Επιστήμη Δεδομένων

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Ενότητες Παρουσίασης

Εισαγωγή

Η εξέλιξη της τεχνολογίας έδωσε τη δυνατότητα εξάπλωσης του Internet.

Με το πέρασμα του χρόνου, η πρόσβαση στο Internet έγινε προσιτή σε ολοένα και περισσότερους ανθρώπους.

Αυτό με τη σειρά του οδήγησε στο να αναπτυχθούν περισσότεροι ιστότοποι και να χρησιμοποιηθούν βάσεις δεδομένων για την αποθήκευση των δεδομένων.

Με τη δημιουργία εμπορικών και κοινωνικών ιστοσελίδων υπήρξαν τα πρώτα άλματα στις απαιτήσεις και ανάγκες για αποθήκευση και διαχείριση μεγάλου όγκου δεδομένων.

Εισαγωγή

Σήμερα, το πλήθος των διαθέσιμων δεδομένων είναι τεράστιο και αυξάνεται εκθετικά κάθε μέρα.

Η μείωση στο κόστος συλλογής και της δυσκολίας στη συλλογή και αποθήκευση των δεδομένων συνετέλεσε σημαντικά στην ανάπτυξη του πεδίου αυτού.

Ο τεράστιος όγκος δεδομένων, που συσσωρεύεται στις βάσεις δεδομένων και στις αποθήκες δεδομένων (data warehouses), δεν μπορεί να αξιοποιηθεί όπως είναι.

Πρέπει αρχικά να γίνουν κάποιες ενέργειες για να δομηθούν κατάλληλα τα δεδομένα, ώστε στη συνέχεια να μπορούν να αξιοποιηθούν.

Στο μάθημα αυτό θα δούμε ποια είναι τα θεμελιώδη στάδια, προκειμένου να μπορεί να εξαχθεί από τα δεδομένα χρήσιμη και αξιοποιήσιμη πληροφορία.

Ενότητες Παρουσίασης

Η Επιστήμη των Δεδομένων

Η Επιστήμη των Δεδομένων (Data Science), είναι ένας καινούριος όρος, ο οποίος ήρθε να αντικαταστήσει προγενέστερους όρους, όπως **Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Data-base)** ή **Εξόρυξη Δεδομένων (Data Mining)**.

Οι τρεις αυτοί όροι χρησιμοποιούνται σχεδόν εναλλακτικά, για να περιγράψουν μία **ημι-αυτοματοποιημένη διαδικασία**, σκοπός της οποίας είναι **να αναλύσει έναν μεγάλο όγκο δεδομένων** που αφορούν ένα συγκεκριμένο πρόβλημα, συνήθως εμπορικού ή επιστημονικού ενδιαφέροντος, για **την παραγωγή προτύπων (patterns)**, όπως συνηθίζεται να λέμε σε ορισμένους τομείς, όπως

1. στη Στατιστική,
2. στη Μηχανική Μάθηση (Machine Learning) και
3. στην Αναγνώριση Προτύπων (Pattern Recognition).

Η Επιστήμη των Δεδομένων

Τα πρότυπα αυτά, τα οποία συναντάμε σε διάφορες μορφές, όπως **συσχετίσεις**, **ανωμαλίες**, **συστάδες**, **κλάσεις** κ.λπ., αποτελούν δομές ή περιστατικά, που εμφανίζονται στα δεδομένα και έχουν κάποια ιδιαίτερη σημασία από στατιστικής πλευράς.

Ένα από τα σημαντικότερα θέματα από πλευράς ουσίας στην Επιστήμη των Δεδομένων αφορά τόσο την εύρεση ή αλλιώς ανακάλυψη (η λέξη ανακάλυψη εμπεριέχει σε μεγάλο βαθμό το γεγονός ότι τα πρότυπα αυτά δεν ήταν αναμενόμενα εκ των προτέρων), **όσο και τον χαρακτηρισμό αυτών των προτύπων.**

Πιο συγκεκριμένα, ένα πρότυπο αναφέρεται και ως μία διευθέτηση ή αλλιώς διάταξη, στην οποία θεωρείται ότι υπάρχει κάποιου είδους οργάνωση της υποκείμενης δομής.

Η Επιστήμη των Δεδομένων

Τα πρότυπα αυτά, ανιχνεύονται κατά κύριο λόγο με τη χρήση μετρήσιμων χαρακτηριστικών γνωρισμάτων ή ιδιοτήτων, τα οποία έχουν εξαχθεί από τα δεδομένα.

Η Επιστήμη των Δεδομένων είναι ουσιαστικά μία καινούρια επιστήμη, η οποία άρχισε να εμφανίζεται σταδιακά, ξεκινώντας από τα τέλη της δεκαετίας του 1980.

Την εποχή εκείνη μεσουρανούσαν τα σχεσιακά συστήματα βάσεων δεδομένων, τα οποία εξυπηρετούσαν τις αποθηκευτικού τύπου ανάγκες των δεδομένων για τις επιχειρήσεις και τους οργανισμούς-με στόχο την καλύτερη οργάνωση και διαχείρισή τους, έτσι ώστε να ικανοποιούνται ταχύτερα μαζικά ερωτήματα, που αφορούσαν την καθημερινή τους λειτουργία.

Η Επιστήμη των Δεδομένων

Τέτοιου είδους **Συστήματα Διαχείρισης Βάσεων Δεδομένων (ΣΔΒΔ)** ακολουθούσαν το λεγόμενο **OLTP (OnLine Transaction Processing)** μοντέλο, το οποίο αφορά στην ηλεκτρονική διεκπεραίωση συναλλαγών ή αλλιώς δοσοληψιών.

Τα εργαλεία αυτά επέτρεπαν ουσιαστικά στον χρήστη να βρει απαντήσεις (επιβεβαίωση) σε ερωτήματα (ειδικού τύπου), τα οποία ήδη ο χρήστης γνώριζε ή να δημιουργήσει κάποιες αναφορές.

