Πρώτη Σειρά Ασκήσεων

Αυτή είναι η πρώτη σειρά ασκήσεων. Η προθεσμία για την παράδοση είναι στις 15 Δεκεμβρίου 11:59 μ.μ. Παραδώστε Notebooks με τον κώδικα και τις αναφορές. Κάνετε export το Notebook σε HTML και παραδώστε και το HTML αρχείο. Για την Ερώτηση 1, μπορείτε να παραδώσετε και pdf με την απόδειξη, ή φωτογραφίες από χειρόγραφα. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Η παράδοση θα γίνει μέσω του ecourse. Λεπτομέρειες στη σελίδα Ασκήσεις του μαθήματος. Η άσκηση είναι **ατομική**.

Ερώτηση 1

Α. Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο Reservoir Sampling που περιγράψαμε στο μάθημα, ώστε να κάνει δειγματοληψία $\boldsymbol{K}$ αντικειμένων ομοιόμορφα τυχαία από ένα ρεύμα N αντικειμένων. Το κάθε αντικείμενο πρέπει να έχει πιθανότητα K/N να εμφανιστεί στο δείγμα. Ο αλγόριθμος σας θα πρέπει να δουλεύει με ένα μόνο πέρασμα στα δεδομένα διαβάζοντας τα αντικείμενα ένα-ένα, χωρίς προηγούμενη γνώση του μεγέθους του ρεύματος (το μέγεθος N), και να χρησιμοποιεί $O(K)$ μνήμη (υποθέστε ότι το μέγεθος του κάθε αντικειμένου είναι σταθερό). Αυτό σημαίνει ότι δεν μπορείτε να αποθηκεύσετε όλο το ρεύμα των δεδομένων στη μνήμη.

1. Περιγράψετε τον αλγόριθμο που διαλέγει ένα ομοιόμορφο δείγμα K αντικειμένων από ένα ρεύμα N αντικειμένων. Η περιγραφή του αλγορίθμου **δεν** πρέπει να είναι σε κώδικα ή ψευδοκώδικα, ούτε η περιγραφή του κώδικα σε φυσική γλώσσα. Στην περίπτωση αυτή η απάντηση σας μηδενίζεται. Η περιγραφή θα πρέπει να εξηγεί τη λογική του αλγορίθμου σε απλά Ελληνικά.
   Για παράδειγμα, αυτή είναι μία περιγραφή σε απλά Ελληνικά του αλγορίθμου για τη δειγματοληψία ενός αντικειμένου που είδαμε στην τάξη:
   Ο αλγόριθμος κρατάει χώρο για ένα αντικείμενο και ένα μετρητή με τον αριθμό των αντικειμένων που έχει δει. Διατρέχει τα αντικείμενα ένα-ένα όπως έρχονται από το ρεύμα. Όταν βλέπει το n-οστό αντικείμενο, το επιλέγει με πιθανότητα $\frac{1}{n}$ και το αποθηκεύει, αντικαθιστώντας το υπάρχον αντικείμενο (αν δεν είναι το πρώτο). Ενημερώνει τον μετρητή. Όταν ολοκληρωθεί το ρεύμα, επιστρέφει το αντικείμενο που έχει αποθηκεύσει.
2. Αποδείξτε ότι ο αλγόριθμος σας παράγει ένα ομοιόμορφα τυχαίο δείγμα, δηλαδή, για κάθε $i, 1 \leq i \leq N$, το i -οστό στοιχείο έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.
3. Γράψτε μία συνάρτηση **sample** σε **Python** που υλοποιεί τον αλγόριθμο σας. Η συνάρτηση σας θα πρέπει να παίρνει σαν όρισμα το όνομα ενός αρχείου και τον αριθμό K, και να επιστρέφει μια λίστα που κρατάει ένα δείγμα με K τυχαίες γραμμές από το αρχείο. Θα πρέπει να διαβάσετε το αρχείο γραμμή-γραμμή και να μην το φορτώσετε στη μνήμη. Χρησιμοποιήστε την συνάρτηση σας μέσα σε ένα πρόγραμμα για να πάρετε 10 τυχαίες γραμμές από το αρχείο *input.txt* που θα σας δίνεται. Εκτυπώστε τις γραμμές στο δείγμα.

