Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική
Σχολή Θετικών Επιστημών
Πανεπιστήμιο Θεσσαλίας

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Ανάλυση Δεδομένων

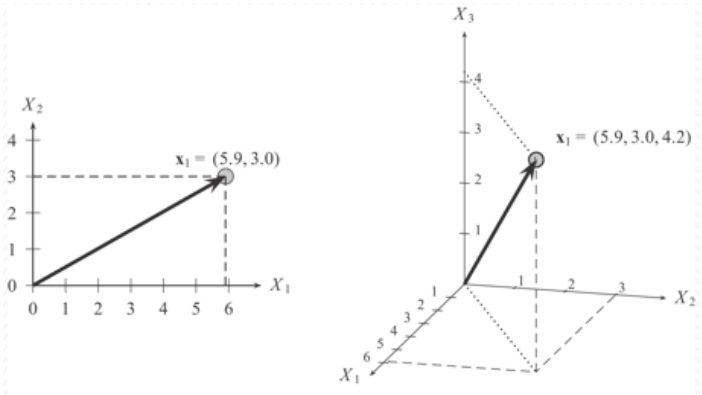
Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD
Adjunct Lecturer

Διάνυσμα

- Μια διατεταγμένη ν-άδα αριθμών.
- Ορίζει ένα πρότυπο στο N-διάστατο χώρο
- Τα στοιχεία του είναι τυχαίες μεταβλητές

- Γενική Μορφή

$$X = [X_1, X_2, \ldots, X_N] \qquad X = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_N \end{bmatrix}$$



Μήτρα δεδομένων

– Τα δεδομένα μπορούν συχνά να αναπαρίστανται συγκεκριμένα ή αφηρημένα με μια μήτρα δεδομένων διαστάσεων n x d, με n γραμμές και d στήλες, η οποία ορίζεται ως

$$\mathbf{D} = \begin{pmatrix} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Γραμμές: Γνωστές και με τους όρους *οντότητες*, *στιγμιότυπα*, *παραδείγματα*, *εγγραφές*, *συναλλαγές*, *αντικείμενα*, *σημεία*, *διανύσματα χαρακτηριστικών* κ.λπ. Ορίζονται ως μια πλειάδα με *d* στοιχεία

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

Στήλες: Γνωστές και με τους όρους *γνωρίσματα*, *ιδιότητες*, *χαρακτηριστικά*, *διαστάσεις*, *μεταβλητές*, *πεδία* κ.λπ. Ορίζονται ως μια πλειάδα με *n* στοιχεία

$$\boldsymbol{X}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Αριθμητική μήτρα δεδομένων

– Αν όλα τα γνωρίσματα είναι αριθμητικά, τότε η μήτρα δεδομένων D είναι μια μήτρα διαστάσεων n x d, ή ισοδύναμα ένα σύνολο με n διανύσματα γραμμής $X_i^T \in \mathbb{R}^d$ ή ένα σύνολο με d διανύσματα στήλης $X_j \in \mathbb{R}^n$

$$\mathbf{D} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} — & \mathbf{x}_1^T & — \\ — & \mathbf{x}_2^T & — \\ & \vdots & \\ — & \mathbf{x}_n^T & — \end{pmatrix} = \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

– Ο μέσος της μήτρας δεδομένων D είναι το διάνυσμα που προκύπτει από τον υπολογισμό του μέσου όρου για όλα τα σημεία:

$$mean(\mathbf{D}) = \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

Μέτρα Κεντρικής Τάσης

- Αριθμητικός Μέσος (Mean)
- ορίζεται ως το άθροισμα των παρατηρήσεων δια του πλήθους αυτών. Είναι δηλαδή η μαθηματική πράξη ανεύρεσης της «μέσης απόστασης» ανάμεσα σε δύο ή περισσότερους αριθμούς.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} t_i = \frac{1}{n}(t_1 + \cdots + t_n)$$ όπου t_i η i παρατήρηση και n το πλήθος των παρατηρήσεων