Η ανάγκη για την καλύτερη εκμετάλλευση των δεδομένων που είχαν δημιουργηθεί μέσω των συστημάτων αυτών – συστήματα τα οποία κάλυπταν καθημερινά τη λειτουργία μιας εταιρίας – οδήγησε στην ανάπτυξη εργαλείων τύπου **OLAP (OnLine Analytical Processing)**.

Η Επιστήμη των Δεδομένων

Με αυτά εργαλεία τύπου **OLAP (OnLine Analytical Processing)** μπορούσαν να απαντηθούν ερωτήματα και αναφορές πιο προηγμένης μορφής, που επέτρεπαν μεγάλες και πολυδιάστατες Βάσεις Δεδομένων να ερωτηθούν με μεγάλη ταχύτητα, ενώ παράλληλα προσέφεραν και οπτικοποίηση των αποτελεσμάτων.

Τα εργαλεία τύπου OLAP θα μπορούσαν να χαρακτηριστούν και ως εργαλεία εξερεύνησης δεδομένων, οδηγούμενα από την οπτικοποίηση των αποτελεσμάτων.

Η Επιστήμη των Δεδομένων

Αυτού του τύπου τα εργαλεία επέτρεπαν στους χρήστες (διευθυντές πωλήσεων, προϊσταμένους τμημάτων προώθησης προϊόντων κ.λπ.) να βρίσκουν και να ανακαλύπτουν καινούρια πρότυπα, με τη διαφορά ότι η ανακάλυψη αυτή γινόταν από τον χρήστη.

Για παράδειγμα, ένας χρήστης θα μπορούσε να υποβάλει ερωτήματα για το ύψος των πωλήσεων των καταστημάτων μιας αλυσίδας καταστημάτων σε μία πόλη, έτσι ώστε να βρει τα καταστήματα με τις χαμηλότερες πωλήσεις.

Η αυτοματοποίηση της διαδικασίας ανακάλυψης προτύπων ή αλλιώς γνώσης έλαβε χώρα μέσω τεχνικών μεθοδολογιών και εργαλείων, τα οποία αναπτύχθηκαν στο πλαίσιο της Επιστήμης των Δεδομένων.

Μέσω αυτών των λύσεων η ανακάλυψη νέων προτύπων ήταν οδηγούμενη (ή ίσως καλύτερα υποβοηθούμενη) από τον τελικό στόχο.

Η Επιστήμη των Δεδομένων

Για παράδειγμα, αντί ο χρήστης να ζητάει μία αναφορά για να δει ποιο είναι το κατάστημα με τις χαμηλότερες πωλήσεις τον προηγούμενο μήνα, θα μπορούσε να ζητάει από το σύστημα ενδιαφέροντα, από στατιστικής πλευράς, γεγονότα που αφορούν γενικότερα στις πωλήσεις των καταστημάτων.

Η μεγάλη άνθιση στον τομέα της Επιστήμης των Δεδομένων έλαβε χώρα σταδιακά και ήταν άμεσα εξαρτημένη από τη δυνατότητα που δόθηκε για συλλογή και καταγραφή τεράστιων ποσοτήτων δεδομένων, διαφορετικών μορφών και τύπων, μέσω της ανάπτυξης γρήγορων δικτυακών υποδομών, πάνω στις οποίες μπορούσαν να υποστηριχτούν αξιόπιστες εμπορικές εφαρμογές.

Η Επιστήμη των Δεδομένων

Στην κατηγορία των εταιριών που πρωτοστάτησαν σε αυτή τη νέα τάξη πραγμάτων ήταν και η Amazon, η οποία ξεκίνησε από την ηλεκτρονική πώληση βιβλίων και άλλων στοιχείων, δημιουργώντας στη συνέχεια ένα πολύ φιλικό προς τον χρήστη σύστημα συστάσεων.

Το σύστημα αυτό χτιζόταν και ρυθμιζόταν με βάση τις αλληλεπιδράσεις των πελατών της, χρησιμοποιώντας μία τεχνική γνωστή ως Συνεργατικό Φιλτράρισμα (Collaborative Filtering).

Το σύστημα αυτό αποτέλεσε τη βάση των Συστημάτων Συστάσεων (Recommender Systems).

Η Επιστήμη των Δεδομένων

Η άνευ όρων παραγωγή δεδομένων σε εικοσιτετράωρη βάση καλύπτει μια τεράστια γκάμα ανθρώπινων δραστηριοτήτων και όχι μόνο, όπως είναι τα δεδομένα από το καλάθι αγορών, τον ιατρικό φάκελο του ασθενούς, τις συζητήσεις ή και ανακοινώσεις στα κοινωνικά μέσα δικτύωσης, τις τραπεζικές ή και χρηματιστηριακές συναλλαγές, τα ίχνη κινούμενων οχημάτων, τα δεδομένα αισθητήρων από κινητήρες αεροσκαφών, η καταγραφή συνομιλιών σε κέντρα εξυπηρέτησης πελατών κ.λπ.

Τα δεδομένα αυτά διαφέρουν πάρα πολύ μεταξύ τους τόσο σε μορφή (εικόνα, βίντεο, κείμενο, πολυδιάστατα ή πραγματικού χρόνου δεδομένα, ακολουθίες DNA και άλλα πολλά) όσο και στην ταχύτητα συλλογής.

Η Επιστήμη των Δεδομένων

Εάν, μάλιστα, δεν υποστούν άμεση ανάλυση, ίσως να είναι ιδιαίτερα δύσκολο να αποθηκευτούν ή να τα επεξεργαστούν οι άνθρωποι, δημιουργώντας έτσι μία καινούρια ερευνητική δράση, γνωστή με τον όρο Μεγάλα Δεδομένα (Big Data).

