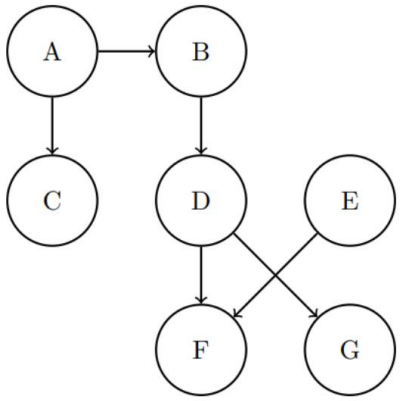
ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ – ΘΕΜΑΤΑ 2020 - 2021

[1] Δίνεται το ακόλουθο δίκτυο Μπεϋζιανών πεποιθήσεων. Γράψτε την κατανομή της από κοινού πιθανότητας ως ένα γινόμενο ανεξάρτητων όρων που εκφράζουν ανεξάρτητες υπό συνθήκη πιθανότητες οι οποίες προκύπτουν από τους πίνακες του Μπεϋζιανού δικτύου.



[2] Σε δεδομένα πραγματικού κόσμου, οι πλειάδες με τιμές που λείπουν για ορισμένα χαρακτηριστικά είναι ένα συνηθισμένο φαινόμενο. Ποιες από τις παρακάτω μεθόδους χρησιμοποιούμε για την αντιμετώπιση αυτού του προβλήματος;

☐ Χρήση μια καθολικής σταθεράς

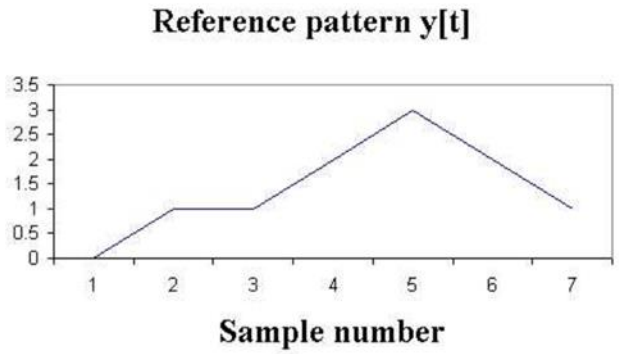
☐ Παράβλεψη της πλειάδας

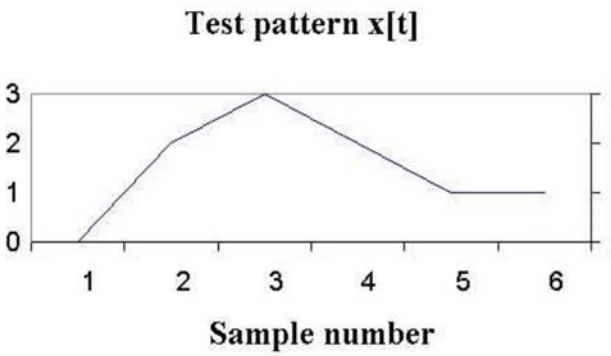
☐ Χρήση μέσου χαρακτηριστικού

☐ Χειροκίνητη συμπλήρωση της τιμής που λείπει

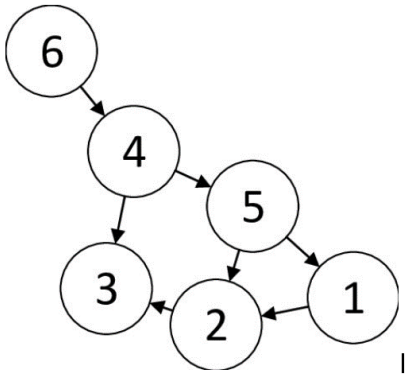
[3] Στην εξόρυξη δεδομένων με στόχο την προστασία της ιδιωτικότητας, περιγράψτε εν συντομία τί είναι η K-ανωνυμία (K-anonymity). Ποια είναι η διαφορά μεταξύ ήμι-αναγνωριστικού και αναγνωριστικού πεδίου;

[4] Δίνονται οι δύο ακόλουθες χρονοσειρές. Οι χρονοσειρές δεν έχουν απαραίτητα το ίδιο μήκος και οι μέγιστες τιμές δεν συμβαίνουν στον ίδιο αριθμό βημάτων. Σύμφωνα με τον αλγόριθμο Δυναμικής Χρονικής Στρέβλωσης (Dynamic Time Warping) σχεδιάστε σε πρόχειρο τον πίνακα απόστασης μεταξύ των δύο χρονοσειρών και στην απάντηση στο διαγώνισμα γράψτε την αλληλουχία των ελάχιστων αποστάσεων μεταξύ των δύο χρονοσειρών.





