Περιεχόμενα

2

Ανασκόπηση των τεχνικών ταξινόμησης

2.1 Εισαγωγή

Όπως αναφέρθηκε στο εισαγωγικό κεφάλαιο, η αυξημένη σημαντικότητα του προβλήματος της ταξινόμησης τόσο σε πρακτικό όσο και σε ερευνητικό επίπεδο, έχει ελκύσει το ενδιαφέρον πολλών ερευνητών από διαφορετικούς επιστημονικούς χώρους. Σκοπός του κεφαλαίου αυτού είναι η ανασκόπηση των βασικότερων μεθοδολογικών προσεγγίσεων οι οποίες έχουν προταθεί για την ανάπτυξη υποδειγμάτων ταξινόμησης. Βέβαια, η ευρύτητα του προβλήματος της ταξινόμησης, καθιστά ιδιαίτερα δύσκολη την πλήρη κάλυψη όλων των μεθοδολογικών προσεγγίσεων που έχουν κατά καιρούς αναπτυχθεί. Για το λόγο αυτό η παρουσίαση επικεντρώνεται στις ευρύτερα διαδεδομένες προσεγγίσεις, βάσει των ερευνητικών και πρακτικών τους εφαρμογών. Οι εξεταζόμενες προσεγγίσεις διακρίνονται σε δύο βασικές κατηγορίες:

1. Στις στατιστικές και οικονομετρικές προσεγγίσεις, οι οποίες αποτελούν τον «παραδοσιακό» τρόπο αντιμετώπισης του προβλήματος της ταξινόμησης.
2. Στις μη παραμετρικές προσεγγίσεις οι οποίες έχουν προταθεί κατά τις τελευταίες δύο δεκαετίες ως καινοτόμες και αποτελεσματικές τεχνικές ανάπτυξης υποδειγμάτων ταξινόμησης.

2.2 Στατιστικές και οικονομετρικές προσεγγίσεις

Η στατιστική είναι ίσως η παλαιότερη των επιστημών αντικείμενο της οποίας είναι η ανάλυση δειγμάτων με απώτερο στόχο την εξαγωγή συμπερασμάτων επί του πληθυσμού. Ως ένα τέτοιο πρόβλημα αντιμετωπίζεται η ταξινόμηση στα πλαίσια της στατιστικής θεωρίας, θεωρώντας ουσιαστικά ότι η κάθε κατηγορία στην οποία πρέπει να γίνει η ταξινόμηση των εναλλακτικών δραστηριοτήτων, αντιστοιχεί σε έναν πληθυσμό. Η μελέτη και ανάλυση δειγμάτων αντικειμένων που ανήκουν σε διαφορετικές κατηγορίες υπήρξε και εξακολουθεί να είναι ένα από τα βασικά θέματα που απασχολεί τους στατιστικούς επιστήμονες. Οι σχετικές τεχνικές που έχουν αναπτυχθεί περιλαμβάνουν τόσο μονοδιάστατες όσο και πολυδιάστατες στατιστικές μεθόδους. Οι πρώτες αναφέρονται στην ανάπτυξη και εφαρμογή στατιστικών ελέγχων περιγραφικού χαρακτήρα και συνεπώς δεν θα αναλυθούν περαιτέρω. Οι βάσεις των πολυδιάστατων στατιστικών μεθόδων τέθηκαν ουσιαστικά από τον Fisher το 1936. Στην ερευνητική του εργασία ο Fisher ανέπτυξε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης, τη γραμμική διακριτική ανάλυση (linear discriminant analysis), η οποία για πολλές δεκαετίες υπήρξε η πλέον διαδεδομένη μεθοδολογία για την ανάπτυξη υποδειγμάτων ταξινόμησης. Αργότερα ο Smith (1947) επέκτεινε την εργασία του Fisher αναπτύσσοντας την τετραγωνική διακριτική ανάλυση (quadratic discriminant analysis) ως μια καταλληλότερη μορφή της διακριτικής ανάλυσης, στην περίπτωση όπου οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών δεν είναι ίσοι.

Στις δεκαετίες που ακολούθησαν τις πρώτες αυτές ερευνητικές εργασίες, ιδιαίτερη έμφαση δόθηκε στην ανάπτυξη οικονομετρικών μεθόδων ταξινόμησης. Γνωστότερες από τις μεθόδους αυτές είναι το γραμμικό υπόδειγμα πιθανότητας (linear probability model), το λογιστικό υπόδειγμα πιθανότητας (logit analysis) και το κανονικό υπόδειγμα πιθανότητας (probit analysis). Οι τρεις αυτές τεχνικές είναι ουσιαστικά ειδικές μορφές της γνωστής στατιστικής παλινδρόμησης σε περιπτώσεις όπου η εξαρτημένη μεταβλητή λαμβάνει διακριτές τιμές. Με εξαίρεση το γραμμικό υπόδειγμα πιθανότητας, το οποίο μπορεί να εφαρμοστεί μόνο στην περίπτωση όπου η ταξινόμηση πραγματοποιείται σε δύο κατηγορίες, τόσο το λογιστικό όσο και το κανονικό υπόδειγμα είναι εφαρμόσιμα ακόμα και στην περίπτωση πολλαπλών κατηγοριών. Τα δύο αυτά υποδείγματα παρουσιάζουν σημαντικά θεωρητικά πλεονεκτήματα έναντι της διακριτικής ανάλυσης, την οποία και σταδιακά αντικατέστησαν.

