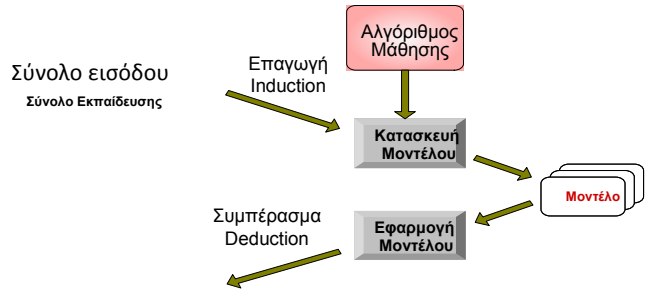
Κατηγοριοποίηση ΙΙ

Κατηγοριοποίηση

Κατηγοριοποίηση (classification)

Το γενικό πρόβλημα της ανάθεσης ενός αντικειμένου σε μία ή περισσότερες *προκαθορισμένες* κατηγορίες (κλάσεις)



Το μοντέλο κατηγοριοποίησης, χρησιμοποιείται ως:

- Περιγραφικό μοντέλο (descriptive modeling): ως επεξηγηματικό εργαλείο
- Μοντέλο πρόβλεψης (predictive modeling): για τη πρόβλεψη της κλάσης άγνωστων εγγραφών

Επανάληψη

Μοντέλο = Δέντρο Απόφασης

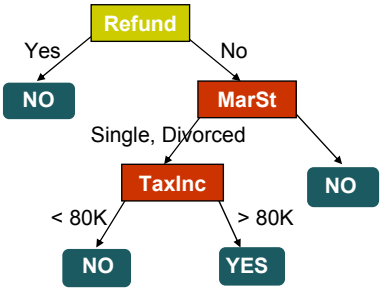
- Εσωτερικοί κόμβοι αντιστοιχούν σε κάποιο γνώρισμα
  - Διαχωρισμός (split) ενός κόμβου σε παιδιά
  - η ετικέτα στην ακμή = συνθήκη/έλεγχος
- Φύλλα αντιστοιχούν σε κλάσεις

Δεδομένα Εκπαίδευσης

κλάση

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Παράδειγμα Δέντρου

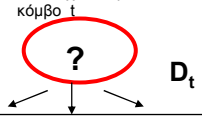


Επανάληψη: Κατασκευή ΔΑ

Κατασκευή του δέντρου (με λίγα λόγια):

1. Ξεκίνα με *έναν* κόμβο που περιέχει *όλες* τις εγγραφές
2. ***Διάσπαση*** του κόμβου (μοίρασμα των εγγραφών) με βάση μια συνθήκη-διαχωρισμού σε κάποιο από τα γνωρίσματα
3. Αναδρομική κλήση του βήματος 2 σε κάθε κόμβο
4. Αφού κατασκευαστεί το δέντρο, κάποιες βελτιστοποιήσεις (tree pruning)

D_t: το σύνολο των εγγραφών εκπαίδευσης που έχουν φτάσει στον κόμβο t



Διάσπαση Κόμβου (Αλγόριθμος του Hunt)

- Αν το D_t *περιέχει εγγραφές που ανήκουν στην ίδια κλάση* y_t, τότε ο κόμβος t είναι κόμβος φύλλο με ετικέτα y_t
- Αν D_t είναι το *κενό σύνολο* (αυτό σημαίνει ότι δεν υπάρχει εγγραφή στο σύνολο εκπαίδευσης με αυτό το συνδυασμό τιμών), τότε D_t γίνεται φύλλο με κλάση αυτή *της πλειοψηφίας* των εγγραφών εκπαίδευσης ή ανάθεση κάποιας default κλάσης
- Αν το D_t *περιέχει εγγραφές που ανήκουν σε περισσότερες από μία κλάσεις*, τότε χρησιμοποίησε έναν έλεγχο-γνωρίσματος για το διαχωρισμό των δεδομένων σε μικρότερα υποσύνολα

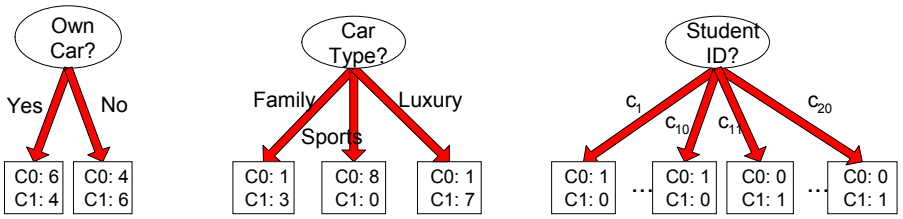
- Είδη διαχωρισμού:
  - Δυαδικός διαχωρισμός
  - Πολλαπλός διαχωρισμός

Επανάληψη: Κατασκευή ΔΑ

Το βασικό θέμα είναι

Ποιο γνώρισμα-συνθήκη διαχωρισμού να χρησιμοποιήσουμε για τη διάσπαση των εγγραφών κάθε κόμβου

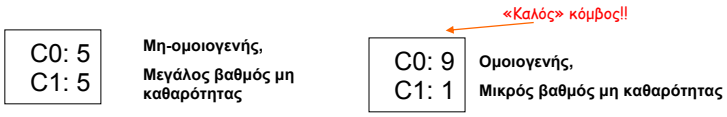
Έστω ότι πριν το διαχωρισμό: 10 εγγραφές της κλάσης C0,
10 εγγραφές της κλάσης C1



