ΞΑΝΑΠΑΝΤΗΣΤΕ ΟΛΑ ΤΑ ΘΕΩΡΗΤΙΚΑ ΓΙΑΤΙ ΜΠΟΡΕΙ ΝΑ ΕΧΟΥΝ ΛΑΘΗ + ΟΣΑ ΔΕΝ ΕΙΝΑΙ ΑΠΑΝΤΗΜΕΝΑ

ΛΥΣΤΕ ΟΣΑ ΘΕΜΑΤΑ ΕΧΟΥΝ ΚΙΤΡΙΝΟ ΧΡΩΜΑ

ΕΠΑΛΗΘΕΥΣΤΕ ΤΑ ΛΥΜΕΝΑ ΘΕΜΑΤΑ

1. 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;.**

   ΕΝΔΕΙΚΤΙΚΗ ΛΥΣΗ
   Η L-diversity ειναι η προσπαθεια βελτιωσης καποιων θεματων του κ-anonymity θετοντας οτι καθε κλαση (χαρακτηριστικο) identifier θα εχει τουλαχιστον L διαφορετικές τιμες που θα είναι well-represented (δεν θα μπορουν να καθοριστούν μονοσήμαντα). Η διαφορά με το κ-anonymity ειναι οτι σε αυτό μπορεί να καταλήξουν να υπάρχουν k records πανω στα οποια παλι να μπορεί να γινει καποια επιθεση σε καποιο ευαισθητο χαρακτηριστικο (λογω της ελλειψης ποικολομορφιας αυτου).

2. 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;.**

   Τα υποσυνολα που μπορουν να προκυψουν απο κ διαφορετικά χαρακτηριστικά ειναι ολα τα υποσυνολα με ακριβως Ν στοιχεια οπου το Ν ειναι απο 1-κ αρα θελουμε για το κ=9 το που ειναι το ΑΜ μου ολες τις πιθανες 9αδες τις 8αδες 7αδες κτλπ... Αυτος ο αριθμός υπολογίζεται να ειναι 511. (2^k - 1)

3.1 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
- [x] 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
- [x] 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

3.2 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).

3.3 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).

3.4 Ποιες από τις παρακάτω προτάσεις είναι ορθές για το υπολογιστικό μοντέλο του MapReduce;.

- [ ] Η είσοδος στους reducers είναι ομαδοποιημένη σύμφωνα με την τιμή του κλειδιού
- [ ] Το πλήθος των ζευγών κλειδιού-τιμής που δίνονται ως είσοδος σε έναν mapper είναι ίσο με το πλήθος των ζευγών κλειδιού-τιμής που παράγει ο mapper στην έξοδό του
- [ ] Η διαδικασία της μείωσης (reduce) μπορεί να ξεκινήσει χωρίς να έχει ολοκληρωθεί η διαδικασία της απεικόνισης (map)
- [ ] Ο τύπος των ζευγών κλειδιών-τιμής που δίνονται ως είσοδος σε ένα reducer δεν είναι υποχρεωτικό να είναι ίδιος με τον τύπο των ζευγών κλειδιών-τιμής της εξόδου του

4. 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.

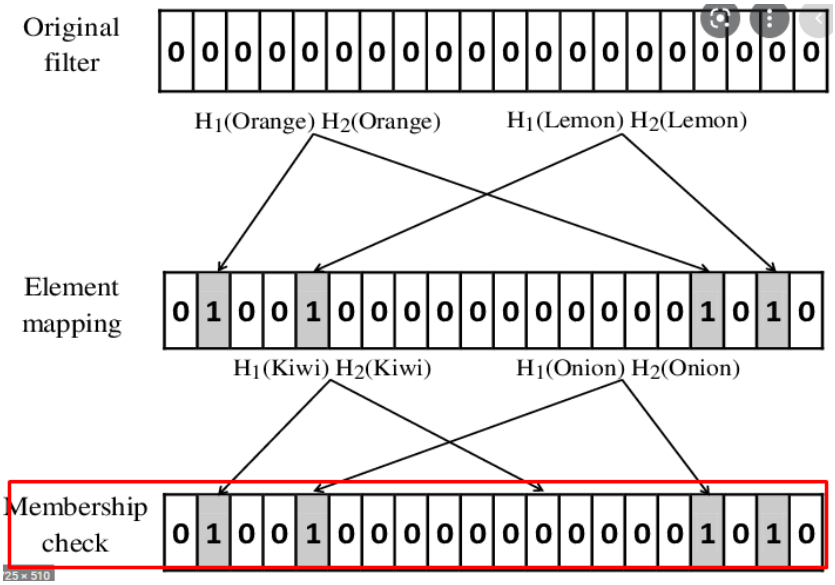
Οι συναρτησεις κατακερματισμου που προκυπτουν ειναι 8x+11 mod 10 και 9x+2

