Ταξινόμηση III

Οι διαφάνειες στηρίζονται στο **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»**, Addison Wesley, 2006

Αποτίμηση Μοντέλου

Μέτρα Εκτίμησης

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.)

Confusion Matrix (Πίνακας Σύγχυσης)

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

	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

TP (true positive) f_{11}
FN (false negative) f_{10}
FP (false positive) f_{01}
TN (true negative) f_{00}

Μέτρα Εκτίμησης

Πιστότητα - Accuracy

Πιστότητα (ακρίβεια;) (accuracy)
Το πιο συνηθισμένο μέτρο

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

= 0

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

Λόγος Λάθους

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

ErrorRate(C) = 1 – Accuracy(C)

Αποτίμηση Μοντέλου

Μπορούμε να χρησιμοποιήσουμε τα λάθη εκπαίδευσης/γενίκευσης (αισιόδοξη ή απαισιόδοξη προσέγγιση)

Δεν είναι κατάλληλα γιατί βασίζονται στα δεδομένα εκπαίδευσης μόνο

Συνήθως, σύνολο ελέγχου

Αποτίμηση Μοντέλου

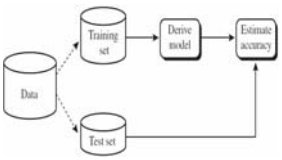
- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
  - Πως να εκτιμήσουμε την απόδοση ενός μοντέλου
  - *Τι θα μετρήσουμε*
- Μέθοδοι για την εκτίμηση της απόδοσης
  - Πως μπορούνε να πάρουμε αξιόπιστες εκτιμήσεις
  - ***Πως** θα το μετρήσουμε*
- Μέθοδοι για την σύγκριση μοντέλων
  - Πως να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

Ισχύουν για όλα τα μοντέλα ταξινόμησης

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

(-) Τα σύνολα ελέγχου και εκπαίδευσης δεν είναι ανεξάρτητα μεταξύ τους (υποσύνολα τυ 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

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

Cross validation

Διαμοίραση των δεδομένων σε k διαστήματα
Κατασκευή του μοντέλου αφήνοντας κάθε φορά ένα διάστημα ως σύνολο ελέγχου και χρησιμοποιώντας όλα τα υπόλοιπα ως σύνολα εκπαίδευσης
(μια εγγραφή χρησιμοποιείται ακριβώς μια φορά για έλεγχο και τον ίδιο αριθμό για εκπαίδευση)
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

Μέθοδοι Αποτίμησης Μοντέλου

Bootstrap

Sample with replacement

Μια εγγραφή που επιλέχθηκε ως δεδομένο εκπαίδευσης, ξαναμπαίνει στο αρχικό σύνολο

Αν N δεδομένα, ένα δείγμα N στοιχείων 63.2% των αρχικών

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

Οι υπόλοιπες εγγραφές (όσες δεν επιλεγούν στο σύνολο εκπαίδευσης) – εγγραφές ελέγχου

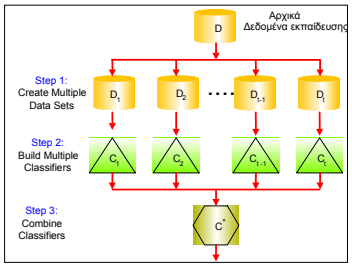
.632 boostrap $$acc_{boot} = \frac{1}{c}\sum_{i}^{b}(0.6328 * errortest + 0.368 * acc_s)$$

Βελτίωση Απόδοσης

Ensemble Methods – Σύνολο Μεθόδων

Κατασκευή ενός συνόλου από ταξινομητές από τα δεδομένα εκπαίδευσης C_1, C_2, ... C_t -> C^*

Υπολογισμός της κλάσης των δεδομένων συναθροίζοντας (aggregating) τις προβλέψεις των t ταξινομητών
Πως: πχ με πλειοψηφικό σύστημα (Voting majority)