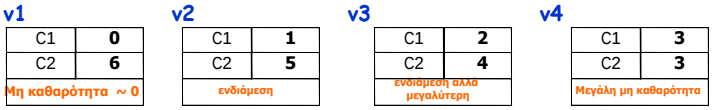
Ποια από τις 3 διασπάσεις να προτιμήσουμε; (Δηλαδή, ποια συνθήκη ελέγχου είναι καλύτερη;)

Επανάληψη: Κατασκευή ΔΑ

Διαισθητικά, προτιμώνται οι κόμβοι με ομοιογενείς κατανομές κλάσεων (homogeneous class distribution) – ιδανικά, όλες οι εγγραφές στην ίδια κλάση



Χρειαζόμαστε μία μέτρηση (I) της μη καθαρότητας ενός κόμβου (node impurity)



I(v1) < I(v2) < I(v3) < I(v4)

Είδαμε 3 διαφορετικούς ορισμούς για την ποιότητα I(v) ενός κόμβου v

Επανάληψη: Κατασκευή ΔΑ

Έχοντας ορίσει το I μπορούμε τώρα να ορίσουμε το «κέρδος» από μια διάσπαση

Έστω μια διάσπαση ενός κόμβου (parent) με N εγγραφές σε k παιδιά u_i

Έστω $N(u_i)$ ο αριθμός εγγραφών κάθε παιδιού ($\sum N(u_i) = N$)

Κοιτάμε το κέρδος, δηλαδή τη διαφορά μεταξύ της ποιότητας του γονέα (πριν τη διάσπαση) και το «μέσο όρο» της ποιότητας ων παιδιών του (μετά τη διάσπαση)

Βάρος (εξαρτάται από τον αριθμό εγγραφών)

$$\Delta = I(parent) - \sum_{i=1}^{k} \frac{N(u_i)}{N} I(u_i)$$

Διαλέγουμε τη διάσπαση με το μεγαλύτερο κέρδος (μεγαλύτερο Δ)

Επανάληψη: Κατασκευή ΔΑ

Μέτρα μη Καθαρότητας

1. Ευρετήριο Gini (Gini Index)
2. Εντροπία (Entropy)
3. Λάθος ταξινομήσεις (Misclassification error)

Επανάληψη: Κατασκευή ΔΑ

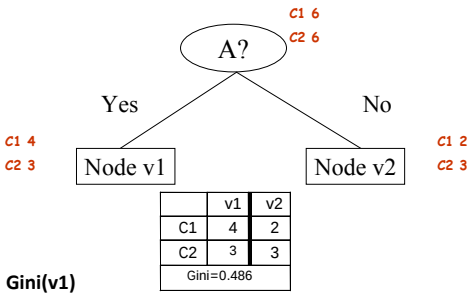
Ευρετήριο Gini

$$GINI(t) = 1 - \sum_{j=1}^{c} [p(j \mid t)]^2$$

p(j|t) σχετική συχνότητα της κλάσης j στον κόμβο t (ποσοστό εγγραφών της κλάσης j στον κόμβο t) και c αριθμός κλάσεων

Αρχικός κόμβος

	Parent
C1	**6**
C2	**6**
Gini = 0.500	



	v1	v2
C1	4	2
C2	3	3
Gini=0.486		

Κέρδος Δ = 0.500 – 0.486

Gini(v1)
= 1 – (4/7)² – (3/7)² = 0.49

Gini(v2)
= 1 – (2/5)² – (3/5)² = 0.48

Gini(Children)
= 7/12 * 0.49 + 5/12 * 0.48 = 0.486

Επανάληψη: Κατασκευή ΔΑ

Εντροπία

$$Entropy(t) = -\sum_{j=1}^{c} p(j \mid t) \log_2 p(j \mid t)$$

p(j|t) σχετική συχνότητα της κλάσης j στον κόμβο t και c αριθμός κλάσεων

C1	**0**
C2	**6**

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6
Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (5/6) = 0.65

- Όταν χρησιμοποιούμε την εντροπία για τη μέτρηση της μη καθαρότητας τότε η διαφορά καλείται **κέρδος πληροφορίας (information gain)**

Επανάληψη: Κατασκευή ΔΑ

Παράδειγμα

Κλάση

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

$$Info_{age}(D) = \frac{5}{14}I(2,3) + \frac{4}{14}I(4,0) + \frac{5}{14}I(3,2) = 0.694$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.246$$

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

Επανάληψη: Κατασκευή ΔΑ

Το κέρδος (Δ) τείνει να ευνοεί διαχωρισμούς που καταλήγουν σε μεγάλο αριθμό από διασπάσεις που η κάθε μία είναι μικρή αλλά καθαρή

$$\Delta = I(parent) - \sum_{i=1}^{k} \frac{N(u_i)}{N} I(u_i)$$

Αντί του κέρδους, χρήση αναλογίας κέρδους

$$GainRATIO_{split} = \frac{\Delta}{SplitINFO}$$

όπου

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

Επανάληψη: Κατασκευή ΔΑ

Λάθος ταξινόμησης (classification error)

$$Error(t) = 1 - \max_{class\ i} P(i \mid t)$$

Όσες ταξινομούνται σωστά

Παράδειγμα

C1	**1**
C2	**5**

P(C1) = 1/6 P(C2) = 5/6

Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

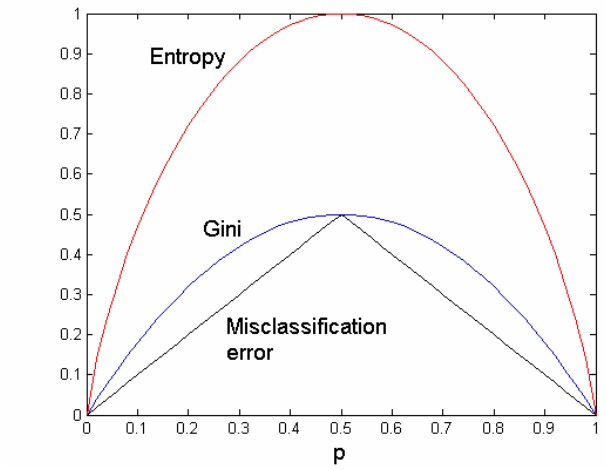
C1	**2**
C2	**4**

P(C1) = 2/6 P(C2) = 4/6

Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

Επανάληψη: Κατασκευή ΔΑ

Για ένα πρόβλημα δύο κλάσεων



p ποσοστό εγγραφών που ανήκει σε μία από τις δύο κλάσεις

(p κλάση +, 1-p κλάση -)