mod 10. **Για χ=1 h1(1) = 9 και h2(1)=1** επομενώς βλέπουμε οτι το bloom

filter έχει 1 σε αυτες τις θέσεις που σημαινει οτι το μονο που μπορουμε

να απαντήσουμε ειναι οτι ΜΠΟΡΕΙ να έχει περάσει. Αν έστω και ένα bloom(hash()) έβγαινε 0, τότε θα λέγαμε ότι δεν έχει περάσει το filter.

ΔΕΣ ΠΑΡΑΔΕΙΓΜΑ



5. Σε ένα μοντέλο διανυσματικής αναπαράστασης κειμένου, δείξτε με ένα παράδειγμα γιατί είναι **προτιμότερη η χρήση του συνημιτόνου για την απόσταση, από την ευκλείδεια απόσταση.**

Η απόσταση συνημιτονου ελεγχει την γωνια μεταξυ δυο vectors ενω η ευκλειδια αποσταση ελεγχει την αποσταση δυο σημειων. Επομενως επειδη διαχειριζομαστε μεγαλες διαστασεις στα κειμενα και στους κειμενικους χωρους η ευκλειδια αποσταση εχει την ταση να φερνει πιο κοντα την μεση αποσταση και την μεγιστη αποσταση μεταξυ τυχαιων σημειων. Επισης σε vectors που ειναι καθετα μεταξυ τους (δηλαδη τελειως ανομοια) η ευκλειδια αποσταση παραγει ομοιοτητα ενω το cosine similarity δίνει 0 (όπως πρεπει).

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 θα είναι η πρώτη μεταβλητή απόφασης διχοτόμησης και γιατι;.

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

Θελουμε τον κανονα που παράγει τις πιο αμειγώς διαχωρισμενες ομάδες επομένως αυτός ο κανόνας φαίνεται να είναι (μέσω υπολογισμών εντροπίας) ο Μέλετησε Ο καθως παραγει τις ομαδες (N-> [N,N,N] και O -> [O,O,N])

Πιο αναλυτικά και χωρίς υπολογισμούς. Θέλουμε την ελάχιστη εντροπία, δηλαδή όσο πιο καθαρό διαχωρισμό μεταξύ Περασε και δεν Περασε (0,N), δηλαδή μια μεταβλητή που δίνει όσο περισσότερα N μαζί ή Ο μαζί. Με το μάτι για την «Προοδο» το Χ και το Μ δεν δίνουν ξεκάθαρη εικόνα για το αν έχουμε Ν ή Ο, παρόλο που το Υ τα χωρίζει τέλεια (δίνει μόνο Ν). Άρα ας μελετήσουμε και τη «Μελέτησε» που τελικά όντως δίνει 3 Ν μαζί και 2 0 με 1 Ν, δλδ την χαμηλότερη εντροπία.

7.

