Υπολογιστική Στατιστική

Κατερίνα Ορφανογιαννάκη

Τμήμα Μαθηματικών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
korfanog@math.uoa.gr

2020-2021

Εισαγωγή

Η εξέλιξη της τεχνολογίας και η κατασκευή υπολογιστών με μεγάλη υπολογιστική ισχή δίνει τη δυνατότητα ανάπτυξης νέων μεθολογιών στη Στατιστική επιστήμη αλλά και την εφαρμογή και βελτίωση μεθοδογιών που είναι ήδη γνωστές. Σε τι χρησιμεύουν αυτές οι νέες μεθοδολογίες; Ενδεικτικά μπορούμε να αναφέρουμε:

- Στην εκτίμηση των παραμέτρων κάποιου μοντέλου. Σε πολλές περιπτώσεις οι εκτιμήτριες δεν γράφονται σε κλειστή μορφή, δηλαδή δεν υπάρχει ένας τύπος που τις υπολογίζει. Σάυτές τις περιπτώσεις χρησιμοποιούμε αριθμητικές μεθόδους υπολογισμού των παραμέτρων.
- Σε περιπτώσεις που δεν ικανοποιούνται οι υποθέσεις περί κανονικότητας του πληθυσμού. Τί γινεται στην περίπτωση που ο πληθυσμός ακολουθεί την εκθετική κατανομή και θέλουμε να κάνουμε στατιστική συμπερασματολογία για τον δειγματικό συντελεστή συσχέτισης μεταξύ δύο μεταβλητών;
- Σε περιπτώσεις που θεωρητικά αποτελέσματα είναι δύσκολο να προκύψουν.

Προσομοίωση

Προσομοίωση = Μέθοδοι Monte Carlo

Στις μεθόδους Monte Carlo ανήκει μία ευρεία συλλογή υπολογιστικών αλγορίθμων που βασίζονται σε επαναλαμβανόμενη τυχαία δειγματοληψία προκειμένου να προκύψουν αριθμητικά αποτελέσματα. Συχνά χρησιμοποιούνται όταν όταν είναι δύσκολη (ή αδύνατη) η χρήση άλλων μαθηματικών μεθόδων. Το κλειδί των μεθόδων αυτών είναι η γένεση τυχαίων μεταβλητών με συγκεκριμένες ιδιότητες ώστε να μιμούνται πραγματικές διαδικασίες.

Μεθοδολογίες που θα εξετάσουμε:

- Ελέγχους Monte Carlo
- Μέθοδο jackknife
- Μέθοδο cross-validation
- Μέθοδο bootstrap
- Αριθμητικές μεθόδους υπολογισμού
- Αλγόριθμο EM

Μέτρα σύγκρισης εκτιμητών

Από την Εκτιμητική υπάρχουν κάποιες ποσότητες που χρησιμεύουν για την αξιολόγηση εκτιμητριών. Αυτές ήταν η μεροληψία (Bias), η διακύμανση (Variance) και το μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE).
Πιο συγκεκριμένα για μια εκτιμήτρια $\hat{\theta}$ μιας άγνωστης παραμέτρου θ ορίζουμε τις εξής ποσότητες:

$$\begin{aligned} Bias(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ Var(\hat{\theta}) &= E\left[(\hat{\theta} - E(\hat{\theta}))^2\right] \\ MSE(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right] \end{aligned}$$

Μέτρα σύγκρισης εκτιμητών

Μπορεί κανείς να επιβεβαιώσει τη σχέση που συνδέει το MSE με τη μεροληψία και τη διακύμανση. Συγκεκριμένα

$$\begin{aligned} MSE(\hat{\theta}) &= E\left[(\hat{\theta}-\theta)^2\right] \\ &= E\left[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2\right] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

Όμως

$$\begin{aligned} [Bias(\hat{\theta})]^2 + Var(\hat{\theta})] &= \left[E(\hat{\theta})-\theta\right]^2 + E\left[(\hat{\theta}-E(\hat{\theta}))^2\right] \\ &= [E(\hat{\theta})]^2 - 2\theta E(\hat{\theta}) + \theta^2 + E(\hat{\theta}^2) - [E(\hat{\theta})]^2 \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