Όλες την μεγαλύτερη τιμή για 0.5 (ομοιόμορφη κατανομή)

Όλες μικρότερη τιμή όταν όλες οι εγγραφές σε μία μόνο κλάση (0 και στο 1)

Επανάληψη: Κατασκευή ΔΑ

- Προτερήματα -
  - + Λογικός χρόνος εκπαίδευσης
  - + Γρήγορη εφαρμογή
  - + Ευκολία στην κατανόηση
  - + Εύκολη υλοποίηση
  - + Μπορεί να χειριστεί μεγάλο αριθμό γνωρισμάτων
- Μειονεκτήματα
  - – Δεν μπορεί να χειριστεί περίπλοκες σχέσεις μεταξύ των γνωρισμάτων
  - – Απλά όρια απόφασης (decision boundaries)
  - – Προβλήματα όταν λείπουν πολλά δεδομένα

Επανάληψη: Θέματα

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

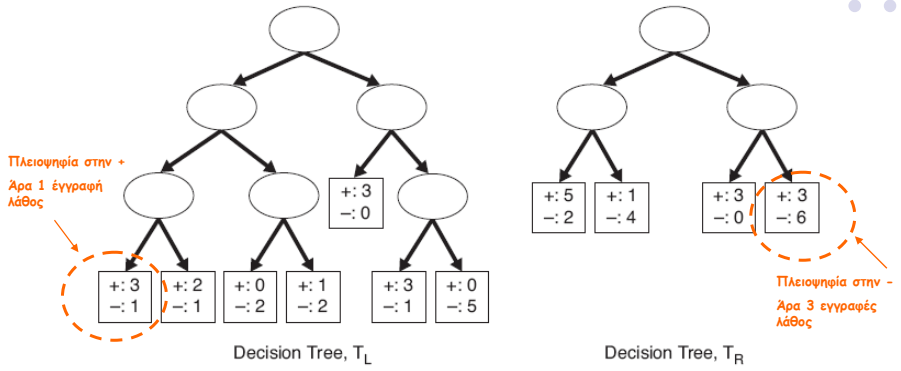
Ως λάθος (σφάλμα/error) μετράμε τις εγγραφές που το μοντέλο τοποθετεί σε λάθος κλάση

Δύο είδη σφαλμάτων

- *Εκπαίδευσης*: σφάλμα κατηγοριοποίησης στα δεδομένα του συνόλου εκπαίδευσης (ποσοστό δεδομένων εκπαίδευσης που κατηγοριοποιούνται σε λάθος κλάση)
- *Γενίκευσης* (generalization): τα αναμενόμενα λάθη κατηγοριοποίησης του μοντέλου σε δεδομένα που δεν έχει δει

Σφάλματα και στα δεδομένα εκπαίδευσης, γιατί χρησιμοποιούμε την πλειοψηφία των εγγραφών σε ένα φύλλο για να αποδώσουμε κλάση

Επανάληψη: Θέματα



Παράδειγμα δύο δέντρων για τα ίδια δεδομένα εκπαίδευσης

Σφάλμα εκπαίδευσης (ρυθμός σφάλματος)

Αριστερό 4/24 = 0.167 Δεξί: 6/24 = 0.25

Επανάληψη: Θέματα

Υπερπροσαρμογή - Overfitting

Μπορεί ένα μοντέλο που ταιριάζει πολύ καλά με τα δεδομένα εκπαίδευσης να έχει μεγαλύτερο λάθος γενίκευσης από ένα μοντέλο που ταιριάζει λιγότερο καλά στα δεδομένα εκπαίδευσης

- Το overfitting έχει ως αποτέλεσμα μοντέλα (δέντρα απόφασης) που είναι πιο περίπλοκα από όσο χρειάζεται
- Τα λάθη εκπαίδευσης δεν αποτελούν πια μια καλή εκτίμηση για τη συμπεριφορά του δέντρου σε εγγραφές που δεν έχει δει ξανά
- Νέοι μέθοδοι για την εκτίμηση του λάθους

Επανάληψη: Θέματα

Εκτίμηση Σφάλματος Γενίκευσης

Ως λάθος μετράμε τις εγγραφές που ο ταξινομητής τοποθετεί σε λάθος κλάση

- Χρήση Δεδομένων Εκπαίδευσης
  - αισιόδοξη εκτίμηση
  - απαισιόδοξη εκτίμηση
- Χρήση Δεδομένων Ελέγχου

Επανάληψη: Θέματα

Σφάλματα στην εκπαίδευση (Σ e(t))
Σφάλματα γενίκευσης (Σ e'(t))