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

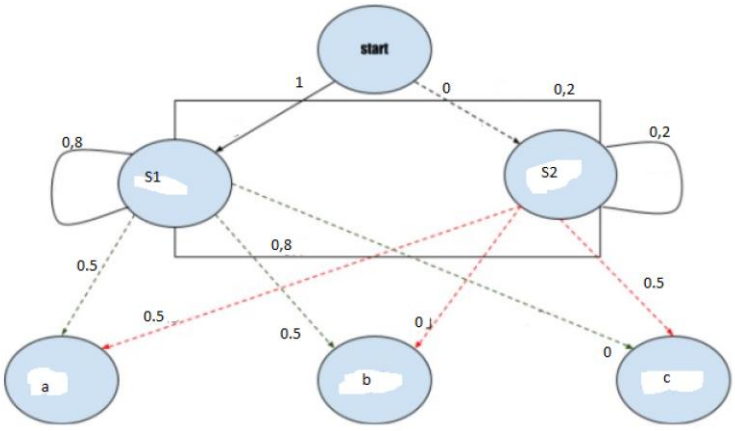
$$P = \begin{pmatrix} 0.x & 1-0.x \\ 0.x & 1-0.x \end{pmatrix}$$

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 ψηφίου 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;

Idea
https://analyticsindiamag.com/a-guide-to-hidden-markov-model-and-its-applications-in-nlp/



Λύση παλιού φοιτητή

P= [[0.8,0.2],[0.8,0.2]] (γραμμή 1 = s1, γραμμή 2 = s2)

Αρχικά Π = {1,0} δλδ ξεκινάω με S1

στο πρωτο βημα για την παραγωγη του b εχουμε πιθανότητες για την κάθε κατασταση [0.25, 0]. Για την παραγωγή του c εχουμε [0,0.025] και για την παραγωγη του a [0.01, 0.0025]. Επομενως η ολική πιθανότητα είναι 0.125 ???? entos I ektos ilis?

Δική μας

S1 S1 S1 = (Να περπατήσω S1 και B) *(Να πάω S1 δεδομένου ότι είμαι S1 * να πάω στο C) * (Να πάω S1 δεδομένου ότι είμαι S1 * να πάω στο C) = 0.5 * (0.8*0) * (0.8*0.5) = 0