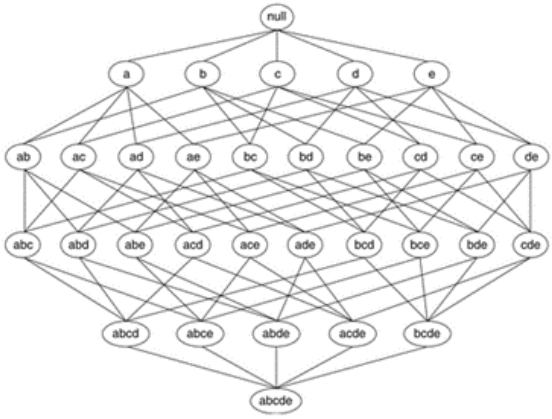
[5] Υπολογίστε τις πιθανότητες σταθερής κατάστασης των κόμβων του παρακάτω γράφου, όπως αυτές προκύπτουν από τον αλγόριθμο PageRank με πιθανότητα τηλεμεταφοράς α ίση με k/20, όπου k είναι το τελευταίο ψηφίο του αριθμού μητρώου σας (ή α=0,15 αν το τελευταίο ψηφίο του αριθμού μητρώου σας είναι 0).



[6] Έχουμε ένα σύνολο δεδομένων και θα χρησιμοποιήσουμε δέντρα αποφάσεων για να προβλέπουμε αν κάποιος σπουδαστής θα περάσει στο μάθημα της Εξόρυξης {Ναι, Όχι}. Έχουμε δύο χαρακτηριστικά, τον βαθμό στην πρόοδο {Υψηλός, Μεσαίος, Χαμηλός} και το αν διάβασε για την τελική εξέταση {Ναι, Όχι}. (χρησιμοποιούμε τα αρχικά των λέξεων) Ποια είναι η εντροπία της υπόθεσης “Πέρασε|Μελέτησε”;

Πρόοδος	Μελέτησε	Πέρασε
X	O	O
X	N	N
M	O	O
M	N	N
Y	O	N
Y	N	N

[7] Έστω ότι έχουμε 5 αντικείμενα (a, b, c, d, e) των οποίων όλοι οι πιθανοί συνδυασμοί απεικονίζονται στο παρακάτω διάγραμμα. Με βάση την αρχή Apriori, αν το στοιχειοσύνολο που αναφέρεται στον παρακάτω πίνακα (με βάση το τελευταίο ψηφίο του αριθμού μητρώου σας) είναι συχνό, τότε και ποια άλλα στοιχειοσύνολα είναι συχνά (δε λαμβάνουμε υπόψη μας την τετριμμένη περίπτωση του κενού στοιχειοσυνόλου);



Τελευταίο ψηφίο ΑΜ	Στοιχειοσύνολο
0	abc
1	abd
2	abe
3	acd
4	ace
5	ade
6	bcd
7	bce
8	bde
9	cde

[8] Με βάση τον πίνακα σύγχυσης, επιλέξτε ποιες επιλογές θα σας δώσουν σωστές προβλέψεις.

n=200	Predicted: NO	Predicted: YES
ACTUAL: NO	60	12
ACTUAL: YES	8	120

☐ Precision ≈ 0.9

☐ Recall ≈ 0.9

☐ True positive rate ≈ 0.85

☐ Accuracy ≈ 0.9

[9] Έστω κρυφό μαρκοβιανό μοντέλο με σύνολο καταστάσεων S = {s_1, s_2}, σύνολο συμβόλων Σ = {a,b,c}, αρχική πιθανότητα καταστάσεων Π = {1, 0}, πίνακα καταστάσεων, πιθανότητες μετάβασης καταστάσεων

$$P = \begin{pmatrix} 0.x & 1-0.x \\ 0.x & 1-0.x \end{pmatrix}$$

και πιθανότητα δημιουργίας συμβόλου σ_i στην κατάσταση s_j όπως παρακάτω

$$\theta^j(\sigma_j) = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

όπου x το τελευταίο ψηφίο του ΑΜ σας (ή 3 αν το τελευταίο ψηφίο του ΑΜ σας είναι 0). Ποια είναι η πιθανότητα παραγωγής της ακολουθίας V = bca από το μοντέλο;

[10] Σε ένα πρόβλημα εξόρυξης γνώσης από κείμενα, εξηγήστε σύντομα σε τί διαφέρει η εξαγωγή χαρακτηριστικών από την επιλογή χαρακτηριστικών, και δώστε κάποια παραδείγματα.

