Εργαστήριο Στοχαστικών Συστημάτων & Σημάτων – Εισαγωγή στο MATLAB

gmΕργαστηριακοί Συνεργάτες: **Πέτρος Μπίθας (**pbithas@ieee.org**) Αντώνης Μπέικος (**antonis.b@hotmail.co.uk**)**

Εξοικείωση Με Το Περιβάλλον Του MATLAB

Η παράδοση και εξέταση των αποτελεσμάτων σας θα γίνει την Δευτέρα-Τρίτη **11-12/11/2013**, αυστηρά στο τμήμα στο οποίο ανήκει ο κάθε φοιτητής.

Καμία εργασία δεν θα γίνει αποδεκτή μετά την καταληκτική ημερομηνία και ώρα δεν θα λαμβάνεται υπόψη στη διαμόρφωση της τελικής βαθμολογίας.

Παραδοτέα:

Πλήρης αναφορά με τον κώδικα, σχήματα, παρατηρήσεις σας κλπ. σε σχέση με κάθε υποερώτημα που τίθεται. Η παράδοση αναφοράς θα πρέπει να γίνεται έντυπα κατά την ημέρα διεξαγωγής του εργαστηρίου.

Βαθμολόγηση:

50%: υλοποίηση (να δίνει αποτέλεσμα)

40%: πλήρης και κατατοπιστικός σχολιασμός στην αναφορά

10%: αποτελεσματικός κώδικας (efficiency)

Υπενθυμίζεται ότι η αντιγραφή απαγορεύεται ρητά. Παρόλα αυτά η συνεργασία και ο γόνιμος προβληματισμός είναι ευπρόσδεκτα. Στην περίπτωση που χρησιμοποιήσετε οποιεσδήποτε πηγές ή αναφορές από βιβλία ή το διαδίκτυο θα ήταν καλό να τις αναφέρετε ως βιβλιογραφία.

Εργαστήριο Στοχαστικών Συστημάτων & Σημάτων – Εισαγωγή στο MATLAB

ΜΕΡΟΣ Α:

Το αρχείο spam_data.dat περιέχει μια συλλογή από λέξεις κλειδιά και σύμβολα που συγκεντρώθηκαν με σκοπό τη σχεδίαση ενός αξιόπιστου φίλτρου κατά της ανεπιθύμητης αλληλογραφίας (spam filter). Για το λόγο αυτό αναλύθηκαν 4601 μηνύματα ηλεκτρονικού ταχυδρομείου αποτελούμενα από 1813 μηνύματα spam και 2788 κανονικά μηνύματα. Από την παραπάνω ανάλυση μια προέκυψε μια λίστα με λέξεις κλειδιά, ειδικούς χαρακτήρες και άλλα χαρακτηριστικά τα οποία καλείστε να μελετήσετε. Διαβάστε τα περιεχόμενα του αρχείου `spam_data.dat` χρησιμοποιώντας την εντολή dlmread του MATLAB ως εξής:

```
data=dlmread('spam_data.dat');
```

Η παραπάνω εντολή θα δημιουργήσει έναν πίνακα `data`[1] 4601 x 8 ως εξής:

Κάθε σειρά του πίνακα αντιπροσωπεύει ένα μήνυμα ενώ οι στήλες(2-6) του πίνακα αντιπροσωπεύουν τη συχνότητα εμφάνισης (%) των λέξεων credit, you, re και των ειδικών χαρακτήρων $ και ; καθενός από τα 4601 μηνύματα. Η συχνότητα εμφάνισης των λέξεων έχει προκύψει ως εξής:

```
100*(πλήθος εμφανίσεων της λέξης στο συγκεκριμένο μήνυμα)/(πλήθος λέξεων στο μήνυμα)
```

ενώ η συχνότητα εμφάνισης των ειδικών χαρακτήρων:

```
100*(πλήθος εμφανίσεων χαρακτήρα στο συγκεκριμένο μήνυμα)/(πλήθος χαρακτήρων στο μήνυμα)
```

Επιπλέον, έχει παρατηρηθεί ότι στα μηνύματα spam μεγάλος αριθμός χαρακτήρων είναι γραμμένοι με κεφαλαία. Και μάλιστα για να διαφεύγουν από τα φίλτρα εντοπισμού στα μηνύματα αυτά έχει επίσης παρατηρηθεί ότι οι λέξεις δεν διαχωρίζονται μεταξύ τους με κενά. Για το λόγο αυτό η στήλη 6 περιέχει το πλήθος των κεφαλαίων χαρακτήρων ενός μηνύματος και η στήλη 7 το πλήθος της μεγαλύτερης μη διακοπτόμενης αλληλουχίας κεφαλαίων χαρακτήρων. Τέλος η τελευταία στήλη περιέχει 1 αν το μήνυμα είναι spam και 0 αν δεν είναι. Για καλύτερη κατανόηση όλων των παραπάνω δίνοται οι 6 πρώτες γραμμές και όλες οι στήλες τους:

# email	Credit	You	Re	$	;	Κεφαλαία	Αριθμός Χαρακτήρων	Κατηγοριοποίηση
1	0.0	1.93	0.0	0.0	0.0	61.0	278.0	1.0
2	0.0	3.47	0.0	0.18	0.0	101.0	1028.0	1.0
3	0.32	1.36	0.06	0.184	0.01	135.0	1259.0	1.0
4	0.0	3.18	0.0	0.0	0.0	40.0	191.0	1.0
5	0.0	3.18	0.0	0.0	0.0	40.0	191.0	1.0
6	0.0	0.0	0.0	0.0	0.0	15.0	54.0	1.0

[1] Το αρχείο spam_data.dat θα πρέπει να βρίσκεται στον ίδιο κατάλογο με το working directory στο Matlab.

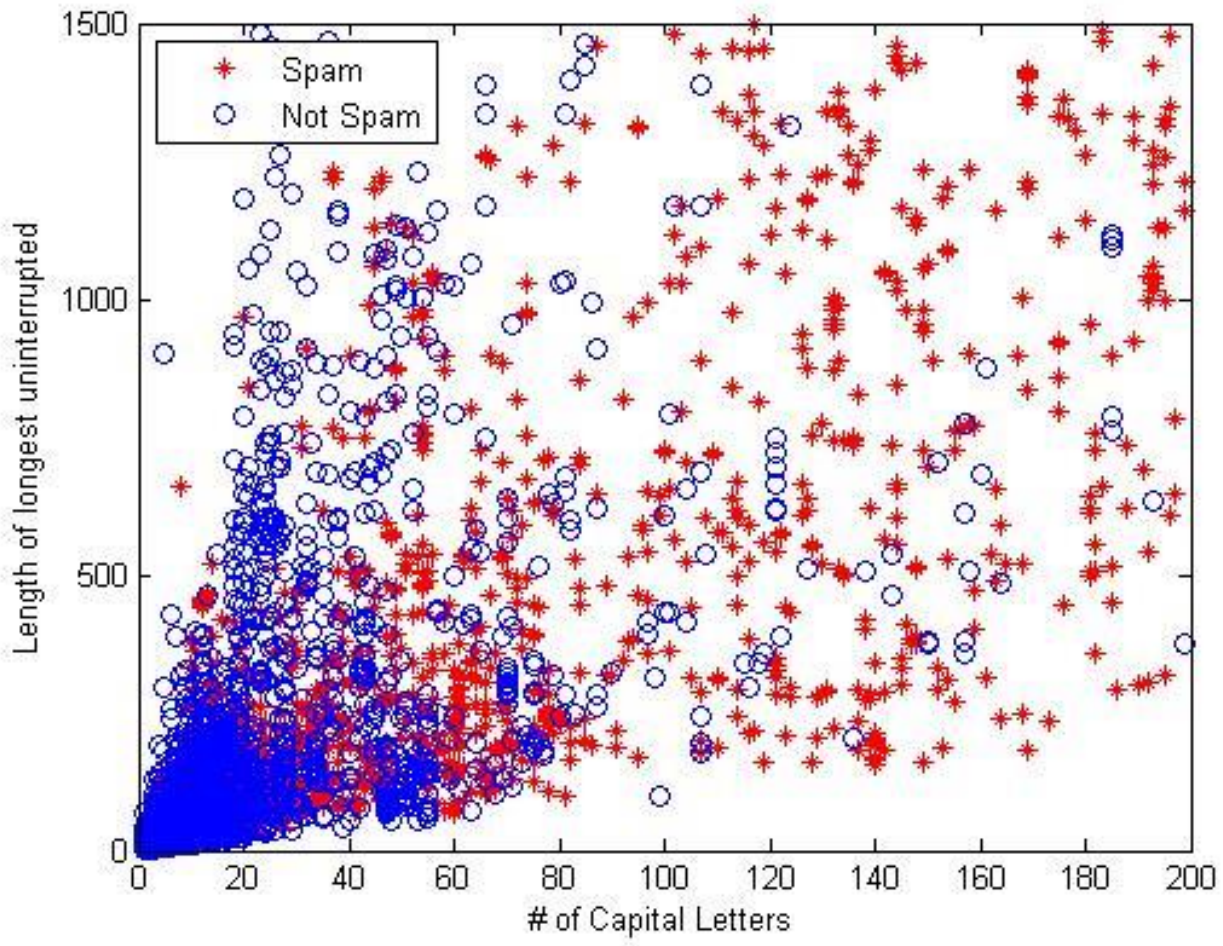
Εργαστήριο Στοχαστικών Συστημάτων & Σημάτων – Εισαγωγή στο MATLAB