- Είναι αντιπροσωπευτικός του συνόλου των παρατηρήσεων
- ΚΑΤΑΛΛΗΛΟΣ
  - όταν η κατανομή των τιμών της X στον πληθυσμό είναι κανονική.
- ΑΚΑΤΑΛΛΗΛΟΣ
  - όταν η κατανομή των τιμών της X στον πληθυσμό απέχει πολύ από την κανονική.
  - Επηρεάζεται πολύ από τις ακραίες παρατηρήσεις

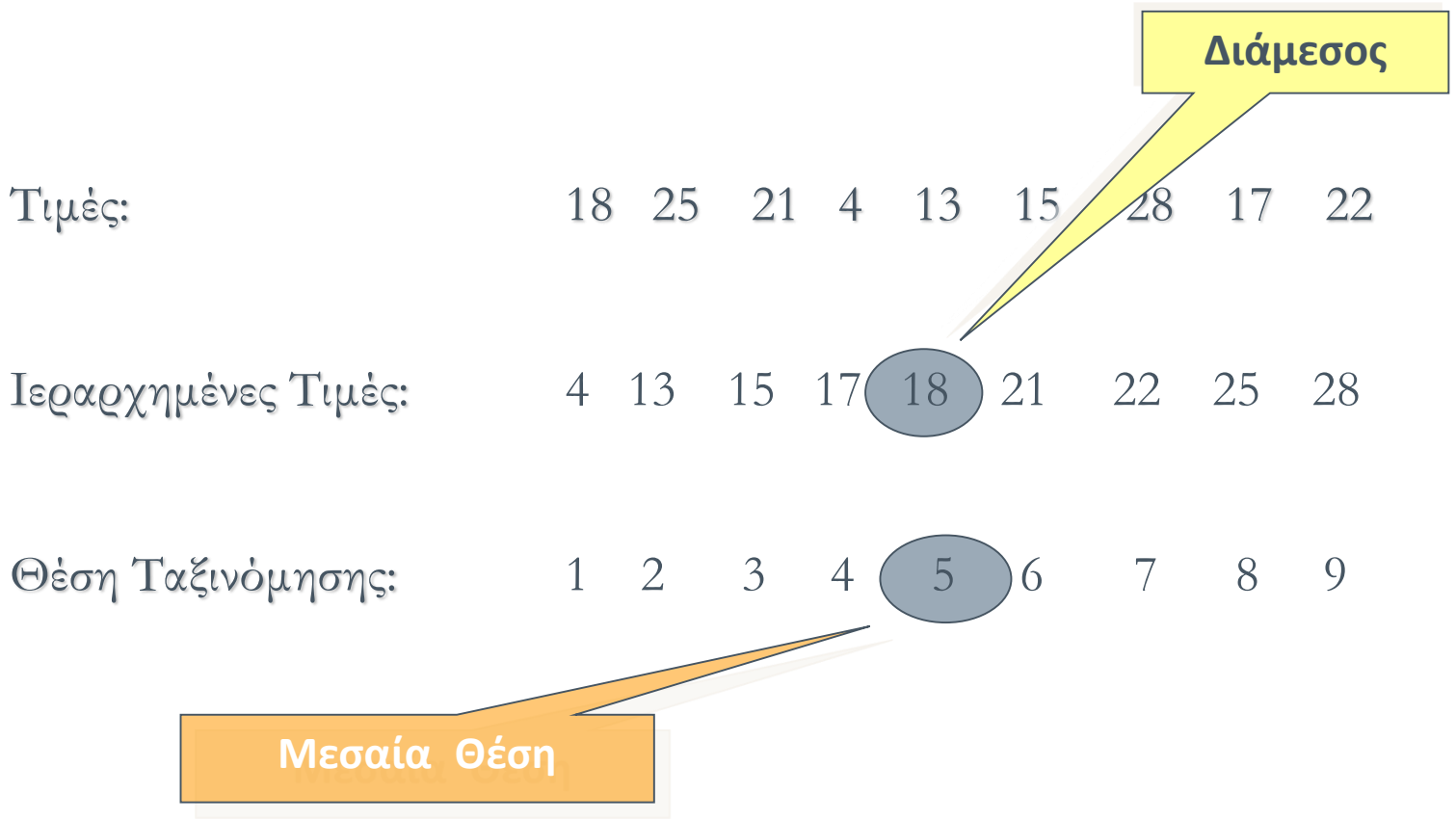
- Σταθμισμένος αριθμητικός μέσος – Weighted Mean:
  - ενός συνόλου n παρατηρήσεων είναι ο αριθμητικός μέσος που προκύπτει σταθμίζοντας κάθε παρατήρηση με συγκεκριμένη βαρύτητα w

