

Χρήση Στατιστικών και Πιθανοθεωρητικών Μοντέλων στην Ανάλυση Δεδομένων

Γκάννα Ε.

April 23, 2025

- Εισαγωγή
- Έλεγχοι Υποθέσεων
- Μέθοδοι Ταξινόμησης
 - Δέντρα Αποφάσεων
 - Νευρωνικά Δίκτυα
 - Γραφική Αναπαράσταση
 - Λογιστική Παλινδρόμηση
- Ανάλυση Χρονοσειρών(Συμβολικές Χρονοσειρές)
- Συμπεράσματα

Περιγραφή

- Επηρεάζεται από την πληροφορική, την επιχειρησιακή έρευνα αλλά και τα μαθηματικά.
- Χρησιμοποιεί αλγορίθμους και στατιστικά μοντέλα για να εξαγάγει 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.

Βασικά Μοντέλα

- **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:**

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

Δημοφιλείς Μέθοδοι

- **Δέντρα Αποφάσεων (Decision Trees):**

- **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):**

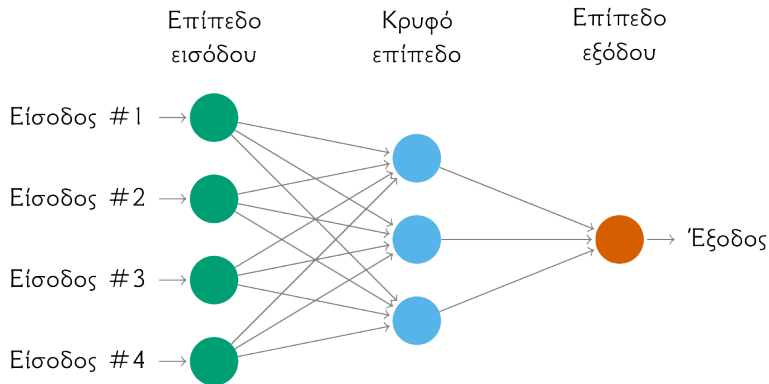
$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad (6)$$

Συνάρτηση Εξόδου:

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

Νευρωνικά δίκτυα

Γραφική αναπαράσταση



Δημοφιλείς Μέθοδοι

- **Λογιστική Παλινδρόμηση (Logistic Regression):**

- **Πιθανότητα επιτυχίας:**

$$p = b_0 + b_1 x$$

- **Σχετική πιθανότητα επιτυχίας:**

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

- **Φυσικός λογάριθμος της σχετικής πιθανότητας:**

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

- **Συνάρτηση εκτίμησης πιθανότητας**

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

(8)

Μαρκοβιανές Αλυσίδες

- **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:**

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

Σύνοψη

- Τα μαθηματικά μοντέλα στατιστικής και πιθανοθεωρίας αποτελούν θεμελιώδη εργαλεία στην ανάλυση δεδομένων και στην διαδικασία λήψης αποφάσεων.
- Η σημασία του ελέγχου υποθέσεων για την επαλήθευση θεωρητικών ισχυρισμών.
- Η συμβολή των μεθόδων ταξινόμησης στην πρόβλεψη και κατηγοριοποίηση δεδομένων.
- Ο ρόλος 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.
- Η ακριβής μαθηματική τους θεμελίωση διασφαλίζει την εγκυρότητα και την αξιοπιστία των αποτελεσμάτων.