Παρά την έντονη κριτική την οποία έχουν δεχθεί αυτές οι «παραδοσιακές» στατιστικές και οικονομετρικές προσεγγίσεις, παραμένουν, ακόμη και σήμερα, ιδιαίτερα διαδεδομένες, τόσο σε ερευνητικό όσο και σε πρακτικό επίπεδο. Το πλήθος των στατιστικών/οικονομετρικών υπολογιστικών προγραμμάτων που είναι διαθέσιμα συμβάλλουν στην εύκολη εφαρμογή των προσεγγίσεων αυτών, στοιχείο το οποίο δικαιολογεί, εν μέρει, την ευρεία τους διάδοση. Παράλληλα, ιδιαίτερα διαδεδομένη είναι και η χρήση τους σε συγκριτικές έρευνες, οι οποίες στόχο έχουν την αξιολόγηση της αποτελεσματικότητας νέων τεχνικών ταξινόμησης που αναπτύσσονται. Προς την κατεύθυνση αυτή, οι στατιστικές/οικονομετρικές προσεγγίσεις αποτελούν ένα σημείο αναφοράς βάσει του οποίου πραγματοποιούνται οι συγκρίσεις των νέων τεχνικών ταξινόμησης. Υπό συγκεκριμένες μάλιστα συνθήκες, αποδεικνύεται ότι αυτό το σημείο αναφοράς αποτελεί το θεωρητικά βέλτιστο αποτέλεσμα ταξινόμησης.

2.2.1 Διακριτική ανάλυση

Η διακριτική ανάλυση αποτέλεσε την πρώτη πολυδιάστατη μέθοδο ταξινόμησης και επί δεκαετίες ήταν και η πλέον διαδεδομένη τεχνική για την αντιμετώπιση σχετικών προβλημάτων. Στη γραμμική της μορφή αναπτύχθηκε από τον Fisher (1936). Χρησιμοποιώντας ως δείγμα εκμάθησης ένα σύνολο

εναλλακτικών δραστηριοτήτων η ταξινόμηση των οποίων είναι γνωστή, σκοπός της μεθόδου είναι η ανάπτυξη μιας σειράς διακριτικών συναρτήσεων οι οποίες μεγιστοποιούν τη διακύμανση μεταξύ των κατηγοριών σε σχέση με τη διακύμανση εντός των κατηγοριών. Στην γενική περίπτωση όπου η ταξινόμηση πραγματοποιείται σε q κατηγορίες, αναπτύσσονται q-1 γραμμικές συναρτήσεις της μορφής:

$$Z_{kl} = a_{kl} + b_{kl1}g_1 + b_{kl2}g_2 + ... + b_{kln}g_n$$

όπου g_1, g_2, …, g_n είναι τα χαρακτηριστικά[1] που περιγράφουν τις εναλλακτικές δραστηριότητες x_1, x_2, …, x_m, a_{kl} είναι μια σταθερά, και b_{kl1}, b_{kl2}, …, b_{kln} είναι οι συντελεστές των χαρακτηριστικών στη διακριτική συνάρτηση. Οι δείκτες k και l αναφέρονται σε ένα ζεύγος κατηγοριών οι οποίες συμβολίζονται αντίστοιχα ως C_k και C_l.

Ο υπολογισμός του σταθερού όρου a_{kl} και του διανύσματος $\boldsymbol{b}_{kl}$ βασίζεται στην υπόθεση ότι οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι και ότι οι επιδόσεις των εναλλακτικών δραστηριοτήτων στα εξεταζόμενα χαρακτηριστικά ακολουθούν την πολυμεταβλητή κανονική κατανομή. Βάσει των υποθέσεων αυτών, οι υπολογισμοί των a_{kl} και του διανύσματος $\boldsymbol{b}_{kl}$ πραγματοποιούνται ως εξής:

$$\boldsymbol{b}_{kl} = \boldsymbol{\Sigma}^{-1} \cdot [\boldsymbol{\mu}_k - \boldsymbol{\mu}_l]$$

$$a_{kl} = -[\boldsymbol{\mu}_k + \boldsymbol{\mu}_l]' \cdot \boldsymbol{b}_{kl}/2$$

όπου:

- $\boldsymbol{\mu}_k$ είναι το διάνυσμα των μέσων τιμών των χαρακτηριστικών για τις εναλλακτικές δραστηριότητες της κατηγορίας C_k, και

[1] Στο παρόν κεφάλαιο χρησιμοποιείται ο όρος «χαρακτηριστικά» για την αναφορά στα κριτήρια αξιολόγησης, ακολουθώντας την ορολογία των παρουσιαζόμενων μεθοδολογικών προσεγγίσεων. Η έννοια πάντως του χαρακτηριστικού διαφέρει από την έννοια του κριτηρίου: το χαρακτηριστικό αποδίδει απλά μια περιγραφή στις εναλλακτικές δραστηριότητες, ενώ το κριτήριο καθορίζει επιπλέον και μια σχέση προτίμησης (βλέπε επόμενο κεφάλαιο).