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Μέτρα Κεντρικής Τάσης

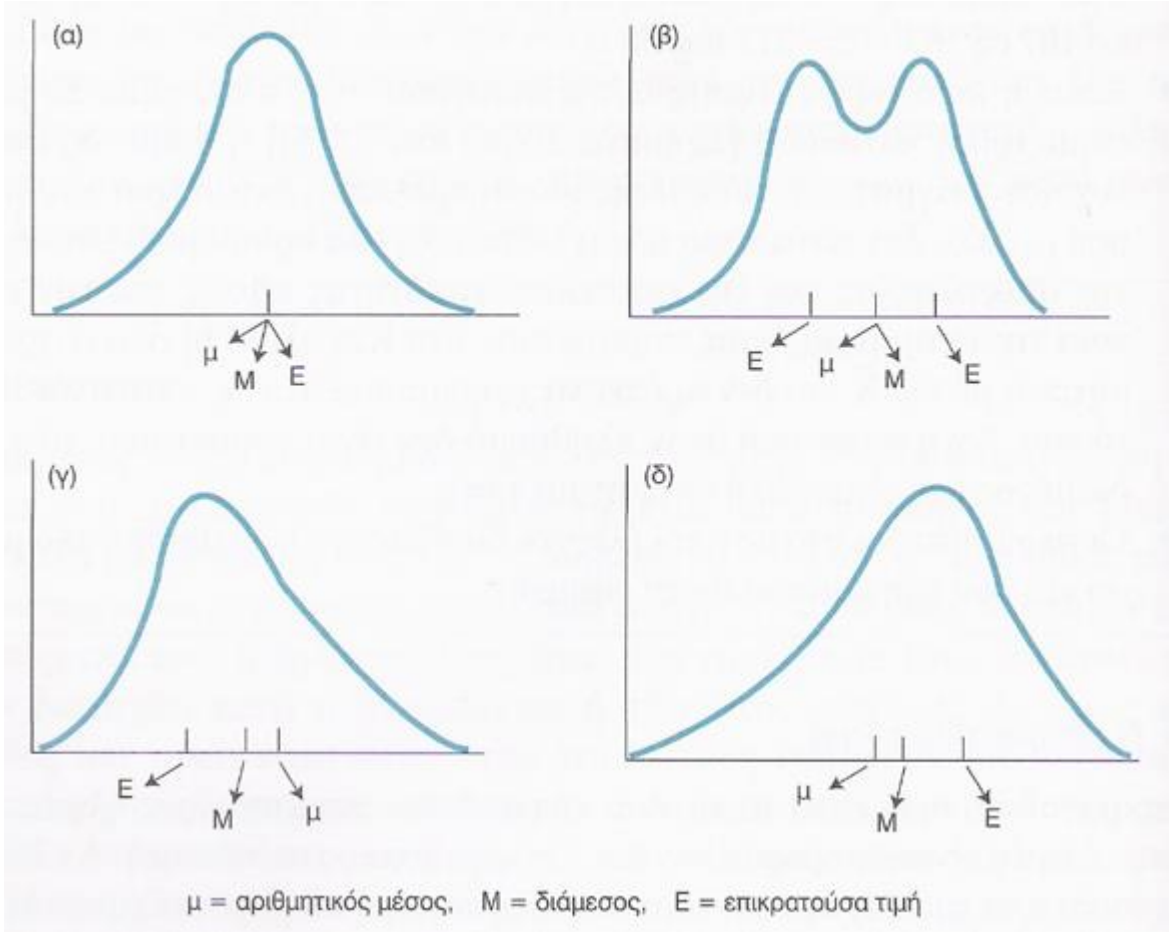
- Διάμεσος (Median)
  - ενός συνόλου n διατεταγμένων κατ' αύξουσα σειρά παρατηρήσεων είναι η κεντρική τιμή αν n περιττός και το ημιάθροισμα των δυο κεντρικών παρατηρήσεων αν n άρτιος