Βελτίωση Απόδοσης

- Έστω t = 25 βασικοί ταξινομητές
  - Αν ο καθένας λάθος, ε = 0.35
  - Έστω ότι ανεξάρτητοι και μόνο 2 κλάσεις
  - Πιθανότητα λανθασμένης πρόβλεψης του συνόλου:

$$\sum_{i=13}^{25}\binom{25}{i}\varepsilon^{i}(1-\varepsilon)^{25-i} = 0.06$$

Βελτίωση Απόδοσης

Bagging (Bootstarp + Aggregation

- Δειγματοληψία με επανένταξη (Sampling with replacement)
- Κατασκευή ταξινομητή για κάθε δείγμα
- Κάθε δείγμα έχει πιθανότητα $(1 - 1/n)^n$ να επιλεγεί

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

Boosting

Δε δίνουμε το ίδιο βάρος σε όλους τους ταξινομητές, αλλά παίρνουμε υπόψη μας την ακρίβειά τους -- C* βάρος με βάση την ακρίβεια του

- Βασική ιδέα:

Έστω C_i, ο C_{i+1} μεγαλύτερο βάθος στις πλειάδες που ταξινόμησε λάθος ο C_i

Πως; «πειράζουμε» την πιθανότητα επιλογής τους στο σύνολο εκπαίδευσης

σωστά, πιθανότητα επιλογής -

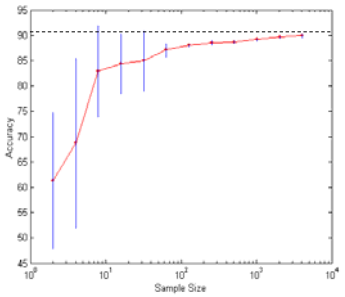
λάθος, πιθανότητα επιλογής +

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 ταξινόμησης
  - Το μέγεθος του συνόλου εκπαίδευσης και του συνόλου ελέγχου

Μέθοδοι Αποτίμησης Μοντέλου

Καμπύλη Μάθησης (Learning Curve)



- 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

Άλλα Μέτρα Εκτίμησης πέραν της Πιστότητας

Πίνακας σύγχυσης

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

Πιστότητα (accuracy) -- υπενθύμιση --

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

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

Άλλα Μέτρα Εκτίμησης πέραν της Πιστότητας

Μειονεκτήματα της πιστότητας

- 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 πιστότητα = 9990/10000 = 99.9 %
  - Η πιστότητα είναι παραπλανητική γιατί το μοντέλο δεν προβλέπει κανένα παράδειγμα της κλάσης 1

Μέτρα Εκτίμησης

Πίνακας Κόστους

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

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

βάρος

C(M) = TP x *C*(Yes|Yes) + FN x *C*(Yes|No) + FP *C*(No|Yes) + TN *C*(No|No)

Αρνητική τιμή κόστους σημαίνει επιπρόσθετη «επιβράβευση» σωστής πρόβλεψης

Στα προηγούμενα, είχαμε

C(YES|YES) = *C*(NO|NO) = 0 και

C(YES|NO) = *C*(NO|YES) = 1

Μέτρα Εκτίμησης

Υπολογισμός του Κόστους της Ταξινόμησης

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

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

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

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

Μέτρα Εκτίμησης

Καθορισμός κλάσης

Κανονικά, ως ετικέτα ενός φύλλου την πλειοψηφούσα κλάση,

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

Για δύο κλάσεις, p(+) > 0.5

Τώρα, την κλάση που ελαχιστοποιεί το: $\text{κλάση φύλλου} = \sum_j p(j)C(j,i)$

Για δύο κλάσεις: p(+) × C(+, +) + p(+) × C(+, -)

p(-) × C(-, -) + p(-) × C(-, +)

Αν C(-, -) = C(+, +) = 0

p(+) C(+, -) > p(-) × C(-, +) =>

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

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

Μέτρα Εκτίμησης

Κόστος vs Πιστότητας (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 πιστότητα 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

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: Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται σωστά $\text{TPR} = \frac{TP}{TP + FN}$

True negative rate or specificity: Το ποσοστό των αρνητικών παραδειγμάτων που ταξινομούνται σωστά $\text{TNR} = \frac{TN}{TN + FP}$

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 ταξινομούνται λάθος (δηλαδή, ως θετικά) $\text{FPR} = \frac{FP}{TN + FP}$