Αισιόδοξη προσέγγιση: Τα σφάλματα γενίκευσης είναι ίσα με τα σφάλματα εκπαίδευσης

$$e'(t) = e(t)$$

Επανάληψη: Θέματα

Occam's Razor

- Δοθέντων δυο μοντέλων με παρόμοια λάθη γενίκευσης, πρέπει να προτιμάται το απλούστερο από το πιο περίπλοκο
- Ένα πολύπλοκο μοντέλο είναι πιο πιθανό να έχει ταιριαστεί (Fitted) τυχαία λόγω λαθών στα δεδομένα
- Για αυτό η πολυπλοκότητα του μοντέλου θα πρέπει να αποτελεί έναν από τους παράγοντες της αξιολόγησής του

Επανάληψη: Θέματα

Απαισιόδοξη προσέγγιση:

k: αριθμός φύλλων,

για κάθε φύλλο t_i προσθέτουμε ένα κόστος $V(t_i)$

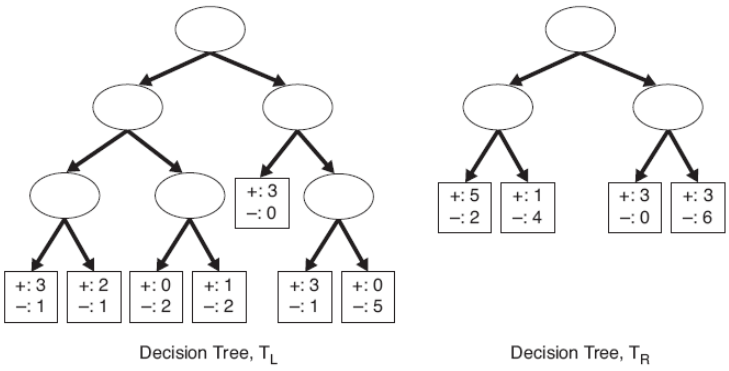
$$e'(T) = \frac{\sum_{i=1}^{k} [e(t_i) + V(t_i)]}{\sum_{i-1}^{k} n(t_i)}$$

Αν για κάθε φύλλο t, V(t) = 0.5: e'(t) = e(t) + 0.5

Συνολικό λάθος: e'(T) = e(T) + $k \times 0.5$ (*k*: αριθμός φύλλων)

Επανάληψη: Θέματα

Παράδειγμα



Decision Tree, T_L Decision Tree, T_R

Παράδειγμα δύο δέντρων για τα ίδια δεδομένα

Με βάση το λάθος εκπαίδευσης και *απαισιόδοξη* προσέγγιση με V(t) = 0.7

Αριστερό (4 + 7*0.7)/24 = 0.371

Δεξί: (6 + 4*0.7)/24 = 0.367

Επανάληψη: Θέματα

Χρήση δεδομένων ελέγχου

Χώρισε τα δεδομένα εκπαίδευσης:

2/3 εκπαίδευση

1/3 (σύνολο επαλήθευσης – validation set) για τον υπολογισμό του σφάλματος

Επανάληψη: Θέματα

Δύο βασικές προσεγγίσεις κλαδέματος (pruning) του δέντρου:

Pre-pruning

Σταμάτημα της ανάπτυξης του δέντρου μετά από κάποιο σημείο

Post-pruning

Η κατασκευή του δέντρου χωρίζεται σε δύο φάσεις:

1. Φάση Ανάπτυξης
2. Φάση Κλαδέματος

Επανάληψη: Θέματα

Pre-Pruning (Early Stopping Rule)

- Τυπικά, σταματάμε την επέκταση ενός κόμβου
  - όταν όλες οι εγγραφές του ανήκουν στην ίδια κλάση ή
  - όταν όλα τα γνωρίσματα έχουν τις ίδιες τιμές

Γρήγορος τερματισμός (ο αλγόριθμος σταματά πριν σχηματιστεί ένα πλήρες δέντρο

Σταμάτα όταν ο αριθμός των εγγραφών είναι μικρότερος από κάποιο προκαθορισμένο κατώφλι

Σταμάτα όταν η επέκταση ενός κόμβου δεν βελτιώνει την καθαρότητα (π.χ., Gini ή information gain) ή το λάθος γενίκευσης περισσότερο από κάποιο κατώφλι

(-) δύσκολος ο καθορισμός του κατωφλιού,
(-) αν και το κέρδος μικρό, κατοπινοί διαχωρισμοί μπορεί να καταλήξουν σε καλύτερα δέντρα

Post-pruning

- Ανάπτυξε το δέντρο πλήρως
- Ψαλίδισε (trim) τους κόμβους από πάνω προς τα κάτω (bottom-up)
- Αν το σφάλμα γενίκευσης μειώνεται με το ψαλίδισμα, αντικατέστησε το υποδέντρο με
  - ένα φύλλο - οι ετικέτες κλάσεις του φύλλου καθορίζεται από την πλειοψηφία των κλάσεων των εγγραφών του υποδέντρου (subtree replacement)
  - ένα από τα κλαδιά του (Branch), αυτό που χρησιμοποιείται συχνότερα (subtree raising)

Χρησιμοποιείται πιο συχνά

Χρήση άλλων δεδομένων για τον υπολογισμό του καλύτερου δέντρου (δηλαδή του σφάλματος γενίκευσης)

Παράδειγμα post-pruning

- Αισιόδοξη προσέγγιση?

  Όχι ψαλίδισμα

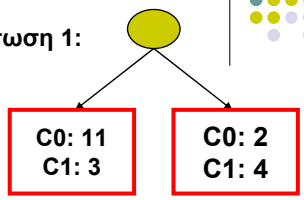
- Απαισιόδοξη προσέγγιση (με 0.5)?