Η Επιστήμη των Δεδομένων στοχεύει σε αυτή τη φάση να καλύψει τις ανάγκες που δημιουργούνται από αυτόν τον νέο τομέα και να προσφέρει λύσεις για την κλιμακούμενη και αποτελεσματική επεξεργασία out-of core (εκτός μνήμης ή εξωτερικής μνήμης) δεδομένων.

Η Επιστήμη των Δεδομένων

Η πρόβλεψη αφορά στη χρήση κάποιων μεταβλητών ή πεδίων μίας βάση δεδομένων, μέσω των τιμών των οποίων μπορεί να εκτιμηθεί η άγνωστη ή μελλοντική τιμή ενός άλλου γνωρίσματος.

Η περιγραφή (σε μορφή σύνοψης ή περιληπτικής παρουσίασης) των δεδομένων εστιάζει στην εύρεση κατανοητών από τον άνθρωπο προτύπων, τα οποία περιγράφουν τα δεδομένα, όπως, δηλαδή, γίνεται κατά την εύρεση συστάδων ή ομάδων αντικειμένων με παρόμοια χαρακτηριστικά.

Ενότητες Παρουσίασης

Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Η Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (ΑΓΒΔ) αποτελείται από συγκεκριμένα στάδια.

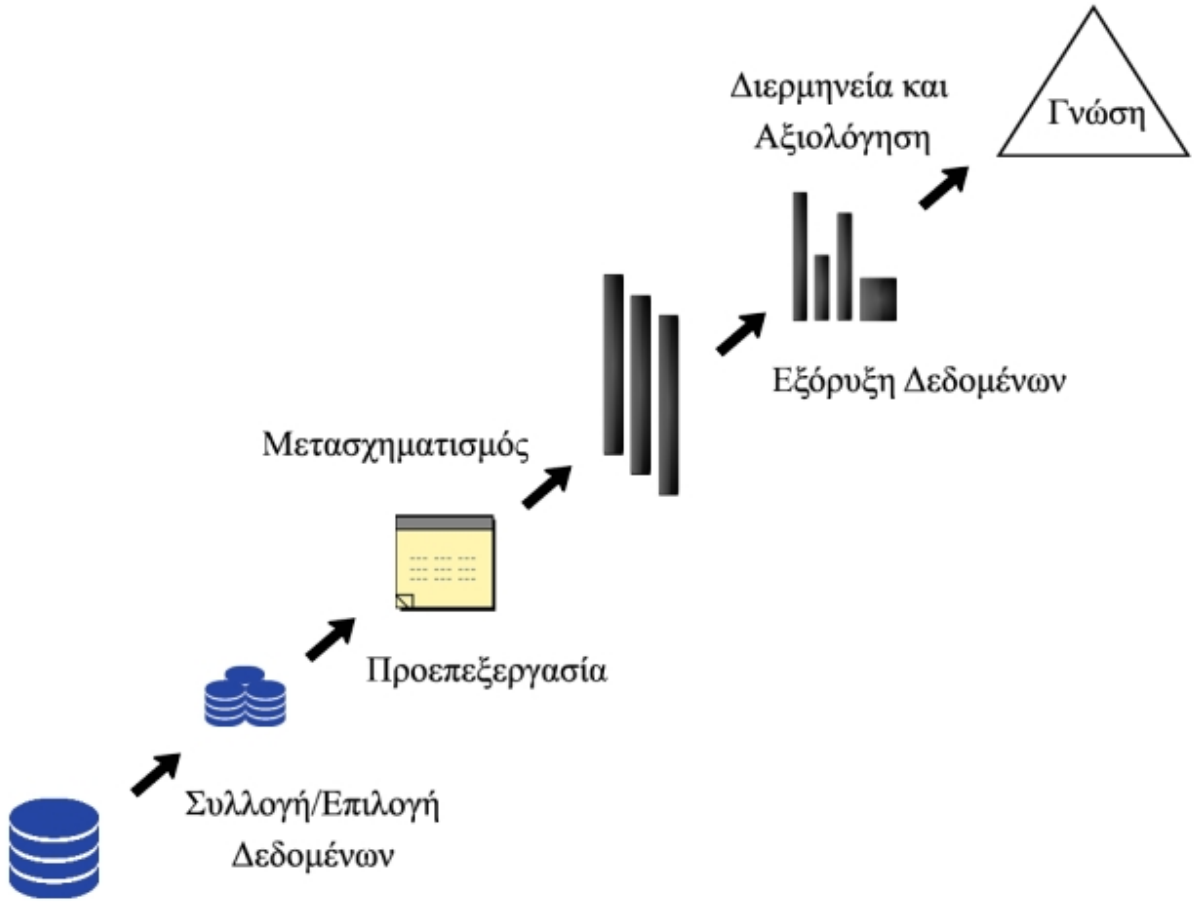
Πρόκειται για την αποκάλυψη ή παραγωγή λειτουργικής γνώσης μέσα από την ανάλυση των δεδομένων.

Αναφέρεται σε ολόκληρη τη διαδικασία, από τη συλλογή δεδομένων μέχρι την αξιοποίηση των αποτελεσμάτων σε πιο πρακτικό επίπεδο.

Τα βασικά στάδια της ΑΓΒΔ (Εικόνα 1.1) είναι τα ακόλουθα:

1. Συλλογή Δεδομένων (Data Collection)
2. Προεπεξεργασία Δεδομένων (Preprocessing)
3. Μετασχηματισμός Δεδομένων (Transformation)
4. Εξόρυξη Δεδομένων (Data Mining)
5. Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation)

Ανακάλυψη Γνώσης από Βάσεις Δεδομένων



Σχήμα: 1

Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

Ο όρος Ανακάλυψη Γνώσης από Βάσεις Δεδομένων συχνά ταυτίζεται με τον όρο Εξόρυξη Δεδομένων, που στην πραγματικότητα αποτελεί ένα μόνο επιμέρους βήμα της.