False negative rate: Το ποσοστό των θετικών παραδειγμάτων που ταξινομούνται λάθος (δηλαδή, ως αρνητικά) $\text{FNR} = \frac{FN}{TP + FN}$

Μέτρα Εκτίμησης

Recall (ανάκληση) – Precision (ακρίβεια)

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

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

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

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

Πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει

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

Μέτρα Εκτίμησης

Recall (ανάκληση) – Precision (ακρίβεια)

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

Πόσα από τα παραδείγματα που ο ταξινομητής έχει ταξινομήσει ως θετικά είναι πραγματικά θετικά

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

Πόσα από τα θετικά παραδείγματα κατάφερε ο ταξινομητής να βρει

Συχνά το ένα καλό και το άλλο όχι

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

Πώς να τα συνδυάσουμε;

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

Μέτρα Εκτίμησης

Αρμονικά, Γεωμετρικά και Αριθμητικά Μέσα

Παράδειγμα

α=1, b=5

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)

- 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 (στον άξονα των y) προς τα FPR (στον άξονα των x)
- Η απόδοση κάθε ταξινομητή αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

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

Αποτίμηση Μοντέλου: ROC

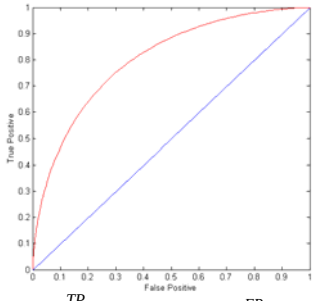
(TP,FP):

- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal

Diagonal line:

- Random guessing

Μια εγγραφή θεωρείται θετική με καθορισμένη πιθανότητα p ανεξάρτητα από τις τιμές των γνωρισμάτων της

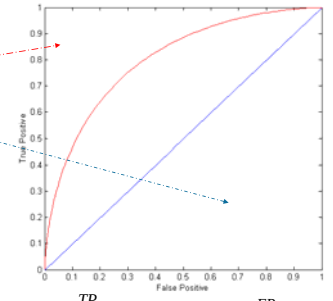


$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{TN + FP}$$

Αποτίμηση Μοντέλου: ROC

Καλοί ταξινομητές κοντά στην αριστερή πάνω γωνία του διαγράμματος

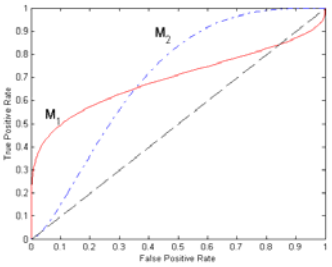
Κάτω από τη διαγώνιο Πρόβλεψη είναι το αντίθετο της πραγματικής κλάσης



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

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
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

Αποτίμηση Μοντέλου

- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
  Πως να εκτιμήσουμε την απόδοση ενός μοντέλου
- Μέθοδοι για την εκτίμηση της απόδοσης
  Πως μπορούνε να πάρουμε αξιόπιστες εκτιμήσεις
- Μέθοδοι για την σύγκριση μοντέλων
  Πως να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

Αποτίμηση Μοντέλου

Έλεγχος Σημαντικότητας (Test of Significance)

- Έστω δύο μοντέλα:
  - Μοντέλο M1: ακρίβεια = 85%, έλεγχος σε 30 εγγραφές
  - Μοντέλο M2: ακρίβεια = 75%, έλεγχος σε 5000 εγγραφές
- Είναι το M1 καλύτερο από το M2;
  - Πόση εμπιστοσύνη (confidence) μπορούμε να έχουμε για την πιστότητα του M1 και πόση για την πιστότητα του M2;
  - Μπορεί η διαφορά στην απόδοση να αποδοθεί σε τυχαία διακύμανση του συνόλου ελέγχου;

Αποτίμηση Μοντέλου

Διάστημα Εμπιστοσύνης για την Ακρίβεια (Confidence Interval)

Η πρόβλεψη μπορεί να θεωρηθεί σε ένα πείραμα Bernoulli

- Ένα Bernoulli πείραμα έχει δύο πιθανά αποτελέσματα
- Πιθανά αποτελέσματα πρόβλεψης: σωστό ή λάθος
- Μια συλλογή από πειράματα έχει δυωνυμική κατανομή Binomial distribution:
  - x ~ Bin(N, p) x: αριθμός σωστών προβλέψεων
  - Πχ: ρίξιμο τίμιου νομίσματος (κορώνα/γράμματα) 50 φορές, αριθμός κεφαλών;
    Expected number of heads = N×p = 50 × 0.5 = 25