Εξηγείστε την αντιστοίχιση μεταξύ της περιγραφής που δώσατε στο 1ο βήμα και του κώδικα σας. Για παράδειγμα, για τον αλγόριθμο που επιλέγει ένα μόνο αντικείμενο, θα μπορούσατε να γράψετε κάτι της μορφής: «*Στις γραμμές 2-3 γίνεται η επιλογή του n-οστού αντικειμένου με πιθανότητα* $\frac{1}{n}$».

Δημιουργείστε ένα Notebook με δύο κελιά κώδικα. Ένα με τα imports και τον ορισμό της συνάρτησης sample και ένα στο οποίο θα χρησιμοποιείτε την συνάρτηση σας στο αρχείο input.txt, θα αποθηκεύετε το αποτέλεσμα και θα το εκτυπώνετε. Μπορείτε να κατεβάσετε το αρχείο input.txt από την σελίδα Ασκήσεις. Προσθέστε και δύο κελιά κειμένου, ένα με την περιγραφή του αλγορίθμου (Βήμα 1), και ένα με την αντιστοίχιση μεταξύ κώδικά και περιγραφής. Μπορείτε να προσθέσετε την απόδειξη (Βήμα 2) σε ένα ξεχωριστό κελί, ή να την γράψετε ξεχωριστά και να παραδώσετε ένα pdf με το κείμενο (ή φωτογραφίες αν είναι χειρόγραφη). Παραδώστε το Notebook και το αρχείο input.txt, και το pdf (ή φωτογραφίες) με την απόδειξη αν υπάρχει.

Β. Στην τάξη περιγράψαμε τον αλγόριθμο Reservoir Sampling για τη δειγματοληψία ενός αντικειμένου από ένα ρεύμα αντικειμένων. Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο ώστε να κάνει **σταθμισμένη δειγματοληψία**. Υποθέτουμε ότι το κάθε αντικείμενο i έχει βάρος w_i. Θα τροποποιήσετε τον αλγόριθμο δειγματοληψίας ώστε από ένα ρεύμα αντικειμένων με βάρη, να επιλέγει ένα αντικείμενο με πιθανότητα ανάλογη του βάρους του αντικειμένου. Δηλαδή, αν τελικά το ρεύμα έχει N αντικείμενα, και το συνολικό βάρος τους είναι $W = \sum_{i=1}^{N} w_i$ το αντικείμενο i θα πρέπει να έχει πιθανότητα w_i/W να επιλεγεί, για κάθε $1 \leq i \leq N$. Όπως και με τον κλασικό Reservoir Sampling αλγόριθμο, το N και το W δεν είναι γνωστά εκ των προτέρων και ο αλγόριθμος θα πρέπει να δουλεύει με σταθερό χώρο μνήμης, ανεξάρτητο του N. Αποδείξτε την ορθότητα του αλγορίθμου σας.

Ερώτηση 2

Σας δίνεται το αρχείο “data.csv” το οποίο μπορείτε να κατεβάσετε από τη σελίδα Ασκήσεις. Το αρχείο έχει τρεις στήλες χωρισμένες με κόμμα, με ονόματα A, B, C, και 1000 γραμμές. Οι τιμές των B και C είναι συνάρτηση αυτών της A. Συγκεκριμένα, για κάθε τιμή x στη στήλη A, η αντίστοιχη τιμή στις στήλες B και C είναι $f_B(x) + \epsilon_B$ και $f_C(x) + \epsilon_C$ αντίστοιχα όπου ϵ_Bκαι ϵ_C είναι συναρτήσεις που παράγουν τυχαίο θόρυβο (διαφορετικό για κάθε x και για κάθε στήλη). Ο στόχος σας είναι να προσδιορίσετε τον **τύπο** των συναρτήσεων f_B και f_C.