Ομοίως όπου το S1 είναι 2ο, τότε μηδενίζονται οι πιθανότητες μετάβασης

S1 S1 S2 = 0

S1 S2 S1 = 0.5*(0.2*0.5)*(0.8*0.5) = 0.02

S1 S2 S2= 0.5*(0.2*0.5)*(0.2*0.5) = 0.005

S2 S1 S1 = 0

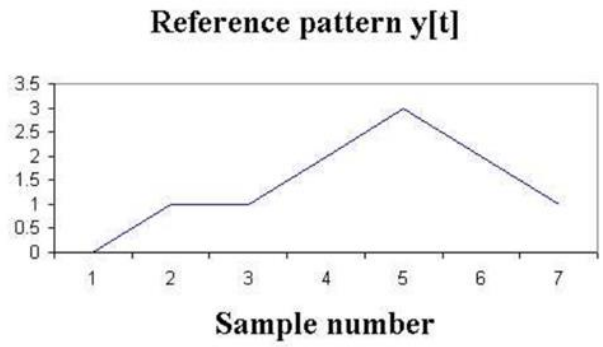
S2 S1 S2 = 0

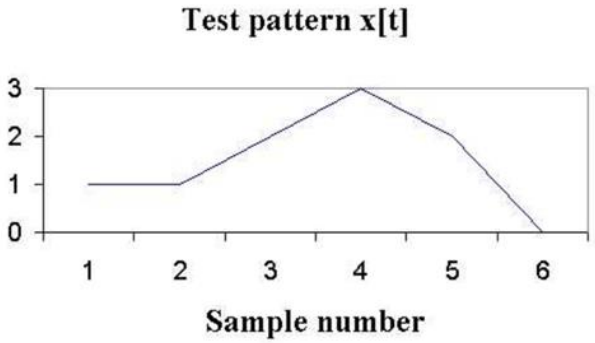
S2 S2 S1 = 0

S2 S2 S2 = 0

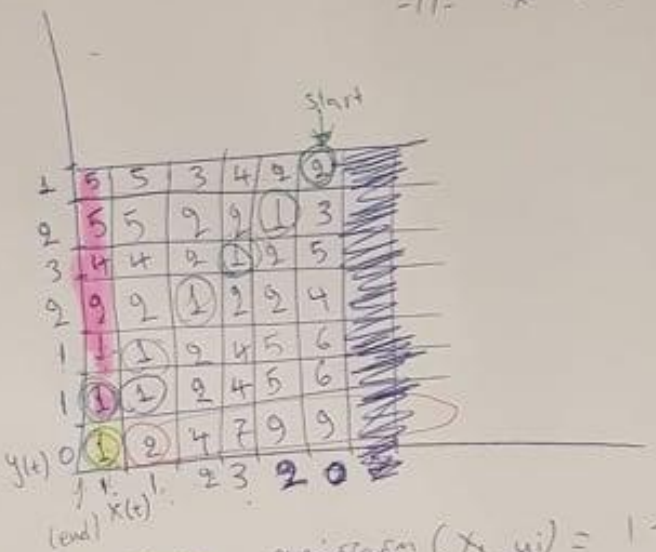
Άρα Ρολικό = Σ(p) = 0.025

8. 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.





Προσοχή! Δε τα ζωγραφίζω ανάποδα, παίρνω συμμετρικά results.
= αντίθετα στον άξονα y τη στιγμή y_t
-11- x τη στιγμή x_t



Εύρεση (1,1) = απόσταση $(x, y_i) = 1-0 = 1$

Εύρεση (1,2) = (-11-) + από κάτω τιμή $= (1-1)+1 = 1$

Ομοίως για στήλη (x, 1) ...

Εύρεση (2,1) = (-11-) + αριστερά τιμή $= (1-0)+1 = 2$

Ομοίως για γραμμή (1, y_i)

Εύρεση (2,2) και παίρνω επόμενό = (-11-) + (min τριπλέτας κάτω αριστερά δεξιών)

$= 0 + \min(1,1,2) = 1$

Μόλις συμπληρώσω τον πίνακα αποστάσεων πάω στο πάνω δεξιά κουτί (start) και κατεβαίνω το min της κάτω αριστερά τριπλέτας ως βέλτιστο μονοπάτι.

Άρα

((1,1), (2,1), (2,2), (3,2), (4,3), (5,4), (6,5), (7,6))

ή

((1,1), (2,2), (3,2), (4,3), (5,4), (6,5), (7,6))

[(1,1), (2,1), (2,2), (3,2), (4,3), (5,4), (6,5), (7,6)]

9. Με **βάση τον πίνακα σύγχυσης,** επιλέξτε ποιες επιλογές θα σας δώσουν σωστές προβλέψεις;.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	TN 50	FP 10
Actual: YES	FN 5	TP 100

1. Accuracy (all **correct** / all) = TP + TN / TP + TN + FP + FN
2. Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN
3. Precision (**true** positives / **predicted** positives) = TP / TP + FP
4. Sensitivity aka Recall (**true** positives / all **actual** positives) = TP / TP + FN
5. Specificity (**true** negatives / all **actual** negatives) =TN / TN + FP

Misclassification rate = δευτερεύουσα διαγώνιος/όλα = (5+10 / όλα)

False positive rate = πόσα προέβλεψε ως Ναι, ενώ πραγματικά είναι Όχι = (10 / όλα τα positive)

Sensitivity = Recall = True positive rate = TP / TP +FN

Accuracy = 1 – Misclassification ή κύρια διαγώνιος / όλα

Precision = TP / TP + FP

FNR = 1 – Sensitivity

FPR = 1- Specificity

- [ ] Misclassification rate ~ 0.91
- [ ] False positive rate ~0.95
- [x] Sensitivity ~0.95
- [x] Accuracy ~0.91

- [ ] Precision ~0.9
- [ ] Recall ~0.9
- [ ] True positive rate ~0.85
- [ ] Accuracy ~0.9

10. 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;.