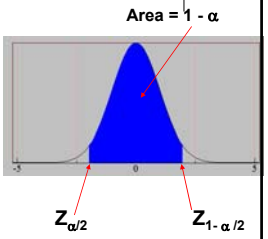
Δοθέντος του x (# σωστών προβλέψεων) ή ισοδύναμα, acc=x/N, και του N (# εγγραφών ελέγχου),

Μπορούμε να προβλέψουμε το p (την πραγματική πιστότητα του μοντέλου);

Αποτίμηση Μοντέλου

Για μεγάλα σύνολα ελέγχου (N > 30), acc έχει κανονική κατανομή με μέσο mean p and variance p(1-p)/N

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



Confidence Interval for p
(Διάστημα εμπιστοσύνης για το p):

$$\frac{2 \times N \times acc + Z_{\alpha/2}^2 \pm \sqrt{Z_{\alpha/2}^2 + 4 \times N \times acc - 4 \times N \times acc^2}}{2(N + Z_{\alpha/2}^2)}$$

Αποτίμηση Μοντέλου

Έστω ένα μοντέλο που έχει accuracy 80% όταν αποτιμάται σε 100 στιγμιότυπα ελέγχου: Ποιο είναι το διάστημα εμπιστοσύνης για την πραγματική του πιστότητα (p) με επίπεδο εμπιστοσύνης (1-α) 95%

N=100, acc = 0.8
1-α = 0.95 (95% confidence)
Από τον πίνακα, $Z_{\alpha/2}$ =1.96
Κάνοντας τις πράξεις 71.1% - 86.7%

1-α	Z
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

Πλησιάζει το 80% όσο το N μεγαλώνει

Αποτίμηση Μοντέλου

- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
  Πως να εκτιμήσουμε την απόδοση ενός μοντέλου
- Μέθοδοι για την εκτίμηση της απόδοσης
  Πως μπορούνε να πάρουμε αξιόπιστες εκτιμήσεις
- Μέθοδοι για την σύγκριση μοντέλων
  Πως να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

Αποτίμηση Μοντέλου

- Δοσμένων δύο μοντέλων, έστω M1 και M2, ποιο είναι καλύτερο;
  - M1 ελέγχεται στο D1 (size=n1), error rate = e_1
  - M2 ελέγχεται στο D2 (size=n2), error rate = e_2
  - Έστω D1 and D2 είναι ανεξάρτητα

Θέλουμε να εξετάσουμε αν η διαφορά d = e_1-e_2 είναι στατιστικά σημαντική

Αν τα n1 και n2 είναι αρκετά are μεγάλα, τότε:

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

Approximate

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

Αποτίμηση Μοντέλου

d = e_1 - e_2

- d ~ $N(d_t, \sigma_t)$ όπου d_t είναι η πραγματική διαφορά
- Since D1 and D2 are independent, their variance adds up:

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

At (1-α) confidence level, $d_t = d \pm Z_{\alpha/2} \hat{\sigma}_t$

Αποτίμηση Μοντέλου

Παράδειγμα

Δοθέντων: M1: n1 = 30, e1 = 0.15 M2: n2 = 5000, e2 = 0.25 d = |e2 - e1| = 0.1

Η εκτιμώμενη variance της διαφοράς στα error rates

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

Για 95% confidence level, Zα/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 είναι στατιστικά μη σημαντική

Άλλοι Ταξινομητές
Ταξινομητές με κανόνες

Ταξινομητές με Κανόνες

Ταξινόμηση των εγγραφών με βάση ένα σύνολο από κανόνες της μορφής "if…then…"

Κανόνας: (*Συνθήκη*) → *y*

όπου

Συνθήκη (Condition) είναι σύζευξη συνθηκών στα γνωρίσματα

y η ετικέτα της κλάσης

LHS: rule antecedent (πρότερο) ή condition (συνθήκη)
RHS: rule consequent (επακόλουθο ή απότοκο)

