Άσκηση 1. Προ-επεξεργασία Δεδομένων

Σκοπός της άσκησης αυτής είναι η εξοικείωση των φοιτητών με σύνολα δεδομένων τα οποία θα πρέπει να υποστούν προ-επεξεργασία πριν χρησιμοποιηθούν σε αλγορίθμους Εξόρυξης Γνώσης.

Τα παραδείγματα που ακολουθούν παρουσιάζουν τον τρόπο που εντοπίζονται, «καθαρίζονται» ή διαγράφονται οι παρατηρήσεις με ελλιπείς τιμές σε ένα σύνολο δεδομένων. Επίσης στις τιμές των συνόλων δεδομένων που θα χρησιμοποιηθούν θα εφαρμοσθούν και κάποιες τεχνικές κανονικοποίησης ή τυποποίησης των δεδομένων.

Οι υπολογισμοί θα γίνουν στο υπολογιστικό περιβάλλον του Matlab. Για το λόγο αυτό δημιουργήστε ένα υποφάκελο με το ΑΕΜ σας στο δίσκο **D:\DataMining\AEM** όπου θα αποθηκεύετε τα αποτελέσματα και τα αρχεία της κάθε άσκησης.

Υπολογισμοί με τιμές NaN

Όταν επιχειρούνται υπολογισμοί με μεταβλητές που περιέχουν τιμές NaN (not a number) (ή Inf) τότε οι τιμές NaN πολλαπλασιάζονται με το τελικό αποτέλεσμα. Αυτό μπορεί να καταστήσει το αποτέλεσμα άχρηστο.

Παράδειγμα 1.1

Θεωρείστε έναν πίνακα 3x3 που είναι μαγικό τετράγωνο. Μαγικό τετράγωνο θεωρείται ο πίνακας στον οποίο το άθροισμα των στηλών είναι ίδιο με το άθροισμα των γραμμών.

```
a = magic(3);
sum(a)
sum(a')'

ans = 15   15   15
```

- Στον πίνακα a το κεντρικό στοιχείο να αντικατασταθεί με την τιμή NaN

```
a(2,2) = NaN

a =
     8     1     6
     3   NaN     7
     4     9     2
```

- Υπολογίστε ξανά το άθροισμα των στηλών του πίνακα a :

```
sum(a)

ans = 15   NaN    15
```

Παρατηρείτε ότι το άθροισμα των στοιχείων στη μεσαία στήλη έχει τώρα την τιμή NaN.

Εάν δε θέλετε να έχετε NaN στα τελικά σας αποτελέσματα θα πρέπει να εξαλείψετε τις τιμές αυτές από τα δεδομένα σας. Το ίδιο ισχύει και όταν έχετε τιμές inf (άπειρο) στα δεδομένα σας.

Απαλοιφή NaNs από τα δεδομένα

Μπορείτε να χρησιμοποιήσετε τη συνάρτηση του Matlab isnan για τον εντοπισμό των NaNs στα δεδομένα και στη συνέχεια να τα εξαλείψετε χρησιμοποιώντας τις τεχνικές του παρακάτω πίνακα.

εντολές	**περιγραφή**
`i = find(~isnan(x));` `x = x(i)`	Βρίσκει τα στοιχεία του διανύσματος που ΔΕΝ είναι NaNs και κρατάει μόνο τα non-NaN στοιχεία.
`x = x(~isnan(x));`	Αφαιρεί τα NaNs από το διάνυσμα x.
`x(isnan(x)) = [];`	Εναλλακτική μέθοδος που αφαιρεί τα NaNs από το διάνυσμα x.
`X(any(isnan(X),2),:) = [];`	Αφαιρεί μόνο τις γραμμές που περιέχουν τιμές NaNs από ένα πίνακα X.