[11] Ας υποθέσουμε ότι έχουμε το ακόλουθο δισδιάστατο σύνολο δεδομένων: { x_1 = (1.5,1.7), x_2 = (2,1.9), x_3 = (1.6,1.8) }. Ταξινομήστε τα σημεία του δοθέντος συνόλου δεδομένων βάσει της ομοιότητας με ένα νέο σημείο δεδομένων x = (1.d4,1.d5) (όπου το d4 είναι το προτελευταίο ψηφίο του Α.Μ. σας ενώ το d5 είναι το τελευταίο - Α.Μ. : d1d2d3d4d5) χρησιμοποιώντας απόσταση Manhattan και δώστε τη σειρά ταξινόμησης του δοθέντος συνόλου δεδομένων (π.χ. x1, x2, x3).

[12] Έστω ότι έχουμε ροή δεδομένων (data stream) ακεραίων αριθμών και ότι καταγράφουμε τα στοιχεία που έχουν περάσει από αυτή με τη χρήση bloom filter και συναρτήσεων κατακερματισμού h1(x) = (kx + 11) mod 10 και h2(x) = (mx + 2) mod 10, όπου k, m το προτελευταίο και το τελευταίο ψηφίο του αριθμού μητρώου σας αντίστοιχα. Έστω ότι την χρονική στιγμή t η τιμή του bloom filter είναι η ακόλουθη: [0 1 1 1 0 1 0 0 0 1]. Να απαντήσετε αν έχει περάσει το στοιχείο x=1 από τη ροή δεδομένων. Σημείωση: Αν το προτελευταίο ψηφίο του αριθμού μητρώου σας είναι 0 θέστε k=4. Αν το τελευταίο ψηφίο του αριθμού μητρώου σας είναι 0 θέστε m = 7.

[13] Έστω ότι έχουμε ένα σύνολο δεδομένων k χαρακτηριστικών, όπου k το τελευταίο ψηφίο του αριθμού μητρώου σας (αν είναι 0 τότε θεωρείστε ότι k=5). Πόσα διαφορετικά υποσύνολα χαρακτηριστικών μπορούμε να εξάγουμε;

[14] Σε ένα αρχείο python δημιουργήστε ένα τυχαίο σύνολο δεδομένων d5.000 τιμών (όπου Α.Μ. : d1d2d3d4d5 και π.χ. αν d5=6 τότε το σύνολο θα περιέχει 6000 τιμές) με ένα χαρακτηριστικό (χρήση εντολής της Python np.random.randint(d3d4d5, size=(d5000, 1))- όπου αν d3=0 χρησιμοποιήστε μόνο το d4d5). α. Ποιο είναι το πλήθος των ακραίων σημείων που βρήκατε; β. Με ποια/ες εντολή/ες μπορούμε να δούμε με γραφικό τρόπο τις τιμές του χαρακτηριστικού του συνόλου δεδομένων καθώς και τις τυχόν ακραίες τιμές που τυχόν υπάρχουν στο σύνολο δεδομένων; γ. Ποιες εντολές μας δίνουν τις αριθμητικές τιμές των ακραίων τιμών του χαρακτηριστικού; δ. Ποιες είναι οι αριθμητικές τιμές των ακραίων τιμών που τυχόν περιέχει το σύνολο των δεδομένων σας;

[15] Ποιες από τις παρακάτω προτάσεις είναι ορθές για το υπολογιστικό μοντέλο του MapReduce;

☐ Ο τύπος των ζευγών κλειδιών-τιμής που δίνονται ως είσοδος σε ένα reducer πρέπει να είναι ίδιος με τον τύπο των ζευγών κλειδιών-τιμής της εξόδου του.

☐ Η είσοδος των shufflers είναι ομαδοποιημένη σύμφωνα με την τιμή του κλειδιού.

☐ Το κατανεμημένο σύστημα αρχείων HDFS είναι κατάλληλο για την αποθήκευση αρχείων μεγάλου μεγέθους.

