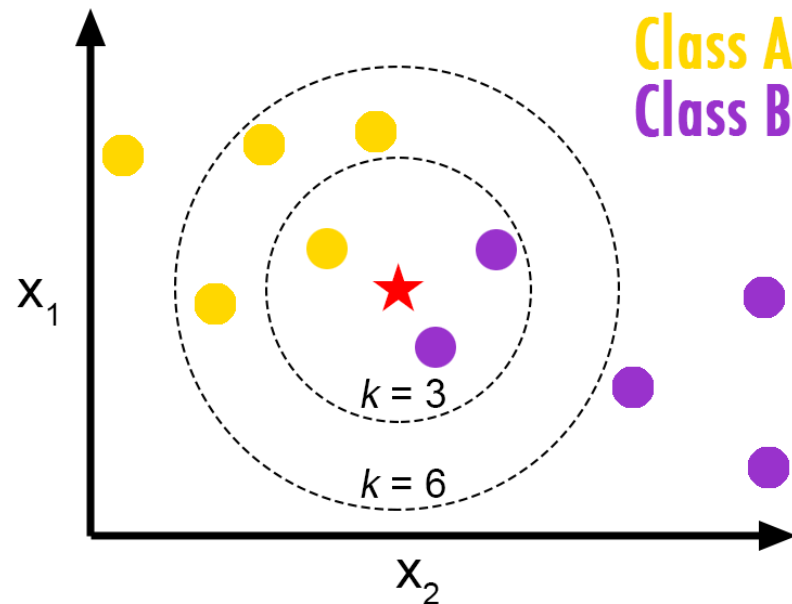


Κεφάλαιο 7

Μηχανική Μάθηση



A) Μάθηση βασισμένη σε
Περιπτώσεις ή Παραδείγματα
Case or Instance Based learning

B) Μετρικές Απόστασης/Ομοιότητας
Similarity Measures



Εισαγωγή

- ❖ Στη **μάθηση βασισμένη σε περιπτώσεις** (*instance-based learning*) τα δεδομένα εκπαίδευσης διατηρούνται αυτούσια.
- ❖ Όταν ένα τέτοιο σύστημα κληθεί να αποφασίσει για την κατηγορία (κλάση) μιας νέας περίπτωσης, εξετάζει εκείνη τη στιγμή τη σχέση της με τα αποθηκευμένα παραδείγματα.
- ❖ Δηλαδή η μέθοδος αυτή αναβάλλει τη μάθηση (τη δημιουργία μοντέλου) έως ότου εμφανιστεί μια νέα περίπτωση
 - ☐ Αναβλητική μάθηση (**lazy learning**).
 - ☐ Αναφέρεται και ως μη-παραμετρική μέθοδος καθώς «μαθαίνουν» από τα δεδομένα απευθείας χωρίς να δημιουργούν κάποιο μαθηματικό μοντέλο
- ❖ Σε αντίθεση, οι μέθοδοι που φτιάχνουν εξ αρχής ένα μοντέλο και το χρησιμοποιούν στη συνέχεια για πρόβλεψη, όπως για παράδειγμα οι αλγόριθμοι γραμμικής ή λογιστικής παρεμβολής, οι αλγόριθμοι κατασκευής δένδρων, κλπ, χαρακτηρίζονται ως
 - ☐ Μέθοδοι ανυπόμονης μάθησης (**eager learners**).
 - ☐ Αναφέρονται και ως παραμετρικοί καθώς δημιουργούν ένα μαθηματικό μοντέλο που συσχετίζει τις εισόδους με τις εξόδους ενός συνόλου δεδομένων.



Ο Αλγόριθμος των k-Πλησιέστερων Γειτόνων

k-Nearest Neighbors ή k-NN

- ❖ Είναι ο χαρακτηριστικός αλγόριθμος αυτής της κατηγορίας.
 - ❑ Γίνεται η παραδοχή ότι τα διάφορα παραδείγματα μπορούν να αναπαρασταθούν ως σημεία σε κάποιον n -διάστατο Ευκλείδειο χώρο R^n
 - ✓ όπου n ο αριθμός των χαρακτηριστικών (ανεξάρτητων μεταβλητών).
 - ❑ Κάθε νέα περίπτωση τοποθετείται στο χώρο αυτό ως νέο σημείο και η τιμή της εξαρτημένης μεταβλητής του (της κλάσης) προσδιορίζεται με βάση την τιμή κλάσης των k πλησιέστερων γειτονικών του σημείων.
 - ❑ Οι πλησιέστεροι γείτονες μπορούν να υπολογιστούν με βάση την Ευκλείδεια απόστασή τους.
 - ✓ Υπάρχουν και άλλες μετρικές απόστασης
- ❖ Έστω ένα σύνολο γνωστών παραδειγμάτων T και έστω (\mathbf{x}_i, y_i) ένα από τα παραδείγματα, με $i=1..|T|$ και \mathbf{x}_i το διάνυσμα των n ανεξάρτητων παραμέτρων. Δηλαδή είναι:

$$\mathbf{x}_i = \langle x_{i1}, x_{i2}, \dots, x_{in} \rangle, \quad x_{ij} \in \mathbb{R}$$
 - ❑ Έστω μια νέα περίπτωση $\mathbf{x}_q = \langle x_{q1}, x_{q2}, \dots, x_{qn} \rangle$ για πρόβλεψη
 - ❑ Η Ευκλείδεια απόσταση μεταξύ των σημείων \mathbf{x}_q και \mathbf{x}_i είναι: $d(\mathbf{x}_q, \mathbf{x}_i) = \sqrt{\sum_{r=1}^n (x_{qr} - x_{ir})^2}$
- ❖ Αν $(\mathbf{x}'_1, y'_1), (\mathbf{x}'_2, y'_2), \dots, (\mathbf{x}'_k, y'_k)$ είναι οι k πλησιέστεροι γείτονες, τότε η πρόβλεψη τιμής είναι (συνέχεια στο επόμενο slide):



Πρόβλεψη τιμής

- ❖ Σε πρόβλημα **ταξινόμησης** (*classification*) σε ένα σύνολο V διακριτών κλάσεων (κατηγοριών), η έξοδος y_q του αλγορίθμου είναι η κατηγορία με τη μεγαλύτερη συχνότητα ανάμεσα στους k -πλησιέστερους γείτονες, δηλαδή:

$$y_q = \arg \max_{v \in V} \sum_{i=1}^k \delta(v, y'_i) \quad \text{όπου} \quad \delta(v, y'_i) = \begin{cases} 1 & \text{αν } y'_i = v \\ 0 & \text{αν } y'_i \neq v \end{cases}$$

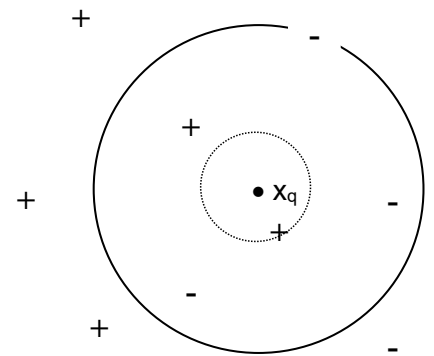
- ❖ Σε πρόβλημα **παρεμβολής** (*regression*), η έξοδος του συστήματος είναι ο μέσος όρος των τιμών της εξαρτημένης μεταβλητής στους k -πλησιέστερους γείτονες:

$$y_q = \frac{\sum_{i=1}^k y'_i}{k}$$



Παράδειγμα

- ❖ Στο Σχήμα απεικονίζονται παραδείγματα δυο κατηγοριών (+ και -).
- ❖ Η νέα περίπτωση x_q χαρακτηρίζεται ως:
 - ❑ Θετική αν ληφθεί υπ' όψη μόνο ο ένας πλησιέστερος γείτονας (1-Nearest Neighbor) και
 - ❑ Αρνητική αν ληφθούν υπ' όψη οι πέντε πλησιέστεροι γείτονες (5-Nearest Neighbors) καθώς η πλειοψηφία αυτών έχει αρνητικό χαρακτηρισμό (εξωτερικός κύκλος στο σχήμα).
- ❖ Ο υπολογισμός της απόστασης στον k-NN επηρεάζεται έντονα από τα χαρακτηριστικά (διαστάσεις) που παίρνουν πολύ μεγάλες τιμές σε σχέση με τα άλλα και δημιουργούν μεγάλες διαφορές Δx .
 - ❑ Έτσι μια διάσταση μπορεί να παίζει καθοριστικό ρόλο στο τελικό αποτέλεσμα χωρίς να είναι πιθανώς σημαντική στην πραγματικότητα.
 - ❑ Η λύση σε αυτό το πρόβλημα είναι η **κανονικοποίηση** (*normalization*) των δεδομένων.
- ❖ Η χρήση της Ευκλείδειας απόστασης δεν είναι υποχρεωτική, αν και είναι η συχνότερη.
 - ❑ Για παράδειγμα, μια εναλλακτική μετρική απόστασης με λιγότερο υπολογιστικό φόρτο είναι η απόσταση **Manhattan**:



