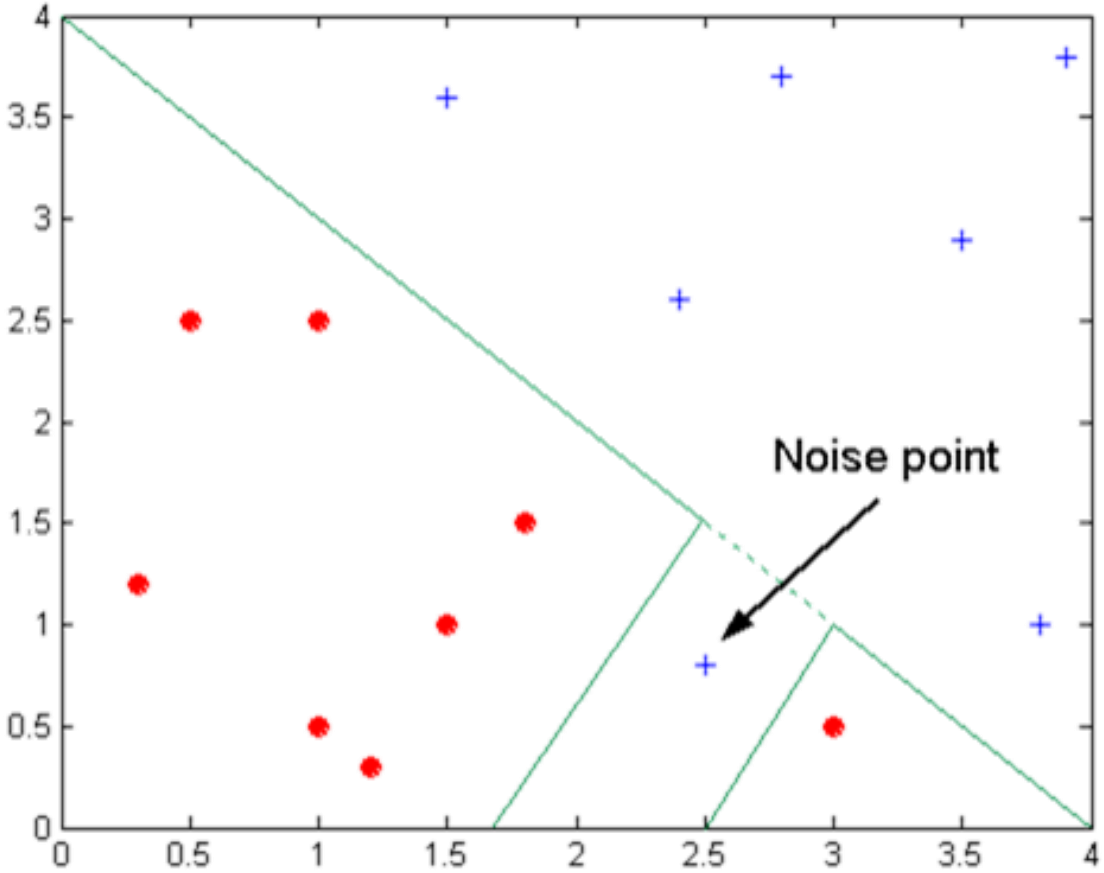
Υπερεκπαίδευση



Έστω ταξινόμηση με τη χρήση δένδρων απόφασης

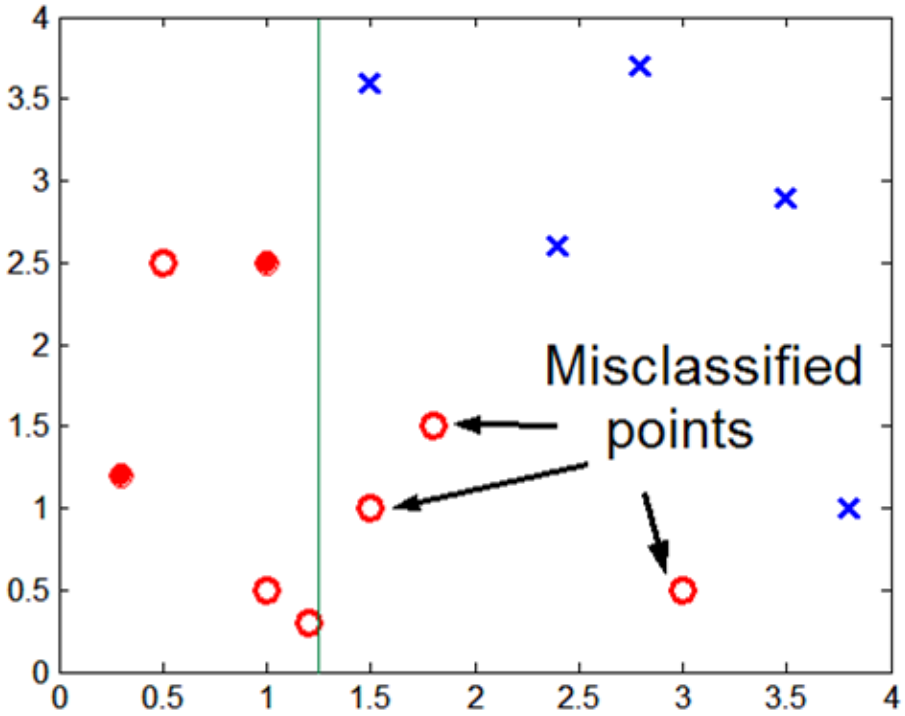
- Σφάλμα εκπαίδευσης
- Σφάλμα γενίκευσης

Υπερεκπαίδευση λόγω θορύβου



- Το όριο απόφασης παραμορφώνεται λόγω θορύβου

Υπερεκπαίδευση λόγω ανεπαρκούς αριθμού παραδειγμάτων



- Η έλλειψη σημείων στο κάτω μισό του διαγράμματος δεν επιτρέπει τη σωστή πρόβλεψη στην περιοχή εκείνη
- Ο ανεπαρκής αριθμός δεδομένων εκπαίδευσης οδηγεί το μοντέλο ταξινόμησης σε πρόβλεψη με βάση δεδομένα που δεν έχουν συνάφεια με την περιοχή εκείνη

Σχετικά με την υπερεκπαίδευση

- Όταν το μοντέλο είναι πιο πολύπλοκο από όσο χρειάζεται, τότε υπάρχει υπερεκπαίδευση.
- Στην περίπτωση αυτή το σφάλμα εκπαίδευσης δεν αποτελεί χρήσιμη μετρική στο πώς ακριβώς θα αποδόσει το μοντέλο σε νέα δεδομένα.
- Ανάγκη για νέες μεθόδους εκτίμησης σφαλμάτων

Occam's Razor

- Δεδομένων δυο μοντέλων με παρόμοια σφάλματα γενίκευσης, προτιμούμε το απλούστερο