Παραδείγματα κανόνων ταξινόμησης:

(Blood Type=Warm) ∧ (Lay Eggs=Yes) → Birds
(Taxable Income < 50K) ∧ (Refund=Yes) → Evade=No

Ταξινομητές με Κανόνες

Παράδειγμα

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Ταξινομητές με Κανόνες

Εφαρμογή Ταξινομητών με Κανόνες

Ένας κανόνας r καλύπτει (covers) ένα στιγμιότυπο (εγγραφή) αν τα γνωρίσματα του στιγμιότυπου ικανοποιούν τη συνθήκη του κανόνα

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

Ο κανόνας R1 καλύπτει το hawk (ή αλλιώς το hawk ενεργοποιεί (trigger) τον κανόνα) => Bird
Ο κανόνας R3 καλύπτει το grizzly bear => Mammal

Ταξινομητές με Κανόνες

Κάλυψη Κανόνα - Coverage:
Το ποσοστό των εγγραφών που ικανοποιούν το LHS του κανόνα

Πιστότητα Κανόνα - Accuracy:
Το ποσοστό των κανόνων που καλύπτουν και το LHS και το RHS του κανόνα

(Status=Single) → No

Coverage = 40%, Accuracy = 50%

Tid	Refund	Marital Status	Taxable Income	Class
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

Ταξινομητές με Κανόνες

Χαρακτηριστικά Ταξινομητών με Κανόνες

- Αμοιβαία αποκλειόμενοι κανόνες (Mutually exclusive rules)
  Ένας ταξινομητής περιέχει αμοιβαία αποκλειόμενους κανόνες αν οι κανόνες είναι ανεξάρτητοι ο ένας από τον άλλο

  Κάθε εγγραφή καλύπτεται από το *πολύ έναν* κανόνα

- Εξαντλητικοί κανόνες (Exhaustive rules)
  Ένας ταξινομητής έχει εξαντλητική κάλυψη (coverage) αν καλύπτει όλους τους πιθανούς συνδυασμούς τιμών γνωρισμάτων

  Κάθε εγγραφή καλύπτεται από *τουλάχιστον έναν* κανόνα

Ταξινομητές με Κανόνες

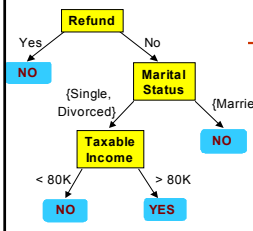
Κατασκευή Ταξινομητών με Κανόνες

- Άμεση Μέθοδος:
  - Εξαγωγή κανόνων απευθείας από τα δεδομένα
  - Π.χ.: RIPPER, CN2, Holte's 1R
- Έμμεση Μέθοδος:
  - Εξαγωγή κανόνων από άλλα μοντέλα ταξινομητών (πχ από δέντρα απόφασης)
  - Π.χ.: C4.5 κανόνες

Ταξινομητές με Κανόνες

Έμμεση Μέθοδος:
Από Δέντρα Απόφασης σε Κανόνες

Ένας κανόνας για κάθε μονοπάτι από τη ρίζα σε φύλλο
Κάθε ζευγάρι γνώρισμα-τιμή στο μονοπάτι αποτελεί ένα όρο στη σύζευξη και το φύλλο αφορά την κλάση (RHS)



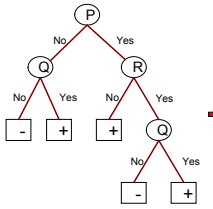
Κανόνες Ταξινόμησης (Classification Rules)

(Refund=Yes) ==> No
(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No
(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes
(Refund=No, Marital Status={Married}) ==> No

- Κανόνες αμοιβαία αποκλειόμενοι και εξαντλητικοί
- Το σύνολο κανόνων περιέχει όση πληροφορία περιέχει και το δέντρο

Ταξινομητές με Κανόνες

Από Δέντρα Απόφασης σε Κανόνες (Παράδειγμα)