$$d(x_q, x_i) = \sum_{r=1}^n |x_{qr} - x_{ir}|$$



Πλησιέστεροι Γείτονες Σταθμισμένης Απόστασης

distance weighted Nearest Neighbors

- ❖ Αποτελεί επέκταση του αλγορίθμου των k-πλησιέστερων γειτόνων
 - ❑ Η συνεισφορά του κάθε γείτονα στην πρόβλεψη της κατηγορίας μιας νέας περίπτωσης είναι αντιστρόφως ανάλογη της απόστασης του από αυτήν.
 - ❑ Δηλαδή οι πλησιέστεροι γείτονες επηρεάζουν περισσότερο από τους μακρινούς.
 - ❑ Αυτό επιτυγχάνεται υιοθετώντας **τιμές βαρών** για κάθε γείτονα.
- ❖ Αν x_q η νέα περίπτωση, το βάρος w_i για τον γείτονα x'_i μπορεί να οριστεί ως το αντίστροφο της απόστασης του από την νέα περίπτωση. Δηλαδή: $w_i = \frac{1}{d(x_q, x'_i)}$
- ❑ όπου $d(x_q, x'_i)$ είναι η απόσταση των γειτόνων από τη νέα περίπτωση
- ❖ Σε πρόβλημα ταξινόμησης η πρόβλεψη τώρα γίνεται:

$$y_q = \arg \max_{v \in V} \sum_{i=1}^k w_i \cdot \delta(v, y'_i) \quad \text{όπου} \quad \delta(v, y'_i) = \begin{cases} 1 & \text{αν } y'_i = v \\ 0 & \text{αν } y'_i \neq v \end{cases}$$

- ❖ Σε προβλήματα παρεμβολής η πρόβλεψη γίνεται:

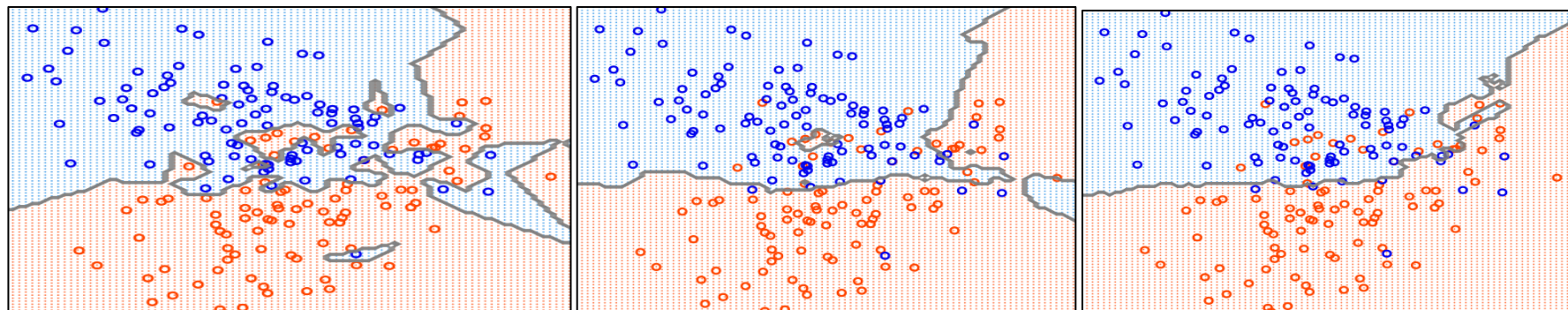
$$y_q = \frac{\sum_{i=1}^k w_i \cdot y'_i}{\sum_{i=1}^k w_i}$$

- ❖ Η στάθμιση με βάρη μας δίνει τη δυνατότητα να λάβουμε υπόψη όλα τα παραδείγματα αντί για τα k πλησιέστερα μόνο, αυξάνοντας βέβαια το χρόνο υπολογισμού.



Σχόλια για τον k-NN

- ❖ Σε προβλήματα ταξινόμησης, ο αλγόριθμος k-NN επιδιώκει να χαράξει ένα σύνορο που να διαχωρίζει τις κλάσεις όσο το δυνατόν καλύτερα.
- ❖ Σε αντίθεση με τα δένδρα ταξινόμησης που χωρίζουν το χώρο των παραδειγμάτων με ευθείες κάθετες στους άξονες των χαρακτηριστικών, ο αλγόριθμος k-NN χαράσσει γενικά ένα πιο πολύπλοκο σύνορο, που καθορίζεται κύρια από την τιμή του k.
- ❖ Στο Σχήμα απεικονίζεται το σύνορο δύο κλάσεων από k-NN για διάφορα k (1, 7 και 20).
 - ❑ Για k=1 ο αλγόριθμος μοντελοποιεί τέλεια τα δεδομένα εκπαίδευσης του εικονιζόμενου προβλήματος, αλλά παράγει ένα μάλλον πολύπλοκο διαχωριστικό σύνορο που υπερπροσαρμόζει τα δεδομένα
 - ✓ Πρακτικά μοντελοποιεί και τυχόν θόρυβο στα δεδομένα, δηλαδή λανθασμένες περιπτώσεις.
 - ❑ Αντίθετα, για μεγάλες τιμές k το σύνολο εξομαλύνεται αρκετά αλλά δε μπορεί να οριοθετήσει σωστά τις κλάσεις και αρχίζει να ταξινομεί λανθασμένα τα δεδομένα εκπαίδευσης.



K=1

K=7

K=20



Σχόλια για τον k-NN (συνεχ.)

- ❖ Αυξάνοντας σταδιακά την τιμή του k προς το πλήθος των παραδειγμάτων, η πρόβλεψη για μια νέα περίπτωση θα καταλήξει να είναι πάντα η πλειοψηφούσα κλάση του συνόλου των περιπτώσεων
 - ❑ Έτσι όμως εκφυλίζεται ο αλγόριθμος καθώς δεν κάνει πρόβλεψη αλλά δίνει πάντα την ίδια απάντηση.
- ❖ Η χρυσή τομή είναι κάπου ενδιάμεσα, δηλαδή το k δεν θα πρέπει να είναι ούτε πολύ μικρό αλλά ούτε και πολύ μεγάλο.
 - ❑ k πολύ μικρό προκαλεί ευαισθησία στα σημεία θορύβου
 - ❑ k πολύ μεγάλο, λαμβάνει υπόψη και δεδομένα που ανήκουν σε άλλες κλάσεις
- ❖ Μια συχνή τιμή του k είναι το \sqrt{N} , όπου N το πλήθος των περιπτώσεων ενώ το $k=10$ συναντάται συχνά ως προεπιλεγμένη (*default*) τιμή σε πολλές περιπτώσεις.
- ❖ Python command (sklearn): [sklearn.neighbors.KNeighborsClassifier\(\)](#)
 - ❑ Παράμετροι:
 - ✓ `n_neighbors`
 - ✓ `weights` (uniform/distance/custom),
 - ✓ `metric` (π.χ. Minkowski με εκθέτη p . [Διαθέσιμες μετρικές στο sklearn](#))



Ποιοτικά χαρακτηριστικά του k-NN

- ❖ Ο αλγόριθμος μεροληπτεί (ευνοεί) περιπτώσεις που είναι "κοντινές"
 - ❑ δηλ. κρίνει τη νέα περίπτωση με βάση την απόσταση της από άλλες περιπτώσεις κάνοντας την παραδοχή ότι η τιμή της εξαρτημένης μεταβλητής θα είναι παρόμοια με εκείνη σε "κοντινές" περιπτώσεις.
- ❖ Τα **πλεονεκτήματα** του k-NN είναι ότι:
 - ❑ Δεν υφίσταται στάδιο εκπαίδευσης και δεν κατασκευάζεται κάποιο μοντέλο.
 - ❑ Κάθε νέα περίπτωση αξιολογείται όταν χρειαστεί, με βάση τις γειτονικές περιπτώσεις της.
 - ❑ Μπορεί να μάθει πολύπλοκες συναρτήσεις επειδή εξειδικεύεται σε κάθε "γειτονιά" με βάση τις περιπτώσεις της.
 - ❑ Δεν έχει απώλεια πληροφορίας γιατί χρησιμοποιεί όλα τα διαθέσιμα δεδομένα.
- ❖ Ως **μειονεκτήματα** του αλγορίθμου θεωρούνται:
 - ❑ Η καθυστέρηση κατά την πρόβλεψη αφού δεν υπάρχει κάποιο μοντέλο (όπως π.χ. στα δένδρα)
 - ✓ Βελτίωση με τη χρήση δενδροειδών δομών δεικτοδότησης (k-d trees¹) που επιτρέπουν το γρήγορο εντοπισμό των κοντινών περιπτώσεων.
 - ❑ Η απόδοση της ίδιας βαρύτητας σε όλα τα χαρακτηριστικά, καθώς δεν είναι σε θέση να ξεχωρίσει κάποια που πιθανώς στην πραγματικότητα είναι ασήμαντα.
 - ✓ Τα ασήμαντα χαρακτηριστικά ενδέχεται να δημιουργήσουν μεγάλη απόσταση μεταξύ των περιπτώσεων και αυτό να εξουδετερώσει μια πιθανή ομοιότητα των περιπτώσεων στις σημαντικές παραμέτρους.