  όχι case 1, ναι case 2

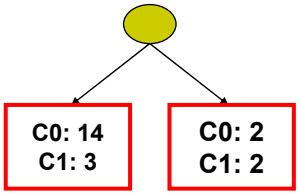
- Χρήση δεδομένων ελέγχου

Εξαρτάται από το σύνολο ελέγχου

Περίπτωση 1:



Περίπτωση 2:



Επανάληψη: Μέτρα Εκτίμησης

Confusion Matrix (Πίνακας Σύγχυσης)

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

		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}

Ακρίβεια - Accuracy

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

Ρυθμός σφάλματος

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

ErrorRate(C) = 1 – Accuracy(C)

Μέτρα Εκτίμησης

Παράδειγμα

		PREDICTED CLASS		
		+	-	
ACTUAL CLASS	+	20	10	30
	-	15	55	70
		35	65	100

Τι σημαίνει;

Αποτίμηση Μοντέλου

Μέτρα Εκτίμησης (κόστος)

Όχι όλα τα σφάλματα το ίδιο σημαντικά -> «βάρη»

Εισάγουμε την έννοια του Πίνακα Κόστους

Πίνακας Κόστους

		PREDICTED CLASS	
	C(i|j)	**Class = +**	**Class = -**
ACTUAL CLASS	**Class = +**	C(+, +)	C(+, -)
	Class = -	C(-, +)	C(-, -)

C(i|j): **κόστος** λανθασμένης κατηγοριοποίησης ενός παραδείγματος της κλάσης i ως κλάση j → βάρος

Για C(+, +), C(-, -) αρνητικό Για C(+, -), C(-, +) θετικό
Αρνητική τιμή κόστους σημαίνει επιπρόσθετη «επιβράβευση» σωστής πρόβλεψης

Μέτρα Εκτίμησης (κόστος)

Παράδειγμα

Πίνακας κόστους		Predicted	
	C(i|j)	+	-
Actual	+	-1	10
	-	3	0

Μέτρα Εκτίμησης (κόστος)

Πίνακας Κόστους

		PREDICTED CLASS	
	C(i|j)	**Class = +**	**Class = -**
ACTUAL CLASS	**Class = +**	C(+, +)	C(+, -)
	Class = -	C(-, +)	C(-, -)

Πίνακας Σύγχυσης

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

$$C(M) = \mathbf{TP} \times C(+, +) + \mathbf{FN} \times C(+, -) + \mathbf{FP}\ C(-, +) + \mathbf{TN}\ C(-, -)$$

Στα προηγούμενα, είχαμε
C(+, +) = C(-, -) = 0 -> όχι επιβράβευση
C(+, -) = C(-, +) = 1 -> κάθε λάθος μετρά 1

Μέτρα Εκτίμησης (κόστος)

Παράδειγμα: Υπολογισμός του Κόστους της Κατηγοριοποίησης

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

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

Model M_1	PREDICTED CLASS			
		+	-	
ACTUAL CLASS	+	150	40	190
	-	60	250	310
		210	290	400

Accuracy = 80%

Cost = 3910

Model M_2	PREDICTED CLASS			
		+	-	
ACTUAL CLASS	+	250	45	295
	-	5	200	205
		255	245	400

Accuracy = 90%

Cost = 4255

Χάνει κάποια θετικά

Μέτρα Εκτίμησης (κόστος)

Κατηγοριοποίηση που λαμβάνει υπό όψιν της το κόστος

Κατασκευή Δέντρου Απόφασης

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

Μέτρα Εκτίμησης (κόστος)

Καθορισμός κλάσης

Κανονικά, ως ετικέτα ενός φύλλου την πλειοψηφούσα κλάση:

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

Τώρα, δίνουμε την κλάση i στον κόμβο που έχει το ελάχιστο:

$$\sum_j p(j)C(j,i)$$

Για όλες τις κλάσεις

Μέτρα Εκτίμησης (κόστος)

Έστω 2 κλάσεις: + και -

Αν όχι κόστος, ετικέτα +, ανν, p(+) > 0.5 (δηλαδή, πλειοψηφία)

Τώρα, αυτήν με το μικρότερο κόστος:

κόστος της κλάσης - : p(+) x C(+, +) + p(+) x C(+, -)

κόστος της κλάσης + : p(-) x C(-, -) + p(-) x C(-, +)

Αν C(-, -) = C(+, +) = 0 (όχι κόστος (επιβράβευση) στα σωστά)

Δίνουμε +, αν

p(+) C(+, -) > p(-) x C(-, +) =>

$$p(+) > \frac{C(-,+)}{C(-,+) + C(+,-)}$$

p(-) = 1 – p(+)

Αν C(-, +) < C(+, -), τότε λιγότερο του 0.5

Μέτρα Εκτίμησης (κόστος)

Κόστος 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

Η ακρίβεια (accuracy) είναι ανάλογη του κόστους αν:

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

N = a + b + c + d

Accuracy = (a + d)/N

Cost = p (a + d) + q (b + c)

= p (a + d) + q (N – a – d)

= q N – (q – p)(a + d)

= N [q – (q-p) × Accuracy]

Μέτρα Εκτίμησης

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

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

Αποτίμηση Μοντέλου

- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
  - Πως να εκτιμήσουμε την απόδοση ενός μοντέλου
  - ***Τι*** *θα μετρήσουμε (π.χ., είδαμε τα σφάλματα, ακρίβεια)*
- Μέθοδοι για την εκτίμηση της απόδοσης
  - Πως μπορούνε να πάρουμε αξιόπιστες εκτιμήσεις
  - ***Πως*** *θα το μετρήσουμε (π.χ., δεδομένα εκπαίδευσης, ελέγχου)*
- Μέθοδοι για την σύγκριση μοντέλων
  - Πως να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