Για να προσδιορίσετε τις συναρτήσεις, φορτώστε τα δεδομένα σε ένα Pandas data frame και δημιουργήστε γραφικές παραστάσεις των B και C ως προς το A, όπως είδαμε την τάξη, καθώς και όποια άλλη γραφική παράσταση χρειάζεστε. Παραδώστε ένα Notebook το οποίο θα περιέχει τον κώδικα για την επεξεργασία των δεδομένων, τις γραφικές παραστάσεις και τους υπολογισμούς που κάνατε, καθώς και μία αναφορά με τα συμπεράσματα σας.

Στο Notebook που θα παραδώσετε θα πρέπει να φαίνεται η προσπάθεια που κάνατε για να εξερευνήσετε τα δεδομένα. Ο στόχος δεν είναι να δοκιμάσετε διάφορες συναρτήσεις μέχρι να βρείτε κάποια που ταιριάζει κάπως με τα δεδομένα, και δεν μας ενδιαφέρει να βρούμε ακριβώς την συνάρτηση, αλλά τον **τύπο** της (λογαριθμική, πολυωνυμική, εκθετική, άλλη?).

Ερώτηση 3

Στην ερώτηση αυτή θα κάνετε διερευνητική ανάλυση (exploratory analysis) δεδομένων με δεδομένα για προϊόντα και reviews από την Amazon. Ο στόχος είναι να κάνετε κάποιες μετρήσεις πάνω στα δεδομένα, να βρείτε ενδιαφέρουσες συσχετίσεις και να ερευνήσετε κάποιες υποθέσεις. Επίσης, να εξασκηθείτε με την χρήση των Pandas για ανάλυση δεδομένων.

Θα χρησιμοποιήσετε (μερικά λόγια για κάποια από τα πεδία ή ένα link με την περιγραφή?)

Σας δίνονται δύο αρχεία με δεδομένα: (1) Το csv αρχείο "Cell_Phones_meta.csv" το οποίο περιέχει με μια συλλογή από προϊόντα-κινητά τηλέφωνα (cell phones) και τα meta-data/χαρακτηριστικά τους. (2) Το json αρχείο "Cell_Phones_and_Accessories_5.json" με τα reviews για (κάποια από) τα προϊόντα στο πρώτο αρχείο, καθώς και για επιπλέον προϊόντα που είναι accessories. Κάθε προϊόν χαρακτηρίζεται μοναδικά από το asin number το οποίο λειτουργεί σαν unique id. Φορτώστε τα αρχεία αυτά σε pandas dataframes και δουλέψτε με αυτά τα δεδομένα. Τα δεδομένα έχουν θόρυβο, οπότε όταν κοιτάτε κάποιο υποσύνολο των δεδομένων, ή δημιουργείτε κάποια νέα dataframes από τα δεδομένα θα πρέπει να φροντίζεται να αφαιρείτε duplicates, κενές τιμές, κλπ. Υπάρχουν σχετικές μεθόδοι στο Pandas.

Η άσκηση αποτελείται από τα παρακάτω κομμάτια. Ο στόχος είναι να υλοποιήσετε τα παρακάτω χρησιμοποιώντας κατά κύριο λόγο μεθόδους της βιβλιοθήκης Pandas (συν δικές σας συναρτήσεις που θα εφαρμόσετε με apply).

Α. Στο κομμάτι αυτό μας ενδιαφέρει να καταλάβουμε την κατανομή που ακολουθεί ο αριθμός των reviews (review count) που έχουν τα προϊόντα (αριθμός διαφορετικών reviewerID που έχουν βαθμολογήσει το κάθε διαφορετικό προϊόν-asin). Χρησιμοποιήστε μεθόδους από Pandas για να υπολογίσετε το review count. Στη συνέχεια, θα κάνετε τις εξής γραφικές παραστάσεις (plots):