Ο βασικός στόχος της Εξόρυξης Δεδομένων (ΕΔ) είναι η εξαγωγή μη τετριμμένης, προηγούμενα άγνωστης και πιθανά χρήσιμης πληροφορίας ή προτύπων από το σύνολο των δεδομένων.

Ο όρος χρησιμοποιείται μάλλον καταχρηστικά αφού στην ουσία δεν γίνεται καμία εξαγωγή δεδομένων.

Αντίθετα, χρησιμοποιούνται τα προεπεξεργασμένα και (ενδεχομένως) μετασχηματισμένα δεδομένα για την εξαγωγή χρήσιμης πληροφορίας, η οποία αξιοποιείται για την απόκτηση γνώσης σε σχέση με κάποιο πρόβλημα. Ακολουθεί σύντομη περιγραφή για το κάθε στάδιο της ΑΓΒΔ.

Συλλογή Δεδομένων

Το πρώτο βήμα της ΑΓΒΔ είναι η συλλογή και η αποθήκευση των δεδομένων.

Η συλλογή των δεδομένων συνήθως γίνεται είτε αυτόματα, π.χ. με χρήση αισθητήρων, είτε μη αυτόματα, π.χ. με χρήση ερωτηματολογίων.

Δυσλειτουργία στους αισθητήρες ή αδυναμία απάντησης κάποιας ερώτησης στα ερωτηματολόγια μπορεί να οδηγήσει σε θορυβώδη ή ελλιπή δεδομένα.

Τα συγκεκριμένα προβλήματα, που ενδεχομένως να προκύψουν κατά τη συλλογή δεδομένων, αναλαμβάνει να τα αντιμετωπίσει το επόμενο στάδιο.

Προεπεξεργασία Δεδομένων

Το δεύτερο και πιο σημαντικό στάδιο της ΑΓΒΔ είναι η προεπεξεργασία του συνόλου δεδομένων, η οποία γίνεται με στόχο τον καθαρισμό τους, δηλαδή την τακτοποίηση εσφαλμένων, προβληματικών ή ελλιπόντων δεδομένων.

Μπορεί να απαιτήσει έως και το 60% της συνολικής προσπάθειας και αυτό διότι, αν τα δεδομένα δεν είναι «καθαρά» και στην κατάλληλη μορφή, δεν έχει νόημα να μιλάμε για ποιότητα αποτελεσμάτων.

Θα εξετάσουμε αναλυτικά στο μάθημα από ποιες διεργασίες συνίσταται η προεπεξεργασία των δεδομένων και πότε χρησιμοποιείται η κάθε μια.

Μετασχηματισμός Δεδομένων

Ο μετασχηματισμός των δεδομένων αποτελεί το τρίτο στάδιο της ΑΓΒΔ.

Ουσιαστικά, πρόκειται για τη μετατροπή των δεδομένων κάτω από ένα κοινό πλαίσιο, για επεξεργασία.

Χρησιμοποιείται κυρίως για την εξομάλυνση των δεδομένων και απομάκρυνση θορύβου, για τη συνάθροιση των δεδομένων, δηλαδή για την παραγωγή σύνοψης τους, για την κανονικοποίηση τους, δηλαδή την κλιμάκωση των χαρακτηριστικών του συνόλου δεδομένων σε ένα συγκεκριμένο και περιορισμένο εύρος τιμών, ή τέλος για τη δημιουργία νέων χαρακτηριστικών από τα ήδη υπάρχοντα.

Ειδικές μορφές μετασχηματισμού αποτελούν η διακριτοποίηση και η συμπίεση.

Εξόρυξη Δεδομένων

Σε αυτό το στάδιο της ΑΓΒΔ εφαρμόζεται κάποιος αλγόριθμος για την παραγωγή ενός μοντέλου.

Έχοντας καθαρίσει και μετασχηματίσει τα δεδομένα, είναι έτοιμα να χρησιμοποιηθούν από κάποιον αλγόριθμο, ώστε να δημιουργηθεί κάποιο μοντέλο, συνήθως κατηγοριοποίησης ή πρόβλεψης.

Θέλουμε να χρησιμοποιήσουμε το μοντέλο αυτό, το οποίο δημιουργήθηκε με βάση κάποια γνωστά δεδομένα, έτσι ώστε να μπορεί να μας δώσει απάντηση για την τιμή ενός χαρακτηριστικού-μεταβλητής στόχου για νέα, άγνωστα δεδομένα.

Διερμηνεία και Αξιολόγηση

Στο τελευταίο στάδιο της ΑΓΒΔ γίνεται η διερμηνεία και η αξιολόγηση των αποτελεσμάτων (όχι του μοντέλου) που παρήχθησαν από την όλη διαδικασία.

Ενότητες Παρουσίασης

Τύποι Μοντέλων

Τα μοντέλα που παράγονται από το στάδιο της Εξόρυξης Δεδομένων διακρίνονται σε δυο βασικούς τύπους:

1. τα μοντέλα πρόβλεψης (predictive) και
2. τα περιγραφικά μοντέλα (descriptive).

Στόχος ενός μοντέλου πρόβλεψης είναι να προβλέψει τιμές για ένα συγκεκριμένο χαρακτηριστικό που παρουσιάζει ενδιαφέρον και που πιθανώς βασίζεται στη συμπεριφορά άλλων χαρακτηριστικών.

Για παράδειγμα, η πρόβλεψη μπορεί να βασίζεται στη χρονολογική κατάταξη των δεδομένων.

Ένα περιγραφικό μοντέλο βρίσκει πρότυπα (patterns) ή σχέσεις (relations) που ενυπάρχουν στα δεδομένα και μελετά τις ιδιότητες τους, ώστε να δοθεί μια αιτιολόγηση της συμπεριφοράς τους.