I) Φτιάξτε ένα διάγραμμα διασποράς (scatter plot) όπου στον οριζόντιο άξονα να φαίνεται το πλήθος των κεφαλαίων χαρακτήρων (Στήλη 6) και στον κατακόρυφο το πλήθος χαρακτήρων της μεγαλύτερης μη διακοπτόμενης αλληλουχίας κεφαλαίων χαρακτήρων (Στήλη 7). Να κωδικοποιήσετε τα δεδομένα που αφορούν τα μηνύματα spam με κόκκινα σημεία και εκείνα που αφορούν τα κανονικά με μπλε. Τι συμπέρασμα βγάζετε;

Υπόδειξη

Για να διευκολυνθείτε, μπορείτε να χωρίσετε τον πίνακα data σε δύο πίνακες spam και notspam οι οποίοι θα περιέχουν τα δεδομένα που αφορούν τα spam και non-spam μηνυμάτων αντίστοιχα.

Μπορείτε να χρησιμοποιήσετε τις εντολές `plot`, `hold on`, `find`, `title`, `ylabel`, `xlabel`, `legend`. Το αποτέλεσμα θα πρέπει να είναι παρόμοιο με το παρακάτω:



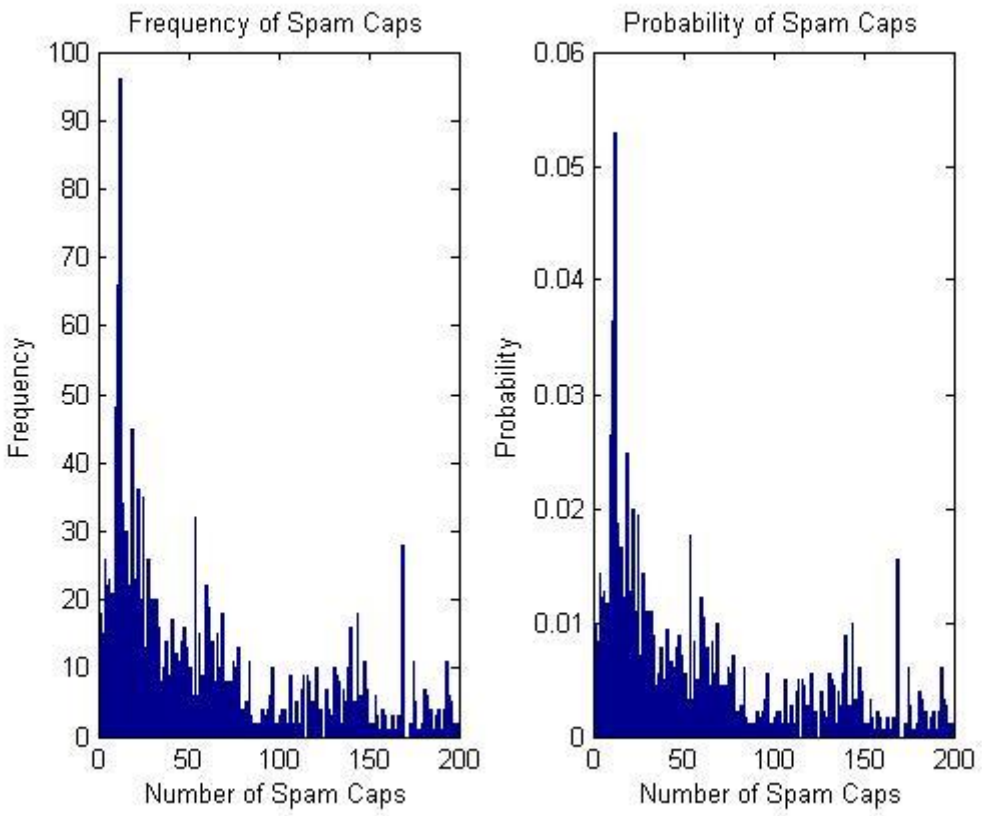
II) Υπολογίστε τα ιστογράμματα (δηλαδή τον αριθμό των δειγμάτων που παίρνουν μια συγκεκριμένη τιμή) για τα ακόλουθα (α) πλήθος κεφαλαίων (spam), (β) πλήθος κεφαλαίων (not spam), (γ) αριθμός χαρακτήρων (spam), (δ) αριθμός χαρακτήρων (not spam). Κανονικοποιήστε κατάλληλα ώστε το ιστόγραμμα να αναπαριστά πιθανότητα εμφάνισης.