- [ ] Παράβλεψη του χαρακτηριστικού
- [ ] Χρήση τυχαίων τιμών
- [x] Χρήση μέσης τιμής του χαρακτηριστικού
- [x] Παράβλεψη της πλειάδας

- [ ] Χρήση μια καθολικής σταθεράς
- [ ] Χρήση μέσου χαρακτηριστικού
- [ ] Χειροκίνητη συμπλήρωση της τιμής που λείπει
- [ ] Παράβλεψη της πλειάδας

11. Ποια/ες από τις ακόλουθες δηλώσεις ισχύουν για τους **ταξινομητές βάσης που χρησιμοποιούν στις μεθόδους συνόλου:**

- [ ] Έχουν χαμηλή διακύμανση.
- [ ] Έχουν χαμηλή μεροληψία, οπότε δεν μπορούν να λύσουν πολύπλοκα προβλήματα.
- [x] Συνήθως παρουσιάζουν overfitting

- [ ] Συνήθως δεν κάνουν overfit.
- [ ] Έχουν υψηλή μεροληψία, οπότε δεν μπορούν να λύσουν πολύπλοκα προβλήματα.
- [ ] Έχουν υψηλή διακύμανση.

12. 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}

Support των itemset ειναι το σε ποσα transactions εμφανιζονται μαζι

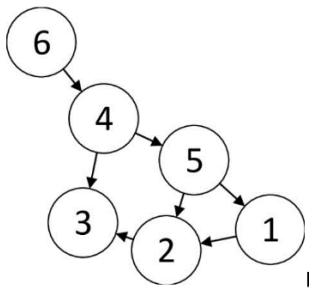
a)

ab 3/10,db 6/10,de 6/10,ce 2/10,be 4/10,cd 4/10,bc 3/10,ae 4/10,ad 4/10,ac 2/10

b) i) Ο κανονας 8 ειναι ad-> b ara support(adb) = 2/10

ii) ara trust = support(adb)/supp(ad) = (2/10)/(4/10) = 2/4 = ½

13. 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).



a = 9/20= 0.45 , n = 6, a/n = 0.075, (1-a) = 11/20 =0.55

6 out 1, 4 out 2, 3 out 0, 5 out 2, 2 out 1, 1 out 1 (εξερχόμενα βέλη)

page 6 = 0.07, page 4= 0.13, page 5 = 0.126, page 1 = 0.124, page 2 = 0.229, page 3=0.321

πολλαπλασιαζοντας με 1-a

προκύπτει οτι ειναι page 6 = 0.0385, page 4= 0.0715, page 5 = 0.0682 page 1 = 0.0558 page 2 = 0.12595 page 3=0.17655

14. 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;.**

Καθως το συνολο κανονων δεν ειναι εξαντλητικό (υπαρχουν instances που δεν θα πυροδοτησουν κανεναν κανονα) πρέπει να προστεθεί μια default class αλλα δεν ειναι αναγκαια η διαταξη των κανονων.

15. 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);.

Πρέπει κ<1/λ και επειδή το κ=98 (λογω του ΑΜ) λ<1/98.

16. 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.

-*****

17. Ας υποθέσουμε ότι έχουμε το ακόλουθο δισδιάστατο σύνολο δεδομένων:(x1(1.5,1), x2(1.2,1.5), x3(1.5,1.7)). 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 **Euclidean απόσταση** και δώστε τη **σειρά ταξινόμησης** του δοθέντος συνόλου δεδομένων (π.χ. x1, x2, x3).