☐ Η διαδικασία της μείωσης (reduce) μπορεί να ξεκινήσει με την ολοκλήρωση της διαδικασίας της απεικόνισης (map).

[16] Στην απλή λογιστική παλινδρόμηση έχουμε τη σχέση $y = b_0 + b_1x$. Τι συμβολίζει το y;

☐ Μια ανεξάρτητη μεταβλητή.

☐ Την εκτιμώμενη κλίση.

☐ Το σημείο τομής του y με την ευθεία παλινδρόμησης.

☐ Τη μέση προβλεπόμενη τιμή.

[17] Θέλουμε να εκτιμήσουμε την τάση συσταδοποίησης ενός χώρου δεδομένων με τη χρήση του στατιστικού Hopkins. Για το σκοπό αυτό, δειγματοληπτούμε p σημεία από το χώρο, το άθροισμα των οποίων από το πλησιέστερο κέντρο τους είναι 55 και επίσης δειγματοληπτούμε άλλα p σημεία ομοιόμορφα τυχαία, το άθροισμα των οποίων από το πλησιέστερο κέντρο τους είναι ίσο με τα δύο τελευταία ψηφία του αριθμού μητρώου σας. Να εξηγήσετε εν συντομία αν υπάρχει δομή στάδων στο συγκεκριμένο χώρο.

[18] Ποια/ες από τις ακόλουθες δηλώσεις ισχύουν για τους ταξινομητές βάσης που χρησιμοποιούν στις μεθόδους συνόλου :

☐ Έχουν υψηλή διακύμανση.

☐ Έχουν υψηλή μεροληψία, οπότε δεν μπορούν να λύσουν πολύπλοκα προβλήματα.

☐ Συνήθως δεν κάνουν overfit.

[19] Έστω ότι θέλουμε να δειγματοληπτήσουμε ροή δεδομένων (data stream), λαμβάνοντας υπόψη την εννοιολογική ολίσθηση, με τη χρήση εκθετικής συνάρτησης μεροληψίας. Ποιο είναι το όριο για τον βαθμό της μεροληψίας, αν το μέγεθος του δείγματος είναι ίσο με τα δύο τελευταία ψηφία του αριθμού μητρώου σας (αν τα δύο τελευταία ψηφία του αριθμού μητρώου σας είναι μικρότερα από 10, προσθέστε σε αυτά τον αριθμό 21);

[20] Στην εξόρυξη δεδομένων με στόχο την προστασία της ιδιωτικότητας, περιγράψτε εν συντομία τί είναι η L-διαφορετικότητα (L-diversity). Ποια είναι η διαφορά μεταξύ L-diversity και K-anonymity;

[21] Έστω ότι έχουμε ένα σύνολο δεδομένων k χαρακτηριστικών, όπου k το τελευταίο ψηφίο του αριθμού μητρώου σας (αν είναι 0 τότε θεωρείστε ότι k=5). Πόσα διαφορετικά υποσύνολα χαρακτηριστικών μπορούμε να εξάγουμε;

[22] Ποιες από τις παρακάτω προτάσεις είναι λανθασμένες για το υπολογιστικό μοντέλο του MapReduce;

☐ Το πλήθος των ζευγών κλειδιού-τιμής που δίνονται ως είσοδος σε έναν reducer μπορεί να είναι ίσος με το πλήθος των ζευγών κλειδιού-τιμής που παράγει ο reducer στην έξοδό του.

☐ Ο τύπος των ζευγών κλειδιών-τιμής που δίνονται ως είσοδος σε ένα shuffler δεν είναι υποχρεωτικό να είναι ίδιος με τον τύπο των ζευγών κλειδιών-τιμής της εξόδου του.

☐ Η απεικόνιση (map) εφαρμόζεται σε τιμές που μπορεί να σχετίζονται με το ίδιο κλειδί.

☐ Η έξοδος των mappers είναι ομαδοποιημένη σύμφωνα με την τιμή του κλειδιού.

[23] Σε ένα μοντέλο διανυσματικής αναπαραστασης κειμένου, δείξτε με ένα παράδειγμα γιατί είναι προτιμότερη η χρήση του συνημιτόνου για την απόσταση, από την ευκλείδεια απόσταση.