1. Ένα ιστόγραμμα του review count με 100 κάδους (bins) χρησιμοποιώντας έτοιμη συνάρτηση της βιβλιοθήκης Pandas
2. Ένα ιστόγραμμα του **λογαρίθμου** του review count με 100 κάδους χρησιμοποιώντας πάλι μεθόδους της βιβλιοθήκης Pandas
3. Ένα scatter plot των review counts και την συχνότητα τους σε λογαριθμική κλίμακα και στους δύο άξονες
4. Υπολογίστε το cumulative frequency vector, δηλαδή για κάθε review count τον αριθμό των προϊόντων που έχουν τουλάχιστον τόσα reviews. Κάνετε μια γραφική παράσταση του cumulative frequency ως συνάρτηση του review count παίρνοντας λογαριθμική κλίμακα και στους δύο άξονες.
5. Το Zipf plot της κατανομής των review_counts. Το Zipf plot κατασκευάζεται έχοντας στον Y άξονα τις τιμές (review count στην περίπτωση μας) και στο X την τάξη (rank) των τιμών. Για παράδειγμα το μέγιστο review count έχει rank 1, το δεύτερο μεγαλύτερο 2, κλπ. Το plot θα είναι σε λογαριθμική κλίμακα και τους δύο άξονες.

Παρουσιάστε τα γραφήματα σας σε ένα grid 2×2 και σχολιάστε την κατανομή

Σημείωση: Δεν υπάρχει σαφές συμπέρασμα για την κατανομή που ακολουθούν οι πόντοι αλλά μπορείτε να κάνετε κάποιες παρατηρήσεις για το σχήμα της κατανομής. Μπορείτε επίσης να προσθέσετε κάποιο δικό σας

plot αν πιστεύετε ότι θα σας βοηθήσει. Τα βήματα 3,4 είναι πιο δύσκολο να υλοποιηθούν χρησιμοποιώντας Pandas (ειδικά το Βήμα 3), μπορείτε αν θέλετε να τα υλοποιήσετε μεταφέροντας τα δεδομένα σε λίστες.

Β. Στο κομμάτι αυτό θα εξετάσουμε πως εξελίσσονται τα ratings των προϊόντων στον χρόνο. Για να πάρετε την ημερομηνία των reviews θα χρησιμοποιήσετε το unixtime πεδίο και θα το μετατρέψετε σε DateTime object. Για το rating θα χρησιμοποιήσετε το πεδίο overall. Χρησιμοποιώντας την ημερομηνία του πρώτου review για κάθε προϊόν θα υπολογίσετε τον αύξοντα αριθμό του μήνα στον οποίο έγινε το κάθε review. (Υποθέστε ότι ο μήνας έχει 30 μέρες. Το πρώτο review είναι την μέρα 0. Όσα reviews είναι λιγότερο από 30 μέρες από το πρώτο review είναι στον πρώτο μήνα, όσα είναι 30-59 μέρες είναι στον δεύτερο, κ.ο.κ.). Κάνετε μια γραφική παράσταση της μέσης τιμής του rating σαν συνάρτηση του μήνα στον οποίο γράφηκε το review. Κάνετε prune τα δεδομένα αν χρειάζεται ώστε να βγαίνουν ενδιαφέρουσες οι γραφικές παραστάσεις. Εξηγήστε τις επιλογές σας. Στη γραφική παράσταση θα πρέπει να φαίνεται και το 95% confidence interval. Συνίσταται να χρησιμοποιήσετε την μέθοδο lineplot της Seaborn. Για την επεξεργασία των ημερομηνιών χρησιμοποιείστε μεθόδους της βιβλιοθήκης datetime. Για την εξαγωγή της ημερομηνίας του πρώτου review και τον υπολογισμό του μήνα του review χρησιμοποιείστε μεθόδους της βιβλιοθήκης Pandas.

Γράψετε τις παρατηρήσεις σας για τις γραφικές παραστάσεις.