Μπορείτε να χρησιμοποιήσετε τις εντολές unique, find, bar.

Εργαστήριο Στοχαστικών Συστημάτων & Σημάτων – Εισαγωγή στο MATLAB

Υπόδειξη

Στο ακόλουθο σχήμα φαίνεται στα αριστερά το ιστόγραμμα του αριθμού των κεφαλαίων που υπάρχουν στα spam email, (δηλαδή πόσα spam email έχουν 1 κεφαλαίο γράμμα, πόσα έχουν 2 κεφαλαία γράμματα κοκ). Το δεξί σχήμα είναι η κανονικοποιημένη εκδοχή του αριστερού όπου απεικονίζεται η πιθανότητα κάποιο από τα spam email να έχει 1 κεφαλαίο γράμμα, η πιθανότητα να έχει 2 κεφαλαία κοκ.



ΜΕΡΟΣ Β:

Α1. Κατασκευάστε συνάρτηση `funct1` η οποία θα δέχεται ως είσοδο ένα MxN λογικό πίνακα και η οποία θα επιστρέφει στην έξοδο κατάλληλες μεταβλητές σύμφωνα με την ακόλουθη κεφαλίδα:

```
function [sRow,sCol,sTot,oneT,zeroT]=funct1(A)
% A: MxN λογικός πίνακας (γεμάτος άσσους και μηδενικά)
% sRow: διάνυσμα με το άθροισμα των στοιχείων της κάθε γραμμής του A
% sCol: διάνυσμα με το άθροισμα των στοιχείων της κάθε στήλης του A
% sTot: άθροισμα όλων των στοιχείων του A
% oneT: πλήθος των λογικών 1 στον πίνακα
% zeroT: πλήθος των λογικών 0 στον πίνακα
```

Α2. Κατασκευάστε συνάρτηση `funct2` η οποία θα δέχεται ως είσοδο δύο πραγματικούς πίνακες `A` και `B` και η οποία αφού πραγματοποιήσει τις κατάλληλες συγκρίσεις θα επιστρέφει στην έξοδο τα διανύσματα `out1`, `out2`, `out3` ως εξής:

`out1`: θα περιέχει τις θέσεις όπου οι πίνακες `A` και `B` έχουν ίδιες τιμές

`out2`: θα περιέχει τις τιμές που είναι κοινές και στους δυο πίνακες (όχι απαραίτητα σε αντίστοιχες θέσεις)

`out3`: θα περιέχει τις τιμές των πινάκων που είναι μεγαλύτερες από το ημιάθροισμα των μεγίστων των δυο πινάκων

Στην περίπτωση που οι πίνακες δεν είναι των ίδιων διαστάσεων φροντίστε ώστε οι έξοδοι να πάρουν ως τιμή τον κενό πίνακα.

πχ

```
A =
     5     7     9
    17     2    11
     3    15     4

B =
     5     6    12
     4     2     3
     3     4    11

out1 =
     1     3     5

out2 =
     5     2     3     4    11

out3 =
    17    15
```

Για τον υπολογισμό του out3 λάβαμε υπόψη το ημιάθροισμα των δύο μεγίστων: `(17+12)/2=14.5`

Α3. Γράψτε συνάρτηση SecMinInd.m η οποία παίρνει ως είσοδο ένα διάνυσμα (vector) και επιστρέφει τη θέση του δεύτερου μικρότερου στοιχείου στο διάνυσμα. Για παράδειγμα, εαν το διάνυσμα V είναι [1, 5, 6, 8, 3, 5,2] τότε η συνάρτηση SecMινInd(V) θα πρέπει να επιστρέφει 7.

Βοηθήματα:

hist: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/hist.html

bar: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/bar.html

find: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/find.html

plot: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/plot.html

unique: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/unique.html

dlmread: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/dlmread.html

exp: http://www.mathworks.com/access/helpdesk/help/techdoc/ref/exp.html

histogram: http://en.wikipedia.org/wiki/Histogram