Σημείωση θα πρέπει να χρησιμοποιείτε την isnan για τον εντοπισμό των NaNs και όχι τη λογική σύγκριση `NaN == NaN`

ΜΗ ΧΡΗΣΙΜΟΠΟΙΕΙΤΕ την εντολή `x(x==NaN) =[ ]` για την απαλοιφή των NaNs από τα δεδομένα σας.

Εάν υπάρχει συχνά η ανάγκη της απαλοιφής των τιμών θα πρέπει να δημιουργήσετε μια μικρή συνάρτηση στο matlab και να την καλείτε.

Παράδειγμα 1.2

```
function X = delNaNsRows(X)
X(any(isnan(X),2),:) = [];
```

Παρεμβολή τιμών σε ελλιπή δεδομένα (Interpolating Missing Data)

Μπορείτε να χρησιμοποιήσετε παρεμβολή για να βρείτε ενδιάμεσα σημεία στα δεδομένα σας. Η απλούστερη λειτουργία για την εκτέλεση παρεμβολής είναι interp1, η οποία είναι μια συνάρτηση παρεμβολής 1-D.

Εξ ορισμού, η μέθοδος παρεμβολής είναι «γραμμική», το οποίο ταιριάζει μια ευθεία γραμμή μεταξύ ενός ζεύγους των υπαρχόντων σημείων δεδομένων για τον υπολογισμό της ενδιάμεσης τιμής. Οι διαθέσιμες μέθοδοι, που μπορείτε να χρησιμοποιήσετε ως ορίσματα στη λειτουργία της `interp1`, είναι οι τα ακόλουθα:

- `'nearest'` `% Παρεμβολή του πλησιέστερου γείτονα (Nearest neighbor)`
- `'linear'` `% Γραμμική παρεμβολή (Linear interpolation)`
- `'spline'` `% Κυβική παρεμβολή spline (Piecewise cubic spline)`

Παράδειγμα 1.3

Τα παρακάτω δύο διανύσματα αντιπροσωπεύουν τα έτη απογραφής του πληθυσμού των Ηνωμένων Πολιτειών από το 1900 έως 1990. Το διάνυσμα αφορά τα έτη από το 1900 έως 1990 με βήμα 10 και το διάνυσμα p αφορά τον πληθυσμό σε εκατομμύρια ανθρώπους

```
t = 1900:10:1990;
p = [75.995  91.972  105.711  123.203  131.669...   %αλλαγή γραμμής
     150.697  179.323  203.212  226.505  249.633];
```

- Εκτελέστε την εντολή

```
interp1(t,p,1975)
```

Η έκφραση `interp1(t,p,1975)` παρεμβάλει δεδομένα μέσα στα στοιχεία της απογραφής για την εκτίμηση του πληθυσμού το έτος 1975. Το αποτέλεσμα είναι η παρακάτω τιμή.

```
ans =
    214.8585
```

- Κάνετε παρεμβολή δεδομένων στα στοιχεία της απογραφής για κάθε έτος από το 1900 έως το 2000 και προβάλετε τα δεδομένα σε γράφημα.

```
x = 1900:1:2000; %δημιουργούμε το διάνυσμα τιμών για τις οποίες θέλουμε
                 % να παρεμβάλουμε δεδομένα του πληθυσμού
y = interp1(t,p,x,'spline');
plot(t,p,'o',x,y)
```

Μερικές φορές είναι βολικότερο να εφαρμόσουμε παρεμβολή δεδομένων σε πίνακα και όχι σε διανύσματα. Έστω ένα τμήμα των δεδομένων απογραφής σε ένα ενιαίο 5 x 2 πίνακα.

```
tab =[ 1950    150.697
```

```
        1960    179.323
        1970    203.212
        1980    226.505
        1990    249.633]
```

Παρεμβάλετε δεδομένα μέσα στα στοιχεία της απογραφής του πίνακα tab για την εκτίμηση του πληθυσμού το έτος 1975.

```
p = interp1(tab(:,1),tab(:,2),1975)
p =
    214.8585
```

