Πολυτεχνείο Κρήτης
Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Μάθημα: **Αυτόνομοι Πράκτορες**

Εργασία Εξαμήνου
Πιθανοτικός Συμπερασμός Πραγματικών Ημερήσιων Κρουσμάτων με Χρήση Dynamic Bayessian Networks

Παπαδόπουλος Γιώργος
2016030132

1 Εισαγωγή

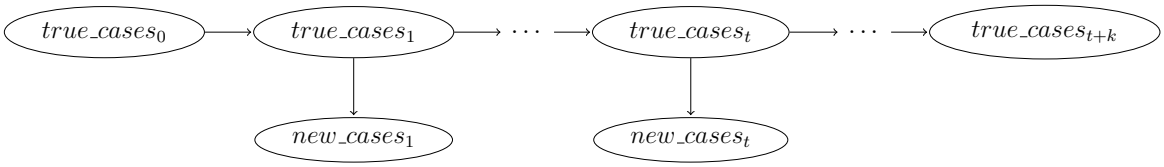
Η συγκεκριμένη εργασία αφορά την εκτίμηση της μεταβολής των πραγματικών κρουσμάτων σε μια χώρα (συγκεκριμένα στην Ελλάδα), δεδομένης της μεταβολής των ημερήσιων επιβεβαιωμένων κρουσμάτων, με χρήση Dynamic Bayessian Networks. Η απόφαση να χρησιμοποιηθούν στατιστικά μεταβολής και όχι ο απόλυτος αριθμός κρουσμάτων, πάρθηκε ώστε να «απλοποιηθεί» το πρόβλημα. Το ποσοστό μεταβολής χωρίστηκε σε 3 κλάσεις:

1. Μείωση κατά 5% ή περισσότερο (συμβολικά $< -5\%$)
2. Αύξηση κατά 5% ή περισσότερο (συμβολικά $> +5\%$)
3. Μεταβολή κατά το πολύ 5% (συμβολικά $\pm 5\%$)

Επιπλέον, στα δεδομένα έγινε smoothing, παίρνοντας τον μέσο όρο 3 ημερών, ώστε να μην επηρεάζεται το μοντέλο από τις απότομες μειώσεις σε ημερήσια κρούσματα τις μέρες που οι δειγματοληπτικοί έλεγχοι ήταν λίγοι (π.χ. Σαββατοκύριακα, αργίες).

2 Ανάλυση δομής του μοντέλου

Η δομή του DBN φαίνεται παρακάτω και αποτελεί ουσιαστικά ένα Hidden Markov Model $1^{ης}$ τάξης (πιο μετά θα δούμε γιατί θα έπρεπε να ήταν μεγαλύτερης τάξης), ενώ στην συνέχεια παρατίθενται και οι πίνακες κατανομών, με την απαραίτητη επεξήγηση.



Έχουμε μία μεταβλητή κατάστασης $\mathbf{X_t}$, η οποία αντιπροσωπεύει τα **πραγματικά κρούσματα** και μία μεταβλητή μαρτυρίας $\mathbf{E_t}$, η οποία αντιπροσωπεύει τα **επιβεβαιωμένα κρούσματα**. Για την εκ των προτέρων κατανομή της μεταβλητής κατάστασης, κάναμε την παραδοχή ότι θα ισούται με πιθανότητα 1, με την 3^{η} κλάση:

$$P(X_0) = \langle 0, 0, 1 \rangle$$

Για την δημιουργία του μοντέλου μετάβασης, το οποίο θα είναι ένας 3×3 πίνακας, χωρίσαμε την μονάδα σε ένατα. «Δώσαμε» την μεγαλύτερη τιμή πιθανότητας (2/9) σε κάθε μετάβαση στην οποία δεν αλλάζει η κλάση και στην συνέχεια με το ποσό πιθανότητας που μας έμεινε, εστιάσαμε στις αλλαγές κλάσεων από και προς την 3^{η} κλάση, που αποτελεί το μεταβατικό βήμα των άλλων 2.