- Δεν επηρεάζεται από ακραίες τιμές



Μέτρα Κεντρικής Τάσης

- Επικρατούσα Τιμή:
  - Ορίζεται ως η παρατήρηση με τη μεγαλύτερη συχνότητα
  - Μπορεί να υπάρχουν περισσότερες της μιας



Μέτρα Διασποράς

- **Διασπορά (variance):**
  - ενός συνόλου n παρατηρήσεων είναι η τιμή εκείνη που δείχνει τον βαθμό απλώματος των δεδομένων από την μέση τιμή
  - Αποτελεί εκτίμηση της διακύμανσης του πληθυσμού

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

- **Τυπική απόκλιση (standard deviation):**
  - ενός συνόλου n παρατηρήσεων είναι η θετική τετραγωνική ρίζα της διακύμανσης
  - Δείτκης Αξιοπιστίας.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

Μέτρα Διασποράς

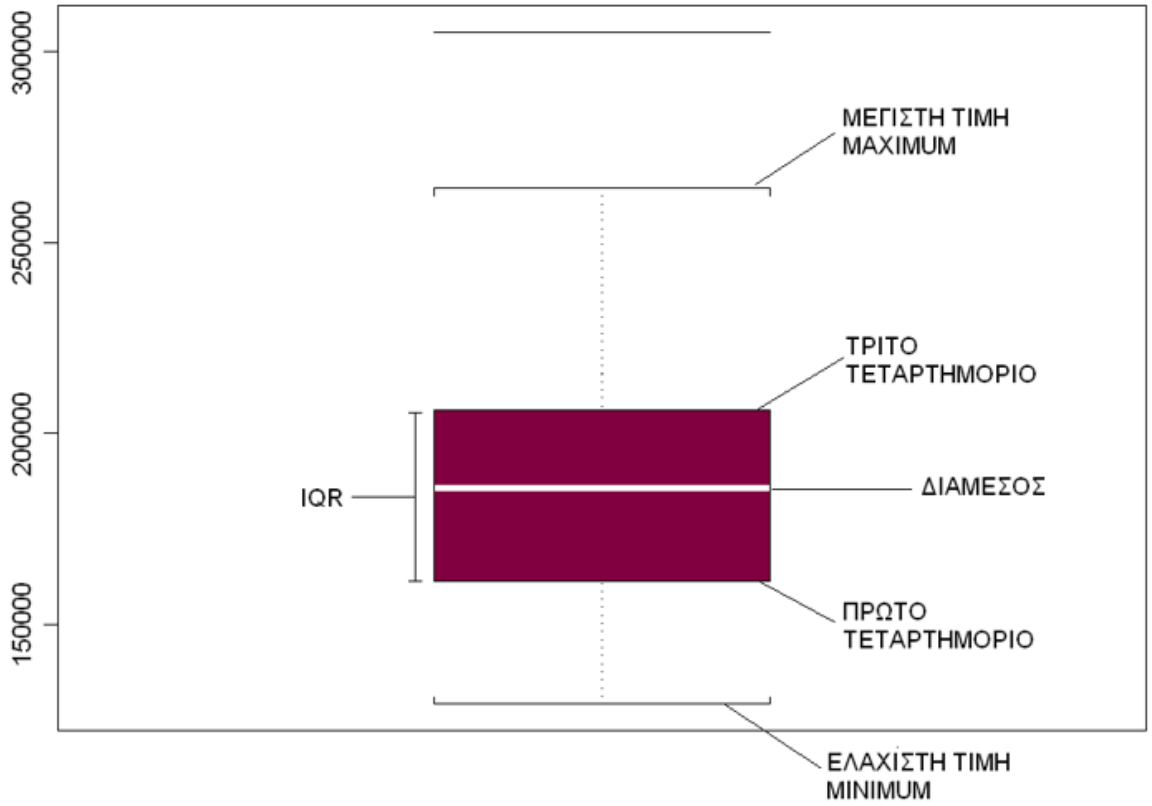
- **Συντελεστής μεταβλητότητας (variation coefficient):**
  - Εκτιμά τη σχέση της τυπικής απόκλισης με το μέγεθος των δεδομένων
  - είναι ο συντελεστής εκείνος ο οποίος μετρά το βαθμό απλώματος των παρατηρήσεων σε σχέση με το μέσο