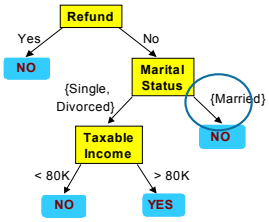


Rule Set

r1: (P=No,Q=No) ==> -
r2: (P=No,Q=Yes) ==> +
r3: (P=Yes,R=No) ==> +
r4: (P=Yes,R=Yes,Q=No) ==> -
r5: (P=Yes,R=Yes,Q=Yes) ==> +

Από Δέντρα Απόφασης σε Κανόνες

Οι κανόνες μπορεί να απλοποιηθούν (απαλοιφή κάποιων όρων στο LHS αν δεν αλλάζει πολύ το λάθος)



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

Αρχικός Κανόνας: **(Refund=No) ∧ (Status=Married) → No**

Απλοποιημένος Κανόνας: **(Status=Married) → No**

Από Δέντρα Απόφασης σε Κανόνες

Αν γίνει απλοποίηση (κλάδεμα):

- Οι κανόνες δεν είναι πια αμοιβαία αποκλειόμενοι
  Μια εγγραφή μπορεί να ενεργοποιήσει παραπάνω από έναν κανόνα
  Λύση (conflict resolution)
  (1) Διάταξη του συνόλου κανόνων (αν μια εγγραφή ενεργοποιεί πολλούς κανόνες, της ανατίθεται αυτός με τη μεγαλύτερη προτεραιότητα) (decision list) ή (2) ο κανόνας με τις πιο πολλές απαιτήσεις (πχ με το μεγαλύτερο αριθμό όρων) (size ordering) ή (3) διάταξη των κλάσεων (αν μια εγγραφή ενεργοποιεί πολλούς κανόνες, της ανατίθεται η τάξη με τη μεγαλύτερη προτεραιότητα) (misclassification cost)
  Χωρίς διάταξη του συνόλου κανόνων – χρήση σχήματος ψηφοφορίας
- Οι κανόνες δεν είναι πια εξαντλητικοί
  Μια εγγραφή μπορεί να μην ενεργοποιεί κάποιον κανόνα
  Λύση
  Χρήση default κλάσης

Άλλοι Ταξινομητές

Ταξινομητές στιγμιοτύπου

Ταξινομητές βασισμένοι σε Στιγμιότυπα

Μέχρι στιγμής

Ταξινόμηση βασισμένη σε δύο βήματα

Βήμα 1: Induction Step – Κατασκευή Μοντέλου Ταξινομητή

Βήμα 2: Deduction Step – Εφαρμογή του μοντέλου για έλεγχο παραδειγμάτων

Eager Learners vs Lazy Learners

πχ Instance Based Classifiers (ταξινομητές βασισμένοι σε στιγμιότυπα)

Μην κατασκευάσεις μοντέλο αν δε χρειαστεί

Ταξινομητές βασισμένοι σε Στιγμιότυπα

Σύνολο Αποθηκευμένων Περιπτώσεων

Atr1	.....	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

- Αποθήκευσε τις εγγραφές του συνόλου εκπαίδευσης
- Χρησιμοποίησε τις αποθηκευμένες εγγραφές για την εκτίμηση της κλάσης των νέων περιπτώσεων

Unseen Case

Atr1	.....	AtrN

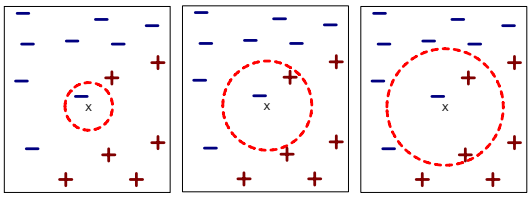
Ταξινομητές βασισμένοι σε Στιγμιότυπα

Παραδείγματα:

- Rote-learner
  - Κρατά (Memorizes) όλο το σύνολο των δεδομένων εκπαίδευσης και ταξινομεί μια εγγραφή αν ταιριάζει πλήρως με κάποιο από τα δεδομένα εκπαίδευσης
- Nearest neighbor – Κοντινότερος Γείτονας
  - Χρήση των k κοντινότερων "closest" σημείων (nearest neighbors) για την ταξινόμηση