– $\boldsymbol{\Sigma}$ είναι ο πίνακας διακύμανσης-συνδιακύμανσης μεταξύ των κατηγοριών (within groups variance-covariance matrix). Συμβολίζοντας ως m το πλήθος των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης, ως $\boldsymbol{g}_j=(g_{j1}, g_{j2}, \ldots, g_{jn})$ το διάνυσμα της περιγραφής της εναλλακτικής δραστηριότητας x_j βάσει των χαρακτηριστικών $\boldsymbol{g}$, και ως q το πλήθος των κατηγοριών, ο πίνακας $\boldsymbol{\Sigma}$ υπολογίζεται ως εξής:

$$\boldsymbol{\Sigma} = \frac{\sum_{k=1}^{q} \sum_{\forall x_j \in C_k} [\boldsymbol{g}_j - \boldsymbol{\mu}_k] \cdot [\boldsymbol{g}_j - \boldsymbol{\mu}_k]'}{m - q}$$

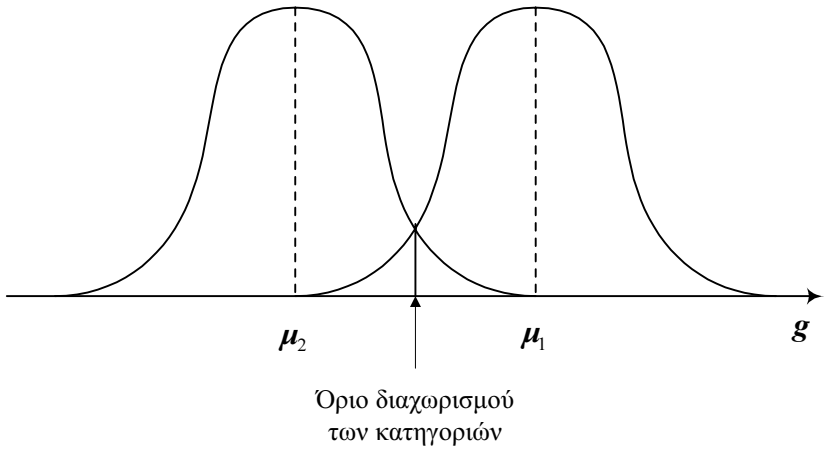
Θα πρέπει βέβαια να τονιστεί οι εκτιμήσεις των συντελεστών και του σταθερού όρου, όπως αυτές προκύπτουν από τις παραπάνω σχέσεις δεν είναι μοναδικές. Ουσιαστικά είναι δυνατή η ανάπτυξη μιας σειράς εναλλακτικών διακριτικών συναρτήσεων των οποίων οι συντελεστές $\boldsymbol{b}'_{kl}$ και οι σταθεροί όροι a'_{kl} μπορούν να προκύψουν ως γραμμικοί μετασχηματισμοί των $\boldsymbol{b}_{kl}$ και a_{kl}. Το γεγονός αυτό αποτελεί και το βασικότερο λόγο για τον οποίο είναι δύσκολο να καθοριστεί η συνεισφορά του κάθε χαρακτηριστικού στην ταξινόμηση των εναλλακτικών δραστηριοτήτων[2].

Η ταξινόμηση κάθε εναλλακτικής δραστηριότητας x_j σε μια εκ των προκαθορισμένων κατηγοριών πραγματοποιείται βάσει των σκορ της δραστηριότητας όπως αυτά υπολογίζονται από την κάθε συνάρτηση. Πιο συγκεκριμένα, μια εναλλακτική δραστηριότητα x_j θα ταξινομηθεί στην κατηγορία C_k εάν για όλες τις άλλες κατηγορίες C_l ισχύει:

$$Z_{kl}(\boldsymbol{g}_j) \geq \ln \frac{K(k \mid l)\pi_l}{K(l \mid k)\pi_k}$$

[2] Σε αντίθεση με την κλασσική πολλαπλή στατιστική παλινδρόμηση, στη διακριτική ανάλυση συνήθως δεν πραγματοποιούνται στατιστικοί έλεγχοι (για παράδειγμα *t*-τεστ) όσον αφορά τη σημαντικότητα των επιμέρους συντελεστών στις διακριτικές συναρτήσεις που αναπτύσσονται, ακριβώς επειδή οι συντελεστές αυτοί δεν είναι μοναδικοί.

Ως $Z_{kl}(\boldsymbol{g}_j)$ συμβολίζεται το σκορ διάκρισης (discriminant score) που αποδίδεται στην εναλλακτική δραστηριότητα x_j από τη διακριτική συνάρτηση Z_{kl}, ως $K(k \mid l)$ συμβολίζεται το κόστος της εσφαλμένης ταξινόμησης μιας εναλλακτικής δραστηριότητας, η οποία ενώ ανήκει στην κατηγορία C_l εντάσσεται στην κατηγορία C_k, ενώ τέλος ως π_k συμβολίζεται η εκ των προτέρων πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα στην κατηγορία C_k. Θεωρώντας τα κόστη εσφαλμένων ταξινομήσεων ίσα όπως και τις εκ των προτέρων πιθανότητες, αυτός ο γραμμικός κανόνας ταξινόμησης μπορεί να αποδοθεί γραφικά μέσω του Σχήματος 2.1, για την απλή περίπτωση της διάκρισης μεταξύ δύο κατηγοριών.