- Κάνετε παρεμβολή δεδομένων στα στοιχεία της απογραφής του πίνακα tab για κάθε έτος από το 1950 έως το 1990 και προβάλετε τα δεδομένα σε γράφημα.

```
x2 = 1950:1:1990;
y = interp1(tab(:,1),tab(:,2),x2,'spline');
plot(tab(:,1),tab(:,2),'o',x2,y)
```

- Επαναλάβετε την παραπάνω διαδικασία με διαφορετικά ορίσματα στη συνάρτηση της παρεμβολής

Παράδειγμα 1.4

Σας δίνεται ο παρακάτω πίνακας με ελλιπείς τιμές και καλείστε να δώσε μια λύση.

```
dataV =    [-0.3999        NaN          -1.0106
              0.6900        0.2573       0.6145
              0.8156        -1.0565      0.5077
              0.7119        NaN          NaN
              NaN           -0.8051      0.5913
              0.6686        0.5287       -0.6436
              1.1908        0.2193       0.3803
              NaN           -0.9219      -1.0091
              -0.0198       NaN          -0.0195
              -0.1567       -0.0592      -0.0482];
```

Εκτελέστε τη εντολή `mean(dataV)`. Τι παρατηρείτε;

Λύση 1. Διαγραφή των γραμμών με NaNs και προβολή του γραφήματος.

```
data= dataV;
data (any(isnan(data),2),:) = [];
figure(1)
plot(data)
```

Λύση 2. Διαγραφή των στηλών με NaNs και προβολή του γραφήματος.

```
data= dataV;
data (:,any(isnan(data),1)) = [];
figure(2)
plot(data)
```

Λύση 3. Αντικατάσταση των NaNs με 0 και προβολή του γραφήματος.

```
data= dataV;
notNaN = ~isnan(data)
data(~notNaN) = 0
figure(3)
plot(data)
```

Λύση 4. Εύρεση των τιμών NaN και αντικατάσταση τους με τη μέση τιμή της στήλης που ανήκουν.

```
data= dataV;
notNaN = ~isnan(data)
data(~notNaN) = 0
totalNo = sum(notNaN)
columnTot = sum(data)
colMean = columnTot./ totalNo

for i = 1:length(colMean)
    data(find(notNaN(:,i)==0),i)=colMean(i);
end
figure(4)
plot(data)
```

Συντομότερος Τρόπος (προϋποθέτει την εγκατάσταση του Financial toolbox του Matlab)

```
colMean = nanmean(A);
[row,col] = find(isnan(A));
A(isnan(A)) = colMean(col);
```

Δείτε τις τιμές του πίνακα μετά την αντικατάσταση των NaN.

```
data =      [-0.3999  -0.2625    -1.0106
            0.6900    0.2573     0.6145
            0.8156    -1.0565    0.5077
            0.7119    -0,2625    -0.0708
            0.4376    -0.8051    0.5913
            0.6686    0.5287     -0.6436
            1.1908    0.2193     0.3803
            0.4376    -0.9219    -1.0091
            -0.0198   -0.2625    -0.0195
            -0.1567   -0.0592    -0.0482];
```

Παράδειγμα 1.5 Iris dataset

Η άσκηση αυτή αναφέρεται στο σύνολο δεδομένων IRIS που υπάρχει στο αρχείο δεδομένων 'iris.dat'. Αυτό το σύνολο δεδομένων συλλέχθηκε από τον βοτανολόγο Anderson και περιέχει τυχαία δείγματα των λουλουδιών που ανήκουν σε τρία είδη της ίριδας (μικρούς κρίνους):

α) το setosa, β) το versicolor και γ) το virginica. Για καθένα από τα τρία διαφορετικά είδη, υπάρχουν 50 παρατηρήσεις που αφορούν το μήκος σέπαλου, το πλάτος σέπαλου, το μήκος πέταλου και το πλάτος πέταλου.