Το σημειο που προκύπτει ειναι (1.9,1.8). Οι αποστάσεις ειναι

χ1-χ = sqrt((1.5-1.9)^2 + (1-1.8)^2)= sqrt(0.16 + 0.64) = 0.894

χ2-χ = 0.761

χ3-χ =0.412

αρα πιο ομοιο με το νεο στοιχειο ειναι το χ3 μετα το χ2 μετα το χ1

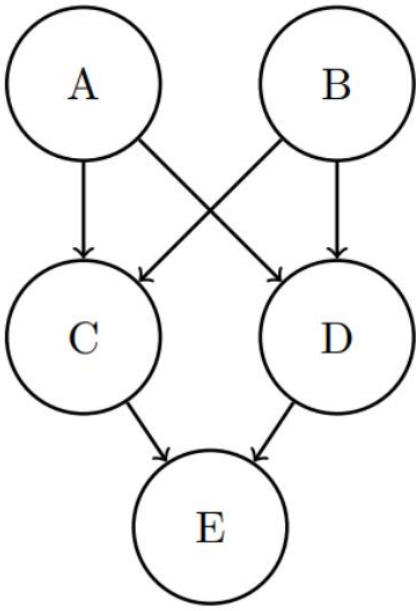
18. **Στην απλή λογιστική παλινδρόμηση** έχουμε τη σχέση y=b0 + b1*x. Τί συμβολίζει το **b0;.**

(3 Points)

- [ ] την εκτιμώμενη κλίση
- [ ] τη μέση προβλεπόμενη τιμή
- [x] το σημείο τομής του y με την ευθεία παλινδρόμησης
- [x] μια ανεξάρτητη μεταβλητή

21. Στην εξόρυξη δεδομένων με στόχο την προστασία της ιδιωτικότητας, τί ορίζουμε ως ήμι-αναγνωριστικό (quasi-identifier); Ποια είναι η διαφορά μεταξύ ήμι-αναγνωριστικού και αναγνωριστικού πεδίου; Περιγράψτε εν συντομία τί είναι η K-ανωνυμία (K-anonymity).

22. 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



P(A,B,C,D,E) = P(A) * P(B) * P(C/A,B) * P(D/A,B) * P(E/C,D) ??? Αυτό μόνο αρκεί;

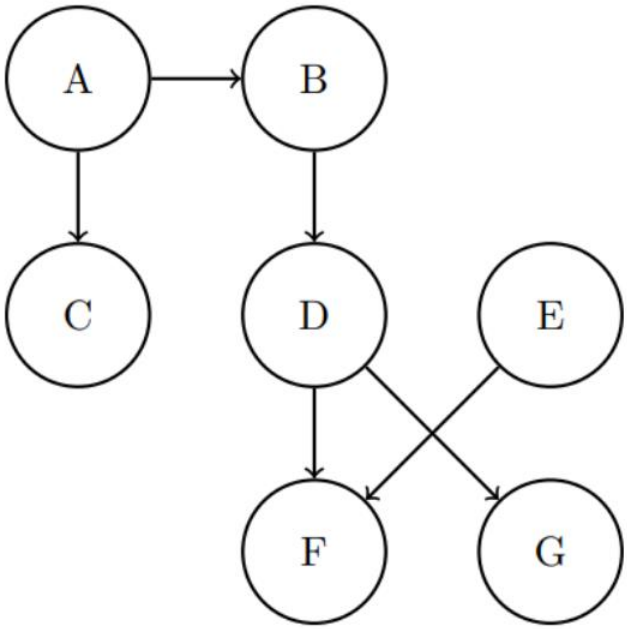
24. Σε ένα πρόβλημα εξόρυξης γνώσης από κείμενα, εξηγήστε τη διαφορά εξηγήστε σύντομα σε τί διαφέρει η εξαγωγή χαρακτηριστικών από την επιλογή χαρακτηριστικών, και δώστε κάποια παραδείγματα

25. Ας υποθέσουμε ότι έχουμε το ακόλουθο δισδιάστατο σύνολο δεδομένων:(x1=(1.5,1.7), x2(2,1.9), x3(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 A.M. σας ενώ το d5 είναι το τελευταίο- A.M. : d1d2d3d4d5) χρησιμοποιώντας απόσταση Manhattan και δώστε τη σειρά ταξινόμησης του δοθέντος συνόλου δεδομένων (π.χ. x1, x2, x3).