Πηγή: Altman et al., 1981

Σχήμα 2.1: Σχηματική απεικόνιση του κανόνα ταξινόμησης της γραμμικής διακριτικής ανάλυσης

Στην περίπτωση όπου οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών δεν είναι ίσοι, τότε αντί της γραμμικής διακριτικής ανάλυσης καταλληλότερη κρίνεται η τετραγωνική διακριτική ανάλυση η οποία αναπτύχθηκε από τον Smith (1947). Η μορφή της τετραγωνικής συνάρτησης που αναπτύσσεται για κάθε ζεύγος κατηγοριών C_k και C_l είναι η ακόλουθη:

$$Z_{kl} = a_{kl} + \sum_{i=1}^{n} b_{kli} g_i + \sum_{i=1}^{n} \sum_{h=1}^{n} c_{klih} g_i g_h$$

Ο υπολογισμός των συντελεστών της συνάρτησης αυτής καθώς και του σταθερού όρου πραγματοποιείται βάσει των ακόλουθων σχέσεων:

$$\boldsymbol{b}_{kl} = -2[\boldsymbol{\mu}_k' \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\mu}_l' \boldsymbol{\Sigma}_l^{-1}]$$

$$\boldsymbol{c}_{kl} = \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_l^{-1}$$

$$a_{kl} = \boldsymbol{\mu}_k' \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_l' \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\mu}_l - \ln\left|\boldsymbol{\Sigma}_l \boldsymbol{\Sigma}_k^{-1}\right|$$

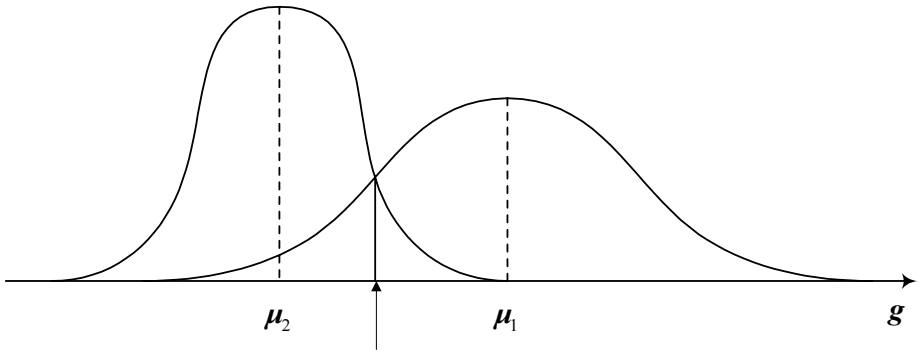
Ως $\boldsymbol{\Sigma}_k$ και $\boldsymbol{\Sigma}_l$ συμβολίζονται οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών C_k και C_l, οι οποίοι υπολογίζονται ως εξής:

$$\boldsymbol{\Sigma}_k = \frac{\sum_{\forall x_j \in C_k} [\boldsymbol{g}_j - \boldsymbol{\mu}_k][\boldsymbol{g}_j - \boldsymbol{\mu}_k]'}{m_k - 1}$$

Έχοντας ως βάση τα σκορ διάκρισης $Z_{kl}(\boldsymbol{g}_j)$ μιας εναλλακτικής δραστηριότητας x_j, για κάθε διακριτική συνάρτηση που αντιστοιχεί σε κάθε ζεύγος κατηγοριών C_k και C_l, ο κανόνας ταξινόμησης είναι παρόμοιος με αυτόν της γραμμικής διακριτικής ανάλυσης και διαμορφώνεται ως εξής (Σχήμα 2.2): η εναλλακτική x_j θα ενταχθεί στην κατηγορία C_k εάν και μόνο εάν για όλες τις άλλες κατηγορίες C_l ισχύει:

$$Z_{kl}(\boldsymbol{g}_j) \geq -2 \ln \frac{K(k \mid l)\pi_l}{K(l \mid k)\pi_k}$$

Στην πράξη βέβαια, τόσο στη γραμμική όσο και στην τετραγωνική διακριτική ανάλυση, ο καθορισμός των εκ των προτέρων πιθανοτήτων π_k και του κόστους εσφαλμένων ταξινομήσεων $K(k \mid l)$ είναι συνήθως μια ιδιαίτερα δύσκολη διαδικασία. Αυτό έχει ως αποτέλεσμα τα όρια που διαχωρίζουν τις κατηγορίες να καθορίζονται μέσω διαδικασιών δοκιμής και λάθους, ώστε να ελαχιστοποιηθεί ο συνολικός αριθμός των εσφαλμένων ταξινομήσεων και παράλληλα να υπάρχει μια ισορροπία στον αριθμό των εσφαλμένων ταξινομήσεων ανά κατηγορία.



(Πηγή: Altman et al., 1981)

Σχήμα 2.2: Σχηματική απεικόνιση της τετραγωνικής διακριτικής ανάλυσης

Εκτός των δυσκολιών που εντοπίζονται στον καθορισμό των παραπάνω πληροφοριών, η διακριτική ανάλυση, τόσο στη γραμμική όσο και στην τετραγωνική της μορφή έχει δεχθεί έντονη κριτική και σε μια σειρά άλλων θεμάτων, τα σημαντικότερα από τα οποία αφορούν την υπόθεση ότι οι μεταβλητές ακολουθούν την πολυμεταβλητή κανονική κατανομή, καθώς και τις υποθέσεις σχετικές με τη μορφή που έχουν οι πίνακες διακύμανσης-συνδιακύμνανσης των κατηγοριών. Αναλυτική παρουσίαση των επιπτώσεων που έχουν οι υποθέσεις αυτές στα αποτελέσματα της διακριτικής ανάλυσης παρουσιάζεται στο βιβλίο των Altman et al. (1981).