Γ. Στο κομμάτι αυτό μας θα ερευνήσετε την εξής υπόθεση για τα δεδομένα: «Όσο πιο ακριβό είναι ένα προϊόν τόσο καλύτερο το μέσο rating του». Για να εξετάσετε την υπόθεση θα πρέπει καταρχάς να εξάγετε μια πραγματική αριθμητική τιμή από τα δεδομένα για την τιμή (price) του προϊόντος. Η στήλη που υπάρχει στα δεδομένα όπως είναι περιέχει πραγματικές τιμές, πραγματικές τιμές με το σύμβολο $ μπροστά, και Strings τα οποία είναι απλά λάθος. Θα πρέπει να βρείτε την πραγματική τιμή για την τιμή του προϊόντος όποτε γίνεται. Αν για το ίδιο asin έχετε πολλές διαφορετικές τιμές, θα πρέπει να το χειριστείτε με κάποιο τρόπο (προσδιορίστε πως). Στη συνέχεια θα πρέπει για κάθε asin να βρείτε το μέσο rating. Μπορείτε να διώξετε προϊόντα που δεν έχουν αρκετά reviews από τα δεδομένα σας.

Για να εξετάσετε την υπόθεση θα κάνετε ένα scatter plot των τιμών και των average ratings, και θα υπολογίσετε το Pearson Correlation Coefficient μεταξύ τιμής και average rating, και το p-value για το Pearson Correlation Coefficient. Κάνετε το ίδιο και παίρνοντας τον λογάριθμο της τιμής. Μπορείτε να εξετάσετε και κάποιο υποσύνολο των δεδομένων που θα επιλέξετε με βάση την τιμή. Γράψτε τις παρατηρήσεις σας και τα συμπεράσματα σας για την αρχική υπόθεση.

Δ. Στο κομμάτι αυτό θα εξετάσετε αν υπάρχει διαφορά μεταξύ των average ratings που παίρνουν προϊόντα από διαφορετικές μάρκες (brands). Από τα δεδομένα υπολογίστε για κάθε μάρκα τον αριθμό των προϊόντων (asin) που έχει, και κρατήστε τις πέντε μάρκες με τα περισσότερα προϊόντα. Υπολογίστε το average rating των προϊόντων και την μέση τιμή των average ratings για κάθε μάρκα. Ο στόχος μας είναι να εξετάσουν αν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές μεταξύ των διαφορετικών brands.

Για το στόχο αυτό δημιουργείστε καταρχάς ένα point plot που θα έχετε την μέση τιμή για κάθε brand και το 95% confidence interval. Προσπαθήστε να εξάγετε κάποιο συμπέρασμα αρχικά από το plot. Στη συνέχεια για κάθε ζευγάρι από brands χρησιμοποιείστε το t-test για να εξετάσετε αν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές μεταξύ των brands.

Περιγράψετε τα αποτελέσματα και τα συμπεράσματα σας.

Ε. Χρησιμοποιώντας μεθοδολογία παρόμοια με το Βήμα Γ, εξετάστε αν ισχύει η εξής υπόθεση: «Υπάρχει συσχέτιση μεταξύ του μήκους του review και του πόσο helpful θεωρείται». Για το helpfulness χρησιμοποιείστε τη στήλη vote. Μπορείτε να κοιτάξετε μόνο τα reviews με μη μηδενικό αριθμό από votes. Κάνετε μετρήσεις και με το λογάριθμο των votes.

Ζ. Χρησιμοποιώντας μεθοδολογία παρόμοια με το Βήμα Δ, εξετάστε αν ισχύει η εξής υπόθεση: «Υπάρχει στατιστικά σημαντική διαφορά μεταξύ των ratings για επιβεβαιωμένες αγορές (verified purchases) και μη».

Η. Διατυπώστε μια δική σας υπόθεση και εξετάστε την χρησιμοποιώντας τα δεδομένα. Η υπόθεση σας θα πρέπει να είναι κάτι μη τετριμμένο και να την εξετάσετε χρησιμοποιώντας (και) κάποιο στατιστικό τεστ.

Παραδώσετε ένα Notebook το οποίο θα περιέχει τον κώδικα για την επεξεργασία των δεδομένων, τις γραφικές παραστάσεις και τους υπολογισμούς που κάνατε, καθώς και τις παρατηρήσεις και τα συμπεράσματα σας. Βάλετε headers ώστε να ξεχωρίζουν τα διαφορετικά κομμάτια της άσκησης.