[24] Έστω ότι έχουμε τον παρακάτω πίνακα συναλλαγών καλαθιού αγορών. Με βάση αυτόν τον πίνακα, να υπολογίσετε: α) την υποστήριξη των στοιχειοσυνόλων και β) i) την υποστήριξη και ii) την εμπιστοσύνη των κανόνων που αναγράφονται στον παρακάτω πίνακα, σύμφωνα με το τελευταίο ψηφίο του αριθμού μητρώου σας.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Τελευταίο ψηφίο ΑΜ	Στοιχειοσύνολο	Κανόνας
0	{a,b}	{a,b} → {e}
1	{b,d}	{b,d} → {a}
2	{d,e}	{d,e} → {c}
3	{c,e}	{c,e} → {b}
4	{b,e}	{b,e} → {a}
5	{c,d}	{c,d} → {e}
6	{b,c}	{b,c} → {d}
7	{a,e}	{a,e} → {c}
8	{a,d}	{a,d} → {b}
9	{a,c}	{a,c} → {d}

[25] Σε ένα πρόβλημα δυαδικής ταξινόμησης έχουμε τα ακόλουθα σύνολα χαρακτηριστικών και τιμών τους: Κλιματισμός = {Λειτουργικός, Χαλασμένος} Κινητήρας = {Λειτουργικός, Χαλασμένος} Χιλιόμετρα = {Πολλά, Μεσαία, Λίγα} Σκουριά = {Ναι, Όχι} Στόχος του ταξινομητή είναι να προβλέψει αν η αξία του οχήματος θα είναι υψηλή ή χαμηλή. Ο ταξινομητής με βάση κανόνες που διαθέτουμε έχει το ακόλουθο σύνολο κανόνων: α) Χιλιόμετρα = Πολλά -> Αξία = Χαμηλή β) Σκουριά = Όχι -> Αξία = Χαμηλή γ) Κλιματισμός = Λειτουργικός, Κινητήρας = Λειτουργικός -> Αξία = Υψηλή δ) Κλιματισμός = Χαλασμένος -> Αξία = Χαμηλή Για τον ταξινομητή αυτό θα χρειαστεί να διατάξουμε τους κανόνες; Θα χρειαστεί να έχουμε μια προεπιλεγμένη κλάση;

[26] Στην εξόρυξη δεδομένων με στόχο την προστασία της ιδιωτικότητας, τί ορίζουμε ως ήμι-αναγνωριστικό (quasi-identifier); Περιγράψτε εν συντομία τί είναι η K-ανωνυμία (K-anonymity).

[27] Ποιες από τις παρακάτω προτάσεις είναι ορθές για το υπολογιστικό μοντέλο του MapReduce;

☐ Η είσοδος των shufflers είναι ομαδοποιημένη σύμφωνα με την τιμή του κλειδιού.

☐ Ο τύπος των ζευγών κλειδιών-τιμής που δίνονται ως είσοδος σε ένα reducer πρέπει να είναι ίδιος με τον τύπο των ζευγών κλειδιών-τιμής της εξόδου του.

☐ Το κατανεμημένο σύστημα αρχείων HDFS είναι κατάλληλο για την αποθήκευση αρχείων μεγάλου μεγέθους.

☐ Η διαδικασία της μείωσης (reduce) μπορεί να ξεκινήσει με την ολοκλήρωση της διαδικασίας της απεικόνισης (map).

[28] Στην απλή λογιστική παλινδρόμηση έχουμε τη σχέση $y = b_0 + b_1x$. Τί συμβολίζει το b_0;

☐ Το σημείο τομής του y με την ευθεία παλινδρόμησης.

☐ Μια ανεξάρτητη μεταβλητή.

☐ Την εκτιμώμενη κλίση.

☐ Τη μέση προβλεπόμενη τιμή.

[29] Στην εξόρυξη δεδομένων με στόχο την προστασία της ιδιωτικότητας, περιγράψτε εν συντομία τί είναι η L-διαφορετικότητα (L-diversity). Τί ορίζουμε ως ένα αναγνωριστικό (dentifier) πεδίο;

[30] Στην απλή λογιστική παλινδρόμηση έχουμε τη σχέση $y = b_0 + b_1x$. Τί συμβολίζει το x;

☐ Την εκτιμώμενη ή προβλεπόμενη τιμή.

☐ Το σημείο τομής του y με την ευθεία παλινδρόμησης.

☐ Μια ανεξάρτητη μεταβλητή.

☐ Την εκτιμώμενη κλίση.