Όταν οι παραπάνω δύο βασικές υποθέσεις ικανοποιούνται, τότε χρησιμοποιώντας τον κανόνα του Bayes, αποδεικνύεται ότι οι δύο μορφές της διακριτικής ανάλυσης αποτελούν τη βέλτιστη μορφή του υποδείγματος ταξινόμησης (η γραμμική στην περίπτωση όπου οι πίνακες διακύμανσης-συνδιακύμανσης των κατηγοριών είναι ίσοι, και η τετραγωνική στην αντίθετη περίπτωση). Για την ακρίβεια, οι δύο μορφές της διακριτικής ανάλυσης είναι ασυμπτωτικά βέλτιστες, καθώς όσο αυξάνει το μέγεθος του εξεταζόμενου δείγματος (δείγμα εκμάθησης) οι επιμέρους κατηγορίες προσεγγίζουν

τους αντίστοιχους πληθυσμούς και τις πραγματικές στατιστικές τους ιδιότητες. Η μαθηματική απόδειξη της διαπίστωσης αυτής παρουσιάζεται αναλυτικά από τους Duda και Hart (1978), καθώς και από τους Patuwo et al. (1993).

Στην περίπτωση όμως όπου οι επιδόσεις των εξεταζόμενων εναλλακτικών δραστηριοτήτων δεν διαθέτουν τις παραπάνω βασικές στατιστικές ιδιότητες, φαινόμενο το οποίο απαντάται στην πλειοψηφία των πρακτικών περιπτώσεων, τότε δεν θα πρέπει απαραίτητα να θεωρηθεί ότι ελαττώνεται και η αποτελεσματικότητα των δύο μορφών της διακριτικής ανάλυσης. Βέβαια, σχετικές έρευνες όπως αυτές των Moore (1973), Krzanowski (1975, 1977), Dillon και Goldstein (1978) έδειξαν ότι σε περιπτώσεις όπου τα εξεταζόμενα δεδομένα περιέχουν διακριτές μεταβλητές, οι οποίες εκ της φύσης τους δεν ακολουθούν την πολυμεταβλητή κανονική κατανομή, τότε η αποτελεσματικότητα της διακριτικής ανάλυσης ελαττώνεται, ιδιαίτερα στην περίπτωση όπου οι παρουσιάζονται υψηλές συσχετίσεις (συντελεστής συσχέτισης μεγαλύτερος από 0,3) μεταξύ των επιδόσεων των εναλλακτικών δραστηριοτήτων στους επιμέρους παράγοντες της αξιολόγησης. Αντίθετα, οι έρευνες των Lanchenbruch et al. (1973), Subrahmaniam και Chinganda (1978) κατέληξαν στο συμπέρασμα ότι ακόμα και σε περιπτώσεις όπου οι επιδόσεις των εναλλακτικών δραστηριοτήτων δεν ακολουθούν την πολυμεταβλητή κανονική κατανομή, τα αποτελέσματα της διακριτικής ανάλυσης, όσον αφορά το σφάλμα της ταξινόμησης, παρουσιάζονται αρκετά ευσταθή, ιδιαίτερα στην περίπτωση της τετραγωνικής διακριτικής ανάλυσης και κυρίως σε περιπτώσεις δεδομένων με μικρό βαθμό ασυμμετρίας (skewness).

2.2.2 Το λογιστικό και το κανονικό υπόδειγμα πιθανότητας

Τα προαναφερθέντα προβλήματα και περιορισμοί της διακριτικής ανάλυσης αποτέλεσαν το βασικό κίνητρο για την ανάπτυξη εναλλακτικών μεθόδων ταξινόμησης, οι οποίες θα πλεονεκτούσαν έναντι της διακριτικής ανάλυσης τόσο σε θεωρητικό επίπεδο, όσο και στην αποτελεσματικότητα των αναπτυσσόμενων υποδειγμάτων. Το γραμμικό υπόδειγμα πιθανότητας (linear probability model), το λογιστικό και το κανονικό υπόδειγμα πιθανότητας (logit και probit analysis, αντίστοιχα), αποτέλεσαν τις βασικότερες τέτοιες εναλλακτικές προσεγγίσεις.

Το γραμμικό υπόδειγμα πιθανότητας ουσιαστικά βασίζεται στην πραγματοποίηση μιας απλής στατιστικής παλινδρόμησης, χρησιμοποιώντας ως εξαρτημένη μεταβλητή τη διακριτή κατηγοριοποίηση των εναλλακτικών δραστηριοτήτων του δείγματος εκμάθησης. Θεωρητικά, το αποτέλεσμα της γραμμικής συνάρτησης που αναπτύσσεται μέσω της παλινδρόμησης υποδεικνύει την πιθανότητα μια εναλλακτική δραστηριότητα να ανήκει σε κάποια εκ των προκαθορισμένων κατηγοριών. Πρακτικά όμως, η πραγματοποίηση της παλινδρόμησης δεν διασφαλίζει ότι το αποτέλεσμα θα βρίσκεται εντός του διαστήματος [0,1], γεγονός που καθιστά δύσκολη την ερμηνεία των αποτελεσμάτων που επιτυγχάνονται ως πιθανότητες. Θεωρώντας πάντως τα αποτελέσματα του γραμμικού υποδείγματος ως πιθανότητες, το όριο βάσει του οποίου πραγματοποιείται η ταξινόμηση στην περίπτωση των δύο κατηγοριών είναι το 0,5. Στην περίπτωση όμως περισσότερων κατηγοριών, η δυσκολία καθορισμού του ορίου διαχωρισμού των κατηγοριών, σε συνδυασμό με τη δυσκολία ερμηνείας των αποτελεσμάτων, καθιστούν το συγκεκριμένο υπόδειγμα ιδιαίτερα δύσχρηστο, τόσο από θεωρητικής όσο και από πρακτικής πλευράς. Για τους λόγους αυτούς η χρησιμοποίηση του γραμμικού υποδείγματος πιθανότητας είναι περιορισμένη, και συνεπώς το συγκεκριμένο υπόδειγμα δεν θα αναπτυχθεί περαιτέρω.