Παραδείγματα και Αντιπαραδείγματα

Αρκετά άτομα δυσκολεύονται να κατανοήσουν και να διαχωρίσουν τις έννοιες ΑΓΒΔ και ΕΔ. Για τον λόγο αυτό κρίνεται σκόπιμο να δοθούν μερικά πρακτικά παραδείγματα και αντιπαραδείγματα, ώστε να ξεκαθαριστεί το τι είναι ΕΔ και τι όχι.

Μερικά παραδείγματα της Εξόρυξης Δεδομένων είναι τα ακόλουθα:

Ο Bill Clinton δήλωσε πως μετά το τρομοκρατικό χτύπημα στις 11 Σεπτεμβρίου 2011, έπειτα από εξέταση πολλών βάσεων δεδομένων, πράκτορες του FBI ανακάλυψαν ότι 5 από τους αυτουργούς υπήρχαν καταχωρημένοι σε αυτές. Ένας από αυτούς κατείχε 30 πιστωτικές κάρτες με χρεωστικό υπόλοιπο της κατηγορία των $250,000 ενώ βρισκόταν στη χώρα για λιγότερο από δυο χρόνια.

Παραδείγματα και Αντιπαραδείγματα

Εταιρίες τηλεπικοινωνιών επιβραβεύουν όχι μόνο όσους ξοδεύουν πολλά, αλλά και κάποιους από τους οικονομικότερους συνδρομητές.

Οι λεγόμενοι «καθοδηγητές» συχνά πείθουν φίλους, συγγενικά πρόσωπα ή συνεργάτες να τους «ακολουθήσουν», όταν αλλάζουν πάροχο.

Επομένως,στόχος είναι να βρεθούν αυτοί οι πελάτες και να μείνουν, προσφέροντας τους δελεαστικές εκπτώσεις και πακέτα προσφορών.

Παραδείγματα και Αντιπαραδείγματα

Χρησιμοποιώντας τις καταγεγραμμένες θερμοκρασίες κατά τη θερινή σεζόν των προηγούμενων 15 ετών, γίνεται προσπάθεια πρόβλεψης του πώς θα κυμανθούν οι θερμοκρασίες το καλοκαίρι των επόμενων χρόνων.

Σκευτείτε κάποια παραδείγματα από την Λογιστική & Χρηματοοικονομική

Τύποι Μοντέλων

Εξόρυξη δεδομένων δεν είναι η απλή επεξεργασία ερωτημάτων, ούτε τα έμπειρα συστήματα ή τα μικρής κλίμακας στατιστικά προγράμματα.

Μερικά αντιπαραδείγματα είναι τα ακόλουθα:

1. Εύρεση αριθμού τηλεφώνου στον τηλεφωνικό κατάλογο.
2. Εύρεση πληροφοριών για τον Αμαζόνιο στο Internet.
3. Μέσος όρος βαθμολογίας μαθημάτων.
4. Αναζήτηση του ιατρικού μητρώου ενός ασθενούς με κάποια ασθένεια, για την ανάλυση του ιστορικού του στο μητρώο.

Κατηγοριοποίηση Μεθόδων Εξόρυξης

Υπάρχει μια μεγάλη ποικιλία μεθόδων εξόρυξης δεδομένων.

Ανάλογα με το είδος των δεδομένων και το είδος της γνώσης που εξάγεται, αυτές κατηγοριοποιούνται σε διαφορετική κατηγορία.

Μερικές βασικές μέθοδοι της Εξόρυξης Δεδομένων παρουσιάζονται παρακάτω.

Κατηγοριοποίηση (classification)

Πρόκειται για μια προγνωστική μέθοδο.

Στόχος είναι η δημιουργία ενός μοντέλου – κατηγοριοποιητή (classifier) με βάση τα υπάρχοντα δεδομένα.

Ουσιαστικά, είναι η μάθηση μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο (συνήθως αναπαρίσταται ως ένα διάνυσμα τιμών για τις χαρακτηριστικές του ιδιότητες) σε μία τιμή μιας κατηγορικής μεταβλητής, η οποία η οποία είναι γνωστή και ως κλάση (ή κατηγορία).

Η μάθηση (learning), μία έννοια που ήδη αναφέρθηκε, αποτελεί συμπεριφορά των ευφυών συστημάτων, τα οποία μελετώνται από τομείς, όπως η Μηχανική Μάθηση ή η Τεχνητή Νοημοσύνη.

Κατηγοριοποίηση (classification)

Εξαιτίας αυτού, όλες αυτές οι περιοχές μελετούν παρόμοια προβλήματα, χωρίς αυτό να σημαίνει ότι δεν υπάρχουν και ξεχωριστά αντικείμενα, που μελετώνται από την κάθε μία ξεχωριστά.

Η κατηγοριοποίηση συχνά συγχέεται με το γενικό όρο της πρόβλεψης.

Στην κατηγοριοποίηση, το αποτέλεσμα που θέλουμε να προβλέψουμε είναι η κλάση των δειγμάτων.

Η κλάση μπορεί να πάρει διακριτές τιμές από ένα πεπερασμένο σύνολο.

Αντίθετα, κατά την πρόβλεψη με χρήση τεχνικών όπως η παλινδρόμηση, η μεταβλητή-στόχος μπορεί να είναι οποιοσδήποτε πραγματικός αριθμός.