Ισχύουν για όλα τα μοντέλα κατηγοριοποίησης (όχι μόνο για τα δέντρα απόφασης)

Μέτρα Εκτίμησης

Confusion Matrix (Πίνακας Σύγχυσης)

	Εκτιμώμενη κλάση (PREDICTED CLASS)		
Πραγματική κλάση (ACTUAL CLASS)		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

Ιδανικά = 0

Ακρίβεια - Accuracy

Το πιο συνηθισμένο μέτρο

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Σωστές προβλέψεις

Ρυθμός σφάλματος

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

Τι συμβαίνει αν μια κατηγορία είναι σπάνια;

ErrorRate(C) = 1 – Accuracy(C)

Άλλα Μέτρα Εκτίμησης

Παράδειγμα

Μειονεκτήματα της ακρίβειας (accuracy)

- Θεωρείστε ένα πρόβλημα με 2 κλάσεις
  - Αριθμός παραδειγμάτων της κλάσης 0 = 9990
  - Αριθμός παραδειγμάτων της κλάσης 1 = 10
- Αν ένα μοντέλο προβλέπει οτιδήποτε ως κλάση 0,
  τότε accuracy = 9990/10000 = 99.9 %

Η accuracy είναι παραπλανητική γιατί το μοντέλο δεν προβλέπει κανένα παράδειγμα της κλάσης 1

Εναλλακτικά Μέτρα Εκτίμησης

Άλλες μετρήσεις με βάση τον Πίνακα Σύγχυσης

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

True positive rate or sensitivity (ευαισθησία): Το ποσοστό των θετικών παραδειγμάτων που κατηγοριοποιούνται σωστά

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

True negative rate or specificity (ιδιαιτερότητα): Το ποσοστό των αρνητικών παραδειγμάτων που κατηγοριοποιούνται σωστά

$$\text{TNR} = \frac{TN}{TN + FP}$$

Εναλλακτικά Μέτρα Εκτίμησης

Άλλες μετρήσεις με βάση τον πίνακα σύγχυσης

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

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

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

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

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

Εναλλακτικά Μέτρα Εκτίμησης

Recall (ανάκληση) – Precision (ακρίβεια)

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

Precision $\text{p} = \frac{TP}{TP + FP}$

Πόσα από τα παραδείγματα που το μοντέλο έχει κατηγοριοποιήσει ως θετικά είναι πραγματικά θετικά

Όσο πιο μεγάλη η ακρίβεια, τόσο μικρότερος ο αριθμός των FP

Recall $\text{r} = \frac{TP}{TP + FN}$

Πόσα από τα θετικά παραδείγματα κατάφερε ο κατηγοριοποιητής να βρει

Όσο πιο μεγάλη η ανάκληση, τόσο λιγότερα θετικά παραδείγματα έχουν κατηγοριοποιηθεί λάθος (=TPR)

Εναλλακτικά Μέτρα Εκτίμησης

Recall (ανάκληση) – Precision (ακρίβεια)

Precision	$p = \frac{TP}{TP + FP}$	Πόσα από τα παραδείγματα που το μοντέλο έχει κατηγοριοποιήσει ως θετικά είναι πραγματικά θετικά
Recall	$r = \frac{TP}{TP + FN}$	Πόσα από τα θετικά παραδείγματα κατάφερε να βρει

Συχνά το ένα καλό και το άλλο όχι

Πχ, ένας κατηγοριοποιητής που όλα τα ταξινομεί ως θετικά, την καλύτερη ανάκληση με τη χειρότερη ακρίβεια

Πώς να τα συνδυάσουμε;

Εναλλακτικά Μέτρα Εκτίμησης

F_1 measure

$$F_1 = \frac{2rp}{r + p} = \frac{2TP}{2TP + FP + FN}$$

$$F_1 = \frac{2}{1/r + 1/p}$$

Αρμονικό μέσο (Harmonic mean)

- Τείνει να είναι πιο κοντά στο μικρότερο από τα δύο
- Υψηλή τιμή σημαίνει ότι και τα δύο είναι ικανοποιητικά μεγάλα
- χρήσιμο ως μέσο ρυθμών (rate)

Εναλλακτικά Μέτρα Εκτίμησης

Αρμονικά, Γεωμετρικά και Αριθμητικά Μέσα

Παράδειγμα

α=1, b=5

Εναλλακτικά Μέτρα Εκτίμησης

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

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

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2TP}{2TP + FN + FP}$$

- Precision - C(Yes|Yes) & C(Yes|No)
- Recall - C(Yes|Yes) & C(No|Yes)
- F-measure όλα εκτός του C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 TP + w_4 TN}{w_1 TP + w_2 FP + w_3 FN + w_4 TN}$$

	w1	w2	w3	w4
Recall	1	1	0	0
Precision	1	0	1	1
F1	2	1	1	0
Accuracy	1	1	1	1

Αποτίμηση Μοντέλου: ROC

ROC (Receiver Operating Characteristic Curve)

Καμπύλη χαρακτηριστικής λειτουργίας δέκτη

- Αναπτύχθηκε στη δεκαετία 1950 για την ανάλυση θορύβου στα σήματα
  - Χαρακτηρίζει το trade-off μεταξύ positive hits και false alarms
- Η καμπύλη ROC δείχνει τα TPR [TruePositiveRate] (στον άξονα των y) προς τα FPR [FalsePositiveRate] (στον άξονα των x)
- Η απόδοση κάθε μοντέλου αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

True Positive Rate
Πόσα από τα θετικά βρίσκει
[πόσα από τα θετικά ταξινομεί σωστά]

