

Χρήση Στατιστικών και Πιθανοθεωρητικών Μοντέλων στην Ανάλυση Δεδομένων

Γκάνα Ε.

April 17, 2025

1 Εισαγωγή

Η επιστήμη των δεδομένων (data science) ως επιστημονικός κλάδος επηρεάζεται από την πληροφορική, την επιχειρησιακή έρευνα αλλά και τα μαθηματικά. Χρησιμοποιεί αλγόριθμους και στατιστικά μοντέλα για να εξάγει πληροφορίες από δεδομένα. Οι τεχνικές που χρησιμοποιούνται βασίζονται σε μοντέλα πιθανοθεωρίας και στατιστικής για την ανάλυση και την πρόβλεψη μεγάλων δεδομένων. Θα επικεντρωθούμε σε τρία βασικά μοντέλα ([2]):

- Έλεγχος Υποθέσεων [1]
- Μέθοδοι Ταξινόμησης [4], [3], [7]
- Ανάλυση Χρονοσειρών [5], [6]

2 Βιβλιογραφική Ανασκόπηση

- **Έλεγχος Υποθέσεων:** Είναι μια διαδικασία για την λήψη απόφασης μεταξύ δύο υποθέσεων. Συνίσταται στην απόρριψη ή μη απόρριψη μιας στατιστικής υπόθεσης (μηδενική υπόθεση), με βάση ένα δείγμα δεδομένων. Μερικοί από τους ελέγχους είναι:

– **t-test:**

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \quad (1)$$

– **Chi-square test:**

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (2)$$

– **ANOVA:**

$$F = \frac{MS_{between}}{MS_{within}} \quad (3)$$

- **Μέθοδοι Ταξινόμησης:** Η μέθοδος ταξινόμησης (classification) είναι μια από τις πιο βασικές και σημαντικές τεχνικές στην επιστήμη δεδομένων, ειδικά στον χώρο της επιβλεπόμενης μάθησης (supervised learning). Στόχος της είναι να βρει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μία βάση δεδομένων και να προβλέψει την κατηγορία(label) στην οποία ανήκει ένα δεδομένο δείγμα με βάση τα χαρακτηριστικά του(features). Μερικές δημοφιλείς μέθοδοι ταξινόμησης είναι:

– **Δέντρα Αποφάσεων (Decision Trees):** Κάθε κόμβος είναι ένα χαρακτηριστικό(feature), κάθε κλαδί είναι μια απόφαση και κάθε φύλλο μια προβλεπόμενη κατηγορία. Συγκεκριμένα για τα δέντρα ταξινόμησης κάποια κριτήρια διαχωρισμού που χρησιμοποιούνται για να επιλεγεί το "βέλτιστο" χαρακτηριστικό σε κάθε κόμβο είναι:

* **Entropy:**

$$H(s) = - \sum_{i=1}^k p_i \log_2 p_i \quad (4)$$

* **Gini index:**

$$G(s) = \sum_{i=1}^k p_i(1 - p_i) \quad (5)$$

- **Νευρωνικά δίκτυα (Neural Networks):** Αποτελούνται από μία συνάρτηση ενεργοποίησης και αντίστοιχα μία συνάρτηση εξόδου. Παρακάτω αναπτύσσονται μαθηματικά κάποιες από τις πιο συχνές περιπτώσεις αυτών (Σιγμοειδής και softmax).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (6)$$

$$g_k(t) = \frac{e^{t_k}}{\sum_{j=1}^K e^{t_j}} \quad (7)$$

- **Λογιστική Παλινδρόμηση (Logistic Regression):** Μπορεί να χρησιμοποιηθεί στις περιπτώσεις όπου η μεταβλητή είναι δίτιμη, δηλαδή η πρόβλεψη είναι το αποτέλεσμα μιας διαδικασίας Bernoulli με:

- * **Πιθανότητα επιτυχίας:**

$$p = b_0 + b_1 x$$

- * **Σχετική πιθανότητα επιτυχίας:**

$$\frac{p}{1-p}$$

- * **Φυσικός λογάριθμος της σχετικής πιθανότητας:**

$$\ln\left(\frac{p}{1-p}\right)$$

Η λογαριθμική αυτή συνάρτηση στην ορολογία των λογαρίθμων ονομάζεται logit της p και συμβολίζεται:

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Θέτοντας $p = z$ και αντιλογαριθμίζοντας:

$$\frac{e^z}{1 + e^z}$$

που διατυπώνεται και ως:

$$\frac{1}{1 + e^{-z}} \quad (8)$$

και αποτελεί την εκτίμηση της πιθανότητας επιτυχίας μιας δίτιμης μεταβλητής έστω Y .