Παλινδρόμηση

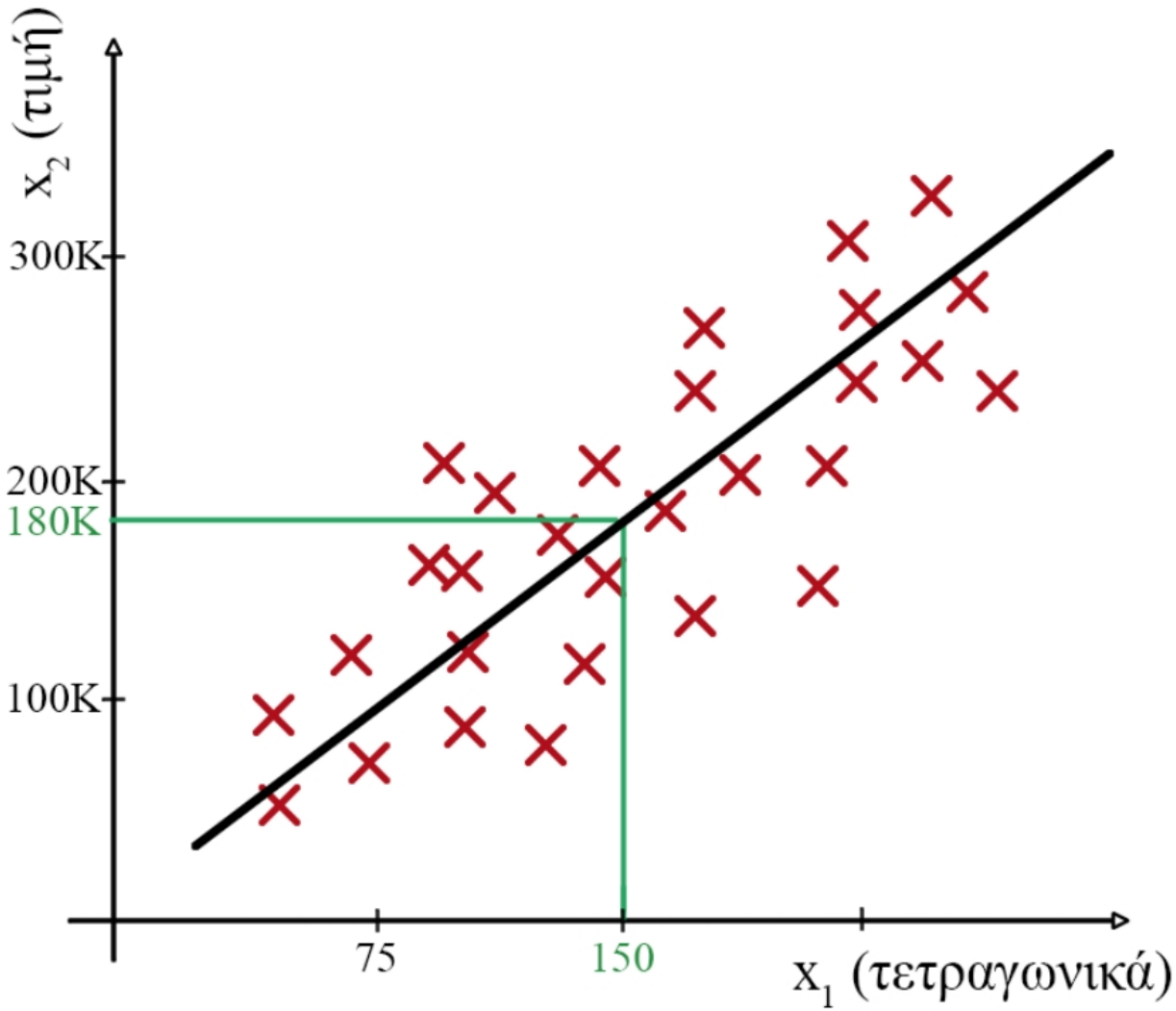
Μια σχετική διαδικασία με την κατηγοριοποίηση είναι η παλινδρόμηση (regression), στόχος της οποίας είναι η μάθηση ή αλλιώς η εκπαίδευση (training) μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μία πραγματική μεταβλητή.

Πρόκειται για μια, επίσης, προγνωστική μέθοδο.

Στόχος είναι με βάση κάποιες ανεξάρτητες μεταβλητές (independent variables) να προβλεφθούν οι τιμές μιας εξαρτημένης μεταβλητής (dependent variable).

Στην Εικόνα 1.2 παρουσιάζουμε ένα απλό παράδειγμα γραμμικής παλινδρόμησης.

Παλινδρόμηση



Σχήμα: 1.2

Παλινδρόμηση

Οι μεταβλητές είναι τα τετραγωνικά ενός σπιτιού και η τιμή πώλησης του σε χιλιάδες Ευρώ.

Η γραμμική παλινδρόμηση προσαρμόζει μια ευθεία στα δείγματα του συνόλου δεδομένων, τα οποία σηματοδοτούνται με κόκκινο Χ.

Η προσαρμογή γίνεται με βάση μια συνάρτηση απόστασης ή συνάρτηση κόστους, την τιμή της οποία θέλουμε να ελαχιστοποιήσουμε.

Έχοντας τη βέλτιστη ευθεία, δηλαδή την ευθεία που ελαχιστοποιεί την τιμή της συνάρτησης κόστους, μπορούμε να δώσουμε μια προσεγγιστικά καλή απάντηση σε ερωτήματα της μορφής: «Σε τι τιμές πωλούνται σπίτια των 150 τετραγωνικών;».

Συσταδοποίηση (clustering)

Η συσταδοποίηση (clustering) είναι μια περιγραφική μέθοδος.

Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα.