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

False Positive Rate
Πόσα από τα αρνητικά θεωρεί θετικά
[πόσα από τα αρνητικά κατηγοριοποιεί λάθος]

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

Αποτίμηση Μοντέλου: ROC

Κάθε σημείο αντιστοιχεί στα μοντέλα που παράγει κάθε κατηγοριοποιητής

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

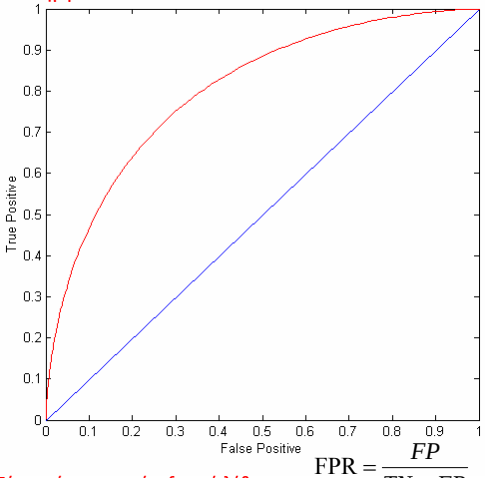
(0,0): το μοντέλο προβλέπει τα πάντα ως αρνητική κατηγορία

(1,1): το μοντέλο προβλέπει τα πάντα ως θετική κατηγορία

(0,1): ιδανικό
Το ιδανικό στην άνω αριστερή γωνία

Διαγώνια γραμμή: Random guessing
Μια εγγραφή θεωρείται θετική με καθορισμένη πιθανότητα p ανεξάρτητα από τις τιμές των γνωρισμάτων της

Πόσα από τα θετικά κατηγοριοποιεί σωστά



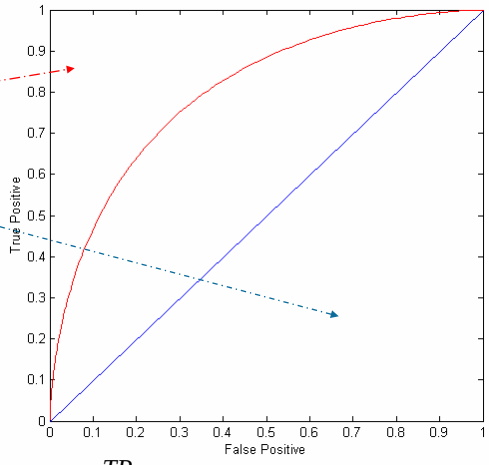
Πόσα από τα αρνητικά ταξινομεί λάθος

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

Αποτίμηση Μοντέλου: ROC

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

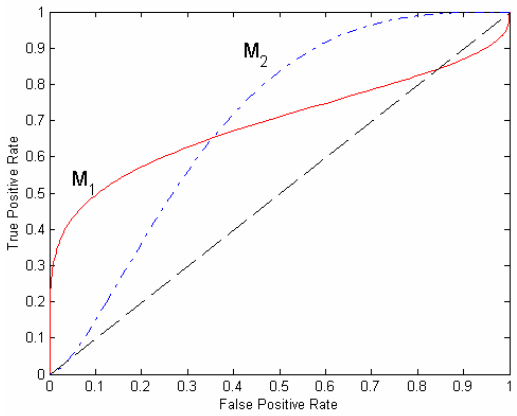
Κάτω από τη διαγώνιο Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



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

Αποτίμηση Μοντέλου: ROC

Σύγκριση δύο μοντέλων



- Κανένα μοντέλο δεν είναι πάντα καλύτερο του άλλου
  - M_1 καλύτερο για μικρό FPR
  - M_2 καλύτερο για μεγάλο FPR
- Η περιοχή κάτω από την καμπύλη ROC
  - Ιδανικό μοντέλο:
    - περιοχή = 1
  - Τυχαία πρόβλεψη:
    - Περιοχή = 0.5

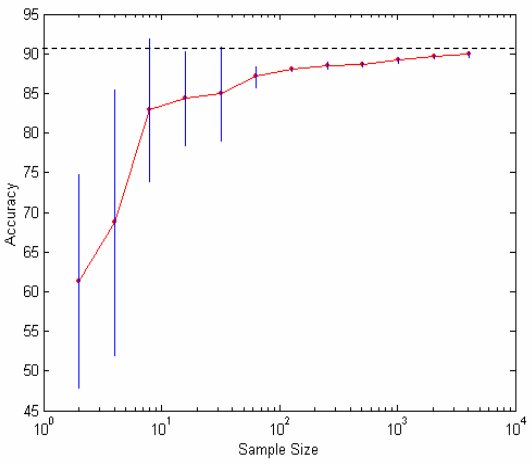
Ένα μοντέλο αυστηρά καλύτερο αν έχει μεγαλύτερη περιοχή κάτω από την καμπύλη

Μέθοδοι Αποτίμησης Μοντέλου

- Πως μπορούμε να πάρουμε αξιόπιστες εκτιμήσεις της απόδοσης
- Η απόδοση ενός μοντέλου μπορεί να εξαρτάται από πολλούς παράγοντες εκτός του αλγορίθμου μάθησης:
  - Κατανομή των κλάσεων
  - Το κόστος της λανθασμένης κατηγοριοποίησης
  - Το μέγεθος του συνόλου εκπαίδευσης και του συνόλου ελέγχου

Μέθοδοι Αποτίμησης Μοντέλου

Καμπύλη Μάθησης (Learning Curve)



- Η καμπύλη μάθησης δείχνει πως μεταβάλλεται η ακρίβεια (accuracy) με την αύξηση του μεγέθους του δείγματος
- Επίδραση δείγματος μικρού μεγέθους:
  - Bias in the estimate
  - Variance of estimate