$$CV = \frac{\sigma}{\mu}$$

- **Τεταρτημόρια**
  - 1ο τεταρτημόριο (Q1): το σημείο κάτω από το οποίο βρίσκεται το 25% των διατεταγμένων τιμών του δείγματος (ή του πληθυσμού, αντίστοιχα)
  - 3ο τεταρτημόριο (Q3): το σημείο πάνω από το οποίο βρίσκεται το 25% των διατεταγμένων τιμών του δείγματος (ή του πληθυσμού, αντίστοιχα)
  - Το δεύτερο τεταρτημόριο (Q2) συμπίπτει με τη διάμεσο.

Θηκόγραμμα (Boxplot)

- Είναι ένας γραφικός τρόπος παρουσίασης πέντε περιληπτικών μέτρων μιας κατανομής ομαδοποιημένων δεδομένων, με συνδυασμό των οποίων είναι δυνατή η άντληση περισσότερων πληροφοριών από αυτήν που περιέχεται στα πέντε αυτά μέτρα.
- Πρόκειται για ένα χρήσιμο διάγραμμα, ειδικά όσον αφορά τη σύγκριση συνεχών κατανομών σε διαφορετικούς πληθυσμούς



Μέτρα Συσχετίσεων

- Συντελεστής Συσχέτισης Pearson (Pearson Correlation Coefficient - PCC)
  - Μέτρηση Απλής Γραμμικής Συσχέτισης
  - Είναι η ενδεδειγμένη εκτιμήτρια (στατιστική) για τη μέτρηση γραμμικής συσχέτισης δύο μεταβλητών X και Y που έχουν μετρηθεί σε φυσική κλίμακα ή σε κλίμακα διαστήματος.

- Αν έχουμε μια σειρά από n μετρήσεις των X και Y γραμμένες ως x_i και y_i για i = 1,2,...,n, τότε ο δειγματικός συντελεστής συσχέτισης μπορεί να χρησιμοποιηθεί για την εκτίμηση του πληθυσμιακού συντελεστή συσχέτισης Pearson *r* μεταξύ X και Y.
- Ο δειγματικός συντελεστής συσχέτισης γράφεται

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

- όπου x και y είναι ο δειγματικός μέσος των X και Y και s_x και s_y είναι οι δειγματικές τυπικές αποκλίσεις των X και Y.