Το λογιστικό και το κανονικό υπόδειγμα πιθανότητας προερχόμενα από το χώρο της οικονομετρίας, παρουσιάζουν σημαντικές ομοιότητες μεταξύ τους. Τα υποδείγματα αυτά, αν και ιδιαίτερα παλιά[3], γνώρισαν ιδιαίτερη διάδοση μετά τη δεκαετία του 1970 και τις εργασίες του πρόσφατα βραβευμένου με Νόμπελ Οικονομίας, Daniel McFadden (1974, 1980) στην ανάπτυξη της θεωρίας της διακριτής επιλογής (discrete choice). Η θεωρία αυτή αποτέλεσε την αναγκαία θεωρητική βάση για την κατανόηση των βασικών εννοιών των εν λόγω προσεγγίσεων καθώς και για την ερμηνεία των υποδειγμάτων που αναπτύσσονται μέσω αυτών.

Τα δύο υποδείγματα οδηγούν στην ανάπτυξη μιας μη γραμμικής συνάρτησης βάσει της οποίας υπολογίζεται η πιθανότητα των εναλλακτικών δρα-

[3] Οι πρώτες ερευνητικές εργασίες σχετικές με την παρουσίαση του κανονικού και του λογιστικού υποδείγματος πραγματοποιήθηκαν στις δεκαετίες του 1930 και 1940, από τους Bliss (1934) και Berkson (1944) αντίστοιχα.

στηριοτήτων να ανήκουν σε κάθε μια από τις υπό εξέταση κατηγορίες. Η διαφορά των δύο υποδειγμάτων έγκειται στη μορφή της συνάρτησης που αναπτύσσεται. Πιο συγκεκριμένα, στο λογιστικό υπόδειγμα, χρησιμοποιείται η γνωστή λογιστική συνάρτηση, ενώ στο κανονικό υπόδειγμα χρησιμοποιείται η αθροιστική συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής. Έτσι, στην περίπτωση της ταξινόμησης σε δύο κατηγορίες, η πιθανότητα να ανήκει μια εναλλακτική δραστηριότητα x_j στην κατηγορία C_2 δίνεται από τις σχέσεις[4]:

Λογιστικό υπόδειγμα: $$P_j = F(a + \boldsymbol{bg}_j) = \frac{1}{1 + e^{-a-\boldsymbol{bg}_j}} \tag{2.1}$$

Κανονικό υπόδειγμα: $$P_j = f(a + \boldsymbol{bg}_j) = \int_{-\infty}^{a+\boldsymbol{bg}_j} \frac{1}{(2\pi)^{1/2}} e^{\frac{-z^2}{2}} dz \tag{2.2}$$

Ο υπολογισμός του σταθερού όρου a και του διανύσματος $\boldsymbol{b}$ το οποίο περιέχει τους συντελεστές των χαρακτηριστικών, πραγματοποιείται χρησιμοποιώντας τεχνικές μέγιστης πιθανοφάνειας, και πιο συγκεκριμένα μεγιστοποιώντας την ακόλουθη συνάρτηση:

$$\ln L = \sum_{\forall x_j \in C_2} \ln(P_j) + \sum_{\forall x_j \in C_1} \ln(1 - P_j)$$

Από τη μορφή της συνάρτησης αυτής γίνεται εμφανές ότι η εκτίμηση των παραμέτρων των δύο υποδειγμάτων ανάγεται σε ένα πρόβλημα μη γραμμικής βελτιστοποίησης, η επίλυση του οποίου είναι πολλές φορές ιδιαίτερα δύσκολη ιδίως στην περίπτωση του κανονικού υποδείγματος. Μάλιστα σε περιπτώσεις όπου είναι δυνατή η ανάπτυξη ενός γραμμικού συνδυασμού των χαρακτηριστικών $g_1, g_2, \ldots, g_n$ που να διαχωρίζει απόλυτα τις κατηγορί-

[4] Εάν στις δύο κατηγορίες C_1 και C_2 αντιστοιχηθεί μια δυαδική 0-1 μεταβλητή ως εξής: $C_1 \rightarrow 0$ και $C_2 \rightarrow 1$, τότε οι σχέσεις (2.1)-(2.2) αποδίδουν την πιθανότητα μια εναλλακτική δραστηριότητα να ανήκει στην κατηγορία C_2. Εάν η αντιστοίχηση πραγματοποιηθεί κατά τον αντίστροφο τρόπο ($C_1 \rightarrow 1$ και $C_2 \rightarrow 0$), τότε οι σχέσεις (2.1)-(2.2) αποδίδουν την πιθανότητα μια εναλλακτική δραστηριότητα να ανήκει στην κατηγορία C_1.

ες μεταξύ τους, τότε η διαδικασία βελτιστοποίησης δεν θα συγκλίνει με αποτέλεσμα να μην είναι δυνατός ο υπολογισμός των παραμέτρων τόσο του κανονικού, όσο και του λογιστικού υποδείγματος (Altman et al., 1981).