Σχέση που συνδέει τα μέτρα σύγκρισης εκτιμητών

Οπότε προκύπτει ότι:

$$MSE(\hat{\theta}) \quad = \quad [Bias(\hat{\theta})]^2 + Var(\hat{\theta})]$$

Ερμηνεία των μέτρων σύγκρισης εκτιμητών

Η μεροληψία μας δείχνει αν κατά μέσο όρο η εκτιμήτρια είναι η ίδια με την πραγματική τιμή (αμερόληπτη) ή αν διαφέρουν (μεροληπτική).
Η διακύμανση μας δείχνει πόση μεταβλητότητα έχει η εκτιμήτρια.
Το μέσο τετραγωνικό σφάλμα προσπαθεί να ισορροπήσει ανάμεσα στις δύο αυτές ποσότητες (καθώς μείωση της μεροληψίας πολλές φορές οδηγεί σε αύξηση της διακύμανσης και το αντίθετο).

Έλεγχοι υποθέσεων

- Η μηδενική και η εναλλακτική υπόθεση: Η μηδενική συνήθως περιέχει μια υπόθεση που θέλουμε να δούμε αν ισχύει. Η εναλλακτική είναι η υπόθεση προς την οποία πρέπει να κινηθούμε αν δούμε ότι η μηδενική υπόθεση δεν ισχύει.
- Η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε για να κάνουμε τον έλεγχο: πρέπει να έχει την ικανότητα να διακρίνει ανάμεσα στις δύο υποθέσεις.
- Το επίπεδο στατιστικής σημαντικότητας a το οποίο είναι η πιθανότητα να απορρίψουμε εσφαλμένα την μηδενική υπόθεση. Συνήθως διαλέγουμε $\alpha = 5\%$.
- Μια κριτική τιμή από την κατανομή που ακολουθεί η ελεγχοσυνάρτηση και η οποία είναι συνάρτηση και του επιπέδου στατιστικής σημαντικότητας. Συγκρίνουμε την κριτική τιμή και με την παρατηρούμενη τιμή της ελεγχοσυνάρτησης ώστε να καταλήξουμε για το αν θα απορρίψουμε τη μηδενική υπόθεση.

Έλεγχοι σημαντικότητας

Όταν μιλάμε για έλεγχο σημαντικότητας (significance testing) δεν καταλήγουμε σε μια απόφαση (απορρίπτω ή δεν απορρίπτω τη μηδενική υπόθεση) αλλά απαντούμε με μια πιθανότητα, το $p - value$, που είναι η πιθανότητα να πάρει η ελεγχοσυνάρτησή μας μια τιμή τόσο ακραία ή και περισσότερο ακραία από την παρατηρούμενη. Ουσιαστικά το $p - value$ μας δείχνει πόσο ισχυρή είναι η μηδενική υπόθεση.

Σχέση μεταξύ ελέγχων υποθέσεων και ελέγχων σημαντικότητας

Απορρίπτουμε τη μηδενική υπόθεση σε επίπεδο στατιστικής σημαντικότητας α αν το $p-value$ είναι μικρότερο από α. Για παράδειγμα αν $\alpha = 5\%$ και το $p-value$ που βρήκαμε είναι 0.10 δεν απορρίπτουμε τη μηδενική υπόθεση.

Έλεγχοι τυχαιοποίησης

Οι έλεγχοι τυχαιοποίησης αποτελούν μια ομάδα ελέγχων για τους οποίους δεν απαιτούνται σχεδόν καθόλου υποθέσεις. Επομένως μπορούν να χρησιμοποιηθούν εκεί που άλλοι κλασικοί έλεγχοι αποτυγχάνουν.