Αποτίμηση Μοντέλου

Πως;

Μπορούμε να χρησιμοποιήσουμε τα

- Σφάλματα εκπαίδευσης
- Σφάλματα γενίκευσης (αισιόδοξη ή απαισιόδοξη προσέγγιση)

Δεν είναι κατάλληλα γιατί βασίζονται στα δεδομένα εκπαίδευσης μόνο

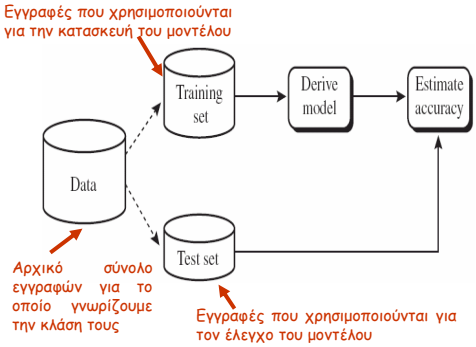
Συχνά, σύνολο ελέγχου -> πιο αναλυτικά

Μέθοδοι Αποτίμησης Μοντέλου

Μέθοδος Holdout

Διαμέριση του αρχικού συνόλου σε δύο ξένα σύνολα:

Σύνολο εκπαίδευσης (2/3) – Σύνολο Ελέγχου (1/3)



- Κατασκευή μοντέλου με βάση το σύνολο εκπαίδευσης
- Αποτίμηση μοντέλου με βάση το σύνολο ελέγχου

Μέθοδοι Αποτίμησης Μοντέλου

Μέθοδος Holdout

(-) Λιγότερες εγγραφές για εκπαίδευση – πιθανόν όχι τόσο καλό μοντέλο, όσο αν χρησιμοποιούνταν όλες

(-) Το μοντέλο εξαρτάται από τη σύνθεση των συνόλων εκπαίδευσης και ελέγχου – όσο μικρότερο το σύνολο εκπαίδευσης, τόσο μεγαλύτερη η variance του μοντέλου

– όσο μεγαλύτερο το σύνολο εκπαίδευσης, τόσο λιγότερο αξιόπιστη η πιστότητα του μοντέλου που υπολογίζεται με το σύνολο ελέγχου – wide confidence interval (διακύμανση)

(-) Τα σύνολα ελέγχου και εκπαίδευσης δεν είναι ανεξάρτητα μεταξύ τους (υποσύνολα του ίδιου συνόλου - πχ μια κλάση που έχει πολλά δείγματα στο ένα, θα έχει λίγα στο άλλο και το ανάποδο)

Μέθοδοι Αποτίμησης Μοντέλου

Τυχαία Λήψη Δειγμάτων – Random Subsampling

Επανάληψη της μεθόδου για τη βελτίωσή της
έστω *k* επαναλήψεις, παίρνουμε το μέσο όρο της ακρίβειας

$$acc_{sub} = \frac{1}{k}\sum_{i=1}^{k} acc_i$$

(-) Πάλι αφαιρούμε δεδομένα από το σύνολο εκπαίδευσης

(-) Ένα ακόμα πρόβλημα είναι ότι μια εγγραφή μπορεί να χρησιμοποιείται (επιλέγεται) ως εγγραφή εκπαίδευσης πιο συχνά από κάποια άλλη

Μέθοδοι Αποτίμησης Μοντέλου

Cross validation (διασταυρωμένη επικύρωση)

Κάθε εγγραφή χρησιμοποιείται τον ίδιο αριθμό φορών στην εκπαίδευση και ακριβώς μια φορά για έλεγχο

- Διαμοίραση των δεδομένων σε *k* ίσα διαστήματα
- Κατασκευή του μοντέλου αφήνοντας κάθε φορά ένα διάστημα ως σύνολο ελέγχου και χρησιμοποιώντας όλα τα υπόλοιπα ως σύνολα εκπαίδευσης
- Επανάληψη *k* φορές

2-fold (δύο ίσα υποσύνολα, το ένα μια φορά για έλεγχο – το άλλο για εκπαίδευση και μετά ανάποδα)

Αν k = N, (N ο αριθμός των εγγραφών) leave-one-out

μεγαλύτερο δυνατό σύνολο εκπαίδευσης
σύνολα ελέγχου αμοιβαία αποκλειόμενα (καλύπτουν όλο το σύνολο)
υπολογιστικά ακριβή
υψηλή διακύμανση του μέτρου (μόνο μια τιμή)

Μέθοδοι Αποτίμησης Μοντέλου

Bootstrap (Αυτοδυναμία)

Sample with replacement – δειγματοληψία με επανένταξη

Μια εγγραφή που επιλέχθηκε ως δεδομένο εκπαίδευσης, *ξαναμπαίνει στο αρχικό σύνολο*
Οι υπόλοιπες εγγραφές (όσες δεν επιλεγούν στο σύνολο εκπαίδευσης) – εγγραφές ελέγχου

Αν *N* δεδομένα, ένα δείγμα *N* στοιχείων *63.2%* των αρχικών, γιατί;

Πιθανότητα ένα δεδομένο να επιλεγεί $1 – (1-1/N)^N$
Για μεγάλο N, η πιθανότητα επιλογής τείνει ασυμπτωτικά στο $1-e^{-1} = 0.632$, πιθανότητα μη επιλογής 0.368

.632 boostrap

$$acc_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0.6328 * errortest_i + 0.368 * acc_s)$$

b: αριθμός επαναλήψεων

acc_s ακρίβεια όταν όλα τα δεδομένα ως σύνολο εκπαίδευσης