¹ <https://www.geeksforgeeks.org/search-and-insertion-in-k-dimensional-tree/>



Ποιοτικά χαρακτηριστικά του k-NN (συνέχ.)

- ❖ Το τελευταίο μειονέκτημα θα μπορούσε να αντιμετωπιστεί με τη χρήση βαρών στα χαρακτηριστικά, ανάλογα με τη σημαντικότητα τους ή με την απαλοιφή των ασήμαντων χαρακτηριστικών.
 - ❑ Μεγάλα βάρη σε κάποια χαρακτηριστικά θα ενισχύουν τις διαφορές στις τιμές τους με αποτέλεσμα αυτά να είναι οι κυρίαρχοι διαμορφωτές της "απόστασης" μεταξύ δύο περιπτώσεων.
- ❖ Η αύξηση του πλήθους των χαρακτηριστικών επιβαρύνει σημαντικά τον αλγόριθμο.
 - ❑ Γενικά σε χώρους λίγων διαστάσεων με πληθώρα δεδομένων ο KNN έχει πολύ καλά αποτελέσματα. Όσο όμως αυξάνεται ο αριθμός των διαστάσεων δημιουργείται πρόβλημα καθώς οι πλησιέστεροι γείτονες σε χώρους πολλών διαστάσεων συνήθως δεν είναι πολύ κοντά.
- ❖ Γενικότερα, το φαινόμενο της δυσκολίας μάθησης με την αύξηση του αριθμού των χαρακτηριστικών ονομάζεται **κατάρα των διαστάσεων** (*curse of dimensionality*) και αντιμετωπίζεται με την **επιλογή χαρακτηριστικών** (*feature selection*)².

- ❑ Το θέμα της επιλογής ή μείωσης των χαρακτηριστικών απασχολεί όλες τις μεθόδους της μηχανικής μάθησης και υπάρχει συγκεκριμένη επιστημονική περιοχή που ασχολείται με αυτό το θέμα.

² Η επιλογή χαρακτηριστικών και η μείωση διαστασιμότητας αναπτύσσεται στο 3^ο μέρος αυτών των σημειώσεων.



Παράδειγμα

- ❖ Ο Πίνακας περιέχει έξι παλιές περιπτώσεις (1-6) και δύο νέες (οι 7 και 8) που πρέπει να ταξινομηθούν σε μια από τις δύο κλάσεις, θετική ή αρνητική.

- ❑ Η ταξινόμηση θα γίνει με δύο τρόπους:

- ✓ α) με τον αλγόριθμο 5-πλησιέστερων γειτόνων, και
- ✓ β) με τον αλγόριθμο 5-πλησιέστερων γειτόνων σταθμισμένης απόστασης

- ❑ για τις αποστάσεις θα χρησιμοποιηθεί η απόσταση Manhattan.

❖ Απάντηση

- ❑ Με τη χρήση 5-πλησιέστερων γειτόνων, για την περίπτωση q_1 :

$$d(\mathbf{e}_1, \mathbf{q}_1) = \sum_{i=1}^2 |x_{1,i} - x_{q,i}| = |0.2 - 0.15| + |0.5 - 0.65| = 0.2,$$

$$d(\mathbf{e}_2, \mathbf{q}_1) = 0.2$$

$$d(\mathbf{e}_3, \mathbf{q}_1) = 0.5$$

$$d(\mathbf{e}_4, \mathbf{q}_1) = 0.2$$

$$d(\mathbf{e}_5, \mathbf{q}_1) = 0.5$$

$$d(\mathbf{e}_6, \mathbf{q}_1) = 1$$

- ❑ Οπότε τα 5 πλησιέστερα παραδείγματα είναι τα $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4, \mathbf{e}_5\}$ με $(y_1, y_2, y_3, y_4, y_5) = (-, -, +, +, +)$ και είναι αυτά που θα συνυπολογιστούν για την ταξινόμηση της περίπτωσης q_1

$$y_q = \operatorname{argmax}((3)_+, (2)_-) = +$$

Παράδειγμα i	e_{i1}	e_{i2}	Κλάση
1	0.15	0.65	-
2	0.20	0.30	-
3	0.40	0.20	+
4	0.40	0.50	+
5	0.70	0.50	+
6	0.75	0.95	-
7 (q_1)	0.20	0.50	$y_1 = ?$
8 (q_2)	0.50	0.20	$y_2 = ?$



Παράδειγμα (συνεχ.)

- Για την περίπτωση q_2 :

$$d(e_1, q_2) = 0.8$$

$$d(e_2, q_2) = 0.4$$

$$d(e_3, q_2) = 0.1$$

$$d(e_4, q_2) = 0.4$$

$$d(e_5, q_2) = 0.5$$

$$d(e_6, q_2) = 1$$

- Οπότε τα 5 πλησιέστερα παραδείγματα είναι και πάλι τα $\{e_1, e_2, e_3, e_4, e_5\}$ με $(y_1, y_2, y_3, y_4, y_5) = (-, -, +, +, +)$ και θα συνυπολογιστούν για την ταξινόμηση της περίπτωσης q_2 :

$$y_q = \operatorname{argmax}((3)_+, (2)_-) = +$$



Παράδειγμα (συνεχ.): Με την χρήση 5-πλησιέστερων γειτόνων σταθμισμένης απόστασης

- ❑ Επιπρόσθετα υπολογίζουμε το βάρος w_i κάθε κοντινού γείτονα e_i βασισμένοι στην απόσταση του από την δοθείσα περίπτωση.

- ❑ Για την περίπτωση q_1 :

$$w_1 = \frac{1}{d(q_1, e_1)} = 1/0.2 = 5, \quad w_2 = 5, \quad w_3 = 2, \quad w_4 = 5, \quad w_5 = 2$$

$$\text{με } (y_1, y_2, y_3, y_4, y_5) = (-, -, +, +, +)$$

- ❑ Οπότε για την ταξινόμηση της περίπτωσης q_1 με χρήση σταθμισμένης απόστασης θα ισχύει

$$y_q = \operatorname{argmax}((5 + 5)_-, (2 + 5 + 2)_+) = \operatorname{argmax}((10)_-, (9)_+) = -$$

- ❑ Για την περίπτωση q_2 :

$$w_1 = \frac{1}{d(q_2, e_1)} = 1/0.8 = 1.25, \quad w_2 = 2.5, \quad w_3 = 10, \quad w_4 = 2.5, \quad w_5 = 2 \quad \text{με } (y_1, y_2, y_3, y_4, y_5) = (-, -, +, +, +)$$

- ❑ Οπότε για την ταξινόμηση της περίπτωσης q_2 με χρήση σταθμισμένης απόστασης θα ισχύει:

$$y_q = \operatorname{argmax}((1.25 + 2.5)_-, (10 + 2.5 + 2)_+) = \operatorname{argmax}((3.75)_-, (14.5)_+) = +$$

- ❑ Παρατηρούμε ότι με τη χρήση σταθμισμένων γειτόνων είχαμε αλλαγή κλάσης της περίπτωσης q_1 αφού τα 2 παραδείγματα της κλάσης (-) αν και λιγότερα από τα 3 παραδείγματα της κλάσης (+), έλαβαν μεγαλύτερο βάρος στην ταξινόμηση της νέας περίπτωσης αφού είναι πιο κοντά σε αυτήν.



Recommendation system

k Nearest Neighbor algorithm is a very basic common approach for implementing recommendation systems. In k Nearest Neighbors, we try to find the most similar k number of users as nearest neighbors to a given user and predict ratings of the user for a given (for example) movie according to the information of the selected neighbors. So the algorithm has a lot of variation based on two points.

One is how to calculate the distance of each user, and another is how to use or analyze the nearest neighbors to predict the ratings of a given user.

I implemented Euclidean Distance and Cosine Similarity as the methods to calculate the distance and tried various ways of analysis to predict the ratings like taking average, weighted average or the majority among nearest neighbors.