Παράδειγμα

5 ασθενείς υποβλήθηκαν σε δυο διαφορετικές θεραπείες Α και Β. Στην συνέχεια οι γιατροί με βάση κάποια ιατρική κλίμακα βαθμολόγησαν την πρόοδο των ασθενών. Τα δεδομένα είναι τα εξής:

Θεραπεία Α (2 ασθενείς): 1 , 2

Θεραπεία Β (3 ασθενείς): 3 , 5 , 9

Υπάρχει διαφορά ανάμεσα στις δύο θεραπείες;

Το πρόβλημα:

Το πρόβλημα έχει να κάνει με το αν δύο μέσες τιμές διαφέρουν ή όχι, δηλαδή:

$H_0 : \mu_{\mathrm{A}} = \mu_{\mathrm{B}}$ έναντι της εναλλακτικής

$H_1 : \mu_{\mathrm{A}} \neq \mu_{\mathrm{B}}$

Η ελεγχοσυνάρτηση: $T = |\bar{a} - \bar{b}|$

Ο έλεγχος τυχαιοποίησης:

Ο έλεγχος τυχαιοποίησης βασίζεται στο γεγονός πως αν πάρουμε όλους τους δυνατούς τρόπους ώστε να μοιράσουμε τους 5 ασθενείς σε 2 ομάδες τότε βρίσκοντας για όλους αυτούς τους συνδυασμούς την τιμή της ελεγχοσυνάρτησης μπορούμε να κρίνουμε αν η τιμή που πήραμε είναι πραγματικά μεγάλη ή απλά οφείλεται στην τύχη.

Όλοι οι δυνατοί συνδυασμοί ασθενών σε 2 ομάδες και η τιμή της ελεγχοσυνάρτησης για καθένα από αυτούς.

Θεραπεία Α	Θεραπεία Β	$\bar{a}$	$\bar{b}$	Τ
1,3	2,5,9	2	5.33	3.33
1,5	2,3,9	3	4.67	1.67
1,9	2,3,5	5	3.33	1.67
1,2	3,5,9	1.5	5.67	4.17
2,3	1,5,9	2.5	5	2.5
2,5	1,3,9	3.5	4.33	0.83
2,9	1,3,5	5.5	3	2.5
3,5	1,2,9	4	4	0
3,9	1,2,5	6	2.67	3.33
5,9	1,,2,3	7	2	5

Ακριβής Έλεγχος Τυχαιοποίησης

- Βήμα 1ο: Θέσε τη μηδενική υπόθεση που δείχνει ότι δεν υπάρχει διαφορά (δεν υπάρχει κάποια δομή στα δεδομένα).
- Βήμα 2ο : Διάλεξε την ελεγχοσυνάρτηση, υπολόγισε την τιμή της t_{obs} για τα δεδομένα που έχεις.
- Βήμα 3ο: Δημιούργησε όλους τους δυνατούς συνδυασμούς δεδομένων και υπολόγισε την τιμή της ελεγχοσυνάρτησης για καθέναν από αυτούς.
- Βήμα 4ο: Υπολόγισε το $p - value$ ως

$$p - value = \frac{\text{αριθμός συνδυασμών με } \ T \geq t_{obs}}{\text{συνολικός αριθμός συνδυασμών}}$$

Γιατί ακριβής Έλεγχος Τυχαιοποίησης;

- Ακριβής: Παίρνουμε όλους τους δυνατούς συνδυασμούς που μπορούν να προκύψουν από τα δεδομένα μας
- Έλεγχος τυχαιοποίησης: μοιράσαμε τυχαία τις παρατηρήσεις στις ομάδες.

Παρατήρηση: Στον έλεγχο τυχαιοποίησης δεν μας ενδιαφέρει να βρούμε την κατανομή της ελεγχοσυνάρτησης θεωρητικά. Στην πράξη την κατασκευάζουμε εμείς. Στον πίνακα έχουμε όλες τις δυνατές τιμές της ελεγχοσυνάρτησης για τα δεδομένα μας. Δεν είναι πάντα εύκολο να δημιουργήσουμε όλα τα δυνατά δείγματα που προκύπτουν από την αναδιάταξη των δεδομένων.