$P(X_t \mid X_{t-1})$		$P(X_t)$		
		$< -5\%$	$> +5\%$	$\pm 5\%$
$P(X_{t-1})$	$< -5\%$	2/9	1/72	5/72
	$> +5\%$	1/72	2/9	5/72
	$\pm 5\%$	6/72	6/72	2/9

Για το μοντέλο παρατήρησης, λειτουργήσαμε με παρόμοιο τρόπο, δίνοντας ακόμα μεγαλύτερα βάρη στις τιμές της διαγωνίου, λόγω του ότι τα ημερήσια test που γίνονται αποτελούν αρκετά μεγάλο δείγμα ώστε να πούμε ότι υπάρχει μικρή πιθανότητα σφάλματος (το οποίο θα εντοπίζεται κυρίως στις «ακραίες» τιμές). Βάσει των παραπάνω ο πίνακας θα έχει ως εξής:

$P(E_t \mid X_t)$		$P(E_t)$		
		$< -5\%$	$> +5\%$	$\pm 5\%$
$P(X_t)$	$< -5\%$	5/18	1/72	2/72
	$> +5\%$	1/72	5/18	2/72
	$\pm 5\%$	3/72	3/72	5/18

Έχοντας ορίσει τις κατανομές που χρειαζόμαστε, είμαστε έτοιμοι να υπολογίσουμε την κατάσταση πεποίθησης $P(X_t \mid e_{1:t})$ (filtering) και την εκ των υστέρων κατανομή μελλοντικής κατάστασης $P(X_{t+k} \mid e_{1:t})$ (prediction).

3 Φιλτράρισμα

Βάσει της θεωρίας, για να υπολογίσουμε την κατάσταση πεποίθησης χρησιμοποιούμε την αναδρομική εξίσωση $f_{1:t+1} = \alpha D_{t+1} T^T f_{1:t}$, όπου $f_{1:t+1}$ η κατάσταση πεποίθησης για την χρονική στιγμή $t+1$, T το μοντέλο μετάβασης και D_t διαγώνιος πίνακας με τιμές $D_t^{i,i} = P(e_t \mid X_t = i)$. Οπότε, βάσει του μοντέλου μετάβασης, θα έχουμε 3 πίνακες D_t^1, D_t^2, D_t^3, έναν για κάθε κλάση που έχουμε και αντίστοιχα θα έχουν τα στοιχεία της $1^{\eta\varsigma}$, $2^{\eta\varsigma}$ και $3^{\eta\varsigma}$ γραμμής του μοντέλου.

$$D_t^1 = D_t^{<-5\%} = \begin{pmatrix} 5/18 & 0 & 0 \\ 0 & 1/72 & 0 \\ 0 & 0 & 2/72 \end{pmatrix}$$

$$D_t^2 = D_t^{>+5\%} = \begin{pmatrix} 1/72 & 0 & 0 \\ 0 & 5/18 & 0 \\ 0 & 0 & 2/72 \end{pmatrix}$$

$$D_t^3 = D_t^{\pm 5\%} = \begin{pmatrix} 3/72 & 0 & 0 \\ 0 & 3/72 & 0 \\ 0 & 0 & 5/18 \end{pmatrix}$$

Εφαρμόζω όλα τα παραπάνω στον κώδικα, ο οποίος είναι σε γλώσσα Matlab, και υπολογίζω τις κατανομές $P(e_t)$, βάσει των δεδομένων που πήρα και επεξεργάστηκα, ενώ σημειώνω και τον πίνακα D που αντιστοιχεί σε κάθε δείγμα/μέρα. Επίσης εκτελώ το φιλτράρισμα κάνοντας και κανονικοποίηση των τιμών, ώστε οι πιθανότητες να αθροίζουν στην μονάδα.