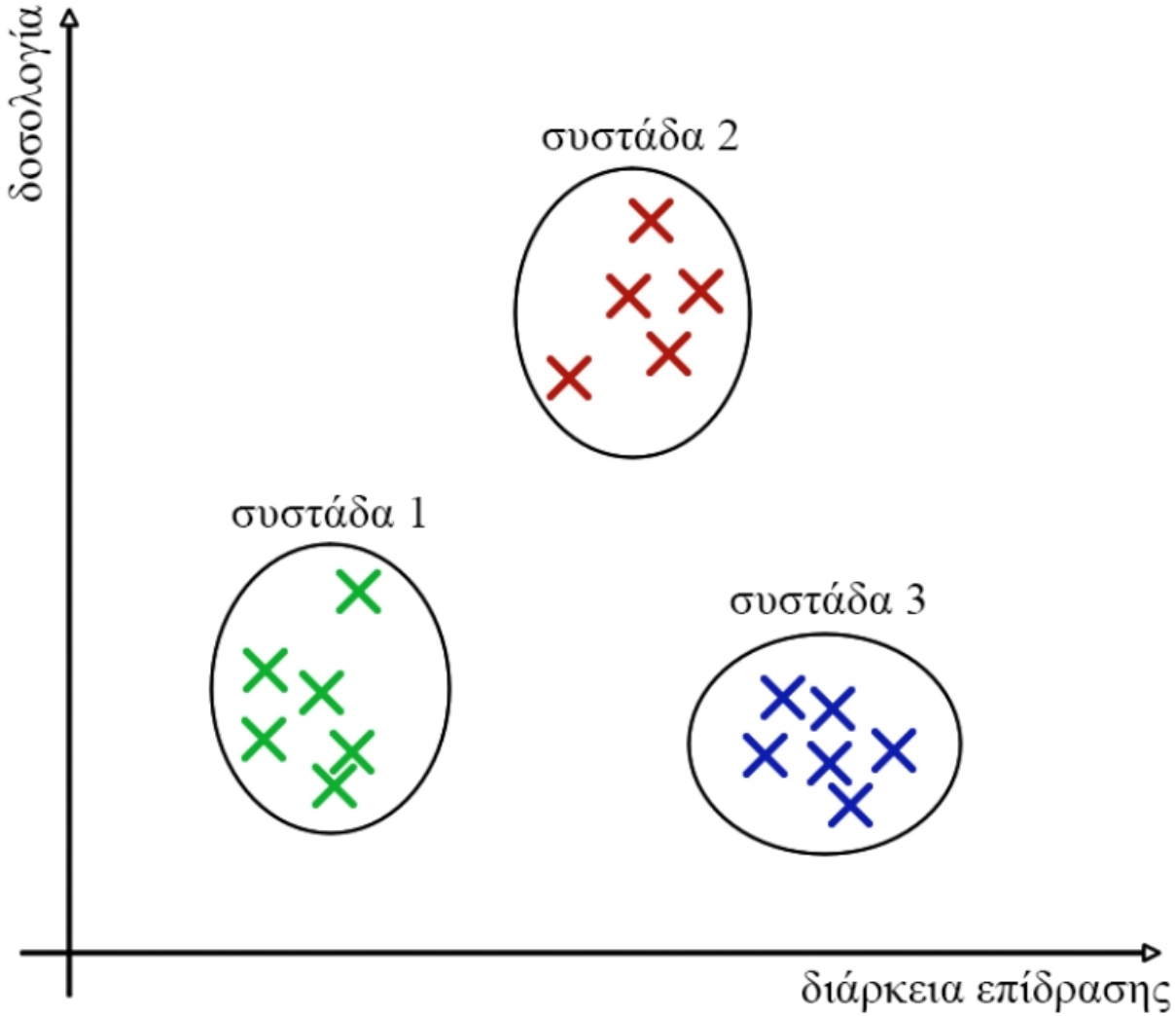
Ουσιαστικά αναζητείται ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων, για να περιγράψει τα δεδομένα.

Οι κατηγορίες μπορεί να είναι αμοιβαία αποκλειόμενες και εξαντλητικές ή να έχουν μία πιο σύνθετη αναπαράσταση, όπως για παράδειγμα ιεραρχικές και επικαλυπτόμενες.

Στο παρακάτω παράδειγμα (Εικόνα 1.3) βλέπουμε το αποτέλεσμα συσταδοποίησης φαρμακευτικών δεδομένων.

Συσταδοποίηση

Έχουν δημιουργηθεί 3 συστάδες με βάση τα χαρακτηριστικά «δοσολογία» και «διάρκεια επίδρασης».



Σχήμα: 1.3

Εξαγωγή και Ανάλυση Συσχετίσεων

Η εξαγωγή κανόνων συσχέτισης (Mining Association Rules) θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων.

Έχει προσελκύσει ιδιαίτερο ενδιαφέρον, καθώς οι κανόνες συσχέτισης παρέχουν έναν συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες.

Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων.

Αυτοί οι συσχετισμοί παρουσιάζονται στη μορφή A -> B, όπου τα A και B αποτελούν σύνολα που αναφέρονται στα χαρακτηριστικά του συνόλου δεδομένων που αναλύουμε.

Εξαγωγή και Ανάλυση Συσχετίσεων

Δεδομένου ενός συνόλου από δεδομένα, ένας κανόνας συσχέτισης A ->B προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου B δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου A.

Κλασικό πεδίο εφαρμογής των κανόνων συσχέτισης είναι η ανάλυση του καλαθιού αγοράς (market basket), όπου τα δεδομένα αφορούν συναλλαγές πελατών σε ένα παντοπωλείο.

Οι συναλλαγές μπορεί να είναι, για παράδειγμα: ψωμί, γάλα, ψωμί, πάνες, μπύρα, αυγά, γάλα, πάνες, μπύρα, σόδα, ψωμί, γάλα, πάνες, μπύρα και ψωμί, γάλα, πάνες, σόδα, και κάποιοι κανόνες συσχέτισης σε αυτές είναι Πάνες -> μπύρα, μπύρα, ψωμί->γάλα, γάλα, ψωμί->αυγά, σόδα.

Ο τελευταίος κανόνας, για παράδειγμα, φανερώνει ότι είναι πολύ πιθανό όποιος αγοράζει γάλα και ψωμί να αγοράσει, επίσης, αυγά και σόδα.

Εξαγωγή και Ανάλυση Συσχετίσεων

Εξάγοντας χρήσιμα συμπεράσματα μέσω των κανόνων συσχέτισης, το τμήμα προώθησης του παντοπωλείου μπορεί να τοποθετήσει κατάλληλα τα προϊόντα στα ράφια, να κάνει την κατάλληλη καμπάνια προώθησης τους και να διαχειριστεί πιο αποδοτικά τα αποθεματικά του.