Now I explain a little about the methods of measuring the distance. First, Euclidean Distance is the ordinary straight line distance between two points in Euclidean Space. If the dimension is two, the distance is just between two points in xy plane space, and we just extend this concept to use for our 17,000 dimensional space to calculate the length of the line.

http://cs.carleton.edu/cs_comps/0910/netflixprize/final_results/knn/index.html

See also Jaccard similarity measure



Ασκήσεις - Παραδείγματα

**ΑΣΚΗΣΗ 1.**

Με βάση τα δεδομένα του παρακάτω πίνακα, υπολογίστε την κλάση της νέας περίπτωσης:

(Ηλικία≤30,Εισόδημα=Μέτριο, Φοιτητής=Ναι, Πιστοληπτική_ Ικανότητα =Μέτρια), χρησιμοποιώντας τον αλγόριθμο των 5 σταθμισμένης απόστασης πλησιέστερων γειτόνων.

Θεωρείστε την εξής απλή συνάρτηση υπολογισμού απόστασης μεταξύ δύο περιπτώσεων:

(Ηλικία=a1, Εισόδημα=a2, Φοιτητής=a3, Πιστοληπτική_Ικανότητα=a4) και

(Ηλικία=b1, Εισόδημα=b2, Φοιτητής=b3, Πιστοληπτική_Ικανότητα=b4):

$$\frac{1}{4} \sum_{i=1}^4 \delta(a_i, b_i) \quad \text{όπου} \quad \delta(a_i, b_i) = \begin{cases} 0 & a_i = b_i \\ 1 & a_i \neq b_i \end{cases}$$

Ηλικία	Εισόδημα	Φοιτητής	Πιστοληπτική Ικανότητα	Αγοράζει Υπολογιστή
≤30	Υψηλό	Όχι	Μέτρια	Όχι
≤30	Υψηλό	Όχι	Άριστη	Όχι
31..40	Υψηλό	Όχι	Μέτρια	Ναι
>40	Μέτριο	Όχι	Μέτρια	Ναι
>40	Χαμηλό	Ναι	Μέτρια	Ναι
>40	Χαμηλό	Ναι	Άριστη	Όχι
31..40	Χαμηλό	Ναι	Άριστη	Ναι
≤30	Μέτριο	Όχι	Μέτρια	Όχι
≤30	Χαμηλό	Ναι	Μέτρια	Ναι
>40	Μέτριο	Ναι	Μέτρια	Ναι
≤30	Μέτριο	Ναι	Άριστη	Ναι
31..40	Υψηλό	Ναι	Μέτρια	Ναι
31..40	Μέτριο	Όχι	Άριστη	Ναι
>40	Μέτριο	Όχι	Άριστη	Όχι

**ΑΣΚΗΣΗ 1.: ΑΠΑΝΤΗΣΗ**

- Για τη νέα περίπτωση: $\mathbf{x}_q = \langle (\text{Ηλικία} \leq 30, \text{Εισόδημα} = \text{Μέτριο}, \text{Φοιτητής} = \text{Ναι}, \text{Πιστοληπτική_Ικανότητα} = \text{Μέτρια}) \rangle$
- Βάσει των δοθέντων παραδειγμάτων και με χρήση 5-πλησιέστερων γειτόνων σταθμισμένης απόστασης θα υπολογίσουμε την κατηγορία
- y_q (Αγοράζει Υπολογιστή = {Ναι, Όχι}) στην οποία θα ταξινομηθεί η περίπτωση \mathbf{x}_q
- Υπολογίζουμε την απόσταση της δοθείσας περίπτωσης με το καθένα από τα παραδείγματα που μας δόθηκε χρησιμοποιώντας την μετρική που μας δόθηκε

$$d(x_q, x_i) = \frac{1}{4} \sum_{i=1}^4 \delta(x_{q,i}, a_i)$$

Ηλικία (a_1)	Εισόδημα (a_2)	Φοιτητής (a_3)	Πιστοληπτική Ικανότητα (a_4)	Αγοράζει Υπολογιστή (y_q)	Απόσταση
≤ 30	Υψηλό	Όχι	Μέτρια	Όχι	0.5
≤ 30	Υψηλό	Όχι	Άριστη	Όχι	0.75
31...40	Υψηλό	Όχι	Μέτρια	Ναι	0.75
>40	Μέτριο	Όχι	Μέτρια	Ναι	0.5
>40	Χαμηλό	Ναι	Μέτρια	Ναι	0.5
>40	Χαμηλό	Ναι	Άριστη	Όχι	0.75
31...40	Χαμηλό	Ναι	Άριστη	Ναι	0.75
≤ 30	Μέτριο	Όχι	Μέτρια	Όχι	0.25
≤ 30	Χαμηλό	Ναι	Μέτρια	Ναι	0.25
>40	Μέτριο	Ναι	Μέτρια	Ναι	0.25
≤ 30	Μέτριο	Ναι	Άριστη	Ναι	0.25
31...40	Υψηλό	Ναι	Μέτρια	Ναι	0.5
31...40	Μέτριο	Όχι	Άριστη	Ναι	0.75
>40	Μέτριο	Όχι	Άριστη	Όχι	0.75



- Σημειώνουμε (με σκίαση) τα 5 κοντινότερα παραδείγματα στην περίπτωση x_q τα οποία και θα χρησιμοποιηθούν.
- Επειδή απόσταση 0.5 έχουν περισσότερα από 1 παραδείγματα, ο 5^{ος} πλησιέστερος γείτονας επιλέγει τυχαία μεταξύ όσων είχαν απόσταση 0.5 από την δοθείσα περίπτωση.

Για τη χρήση σταθμισμένων γειτόνων υπολογίζουμε το βάρος των 5 κοντινότερων παραδειγμάτων σύμφωνα με τον τύπο $w_i = 1/d(x_q, x_i)$

Ηλικία (x_1)	Εισόδημα (x_2)	Φοιτητής (x_3)	Πιστοληπτική Ικανότητα (x_4)	Αγοράζει Υπολογιστή (y_q)	$d(x_q, x_i)$	$w_i = 1/d(x_q, x_i)$
≤ 30	Υψηλό	Όχι	Μέτρια	Όχι	0.5	-
≤ 30	Υψηλό	Όχι	Άριστη	Όχι	0.75	-
31...40	Υψηλό	Όχι	Μέτρια	Ναι	0.75	-
>40	Μέτριο	Όχι	Μέτρια	Ναι	0.5	-
>40	Χαμηλό	Ναι	Μέτρια	Ναι	0.5	-
>40	Χαμηλό	Ναι	Άριστη	Όχι	0.75	-
31...40	Χαμηλό	Ναι	Άριστη	Ναι	0.75	-
≤ 30	Μέτριο	Όχι	Μέτρια	Όχι	0.25	4.00
≤ 30	Χαμηλό	Ναι	Μέτρια	Ναι	0.25	4.00
>40	Μέτριο	Ναι	Μέτρια	Ναι	0.25	4.00
≤ 30	Μέτριο	Ναι	Άριστη	Ναι	0.25	4.00
31...40	Υψηλό	Ναι	Μέτρια	Ναι	0.5	2.00
31...40	Μέτριο	Όχι	Άριστη	Ναι	0.75	-
>40	Μέτριο	Όχι	Άριστη	Όχι	0.75	-

Οπότε για την κλαση y_q , Αγοράζει Υπολογιστή = {Ναι, Όχι} θα είναι:

$$y_q = \operatorname{argmax}((4+4+4+2)_{\text{ΝΑΙ}}, (4)_{\text{ΟΧΙ}}) = \operatorname{argmax}((14)_{\text{ΝΑΙ}}, (4)_{\text{ΟΧΙ}}) = \text{Ναι}$$

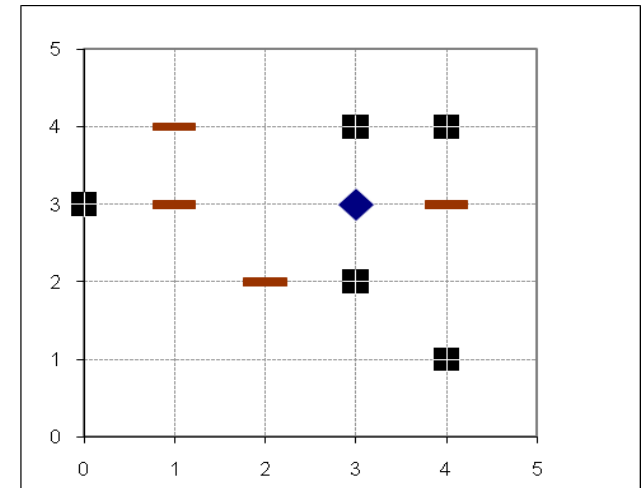
**ΑΣΚΗΣΗ 2.**

Στο διπλανό σχήμα να κατηγοριοποιηθεί το στιγμιότυπο (3,3) (Ο ρόμβος):

α) Με τον αλγόριθμο 9- πλησιέστερων γειτόνων

β) Με τον αλγόριθμο πλησιέστερων γειτόνων σταθμισμένης απόστασης.