Ταξινομητές Κοντινότερου Γείτονα

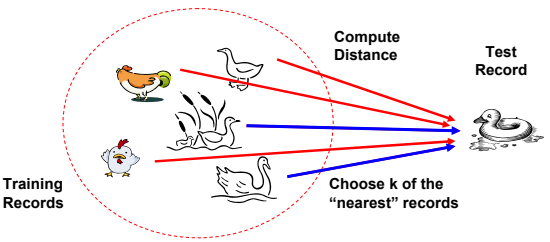
k-κοντινότεροι γείτονες μιας εγγραφής x είναι τα σημεία που έχουν την k-οστή μικρότερη απόσταση από το x



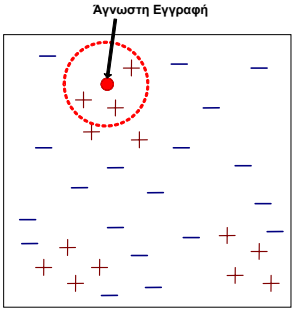
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Ταξινομητές Κοντινότερου Γείτονα

Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck



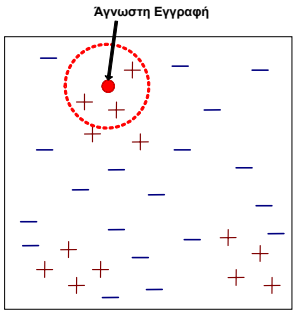
Ταξινομητές Κοντινότερου Γείτονα



Χρειάζεται

1. Το σύνολο των αποθηκευμένων εγγραφών
2. Distance Metric Μετρική απόστασης για να υπολογίσουμε την απόσταση μεταξύ εγγραφών
3. Την τιμή του *k*, δηλαδή τον αριθμό των κοντινότερων γειτόνων που πρέπει να ανακληθούν

Ταξινομητές Κοντινότερου Γείτονα



Για να ταξινομηθεί μια άγνωστη εγγραφή:

- Υπολογισμός της απόστασης από τις εγγραφές του συνόλου
- Εύρεση των *k* κοντινότερων γειτόνων
- Χρήση των κλάσεων των κοντινότερων γειτόνων για τον καθορισμό της κλάσης της άγνωστης εγγραφής - π.χ., με βάση την πλειοψηφία (majority vote)

Ταξινομητές Κοντινότερου Γείτονα

- Απόσταση μεταξύ εγγραφών:
  - Πχ ευκλείδεια απόσταση

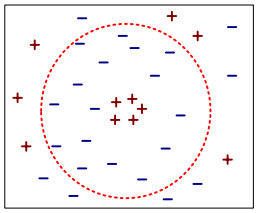
$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Καθορισμός τάξης
  - Απλά τη πλειοψηφική κλάση
  - Βάρος σε κάθε ψήφο με βάση την απόσταση
    - weight factor, $w = 1/d^2$

Ταξινομητές Κοντινότερου Γείτονα

Επιλογή της τιμής του k:

- k πολύ μικρό, ευαισθησία στα σημεία θορύβου
- k πολύ μεγάλο, η γειτονιά μπορεί να περιέχει σημεία από άλλες κλάσεις



Ταξινομητές Κοντινότερου Γείτονα

- Θέματα Κλιμάκωσης
  - Τα γνωρίσματα ίσως πρέπει να κλιμακωθούν ώστε οι αποστάσεις να μην κυριαρχηθούν από κάποιο γνώρισμα
  - Παράδειγμα:
    - height of a person may vary from 1.5m to 1.8m
    - weight of a person may vary from 90lb to 300lb
    - income of a person may vary from $10K to $1M

- Δεν κατασκευάζεται μοντέλο, μεγάλο κόστος για την ταξινόμηση
- Πολλές διαστάσεις (κατάρα των διαστάσεων)
- Θόρυβο (ελάττωση μέσω k-γειτόνων)

Περίληψη

- Ορισμός Προβλήματος Ταξινόμησης
- Μια Κατηγορία Ταξινομητών: Δέντρο Απόφασης
- Μέθοδοι ορισμού της μη καθαρότητας ενός κόμβου
- Θέματα στην Ταξινόμηση: over and under-fitting, missing values, εκτίμηση λάθους
- Αποτίμηση μοντέλου
- Ταξινομητές Στιγμιότυπου (k-κοντινότεροι γείτονες)