Παράδειγμα II

A	3,4,4.3,4.5,4.8,5,6,5,7,8,9,10,12,13,14,15,16,17,22,23
B	20,24,26,28,29,31,32,34,34,35

Έστω ότι έχουμε 2 ομάδες και θέλουμε να ελέγξουμε αν οι δύο μέσοι είναι ίσοι (μηδενική υπόθεση) έναντι της εναλλακτικής ότι η ομάδα B έχει μεγαλύτερη μέση τιμή.

Ελεγχοσυνάρτηση:

$$T = \sum_{i=1}^{10} X_i^B$$

Από τον ακριβή έλεγχο τυχαιοποίησης στον προσεγγιστικό

Σε περιπτώσεις στις οποίες ο πλήρης υπολογισμός όλων των συνδυασμών δεν είναι εφικτός, μια λύση είναι να δημιουργήσουμε όχι όλους αλλά έναν αριθμό από τους δυνατούς συνδυασμούς με τυχαίο τρόπο και να χρησιμοποιήσουμε αυτές τις τιμές της ελεγχοσυνάρτησης ως μια εκτίμηση, πλέον, της κατανομής της ελεγχοσυνάρτησης για να κάνουμε τον έλεγχο.

Ο προσεγγιστικός έλεγχος τυχαιοποίησης διαφέρει από τον ακριβή στο ότι αντί να πάρουμε όλους τους συνδυασμούς παίρνουμε ένα τυχαίο δείγμα από αυτούς.

Προσεγγιστικός έλεγχος τυχαιοποίησης

- Βήμα 1ο: Θέσε τη μηδενική υπόθεση
- Βήμα 2ο: Διάλεξε την ελεγχοσυνάρτηση, υπολόγισε την τιμή της t_{obs} για τα δεδομένα που έχεις.
- Βήμα 3ο: Δημιούργησε k συνδυασμούς δεδομένων (αντί για όλους τους δυνατούς) και υπολόγισε την τιμή της ελεγχοσυνάρτησης για καθέναν από αυτούς
- Βήμα 4ο: Εκτίμησε το $p - value$ ως

$$p - value = \frac{m + 1}{\kappa + 1} \quad ,$$

όπου $m =$ αριθμός συνδυασμών με $T \geq t_{obs}$

Ερμηνεία για τον τρόπο υπολογίσμού του $p - value$

Παρονομαστής: Από τα k δείγματα παίρνουμε k τιμές από την κατανομή της ελεγχοσυνάρτησης. Όμως και η παρατηρηθήσα τιμή ανήκει στην κατανομή της ελεγχοσυνάρτησης. Επομένως έχουμε διαθέσιμες $k + 1$ τιμές.
Αριθμητής: Η παρατηρηθήσα τιμή (η $k + 1$) είναι σίγουρα ίση με την τιμή της ελεγχοσυνάρτησης. Άρα έχουμε m τιμές που προκύπτουν από τα k δείγματα και την παρατηρηθήσα.

Πως διαλλέγουμε το k;

Αν το $p - value$ που παίρνουμε είναι κοντά στο επίπεδο στατιστικής σημαντικότητας, πρέπει να αυξήσουμε το k μέχρι το 95% διάστημα εμπιστοσύνης ώστε να είναι ξεκάθαρη η απόφαση που θα πάρουμε, δηλαδή να μην περιέχει την τιμή του επιπέδου στατιστικής σημαντικότητας.
Αν το k είναι πολύ μικρό (πχ 20) τότε η ελεγχοσυνάρτηση μας έχει μεγάλη μεταβλητότητα και αυξάνει η πιθανότητα σφάλματος.
Αν το k είναι σχετικά μεγάλο μπορεί να χρησιμοποιηθεί ως εκτίμηση του $p - value$ ο λόγος m/k.