Και για τα 2 ερωτήματα να χρησιμοποιήσετε την απόσταση Manhattan.



**ΑΣΚΗΣΗ 2: ΑΠΑΝΤΗΣΗ**

Η άγνωστη περίπτωση είναι η $\mathbf{x}_q = (3,3)$

Καταγράφουμε τα παραδείγματα που απεικονίζονται στο σχήμα:

α) Βάσει των παραδειγμάτων και με χρήση 9-πλησιεστερων γειτόνων θα υπολογίσουμε την κατηγορία y_q (σύμβολο= {, }={ΚΟΚΚΙΝΟ,ΜΑΥΡΟ}) στην οποία θα ταξινομηθεί η περίπτωση \mathbf{x}_q χρησιμοποιώντας ως μετρική $d(x,y)$ την μετρική Manhattan.

$$d(\mathbf{x}_1, \mathbf{x}_q) = \sum_{i=1}^2 |x_{1,i} - x_{q,i}| = |3 - 0| + |3 - 3| = 3$$

Αντίστοιχα, για τα υπόλοιπα παραδείγματα:

$$d(\mathbf{x}_2, \mathbf{x}_q) = 2$$

$$d(\mathbf{x}_5, \mathbf{x}_q) = 1$$

$$d(\mathbf{x}_8, \mathbf{x}_q) = 1$$

$$d(\mathbf{x}_3, \mathbf{x}_q) = 3$$

$$d(\mathbf{x}_6, \mathbf{x}_q) = 1$$

$$d(\mathbf{x}_9, \mathbf{x}_q) = 2$$

$$d(\mathbf{x}_4, \mathbf{x}_q) = 2$$







$$d(\mathbf{x}_7, \mathbf{x}_q) = 3$$

Από τα δεδομένα έχουμε ότι: $(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9) = (\blacksquare, \text{---}, \text{---}, \text{---}, \blacksquare, \blacksquare, \blacksquare, \text{---}, \blacksquare)$

Αφού τα παραδείγματα είναι όσα και οι κοντινότεροι γείτονες που αναζητάμε, θα συνυπολογιστούν όλα για την ταξινόμηση της περίπτωσης \mathbf{x}_q

$$y_q = \arg \max((4)_{\text{ΚΟΚΚΙΝΟ}}, (5)_{\text{ΜΑΥΡΟ}}) = B \rightarrow \blacksquare$$

Άρα η περίπτωση \mathbf{x}_q θα ταξινομηθεί ως \blacksquare

Παράδειγμα	$x_{i,1}$	$x_{i,2}$	y_i
\mathbf{x}_1	0	3	
\mathbf{x}_2	1	3	
\mathbf{x}_3	1	4	
\mathbf{x}_4	2	2	
\mathbf{x}_5	3	2	
\mathbf{x}_6	3	4	
\mathbf{x}_7	4	1	
\mathbf{x}_8	4	3	
\mathbf{x}_9	4	4	



β) Για την χρήση πλησιεστερων γειτόνων σταθμισμένης απόστασης επιπρόσθετα υπολογίζουμε το βάρος w_i κάθε κοντινού γείτονα x_i βασισμένοι στην απόσταση του από την δοθείσα περίπτωση x_q

$$w_1 = \frac{1}{d(\mathbf{x}_1, \mathbf{x}_q)} = 1/3 = 0.33$$

$$w_2 = 0.5$$

$$w_5 = 1$$

$$w_8 = 1$$

$$w_3 = 0.33$$

$$w_6 = 1$$

$$w_9 = 0.5$$

$$w_4 = 0.5$$

$$w_7 = 0.33$$

και από τα δεδομένα $(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9) = (\blacksquare, \text{---}, \text{---}, \text{---}, \blacksquare, \blacksquare, \blacksquare, \text{---}, \blacksquare)$

Οπότε για την ταξινόμηση της περίπτωσης x_q με χρήση σταθμισμένης απόστασης έχουμε

$$\begin{aligned} y_q &= \arg \max((0.5 + 0.33 + 0.5 + 1)_{\text{KOKKINO}}, (0.33 + 1 + 1 + 0.33 + 0.5)_{\text{MAYPO}}) \\ &= \arg \max((2.33)_{\text{KOKKINO}}, (3.16)_{\text{MAYPO}}) = \text{MAYPO} \end{aligned}$$

Παρατηρούμε ότι με τη χρήση σταθμισμένων γειτόνων η περίπτωση x_q ταξινομήθηκε και πάλι ως \blacksquare

**ΑΣΚΗΣΗ 3: (Homework)**

Για τα δεδομένα του διπλανού πίνακα, προβλέψτε την άγνωστη περίπτωση **<100, West, 5>** με τη χρήση του αλγορίθμου 5 – κοντινότερων γειτόνων σταθμισμένης απόστασης:

α) με υπολογιστικές μεθόδους,

β) με χρήση ενός εργαλείου προγραμματισμού (Weka ή Python ή R) και δοκιμή διαφόρων τιμών των παραμέτρων (π.χ. αριθμός γειτόνων, συνάρτηση απόστασης), και

γ) σύγκριση των δυο αποτελεσμάτων (υπολογιστικό/προγραμματιστικό)

Σημείωση: Κανονικοποιείτε πρώτα τις τιμές των χαρακτηριστικών

Data set in Weka format (.arf)

```
% Title: Estimation of a House Value
```

```
% Attribute Information:
```

```
% 1. size in square meters
```

```
% 2. area in which the house is located (East, Centre, West)
```

```
% 3. floor is an integer number
```

```
% 4. class in thousands of euros
```

```
@RELATION HouseValue
```

```
@ATTRIBUTE size REAL
```

```
@ATTRIBUTE area {East, Centre, West}
```

```
@ATTRIBUTE floor REAL
```

```
@ATTRIBUTE class REAL
```

```
@DATA
```

```
100,East,2,200
```

```
120,East,3,250
```

```
80,East,4,150
```

```
90,Centre,2,250
```

```
100,Centre,3,280
```

```
120,Centre,4,360
```

```
110,West,2,100
```

```
130,West,3,120
```

```
90,West,4,90
```

```
150,West,2,120
```

No	Size (sq meters)	Area	Floor	Value (K €)
1	100	East	2	200
2	120	East	3	250
3	80	East	4	150
4	90	Centre	2	250
5	100	Centre	3	280
6	120	Centre	4	360
7	110	West	2	100
8	130	West	3	120
9	90	West	4	90
10	150	West	2	120

**ΑΣΚΗΣΗ 3:** με παραλλαγή του Data set

Προβλέψτε την άγνωστη περίπτωση 7

Σημείωση: Κανονικοποιείτε πρώτα τις τιμές των χαρακτηριστικών

Τα δεδομένα size είναι βαθμωτά (ordinal values).

No	Size (sq meters)	Area	Floor	Value (Κ €)
1	100-120	East	2	250
2	<100	East	4	150
3	<100	Centre	2	250
4	100-120	Centre	4	350
5	100-120	West	2	100
6	>120	West	4	150
7	<100	West	3	?



Παράρτημα



Weka lazy classifiers

IBk

K-nearest neighbors classifier. Can select appropriate value of K based on cross-validation. Can also do distance weighting.

KStar (K*)

K* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function.

Locally weighted learning (LWL)

Locally weighted learning, uses an instance-based algorithm to assign instance weights which are then used by a specified Weighted Instances Handler.

Can do classification (e.g. using naive Bayes) or regression (e.g. using linear regression).

For more info, see:

Eibe Frank, Mark Hall, Bernhard Pfahringer: Locally Weighted Naive Bayes. In: 19th Conference in Uncertainty in Artificial Intelligence, 249-256, 2003.

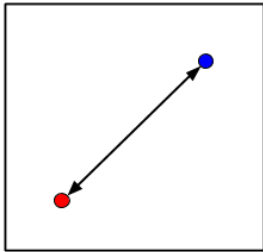
C. Atkeson, A. Moore, S. Schaal (1996). Locally weighted learning. AI Review..