Οπτικοποίηση

Η οπτικοποίηση των δεδομένων συχνά βοηθάει στην καλύτερη κατανόηση όχι μόνο των ίδιων των δεδομένων, αλλά και των συσχετίσεων που μπορεί να υπάρχουν μεταξύ τους.

Σε επόμενα μαθήματα θα περιγράψουμε τους βασικούς τρόπους οπτικοποίησης στην R.

Ωστόσο, οπτικοποίηση μπορεί να γίνει μόνο για συγκεκριμένο αριθμό διαστάσεων.

Αυτό σημαίνει ότι για σύνολα δεδομένων με πολλά χαρακτηριστικά, η οπτικοποίηση τους είναι ανέφικτη ή εναλλακτικά αρκούμαστε στην οπτικοποίηση ενός μικρού μέρους αυτών.

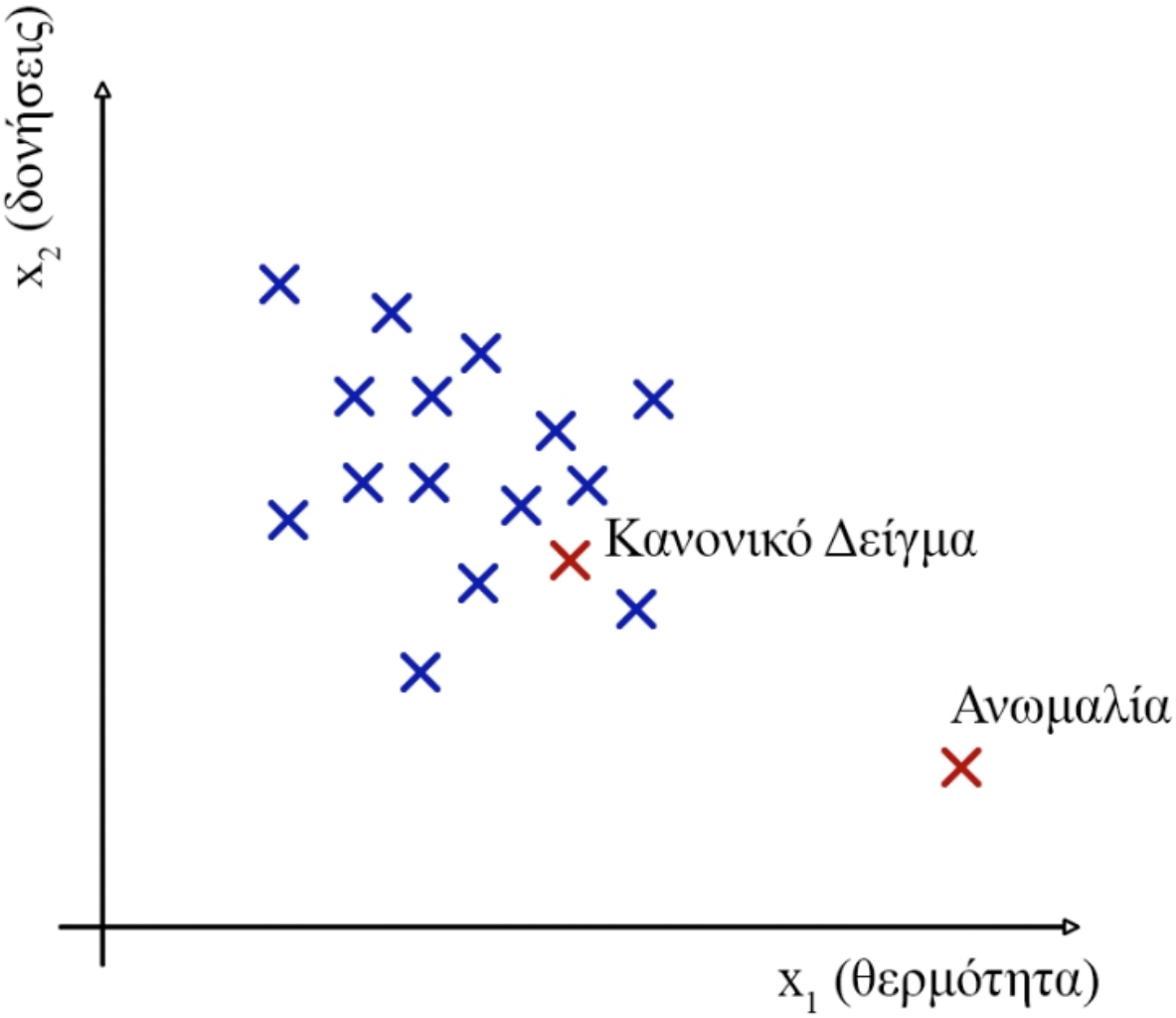
Σε κάθε περίπτωση, οι οπτικοποιήσεις θα πρέπει να συνοδεύονται και από τους αντίστοιχους στατιστικούς ελέγχους, προκειμένου να βεβαιωθούμε για την εγκυρότητα των συσχετίσεων που απεικονίζονται.

Ανίχνευση Ανωμαλιών

Η ανίχνευση ανωμαλιών εστιάζει στην ανακάλυψη αποκλίσεων στα δεδομένα σε σχέση με αντίστοιχα δεδομένα, τα οποία έχουν συλλεχθεί στο παρελθόν ή με τυπικές τιμές των δεδομένων αυτών.

Η Εικόνα 1.4 παρουσιάζει ένα τέτοιο παράδειγμα, στο οποίο με κόκκινο φαίνονται ένα κανονικό δείγμα, κοντά στα υπόλοιπα με φυσιολογικές τιμές δείγματα, και ένα ανώμαλο δείγμα, του οποίου η τιμή απέχει αρκετά από τα υπόλοιπα.

Ανίχνευση Ανωμαλιών



Σχήμα: 1.4