Σε αντίθεση με την περίπτωση της διακριτικής ανάλυσης, τόσο στο λογιστικό όσο και στο κανονικό υπόδειγμα η σημαντικότητα των επιμέρους χαρακτηριστικών στην πραγματοποίηση της ταξινόμησης είναι δυνατόν να εκτιμηθεί μέσω γνωστών στατιστικών ελέγχων όπως το *t*-τεστ, κατά παρόμοιο τρόπο με την πολλαπλή παλινδρόμηση.

Η ταξινόμηση των εναλλακτικών δραστηριοτήτων πραγματοποιείται βάσει των πιθανοτήτων που υπολογίζονται μέσω των δύο υποδειγμάτων. Πιο συγκεκριμένα, κάθε εναλλακτική δραστηριότητα ταξινομείται στη κατηγορία όπου η αντίστοιχη πιθανότητα είναι μεγαλύτερη. Έτσι εάν η πιθανότητα που υπολογίζεται από τις σχέσεις (2.1) και (2.2), να ανήκει μια εναλλακτική δραστηριότητα στην κατηγορία C_2 είναι μεγαλύτερη από 0,5, τότε η εναλλακτική δραστηριότητα εντάσσεται στην κατηγορία C_2, διαφορετικά εντάσσεται στην κατηγορία C_1.

Στην περίπτωση όπου η ταξινόμηση αφορά περισσότερες από δύο κατηγορίες τότε το λογιστικό και το κανονικό υπόδειγμα εφαρμόζονται υπό δύο μορφές: την πολλαπλή ονομαστική (multinomial) και τη διατεταγμένη (ordered). Η διαφορά μεταξύ των δύο περιπτώσεων έγκειται στον τρόπο με τον οποίο ορίζονται οι κατηγορίες, βάσει της λογικής που αναπτύχθηκε στο προηγούμενο κεφάλαιο. Έτσι τα διατεταγμένα υποδείγματα είναι χρήσιμα στην περίπτωση προβλημάτων διατεταγμένης ταξινόμησης, ενώ στην περίπτωση όπου ο ορισμός των κατηγοριών πραγματοποιείται ονομαστικά (nominal), τότε χρησιμοποιούνται τα αντίστοιχα ονομαστικά υποδείγματα.

Τα διατεταγμένα λογιστικά και κανονικά υποδείγματα οδηγούν στον υπολογισμό ενός διανύσματος συντελεστών $\boldsymbol{b}$ και ενός διανύσματος σταθερών όρων $\boldsymbol{a}$, βάσει των οποίων η πιθανότητα P_{kj} να ανήκει η εναλλακτική δραστηριότητα x_j στην κατηγορία C_k υπολογίζεται κατά τον τρόπο που παρουσιάζεται στον Πίνακα 2.1.

Πίνακας 2.1: Το διατεταγμένο λογιστικό και κανονικό υπόδειγμα πιθανότητας

Διατεταγμένο λογιστικό υπόδειγμα (ordered logit model)	$P_{1j} = F(a_1 + \boldsymbol{g}_j'\boldsymbol{b})$ $P_{2j} = F(a_2 + \boldsymbol{g}_j'\boldsymbol{b}) - F(a_1 + \boldsymbol{g}_j'\boldsymbol{b})$ . . . $P_{kj} = 1 - (P_{1j} + P_{2j} + ... + P_{k-1,j})$
Διατεταγμένο κανονικό υπόδειγμα (ordered probit model)	$P_{1j} = \int_{-\infty}^{a_1+\boldsymbol{g}_j'\boldsymbol{b}} f(z)dz$ $P_{2j} = \int_{a_1+\boldsymbol{g}_j'\boldsymbol{b}}^{a_2+\boldsymbol{g}_j'\boldsymbol{b}} f(z)dz$ . . . $P_{kj} = \int_{a_{k-1}+\boldsymbol{g}_j'\boldsymbol{b}}^{+\infty} f(z)dz$

Οι συναρτήσεις F και f είναι αντίστοιχα η λογιστική συνάρτηση και η αθροιστική συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής, όπως αυτές ορίστηκαν στην περίπτωση των δύο κατηγοριών. Οι σταθεροί όροι ορίζονται έτσι ώστε: $a_{k-1} > a_{k-2} > \ldots > a_2 > 0$ (a_1=0). Ο υπολογισμός των παραμέτρων των δύο υποδειγμάτων πραγματοποιείται μέσω τεχνικών μέγιστης πιθανοφάνειας, κατά τρόπο παρόμοιο με την περίπτωση των δύο κατηγοριών.

Σε αντίθεση με τα διατεταγμένα λογιστικά και κανονικά υποδείγματα, τα ονομαστικά οδηγούν στην ανάπτυξη ενός συνόλου διανυσμάτων συντελεστών $\boldsymbol{b}_k$ και σταθερών όρων $\boldsymbol{a}_k$, για κάθε κατηγορία C_k (k=1, 2, …, q).

Στην περίπτωση του ονομαστικού λογιστικού υποδείγματος, η πιθανότητα P_{kj} να ανήκει η εναλλακτική δραστηριότητα x_j στην κατηγορία C_k, υπολογίζεται βάσει της σχέσης:

$$P_{kj} = \frac{e^{\boldsymbol{g}_j'\boldsymbol{b}_k + a_k}}{\sum_{l=1}^{q} e^{\boldsymbol{g}_j'\boldsymbol{b}_k + a_l}}$$

Για λόγους κανονικοποίησης των εκτιμήσεων των παραμέτρων του υποδείγματος, τίθενται $\boldsymbol{b}_1$=0 και $a_1 = 0$, ενώ όλα τα υπόλοιπα $\boldsymbol{b}_k$ και a_k ($k = 2, \ldots, q$) υπολογίζονται μέσω τεχνικών μέγιστης πιθανοφάνειας.