Python

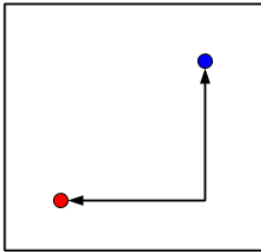
[Tutorial To Implement k-Nearest Neighbors in Python From Scratch](#)

Μετρικές Απόστασης/Ομοιότητας

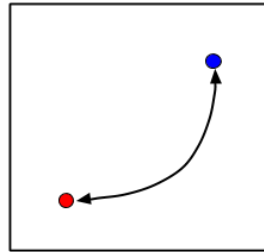
Euclidean



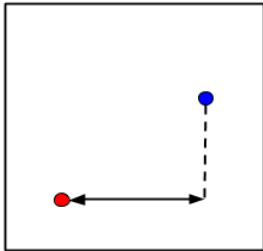
Manhattan



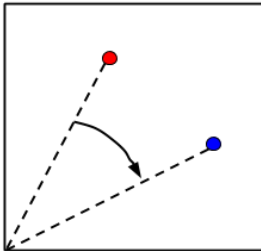
Minkowski



Chebychev



Cosine Similarity



Hamming





Μετρικές Απόστασης/Ομοιότητας

- ❖ Στα προηγούμενα, τα παραδείγματα ήταν ήδη σημεία σε Ευκλείδειο χώρο και η έννοια της απόστασης ήταν φυσιολογική.
- ❖ Σε χαρακτηριστικά όμως με μη αριθμητικές τιμές απαιτείται προφανώς ένας άλλος τρόπος υπολογισμού της απόστασης τους.
- ❖ Απόσταση ονομαστικών τιμών
 - ❑ Στις **ονομαστικές τιμές χωρίς τάξη** (*nominal values*) μπορούμε να θέσουμε την απόσταση ίση με μηδέν (0) αν δύο τιμές είναι ίδιες και ίση με ένα (1) όταν είναι διαφορετικές.
 - ❑ Επιπλέον, μπορούμε αν χρειάζεται να καθορίσουμε κατά σύμβαση το τι θεωρείται ίδιο και τι όχι.
 - ✓ Για παράδειγμα, σε ένα χαρακτηριστικό "επάγγελμα", τα επαγγέλματα "δάσκαλος" και "καθηγητής" θα μπορούσε κατά σύμβαση να θεωρούνται το ίδιο (απόσταση 0) ενώ τα "δάσκαλος" και "πωλητής" διαφορετικά (απόσταση 1).³
 - ❑ Στις ονομαστικές τιμές με τάξη δηλ. τις **βαθμωτές** (*ordinal values*), μπορούμε να αναθέσουμε αριθμητικές τιμές έτσι ώστε να μπορεί να υπολογιστεί κανονικά η απόσταση.
 - ✓ Για παράδειγμα, σε ένα χαρακτηριστικό "σπουδές", οι πιθανές τιμές "γυμνάσιο", "λύκειο", "πανεπιστήμιο" έχουν εκ φύσεως απόσταση μεταξύ τους.
 - ✓ Θα μπορούσαμε να ορίσουμε τη διαφορά ανάμεσα σε "γυμνάσιο" και "λύκειο" σε ένα (1), μεταξύ "γυμνάσιο" και "πανεπιστήμιο" σε δύο (2), κ.ο.κ.

³ Στον KNN της python δεν υποστηρίζονται nominal χαρακτηριστικά και πρέπει να προηγηθεί μετατροπή τους σε αριθμητικές τιμές (π.χ. One Hot Encoding).

Στο WEKA η μετατροπή γίνεται εσωτερικά χωρίς την παρέμβαση του χρήστη.



One Hot Encoding

- ❖ Is a process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions.
 - ❑ One hot encoding is a crucial part of feature engineering for machine learning.
- ❖ Some machine learning algorithms can work directly with categorical data depending on implementation, such as a Bayes and decision tree⁴, but most require any inputs or outputs variables to be a number, or numeric in value.
 - ❑ This means that any categorical data must be mapped to integers.
- ❖ One hot encoding is one method of converting data to integers.
 - ❑ With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.
 - ❑ Each integer value is represented as a binary vector.
 - ❑ All the values are zero, and the index is marked with a 1.

Type		Type	AA_Onehot	AB_Onehot	CD_Onehot
AA	Onehot encoding →	AA	1	0	0
AB		AB	0	1	0
CD		CD	0	0	1

Source: <https://www.educative.io/blog/one-hot-encoding>

⁴ Ο CART στο sklearn δεν υποστηρίζει categorical learning



❖ Απόσταση σε ιεραρχίες

- ❑ Σε χαρακτηριστικά που οι τιμές τους ανήκουν σε μια **ιεραρχία (ταξινόμια)**, μπορούν να χρησιμοποιηθούν μετρικές σημασιολογικής απόστασης.
 - ✓ Για παράδειγμα, η απόσταση μεταξύ δύο τιμών στο ίδιο μονοπάτι θα μπορούσε να είναι τα βήματα της μεταξύ τους διαδρομής, ενώ αν βρίσκονται σε διαφορετικά μονοπάτια θα μπορούσε να είναι το άθροισμα των βημάτων της διαδρομής του καθενός ως τον κοινό τους γονέα.
 - ✓ Αν υπάρχει πολλαπλή κληρονομικότητα και υφίστανται περισσότεροι του ενός κοινοί γονείς τότε η απόσταση θα μπορούσε να είναι το ελάχιστο των αθροισμάτων μέχρι τους κοινούς γονείς

❖ Απόσταση κωδικοποίησης

- ❑ Σε (διανυσματικά) χαρακτηριστικά που η διαφορά σχετίζεται με την κωδικοποίηση των τιμών, θα μπορούσαμε να ορίζουμε τη διαφορά τους εξετάζοντας για παράδειγμα την **απόσταση Hamming (Hamming distance)**.
- ❑ Στη θεωρία πληροφορίας, ως απόσταση Hamming μεταξύ δύο συμβολοσειρών ίσου μήκους, ορίζεται ο αριθμός θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά.
 - ✓ Για παράδειγμα η απόσταση της λέξης "Χαμηλό" από την λέξη "Υψηλό" είναι 3 αφού διαφέρουν κατά 3 χαρακτήρες.
 - ✓ Μεταξύ του $x=[1,2,3,4]$ και $y=[1,2,5,7]$ είναι 2.

❑ Python:

`scipy.spatial.distance.hamming(array1, array2)`

{the function returns the percentage of corresponding elements that differ between the two arrays}

to obtain the Hamming distance we can simply multiply by the length of one of the arrays:

`scipy.spatial.distance.hamming(array1, array2) * len(array1)`

[How to Calculate Hamming Distance in Python \(With Examples\)](#)

```
from scipy.spatial.distance import hamming

#define arrays
x = [0, 1, 1, 1, 0, 1]
y = [0, 0, 1, 1, 0, 0]

#calculate Hamming distance between the two arrays
hamming(x, y) * len(x)

2.0
```




❖ Απόσταση Minkowski

- ❑ Οι σχέσεις της Ευκλείδειας απόστασης και της απόστασης Manhattan είναι ειδικές περιπτώσεις μιας γενικής μορφής απόστασης που ονομάζεται *απόσταση Minkowski*.
- ❑ Έστω ένα σύνολο δεδομένων D και δύο δεδομένα αυτού x, y , που περιγράφονται από m χαρακτηριστικά (x_1, x_2, \dots, x_m) και (y_1, y_2, \dots, y_m) , αντίστοιχα.
- ❑ Η απόσταση Minkowski δίνεται από τη σχέση: $d(x, y) = \sqrt[q]{\sum_i (x_i - y_i)^q}$
 - ✓ ταυτίζεται με τη Μανχάταν για $q=1$ και την Ευκλείδεια για $q=2$.
 - ✓ Επίσης για $q=1$ και δυαδικά δεδομένα ταυτίζεται με την απόσταση Hamming.
- ❑ Αν τα χαρακτηριστικά έχουν βάρη τότε η σχέση Minkowski γίνεται: $d(x, y) = \sqrt[q]{\sum_i w_i \cdot (x_i - y_i)^q}$

❖ Απόσταση Mahalanobis

- ❑ Μια πιο πολύπλοκη μετρική που λαμβάνει υπόψη την συνδιακύμανση⁵ μεταξύ των διαστάσεων.
- ❑ Is a measure of the distance between a point P and a distribution D .
 - ✓ It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D .
 - ✓ This distance is zero for P at the mean of D and grows as P moves away from the mean along each principal component axis.