Μέτρα Συσχετίσεων

- Συντελεστής Συσχέτισης Spearman
- Είναι η ενδεδειγμένη στατιστική όταν μια τουλάχιστον από τις μεταβλητές X και Y είναι μεταβλητή διάταξης.
- Χρήσιμος όταν η μία ή και οι δύο μεταβλητές είναι μη κανονικές.
- Μπορεί ακόμη να χρησιμοποιηθεί και στην περίπτωση που οι X και Y είναι ποιοτικές μεταβλητές αλλά είμαστε σε θέση να διατάξουμε τις κατηγορίες κάθε μιας
- Ο συντελεστής συσχέτισης Spearman ορίζεται όπως ο συντελεστής συσχέτισης Pearson μεταξύ των μεταβλητών κατάταξης.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

- Απλούστευση του τύπου (όταν δεν υπάρχει ισοψηφία)

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

Συντελεστής Συσχέτισης Spearman - Παράδειγμα

IQ, X_i	Ώρες τηλεόρασης ανά εβδομάδα, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

IQ, X_i	Ώρες τηλεόρασης ανά εβδομάδα, Y_i	κατάταξη x_i	κατάταξη y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$$

$$\sum d_i^2 = 194.$$

$$\rho = 1 - \frac{6 \times 194}{10(10^2-1)} = -0.175757575$$

Δεν υπάρχει συσχέτιση μεταξύ του δείκτη νοημοσύνης και των ωρών ενασχόλησης με την τηλεόραση

Η έννοια της απόστασης

- Η απόσταση είναι μια θεμελιώδης έννοια στην πολυμεταβλητή ανάλυση και όχι μόνο για την ανάλυση δεδομένων.

- Σκοπός της απόστασης είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις, να ποσοτικοποιήσει δηλαδή αν μοιάζουν ή όχι οι παρατηρήσεις.

- ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

  - Έστω X το άγνωστο πρότυπο και y είναι το πρωτότυπο μιας κατηγορίας

  - Σημαντικό να υπολογίσουμε πόσο απέχουν μεταξύ τους

Ευκλείδια απόσταση

- Ευκλείδια Απόσταση
  - Η συνάρτηση μετράει τη "συνήθη"(Ευκλείδεια) απόσταση μεταξύ δύο σημείων στον επίπεδο, n-διάστατο χώρο κάνοντας επανειλημμένη χρήση του Πυθαγόρειου θεωρήματος.

$$d_e(x,y) = \left(\sum_{i=1}^{N} (x_i - y_i)^2 \right)^{1/2}$$

,όπου x_i, y_i είναι η i συνιστώσα του διανύσματος του αγνώστου προτύπου και του αντίστοιχου προτύπου

- Ας υποθέσουμε πως ενδιαφερόμαστε για δύο μεταβλητές το βάρος και το ύψος, δηλαδή για κάθε παρατήρηση έχουμε μετρήσεις για αυτές τις δύο μεταβλητές.
- Αν συμβολίσουμε τις δύο παρατηρήσεις ως y = (y1, y2) και x = (x1, x2) τότε μια πρώτη προσέγγιση για την επιλογή μιας απόστασης ανάμεσα στις δύο παρατηρήσεις θα ήταν η ευκλείδεια απόσταση

$$d\,(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Συναρτήσεις Απόστασης Προτύπων

- Έστω X το άγνωστο πρότυπο και $y \in S$ είναι το πρωτότυπο μιας κατηγορίας

- Γενικευμένη Ευκλείδια Απόσταση
  - Αποτελεί γενίκευση της Ευκλείδιας Απόστασης (s=2)

$$d_e(x,y) = \left(\sum_{i=1}^{N} (x_i - y_i)^s \right)^{\frac{1}{s}}, s = 2,4,6, \ldots$$

,όπου x_i, y_i είναι η i συνιστώσα του διανύσματος του αγνώστου προτύπου και του αντίστοιχου προτύπου

Απόσταση Mahalanobis

– Mahalanobis: Ινδός Μαθηματικός

$D(x, y, s) = (x - y)^{\mathrm{T}} s^{-1} (x - y)$, όπου S είναι ο πίνακας συνδιασποράς

Συνδιασπορά ή συνδιακύμανση (covariance) δύο μεταβλητών

$$\mathrm{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Ερμηνεία: Το μέσο ποσό της ταυτόχρονης μεταβλητότητας των x kai y από τη μέση τιμή τους

– Πότε χρησιμοποιείται ?