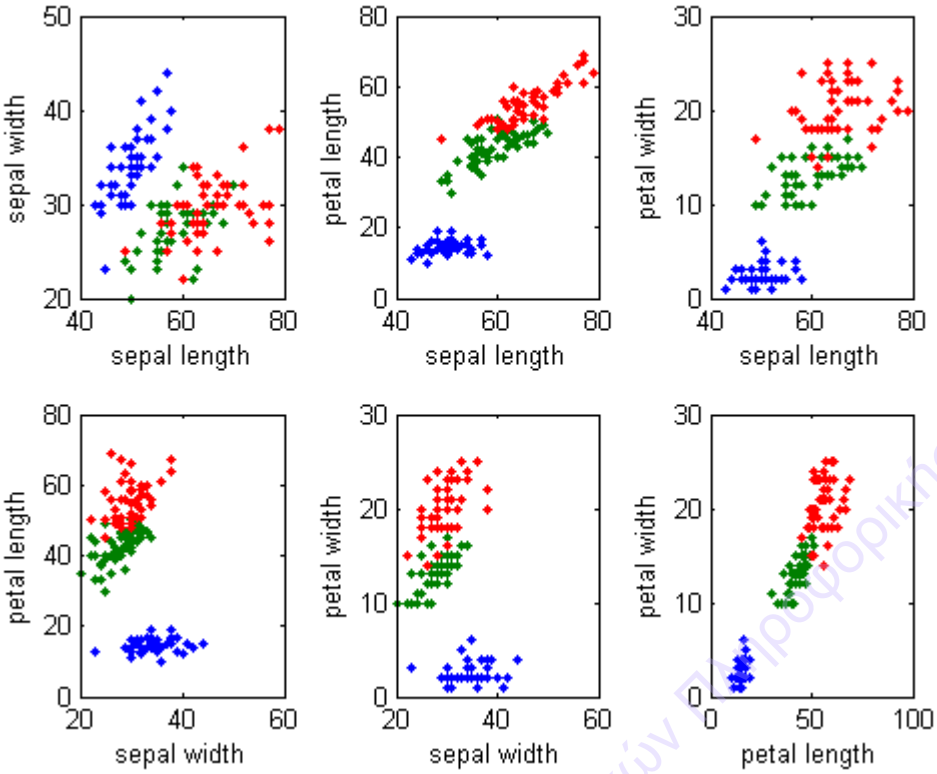
Το σύνολο δεδομένων χωρίζεται σε τρεις ομάδες με τα ονόματα των ειδών: setosa, versicolor, και virginica.

```
load iris.dat
setosa = iris((iris(:,5)==1),:);        % data for setosa
versicolor = iris((iris(:,5)==2),:);    % data for versicolor
virginica = iris((iris(:,5)==3),:);     % data for virginica
obsv_n = size(iris, 1);                 % total number of observations
```

Γραφική Αναπαράσταση των δεδομένων σε 2-D

Τα δεδομένα είναι 4-διαστάσεων. Οι πρώτες 4 στήλες, αντιπροσωπεύουν όπως είπαμε το μήκος σέπαλου, το πλάτος σέπαλου, το μήκος πέταλου και το πλάτος πέταλου ενώ η $5^{η}$ στήλη αναφέρεται στην κλάση που ανήκει η κάθε παρατήρηση. Μπορούμε να απεικονίσουμε σε γραφήματα 2 διαστάσεων τις τρείς ομάδες λουλουδιών (setosa, versicolor και virginica), με συνδυασμούς δύο χαρακτηριστικών κάθε φορά (για παράδειγμα, σέπαλο μήκος vs. σέπαλο πλάτος). Αυτό γίνεται χρησιμοποιώντας το ακόλουθο απόσπασμα κώδικα.

```
Characteristics = {'sepal length','sepal width','petal length','petal width'};
pairs = [1 2; 1 3; 1 4; 2 3; 2 4; 3 4];
h = figure;
for j = 1:6,
    x = pairs(j, 1);
    y = pairs(j, 2);
    subplot(2,3,j);
    plot([setosa(:,x) versicolor(:,x) virginica(:,x)],...
         [setosa(:,y) versicolor(:,y) virginica(:,y)], '.');
    xlabel(Characteristics{x});
    ylabel(Characteristics{y});
end
```



Ερώτηση: Ποιο γράφημα διαχωρίζει καλύτερα τις τρεις κλάσεις λουλουδιών?