⁵ <https://en.wikipedia.org/wiki/Covariance>. Υπάρχει και στο Γ' μέρος αυτού του κεφαλαίου



Δείκτες SMC και Jaccard

- ❑ Η χρήση χαρακτηριστικών με δυαδικές τιμές είναι συνηθισμένη στη μηχανική μάθηση καθώς είναι ο τρόπος κωδικοποίησης της "παρουσίας" (1) ή "απουσίας" (0) ενός στοιχείου από ένα σύνολο
 - ✓ Για παράδειγμα, έτσι μπορεί να κωδικοποιηθεί η ύπαρξη ή όχι μιας λέξης σε ένα κείμενο, ενός προϊόντος σε ένα καλάθι αγορών, η απάντηση σε μια ερώτηση true/false, κτλ.
- ❑ Για τη μέτρηση της ομοιότητας δύο δεδομένων A και B που περιγράφονται από n δυαδικά χαρακτηριστικά (δηλ. τιμή 0 ή 1), εξετάζουμε τα χαρακτηριστικά ίδιας θέσης στα A και B.
- ❑ Υπάρχουν τέσσερις περιπτώσεις που η καταμέτρησή τους οδηγεί στα ακόλουθα μεγέθη:
 - ✓ M_{11} : το πλήθος των χαρακτηριστικών στα οποία τα A και B έχουν τιμή 1.
 - ✓ M_{01} : το πλήθος των χαρακτηριστικών στα οποία το A έχει τιμή 0 και το B τιμή 1.
 - ✓ M_{10} : το πλήθος των χαρακτηριστικών στα οποία το A έχει τιμή 1 και το B τιμή 0.
 - ✓ M_{00} : το πλήθος των χαρακτηριστικών στα οποία τα A και B έχουν τιμή 0.
- ❑ Ένας απλός δείκτης που ορίζεται με βάση τις παραπάνω ποσότητες είναι ο **Συντελεστής Απλού Ταιριάσματος** (*simple matching coefficient – SMC*) που ορίζεται ως το ποσοστό των χαρακτηριστικών που "ταιριάζουν":

$$MC = \frac{M_{00} + M_{11}}{M_{01} + M_{10} + M_{00} + M_{11}}$$



❖ Δείκτες SMC και Jaccard (συνέχεια)

- ❑ Ο SMC είναι κατάλληλος για περιπτώσεις που ενδιαφέρει η συμφωνία τόσο στις "παρουσίες" όσο και στις "απουσίες", όπως σε ένα ερωτηματολόγιο true/false.
- ❑ Αυτό όμως δεν ισχύει σε κάποια προβλήματα,
 - ✓ όπως αυτά που σχετίζονται με καλάθια αγορών ή κείμενα ή προτιμήσεις π.χ. ταινιών, όπου τα δεδομένα είναι διανύσματα με πλήθος παραμέτρων ίσο με το μέγεθος ενός λεξικού ή το πλήθος προϊόντων ή των ταινιών ενός παρόχου, και μόνο μερικές (λίγες) από τις παραμέτρους αυτές έχουν τιμή 1.
 - ✓ Η εφαρμογή του SMC σε αυτές τις περιπτώσεις (με ασύμμετρα δυαδικά χαρακτηριστικά) θα δίνει πάντα τιμή κοντά στη μονάδα καθώς ο αριθμητής και ο παρονομαστής στην παραπάνω σχέση θα κυριαρχείται από την μεγάλη τιμή του M_{00} (αφού τα μηδενικά θα είναι πάντα πολύ περισσότερα της μονάδας).
- ❑ Ο δείκτης **Jaccard** (**Jaccard Index**) μετρά την ομοιότητα δεδομένων με δυαδικά χαρακτηριστικά δίνοντας έμφαση μόνο στο ταίριασμα των "παρουσιών" (1-1) και όχι των "απουσιών" (0-0):

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

- ❑ Δεδομένου ότι η τιμή ομοιότητας κυμαίνεται μεταξύ 0 και 1, η απόσταση d (**απόσταση Jaccard**) των A και B στο παραπάνω πλαίσιο θα ορίζεται ως: $d = 1 - J$
- ❑ The Jaccard similarity⁶ index measures the similarity between two sets of data.
 - ✓ Jaccard Similarity = (number of observations in both sets) / (number in either set)
 - ✓ $J(A, B) = |A \cap B| / |A \cup B|$
 - ✓ [How to Calculate Jaccard Similarity in Python](#)

⁶ Η απόσταση Hamming υπολογίζει τη διαφορά μεταξύ 2 διανυσμάτων ενώ η Jaccard την ομοιότητα (απόσταση) τους.



❖ Ομοιότητα συνημιτόνου

- ❑ Η **ομοιότητα συνημίτονου** (*cosine similarity*) είναι ένα μέτρο ομοιότητας δυο μη μηδενικών διανυσμάτων \mathbf{A} (A_1, A_2, \dots, A_k) και \mathbf{B} (B_1, B_2, \dots, B_k) που υπολογίζει το συνημίτονο της μεταξύ τους γωνίας:

$$\text{similarity} = \cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{k=1}^n A_k \cdot B_k}{\sqrt{\sum_{k=1}^n A_k^2} \cdot \sqrt{\sum_{k=1}^n B_k^2}}$$

- ❑ όπου $\mathbf{A} \cdot \mathbf{B}$ το εσωτερικό γινόμενο των διανυσμάτων και $\|\mathbf{A}\|$ και $\|\mathbf{B}\|$ τα μήκη τους
- ❑ Άρα το μέγεθος αυτό εξετάζει τον προσανατολισμό των διανυσμάτων και όχι το μήκος τους.
 - ✓ Αν τα διανύσματα έχουν τον ίδιο προσανατολισμό η παραπάνω σχέση δίνει 1,
 - ✓ αν είναι κάθετα μεταξύ τους δίνει 0 (το οποίο θεωρείται ως μη-συσχέτιση) και
 - ✓ αν είναι αντίθετα μεταξύ τους δίνει -1.
- ❑ Προφανώς ενδιάμεσες τιμές μεταφράζονται ως ενδιάμεση ομοιότητα (ή ανομοιότητα) ενώ αν τα διανύσματα είναι κανονικοποιημένα (διαιρεμένα με το μήκος τους) τα $\|\mathbf{A}\|$ και $\|\mathbf{B}\|$ είναι πάντα μονάδα οπότε αρκεί το εσωτερικό γινόμενο για τον παραπάνω υπολογισμό.
- ❑ Εφαρμογές: Σύγκριση κειμένων διαφορετικού μεγέθους, σύγκριση καταναλωτικής συμπεριφοράς καταναλωτών (πόσα προϊόντα αγοράζουν ή χρήματα που ξοδεύουν για κάθε κατηγορία), σύγκριση συνδρομητών π.χ. του Netflix, δηλ. πόσες ταινίες έχουν δει από κάθε κατηγορία (ποσοστό), κλπ.

❖ Γενικά το θέμα της απόστασης των δεδομένων είναι σύνθετο και απαιτεί διεξοδική μελέτη γιατί εξαρτάται ισχυρά από το είδος των δεδομένων και τη φύση του προβλήματος



❖ Εφαρμογή σε κείμενα

- ❑ Η ομοιότητα συνημίτονου δεν έχει περιορισμό στις τιμές που έχουν οι συνιστώσες των διανυσμάτων, δηλαδή αν παίρνουν δυαδικές τιμές ή όχι.
- ❑ Έτσι, η χρήση διανυσμάτων είναι συνήθης για την κωδικοποίηση κειμένων στην **ανάκτηση πληροφορίας** (*information retrieval*) και στην **ανακάλυψη γνώσης από κείμενα** (*text mining*).
- ❑ Ειδικότερα, τα κείμενα μπορούν να αναπαρασταθούν με δυαδικά διανύσματα, όπου κάθε συνιστώσα αντιστοιχεί σε μια λέξη ενός λεξιλογίου
 - ✓ τιμή 1 υπονοεί την παρουσία της λέξης στο κείμενο ενώ η τιμή 0 την απουσία της.
- ❑ Εναλλακτικά, οι τιμές μπορούν να είναι πραγματικοί αριθμοί και να αφορούν σε συχνότητα εμφάνισης των λέξεων (**term frequency - tf**),
 - ✓ Δηλαδή το πλήθος εμφανίσεων μιας λέξης δια το πλήθος του συνόλου των λέξεων.
 - ✓ Αυτή η περιγραφή είναι γνωστή και ως **bag-of-words model** και η συχνότητα εμφάνισης μια λέξης μας λέει πόσο σημαντική είναι στο συγκεκριμένο κείμενο.
- ❑ Βέβαια, μόνο του το **tf** δεν αρκεί για να αναδείξει σημαντικές λέξεις σε ένα έγγραφο.
 - ✓ Μια λέξη είναι σημαντική για ένα έγγραφο αν εμφανίζεται συχνά σε αυτό αλλά όχι συχνά στα άλλα
- ❑ Κάτι τέτοιο υπολογίζει το γινόμενο **tf-idf**, όπου το **idf** (**inverse document frequency**) είναι η αντίστροφη συχνότητα εμφάνισης του όρου στα έγγραφα και για έναν όρο (λέξη) w υπολογίζεται από τη σχέση:

$$idf = \log \left(\frac{\text{πλήθος εγγράφων}}{\text{πλήθος εγγράφων με τον όρο } w} \right)$$



- ❑ Το γινόμενο ***tf·idf*** χρησιμοποιείται πολύ συχνά στις διανυσματικές αναπαραστάσεις εγγράφων.
- ❑ Με την ομοιότητα συνημίτονου, μπορεί να ποσοτικοποιηθεί η ομοιότητα (άρα και η απόσταση) μεταξύ κειμένων, ανεξάρτητα από το μέγεθός τους.
 - ✓ Καθώς οι τιμές αυτές δεν μπορεί να είναι αρνητικές, η ομοιότητα θα κυμαίνεται μεταξύ 0 και 1 (και η γωνία μεταξύ 0° και 90°).
- ❑ Σε τέτοια προβλήματα, το πλήθος των διαστάσεων είναι εξαιρετικά μεγάλο (χιλιάδες ή δεκάδες χιλιάδες) καθώς συνήθως ισούται με το πλήθος των λέξεων (ή των ριζών λέξεων) ενός λεξικού.
 - ✓ Αυτό δημιουργεί προβλήματα απόδοσης και γι' αυτό έχουν προταθεί διάφορες τεχνικές μείωσης των διαστάσεων.