Τι συμβαίνει για μεγάλο k;

Για μεγάλο k μπορεί να χρησιμοποιηθεί ως εκτίμηση του $p - value$ ο λόγος m/k.
Όμως τι συμβαίνει με το m; Το m είναι τυχαία μεταβλητή που ακολουθεί τη διωνυμική κατανομή για k επαναλήψεις και πιθανότητα επιτυχίας p_{true}, οπότε $E(m) = kp_{true}$.
Όμως για μεγάλο k η διωνυμική κατανομή προσεγγίζει ακρετά γρήγορα την κανονική κατανομή. Οπότε ένα προσεγγιστικό 95

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/k}.$$

Τι συμβαίνει για μεγάλο k; Παρατηρήσεις:

- Η προσέγγιση της διωνυμικής από την κανονική δεν είναι σωστή αν το $p - value$ είναι πολύ κοντα στο 0.
- Στην περίπτωση όμως που το $p - value$ είναι πολύ κοντα στο 0 τότε σχεδόν βέβαια απορρίπτουμε τη μηδενική υπόθεση.
- Τα διαστήματα εμπιστοσύνης είναι χρήσιμα όταν είμαστε κοντά στο επίπεδο σημαντικότητας και θέλουμε να δούμε αν δεχόμαστε ή απορρίπτουμε τη μηδενική υπόθεση.

Παράδειγμα III

20 φοιτητές και 20 φοιτήτριες έγραψαν ένα διαγώνισμα στα Μαθηματικά και βαθμολογήθηκαν με άριστα το 60. Υπάρχει διαφορά στους μέσους όρους των δύο φύλων· Οι βαθμολογίες:

Αγόρια	39, 44, 43, 47, 39, 46, 43, 43, 47, 41,
	53, 60, 46, 45, 47, 53, 53, 57, 50, 46
Κορίτσια	53, 40, 53, 33, 52, 60, 49, 51, 44, 36,
	44, 49, 39, 53, 51, 37, 48, 47, 42, 37

Η ελεγχοσυνάρτηση: $T = \left|\bar{x}_A - \bar{x}_K\right|$

Οι τιμές της ελεγχοσυνάρτησης για το παράδειγμα:

0.00, 0.00, 0.00, 0.00, 0.05, 0.05, 0.05, 0.10, 0.10, 0.15, 0.15,
0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.25, 0.25, 0.25, 0.25, 0.25,
0.25, 0.30, 0.30, 0.30, 0.35, 0.35, 0.35, 0.35, 0.35, 0.40, 0.40,
0.40, 0.40, 0.45, 0.45, 0.45, 0.55, 0.55, 0.55, 0.60, 0.65, 0.70,
0.70, 0.70, 0.70, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.75, 0.85,
0.85, 0.85, 0.85, 0.85, 0.90, 0.90, 0.90, 0.95, 0.95, 0.95, 0.95,
1.00, 1.05, 1.05, 1.05, 1.10, 1.10, 1.15, 1.15, 1.15, 1.25, 1.30,
1.30, 1.35, 1.35, 1.35, 1.40, 1.45, 1.45, 1.50, 1.50, 1.50, 1.55,
1.55, 1.70, 1.95, 2.00, 2.00, 2.00, 2.05, 2.05, 2.05, 2.15, 2.55

Συμπεράσματα

- Η επιλογή της ελεγχοσυνάρτησης πρέπει να είναι τέτοια που να διακρίνει ανάμεση στην H_0 και την H_1.
- Δεν χρειάζεται να κάνουμε καμία υπόθεση σχετικά με τον μηχανισμό που δημιούργησε τα δεδομένα. Η μόνη υπόθεση που κάνουμε είναι ότι μπορούμε να αλλάζουμε ¨ταμπέλες¨ στις παρατηρήσεις μας.
- Η μηδενική υπόθεση δηλώνει ουσιαστικά απουσία δομής στα δεδομένα.
- Σε πολλές περιπτώσεις οι έλεγχοι τυχαιοποίησης έχουν την ίδια ή μεγαλύτερη ισχή από υπάρχοντες παραμετρικούς ελέγχους.