  – Όταν οι τιμές δεν είναι συγκρίσιμες (Πρόβλημα !!!)

  – Για να φέρουμε κάθε μεταβλητή σε συγκρίσιμη κλίμακα διαιρούμε κάθε μεταβλητή με την τυπική της απόκλιση κι επομένως αφού όλες οι μεταβλητές πια θα αναφέρονται σε μονάδες τυπικής απόκλισης έχουμε εξαλείψειτο πρόβλημα.

Απόσταση Manhattan

– Η απόσταση Manhattan μοιάζει πολύ με την ευκλείδεια απόσταση με τη διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις.

$$d(x, y) = \sum_{i=1}^{N} |x_i - y_i|$$

– Συνήθως λόγω της ομοιότητας της με την ευκλείδεια απόσταση δίνει περίπου ίδια αποτελέσματα εκτός από την περίπτωση που υπάρχουν outliers όπου επειδή τους δίνει μικρότερο βάρος (εξαιτίας της απόλυτης τιμής) μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα.

Απόσταση Minkowski (p-norm)

– Η απόσταση Minkowski γενικεύει την Ευκλείδεια απόσταση και την απόσταση Manhattan.

$$d\,(x,y) = \left[\sum_{i=1}^{N} |x_i - y_i|^q\right]^{1/q}$$

– Η τιμή της παραμέτρου q μπορεί να χρησιμοποιηθεί για να δώσει ιδιαίτερο βάρος σε κάποιες αποκλίσεις.

– Προφανώς αν q=1 προκύπτει η απόσταση Manhattan ενώ αν q=2 η ευκλείδεια απόσταση.

Απόσταση σε Ευκλείδιους Χώρους

1-νορμική απόσταση $d_1 = \sum_{i=1}^{n} |x_i - y_i|$

2- νορμική απόσταση $d_2 = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$

Ευκλείδεια απόσταση

p-νορμική απόσταση $d_p = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$

Απόσταση Minkowski

∞- νορμική απόσταση $d_\infty = \lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$

Μέθοδοι Αξιολόγησης Αποτελεσμάτων

- Διάφορες τεχνικές αξιολόγησης των αποτελεσμάτων της αναγνώρισης προτύπων έχουν προταθεί με σκοπό να αποδώσουν με κάποιες μετρικές το πόσο αξιόπιστα είναι τα αποτελέσματα μας

- Υπάρχουν πολλοί τρόποι αξιολόγησης ενός αλγόριθμου, ενώ υπάρχουν επίσης πολλές μετρικές

confusion πίνακας

- Ένας confusion πίνακας, όπως φαίνεται και παρακάτω, περιέχει πληροφορίες σχετικά με πραγματικά και προβλέψιμα αποτελέσματα, δοσμένα από έναν ταξινομητή.

	Classified positive	Classified negative
Actual positive	TP	FN
Actual negative	FP	TN

- Όπου στον άνωθεν πίνακα, ισχύουν τα εξής:
- TP: αριθμός σωστών ταξινομήσεων των θετικών παραδειγμάτων (true positive)
- TN: αριθμός σωστών ταξινομήσεων των αρνητικών παραδειγμάτων (true negative)
- FP: αριθμός λανθασμένων ταξινομήσεων των θετικών παραδειγμάτων (false positive)
- FN: αριθμός λανθασμένων ταξινομήσεων των αρνητικών παραδειγμάτων (false negative)