4 Πρόβλεψη

Τελευταίο βήμα (που πρόλαβα να κάνω) στον αλγόριθμο, ήταν η πρόβλεψη, για την οποία δημιούργησα ένα νέο excel αρχείο, που μέσα περιέχει δεδομένα για τις μέρες που πέρασαν από όταν ξεκίνησα το project, όταν και είχα δημιουργήσει το 1ο excel αρχείο. Για να γίνει πιο κατανοητή η αρχή και το τέλος των δεδομένων που επεξεργάζομαι, παρακάτω παρουσιάζω τις ημέρες στις οποίες αντιστοιχούν τα αρχικά δεδομένα, τα smoothed και οι ποσοστιαίες μεταβολές.

Δεδομένα στην παραγωγή του μοντέλου

	Αρχικά	Smoothed	% μεταβολή
από	13/03/2020	14/03/2020	14/03/2020
έως	15/02/2022	14/02/2022	14/02/2022

Δεδομένα στην πρόβλεψη

	Αρχικά	Smoothed	% μεταβολή
από	13/02/2022	14/02/2022	15/02/2022
έως	03/03/2022	02/03/2022	02/03/2022

Για το prediction, ξεκινάω από την πρώτη ημέρα πρόβλεψης, πολλαπλασιάζοντας την τιμή της δεσμευμένης πιθανότητας $P(X_{t+1}|e_{1:t})$ με το μοντέλο μετάβασης T, όπου t η τελευταία μέρα για την οποία έχουμε δεδομένα παρατήρησης, βγάζοντας έτσι την δεσμευμένη πιθανότητα $P(X_{t+2}|e_{1:t})$. Συνεχίζω στις επόμενες μέρες με το ίδιο μοτίβο, πολλαπλασιάζοντας την δεσμευμένη πιθανότητα της αμέσως προηγούμενης ημέρας με το μοντέλο μετάβασης, μέχρι να φτάσω στην τελευταία ημέρα.

5 Αποτελέσματα

Τα παραπάνω αποτελέσματα τα συγκρίνουμε με τα καταγεγραμμένα δεδομένα για τα επιβεβαιωμένα νέα κρούσματα των συγκεκριμένων ημερών και βλέπουμε ότι με ποσοστό 25% υπάρχει ταύτιση στην κλάση των πιθανοτήτων. Η πιθανότητα αυτή είναι ανάλογη του 33% που είναι η πιθανότητα να επιλεχθεί τυχαία μια από τις 3 κλάσεις, οπότε συμπεραίνουμε ότι η επιτυχία είναι καθαρά **τυχαία**.

6 Βελτιώσεις/Επίλογος

Εδώ είναι που ξεκινάει και η συζήτηση για περαιτέρω βελτίωση του μοντέλου, με πρώτη και βασικότερη αλλαγή που θα έκανα, να ήταν η αύξηση της τάξης του Hidden Markov Model, ώστε να μπορεί να «εντοπίζει» την κλίση της καμπύλης νέων κρουσμάτων, να έχει δηλαδή «μνήμη» του τι συνέβη τις προηγούμενες μέρες. Επιπλέον θα μπορούσαν να αυξηθούν οι κλάσεις στις οποίες χωρίζονται οι ποσοστιαίες μεταβολές ώστε να εντοπίζονται ακόμα μικρότερες μεταβολές στα κρούσματα και σε συνδυασμό με την προηγούμενη πρόταση να ενισχύεται η «μνήμη» του μοντέλου. Επιπλέον θα μπορούσαμε να αυξήσουμε τις μεταβλητές μαρτυρίας, χρησιμοποιώντας και άλλες καταγεγραμμένες μετρικές όπως οι ημερήσιοι θάνατοι, το reproduction rate (πόσους θα κολλήσει ένας νοσούντας), τις εβδομαδιαίες εισαγωγές ασθενών σε ΜΕΘ, το ποσοστό θετικότητας των ημερήσιων δειγματοληπτικών ελέγχων κ.λπ.. Να σημειώσουμε επίσης ότι αυτό θα δημιουργήσει επίσης την δυνατότητα αύξησης και τον μεταβλητών κατάστασης, κάνοντας έτσι το μοντέλο πιο «πλούσιο».

7 Παραπομπές

`https://github.com/owid/covid-19-data/tree/master/public/data`