- **Ανάλυση Χρονοσειρών:** Με τον όρο συμβολικές χρονοσειρές εννοούμε ένα σύνολο από διαδοχικά δεδομένα ποιοτικού χαρακτήρα (σύμβολα). Η διαδοχή μπορεί να αναφέρεται είτε σε χρονικό επίπεδο είτε σε επίπεδο διάταξης στον χώρο. Μια συμβολική χρονοσειρά μπορεί να θεωρηθεί πραγματοποίηση μιας Μαρκοβιανής Αλυσίδας.

- **Μαρκοβιανή Ιδιότητα:**

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) =$$

$$P(X_{n+1} = x | X_n = x_n) \quad (9)$$

- **Ιδιότητα Χρονικής Ομογένειας:**

$$P(X_{n+1} = x | X_n = y) = P(X_n = x | X_{n-1} = y) \quad (10)$$

- **Πιθανότητα Μετάβασης σε n -βήματα:**

$$p_{ij}^{(n)} = P(X_n = j | X_0 = i) \quad (11)$$

- **Πίνακας Μεταβάσεων n -οστού βήματος:** Αποτελείται από όλες τις πιθανότητες μετάβασης n -οστού βήματος και είναι στοχαστικός δηλαδή:

$$p_{ij}^{(n)} > 0$$

και

$$\sum_j p_{ij}^{(n)} = 1 \quad (12)$$

3 Συμπεράσματα

Η χρήση λοιπόν των μαθηματικών μοντέλων στατιστικής και πιθανοθεωρίας αποτελεί θεμελιώδες εργαλείο στην ανάλυση δεδομένων και στην διαδικασία λήψης αποφάσεων. Αναπτύξαμε την σημασία του ελέγχου υποθέσεων για την επαλήθευση θεωρητικών ισχυρισμών, τη συμβολή των μεθόδων ταξινόμησης στην πρόβλεψη και κατηγοριοποίηση δεδομένων, καθώς και το ρόλο της ανάλυσης χρονοσειρών στην κατανόηση της χρονικής εξέλιξης φαινομένων. Η ενσωμάτωση αυτών

των μεθόδων σε εφαρμογές της επιστήμης των δεδομένων επιτρέπει την εξαγωγή χρήσιμων συμπερασμάτων από πολύπλοκα και πολυδιάστατα σύνολα δεδομένων. Παράλληλα, η ακριβής μαθηματική τους θεμελίωση διασφαλίζει την εγκυρότητα και την αξιοπιστία των αποτελεσμάτων.

4 Αναφορές

References

- [1] Andy Field. *Η διερεύνηση της στατιστικής με τη χρήση του SPSS της IBM*. Προπομπός, 2016. Επιμέλεια: Γιάννης Γκίτσος, Μαρία Μάυρη.
- [2] C. Weihs and K. Ickstadt. Data science: the impact of statistics. *International Journal of Data Science and Analytics*, 2018.
- [3] Κ.Π. Δρόσου. PhD thesis.
- [4] Μάριος-Αλέξανδρος Μιχαήλ. *Βελτιστοποίηση ανάλυσης δεδομένων με χρήση πολυκριτήριων μεθόδων ανάλυσης αποφάσεων και μηχανικής μάθησης*. PhD thesis, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πολυτεχνείο Κρήτης, 2020.
- [5] Κ.Ι. Πάντου. *Ανάλυση χρονοσειρών*. PhD thesis, Πανεπιστήμιο Θεσσαλίας, 2020.
- [6] Κ.Μ. Ροδίτου. *Διερεύνηση τάξης μαρκοβιανής αλυσίδας σε συμβολικές χρονοσειρές*. PhD thesis, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2007.
- [7] Α. Σκουφά. *Λογιστική παλινδρόμηση*. PhD thesis, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, 2009.