TP, TN, FP, FN στην εύρεση υπογράφων

- TP: θεωρήσαμε τους υπογράφους (ή κόμβους) που ανήκουν σε ιδεατά clusters και τα βρήκε επιτυχώς ο αλγόριθμος.
- TN: θεωρήσαμε τους κόμβους που δεν ανήκουν σε ιδεατά clusters και τα βρήκε επιτυχώς ο αλγόριθμος.
- FP: θεωρήσαμε τους κόμβους που δεν ανήκουν σε ιδεατά clusters και ο αλγόριθμος τα θεώρησε λανθασμένα ότι ανήκουν σε ιδεατά clusters.
- FN: θεωρήσαμε τους κόμβους που ανήκουν σε ιδεατά clusters και δε τα βρήκε ο αλγόριθμος.

confusion πίνακας

		predicted condition	
	total population	prediction positive	prediction negative
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**

confusion πίνακας

		predicted condition			
	total population	prediction positive	prediction negative	Prevalence $= \frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$	
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection $= \frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$	False Negative Rate (FNR), Miss Rate $= \frac{\Sigma \text{ FN}}{\Sigma \text{ condition positive}}$
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm $= \frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$	True Negative Rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ condition negative}}$
	Accuracy $= \frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$	Positive Predictive Value (PPV), Precision $= \frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) $= \frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False Discovery Rate (FDR) $= \frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) $= \frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	

Ακρίβεια θετικής πρόβλεψης (precision):

- Στην ανάκτηση πληροφορίας ο όρος precision αναφέρεται στον αριθμό των σωστά προβλεπόμενων positive παρατηρήσεων προς όλες τις παρατηρήσεις που θεωρήθηκαν σαν positive στα αποτελέσματα.
- Στην περίπτωση κατηγοριοποίησης το μέτρο υπολογίζεται απ τον τύπο:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \text{ ή } \text{Pr} = \frac{\text{Tp}}{\text{Tp} + \text{Fp}}$$

Πραγματικό ποσοστό θετικών

- Πραγματικό ποσοστό θετικών (True positive rate or Recall or Sensitivity):

- Το μέτρο recall μετράει το ποσοστό από τις προβλεπόμενες positive παρατηρήσεις που είναι πραγματικές positive, δηλαδή:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \text{ ή } \text{Rec} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}}$$

Πραγματικό ποσοστό αρνητικών

- Πραγματικό ποσοστό αρνητικών (True Negative Rate or Specificity)

- Αυτό το μέτρο μετράει το ποσοστό των πραγματικών negative παρατηρήσεων που χαρακτηρίστηκαν ως negative:

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \text{ ή } Sp = \frac{Tn}{Tn + Fp}$$

Συνολική ακρίβεια (Accuracy)

› Η συνολική ακρίβεια του μοντέλου υπολογίζεται ως ο λόγος των σωστά προβλεπόμενων παρατηρήσεων προς όλες τις παρατηρήσεις. Άρα έχουμε για το μέτρο accuracy τον τύπο

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Instances}} \text{ ή } \text{Acc} = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}$$

F-measure ή F-score

- Το μέτρο F-score αποτελεί τον αρμονικό μέσο των Precision και Recall (ή και των Specificity και Recall) με τιμές ανάμεσα σε 0 και 1 (1 για την τέλεια ακρίβεια και 0 για την χειρότερη).
- Ο τύπος του F-score είναι ο εξής:

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

F-measure ή F-score

- Για β=1 έχουμε:

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

- όπου η παράμετρος β ορίζει τι θεωρούμε πιο σημαντικό στις μετρήσεις από τα Precision, Recall.
- Εφόσον θεωρηθούν εξίσου σημαντικά, το β ισούται με 1 και βρίσκουμε το F1 - score .
- Σε άλλες περιπτώσεις μπορεί να είναι προτιμότερο να θεωρήσουμε άλλα F-scores, όπως τα F2-score ή F0.5-score , τα οποία δίνουν περισσότερο βάρος στα μέτρα Precision και Recall αντίστοιχα.