Παράδειγμα 1.6 . Εισαγωγή NaNs στο Iris dataset

Σε ποσοστό 60% αντικαθιστούμε τις τιμές του πίνακα iris με τιμές NaN σύμφωνα με τον παρακάτω κώδικα.

```
[ro,co]=size(iris);
p = 60; %ποσοστό των τιμών NaN που θα εισαχθούν στον πίνακα
irisV=iris; %κρατάμε τον αρχικό πίνακα ανέπαφο και δουλεύουμε με τον irisV
r1 = randperm(ro); % αναδιατάσουμε τυχαία τις γραμμές του πίνακα iris
irisV(r1(1:p),1)=NaN; % αντικαθιστούμε με NaN το 60% των τιμών της 1ης στήλης
r1 = randperm(ro);
irisV(r1(1:p),2)=NaN; % αντικαθιστούμε με NaN το 60% των τιμών της 2ης στήλης
r1 = randperm(ro);
irisV(r1(1:p),3)=NaN; % αντικαθιστούμε με NaN το 60% των τιμών της 3ης στήλης
r1 = randperm(ro);
irisV(r1(1:p),4)=NaN; % αντικαθιστούμε με NaN το 60% των τιμών της 4ης στήλης
```

Παράδειγμα 1.7 . Αντικατάσταση NaNs στο Iris dataset

Επαναλάβετε τα βήματα του παραδείγματος 1.4 για το νέο dataset irisV

1. Διαγραφή των γραμμών με NaNs και προβολή του γραφήματος.
2. Διαγραφή των στηλών με NaNs και προβολή του γραφήματος
3. Αντικατάσταση των NaNs με 0 και προβολή του γραφήματος.
4. Εύρεση των τιμών NaN και αντικατάσταση τους με τη μέση τιμή της στήλης που ανήκουν και προβολή του γραφήματος (χρησιμοποιώντας τις διαστάσεις μήκος πέταλου ως άξονας χ και πλάτος πέταλου ως άξονας Υ)

Παράδειγμα 1.8 . Κανονικοποίηση Δεδομένων

Δημιουργείστε τις παρακάτω συναρτήσεις στο φάκελό σας.

```
%% Γραμμική κανονικοποιηση
function yV = LinearTransform(xV)
xV = xV(:);
xmin = min(xV);
xmax = max(xV);
d = xmax - xmin;
yV = (xV - xmin) / d;

%% zscore κανονικοποιηση
function yV = zscoreTransform(xV)
xV = xV(:);
mx = mean(xV);
xsd = std(xV);
yV = (xV - mx) / xsd;
```

παραδείγματος 4, μετά την εφαρμογή της ΛΥΣΗΣ 4, δηλαδή να χρησιμοποιηθεί ο πίνακας που βρίσκεται στο τέλος της σελίδας 5

- Κανονικοποιήστε τις τιμές του πίνακα data του ~~παραδείγματος 2~~ με τους δύο παραπάνω τρόπους και κάντε τα αντίστοιχα γραφήματα. Υπάρχει διαφορά μεταξύ τους;

- Κανονικοποιήστε τις τιμές του πίνακα iris (τις 4 πρώτες στήλες) του παραδείγματος 1.5 με τους δύο παραπάνω τρόπους και κάντε τα αντίστοιχα γραφήματα. Υπάρχει διαφορά μεταξύ τους;