ΛΥΣΗ

$$|x_1 - x_2| + |y_1 - y_2|$$

27 . Δίνεται το ακόλουθο δίκτυο Μπεϋζιανών πεποιθήσεων. Υποθέτουμε ότι κάθε κόμβος μπορεί να πάρει 4 τιμές. Πόσες γραμμές έχει ο πίνακας του δικτύου για καθένα από του παράγοντες A, D και F;

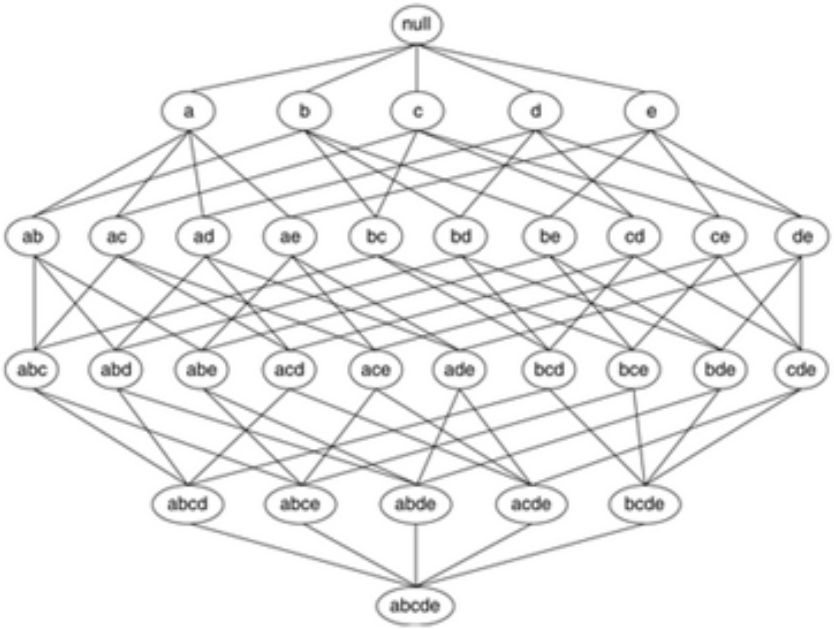


Πίνακας δικτύου για A = 4 ΓΡΑΜΜΕΣ

Πίνακας δικτύου D : Ο D μελετώντας για ένα βήμα εξαρτάται από τον B, άρα 4^2 = 16 γραμμές

Πίνακας δικτύου F: Ο F μελετώντας για ένα βήμα εξαρτάται από τους D, E άρα 4^3 = 64 (όλοι οι συνδυασμοί F,D,E) – ΠΡΟΣΟΧΗ ΑΝ ΕΧΩ ΛΑΘΟΣ ΤΟΤΕ ΘΕΛΕΙ ΑΠΑΝΤΗΣΗ 16+16=32

28. 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

ΕΣΤΩ ΑΜ =4 , άρα στοιχεισύνολο ace.

All subsets of a frequent itemset must be frequent(Apriori property).

If an itemset is infrequent, all its supersets will be infrequent.

Εντοπίζω το ace στον πίνακα και βρίσκω ποια σύνολα φτιάχνουν το ace.

ΆΡΑ ac, ae , ce και a,c,e

31. Σε ένα μοντέλο διανυσματικής αναπαραστασης κειμένου, κάποιες καλές έννοιες (concepts) θα είναι.

- [ ] Ορθογώνιες μεταξύ τους
- [ ] Ικανές να υπολογίσουν αυτόματα τα βάρη τους σε κάθε έγγραφο
- [ ] Βασισμένες σε γλωσσολογικές μελέτες
- [ ] Κατανοητές από τους ανθρώπους

32. Σε ένα σώμα N εγγράφων, διαλέγουμε τυχαία ένα. Αυτό περιέχει συνολικά T όρους, και ο όρος “Καλημέρα” εμφανίζεται K φορές. Ποια είναι η σωστή τιμή για το γινόμενο TF x IDF αν ο όρος “Καλημέρα” εμφανίζεται στο 1/3 των εγγράφων.

- [ ] KT * Log(3)
- [ ] K * Log(3) / T
- [ ] T * Log(3) / K
- [ ] Log(3) / KT

Δες και άλλες ασκήσεις Apriori και FP-growth.