Μεταξύ του κανονικού και του λογιστικού υποδείγματος, τόσο σε ερευνητικό όσο και σε πρακτικό επίπεδο συνήθως προτιμάται το δεύτερο. Συγκριτικά, το λογιστικό υπόδειγμα απαιτεί σημαντικά απλούστερες υπολογιστικές διαδικασίες βελτιστοποίησης για την εκτίμηση των παραμέτρων του, ενώ παράλληλα δεν έχει εντοπιστεί, σε ερευνητικό και πρακτικό επίπεδο κάποιο επιπλέον αξιοπρόσεκτο όφελος από τη χρήση του κανονικού υποδείγματος σε ότι αφορά την ακρίβεια των αποτελεσμάτων που παρέχει.

Τις τελευταίες δύο δεκαετίες, τα παραπάνω υποδείγματα γνώρισαν σημαντική διάδοση. Σε ορισμένους μάλιστα ερευνητικούς χώρους, όπως σε αυτόν της χρηματοοικονομικής διοίκησης, μια σειρά προβλημάτων της οποίας θα εξεταστούν στο Κεφάλαιο 6, το λογιστικό (κατά κύριο λόγο) και το κανονικό υπόδειγμα έχουν «αντικαταστήσει» ουσιαστικά τη διακριτική ανάλυση. Συχνά, ως βασικός λόγος της προτίμησης των ερευνητών προς τα υποδείγματα αυτά προβάλλεται το γεγονός ότι δεν υπόκεινται σε στατιστικούς περιορισμούς. Ουσιαστικά όμως, όπως φαίνεται και από την μορφή που έχουν τα δύο αυτά υποδείγματα [βλ. σχέσεις (2.1), (2.2) και Πίνακα 2.1], υποθέτουν ότι η πιθανότητα εμφάνισης της κάθε επιλογής (κατηγορίας) ακολουθεί μια συγκεκριμένη στατιστική κατανομή. Έτσι παρά το γεγονός ότι δεν πραγματοποιούνται υποθέσεις σχετικές με τις στατιστικές ιδιότητες των εξεταζόμενων δεδομένων, εξακολουθούν να υπάρχουν άλλες μορφές στατιστικών υποθέσεων.

Τέλος, αξιοσημείωτο είναι το γεγονός ότι η πλειοψηφία των ερευνών που έχουν πραγματοποιηθεί δεν έχουν δείξει κάποια σημαντική βελτίωση στην αποτελεσματικότητα των υποδειγμάτων ταξινόμησης που αναπτύσσονται μέσω του λογιστικού υποδείγματος, έναντι των υποδειγμάτων που αναπτύσσονται μέσω της διακριτικής ανάλυσης (στη γραμμική της μορφή). Χαρακτηριστικές είναι οι σχετικές έρευνες που πραγματοποιήθηκαν από τον Krzanowski (1975) καθώς και από τους Press και Wilson (1978).

2.3 Μη παραμετρικές προσεγγίσεις

Όπως προαναφέρθηκε κατά την περιγραφή της διακριτικής ανάλυσης, στην περίπτωση όπου οι επιδόσεις των εξεταζόμενων εναλλακτικών δραστηριοτήτων στα κριτήρια αξιολόγησης, ακολουθούν την πολυμεταβλητή κανονική κατανομή και οι πίνακες διακύμανσης-συνδιακύμανσης των επιδόσεων αυτών για κάθε κατηγορία δραστηριοτήτων είναι γνωστοί, τότε η διακριτική ανάλυση (στη γραμμική ή τετραγωνική της μορφή) οδηγεί στην ανάπτυξη του βέλτιστου υποδείγματος ταξινόμησης.

Στην πράξη όμως, οι στατιστικές ιδιότητες των εξεταζόμενων εναλλακτικών δραστηριοτήτων είναι συνήθως άγνωστες, καθώς πολλές φορές ο εντοπισμός του αντίστοιχου πληθυσμού είναι αδύνατος. Το γεγονός αυτό ώθησε πληθώρα ερευνητών στην ανάπτυξη μιας σειράς εναλλακτικών, μη παραμετρικών προσεγγίσεων ταξινόμησης. Οι προσεγγίσεις αυτές δεν βασίζονται σε στατιστικές υποθέσεις και συνεπώς αναμένεται ότι μπορούν να προσαρμόζονται ικανοποιητικά, ανάλογα με τα χρησιμοποιούμενα σύνολα δεδομένων, είτε ως γραμμικά υποδείγματα ταξινόμησης είτε ως μη γραμμικά υποδείγματα. Συνεπώς, τέτοιου είδους προσεγγίσεις παρέχουν αυξημένη ευελιξία στο χρήστη-αποφασίζοντα, «απαλλάσσοντάς» τον από την ανάγκη εντοπισμού και ανάλυσης των στατιστικών ιδιοτήτων των δεδομένων που αφορούν το εξεταζόμενο πρόβλημα. Στις παραγράφους που ακολουθούν παρουσιάζονται οι σημαντικότερες από αυτές τις μη παραμετρικές προσεγγίσεις και ο τρόπος με τον οποίο συμβάλουν στην αντιμετώπιση του προβλήματος της ταξινόμησης.