❖ Παράδειγμα

- ❑ Έστω δυο κείμενα που περιγράφονται με λεξιλόγιο επτά (7) λέξεων.
 - ✓ Άρα κάθε κείμενο αναπαρίσταται με ένα διάνυσμα στο χώρο των επτά (7) διαστάσεων.
- ❑ Έστω $A=(3, 2, 1, 0, 0, 2, 0)$ και $B=(2, 0, 1, 3, 0, 1, 1)$ τα διανύσματα με τις συχνότητες εμφάνισης των 7 όρων (λέξεων) στα 2 έγγραφα.
- ❑ Θέλουμε να υπολογίσουμε την ομοιότητα των κειμένων.

$$A \cdot B = 3 \cdot 2 + 2 \cdot 0 + 1 \cdot 1 + 0 \cdot 3 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 1 = 9$$

$$\|A\| = \sqrt{3^2 + 2^2 + 1^2 + 0^2 + 0^2 + 2^2 + 0^2} = \sqrt{18} = 4.24$$

$$\|B\| = \sqrt{2^2 + 0^2 + 1^2 + 3^2 + 0^2 + 1^2 + 1^2} = \sqrt{16} = 4$$

$$\cos(A, B) = 9/(4.24 \cdot 4) = 0.53$$

- ❑ [comparison and contrasts of some similarity and distance measures](#)



Κανονικοποίηση (Normalization)⁷

- Η κανονικοποίηση (normalization) είναι μια διαδικασία μετασχηματισμού δεδομένων, κατά την οποία αριθμητικές τιμές αντικαθίστανται με άλλες, πιο «κατάλληλες».
- Η κανονικοποίηση των δεδομένων γίνεται ώστε να αντιμετωπιστούν δυσκολίες ορισμένων μεθόδων μάθησης.
- Για παράδειγμα, τα Νευρωνικά Δίκτυα λειτουργούν καλύτερα όταν οι τιμές εισόδου κυμαίνονται στην περιοχή [0..1].
- Επίσης, η μέθοδος των k-Πλησιέστερων Γειτόνων, η οποία υπολογίζει αποστάσεις μεταξύ των παρατηρήσεων, αντιμετωπίζει πρόβλημα όταν ορισμένες μεταβλητές εισόδου έχουν μικρές τιμές, ενώ άλλες έχουν μεγάλες τιμές.
- Το πρόβλημα συνίσταται στο γεγονός ότι οι μεταβλητές με τις μεγάλες τιμές καθορίζουν ουσιαστικά την απόσταση των παρατηρήσεων, ενώ οι μεταβλητές με τις μικρές τιμές επηρεάζουν την απόσταση ελάχιστα και τελικά, δεν παίζουν κανένα ρόλο στον υπολογισμό του αποτελέσματος.

Υπάρχουν διάφορες μέθοδοι κανονικοποίησης των αριθμητικών τιμών:

Κανονικοποίηση ελάχιστου-μέγιστου.

- Οι αριθμητικές τιμές αντιστοιχίζονται με άλλες, οι οποίες κυμαίνονται εντός μιας προκαθορισμένης περιοχής τιμών.
- Η αντιστοίχιση γίνεται με γραμμικό μετασχηματισμό.
- Αν θεωρήσουμε μια μεταβλητή A, όπου η μεγαλύτερη τιμή της είναι η max_A και η μικρότερη τιμή της είναι η min_A , μπορούμε να αντιστοιχίσουμε όλες τις τιμές με άλλες που κυμαίνονται εντός μιας περιοχής με κατώτερο όριο την new_min_A και ανώτερο όριο την new_max_A

$$new_max_A \text{ σύμφωνα με τη σχέση: } x' = \frac{x - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

όπου x η εκάστοτε τιμή της μεταβλητής A και x' η νέα τιμή.

- Η μέθοδος αυτή έχει το πλεονέκτημα ότι ο χρήστης καθορίζει την περιοχή τιμών.
- Αν ορίσουμε σαν $new_min_A = 0$ και $new_max_A = 1$ η σχέση γίνεται: $x' = \frac{x - min_A}{max_A - min_A}$ και το x' παίρνει τιμές στο διάστημα [0..1].
- Με τη μέθοδο αυτή διατηρείται η αναλογία μεταξύ των τιμών που υπήρχε στα αρχικά δεδομένα.

⁷ Πηγή: Βιβλίο Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων, Ε. Κύρκος, (ISBN: 978-960-603-109-0) Ελληνικά Ακαδημαϊκά Συγγράμματα (www.kallipos.gr)



Κανονικοποίηση δεκαδικής κλιμάκωσης.

- Η μέθοδος αυτή πραγματοποιεί υποδεκαπλασιασμό των τιμών, διαιρώντας τες με μια δύναμη του 10.
- Η δύναμη του 10 υπολογίζεται με τέτοιο τρόπο ώστε η απόλυτη τιμή του νέου μέγιστου να είναι μικρότερη από 1.
- Ο μετασχηματισμός γίνεται σύμφωνα με τη Σχέση: $x' = \frac{x}{10^k}$

Κανονικοποίηση z-score.

- Η μέθοδος πραγματοποιεί μετασχηματισμό των αριθμητικών τιμών, χρησιμοποιώντας τη μέση τιμή και την τυπική απόκλιση τους.
- Για μία μεταβλητή A, με μέση τιμή μ_A και τυπική απόκλιση σ_A , ο μετασχηματισμός των τιμών γίνεται σύμφωνα με τη Σχέση:

$$x' = \frac{x - \mu_A}{\sigma_A}$$

όπου x η εκάστοτε τιμή της μεταβλητής A και x' η νέα τιμή.

- Η μέθοδος αυτή είναι ιδιαίτερα κατάλληλη σε περιπτώσεις όπου τα δεδομένα περιέχουν ακραίες τιμές, γιατί η κανονικοποίηση ελάχιστου-μέγιστου θα συγκέντρωνε τη μεγάλη πλειοψηφία των τιμών σε ένα ελάχιστο τμήμα της περιοχής τιμών και θα χρησιμοποιούσε το υπόλοιπο τμήμα της περιοχής για τις εξαιρέσεις.
- Επίσης, η μέθοδος δίνει τιμές των οποίων η μέση τιμή ισούται με 0.

Τυπική Απόκλιση και Διακύμανση (ή Διασπορά)

- Στη στατιστική, η τυπική απόκλιση (σ) είναι ένα μέτρο που χρησιμοποιείται για να υπολογιστεί το ποσό της μεταβολής ή της διασποράς ενός συνόλου τιμών δεδομένων.
- Η τυπική απόκλιση μιας τυχαίας μεταβλητής, ενός στατιστικού πληθυσμού, ενός συνόλου δεδομένων ή της κατανομής πιθανότητας είναι η τετραγωνική ρίζα της διακύμανσης της (ή αλλιώς διασποράς).
- Για ένα πεπερασμένο σύνολο αριθμών, η απόκλιση βρίσκεται λαμβάνοντας την τετραγωνική ρίζα του μέσου όρου των τετραγώνων των αποκλίσεων των τιμών από τη μέση τιμή τους.



- Δηλαδή βρίσκουμε τη μέση τιμή (μέσο όρο) του συνόλου των αριθμών και στη συνέχεια υπολογίζονται τα τετράγωνα των αποκλίσεων του κάθε στοιχείου από τη μέση τιμή:
- **Η διακύμανση είναι ο μέσος των τιμών αυτών και η τυπική απόκλιση είναι η τετραγωνική ρίζα της διακύμανσης.**
- Εκτός από την έκφραση της μεταβλητότητας του πληθυσμού, η τυπική απόκλιση συνήθως χρησιμοποιείται για τη μέτρηση της εμπιστοσύνης στα στατιστικά συμπεράσματα.

...the end

Questions?