- Στα πολύπλοκα μοντέλα υπάρχει μεγάλη πιθανότητα να έχουν υπερεκπαιδευτεί λόγω σφαλμάτων ή θορύβου στα δεδομένα

Σφάλματα Γενίκευσης

- Σφάλμα επαν-αντικατάστασης: σφάλμα στο σετ εκπαίδευσης (Σ e(t))
- Σφάλμα γενίκευσης: σφάλμα στο σετ ελέγχου (Σ e'(t))
- Μέθοδοι υπολογισμού του σφάλματος γενίκευσης (όπου μπορεί να εφαρμοστεί):
  - Αισιόδοξη προσέγγιση: e'(t) = e(t)
  - Απαισιόδοξη προσέγγιση:
    - Για κάθε τερματικό κόμβο: e'(t) = (e(t)+ξ) ξ: ποινή γενίκευσης (έστω 0.5)
    - Το συνολικό σφάλμα: e'(T) = e(T) + N $\times$ 0.5 (N: ο αριθμός των τερματικών κόμβων)
    - Π.χ. Για ένα δένδρο με 30 τερματικούς κόμβους και 10 σφάλματα στην εκπαίδευση (σε 1000 σημεία εκπαίδευσης):
      - Σφάλμα εκπαίδευσης= 10/1000 = 1%
      - Σφάλμα γενίκευσης= (10 + 30$\times$0.5)/1000 = 2.5%
  - Ελατωμμένο σφάλμα μείωσης – Reduced error pruning (REP):
    - χρησιμοποιεί ένα σετ επικύρωσης για τον υπολογισμό του σφάλματος γενίκευσης

Πηγές

- Introduction to Data Mining, Tan, Steinbach, Kumar.