Lab 6: Data Mining.
Serii temporale

Analiza seriilor temporale are ca scop să modeleze şi să explice dependenţa unor date de momente de timp succesive. Exemple tipice de serii temporale sunt: temperatura inregistrată zilnic, curs de schimb valutar, preţul unor acţiuni, valori ale unor mărimi fiziologice înregistrate de dispozitive de monitorizare (tensiune arterială, puls etc) etc.

Principalele prelucrări care pot fi efectuate asupra unei serii de timp sunt:

- *Vizualizare:*
  - Vizualizarea evoluției în timp a mărimii analizate (R: pentru obiecte de tip ts se poate folosi ts.plot sau plot.ts; cele două funcții diferă în principal prin modul de vizualizare a mai multor serii)
- *Pre-procesare*
  - completarea valorilor absente prin interpolare;
- *Analiză:* identificare
  - Tendință (trend)
    - se poate determina din seria inițială prin eliminarea zgomotului folosind o tehnică de *netezire* (smoothing): medie mobilă (Moving Average) sau netezire exponențială. In R se poate folosi funcția filter pentru a define filtrul de netezire
    - se poate elimina prin "*diferentiere*": noua serie se obține din cea anterioară prin calculul diferențelor dintre elementele vecine; pentru diferențe de ordin mai mare decât 1 se aplică repetat strategia; în practică diferențele de ordin 1 permit eliminarea tendinței liniare, iar cele de ordin 2 permit eliminarea tendinței pătratice. R: funcția diff
  - Caracter sezonier (seasonal)
    - Se poate determina calculând media valorilor corespunzătoare unei perioade. De exemplu pentru date înregistrate lunar pentru care există caracter sezonier la nivel de an (există un tipar de comportament de-a lungul unui an) se calculează mediile pentru toate lunile şi se obține o estimare a tiparului corespunzător unui an.
    - Se poate elimina prin calcul diferențe între valori corespunzătoare unor întârzieri în timp egale cu perioada caracterului sezonier. De exemplu pentru date înregistrate lunar pentru care există caracter sezonier la nivel de an (există un tipar de comportament de-a lungul unui an) se poate folosi intârziere (lag) L=12.
  - Zgomot (noise, residuals)
    - Se obține prin eliminarea tendinței și a caracterului sezonier
  - Staționaritate (stationary): seria este staționară dacă caracteristicile sale statistice locale nu se schimbă de-a lungul seriei (nu există tendință și nici character sezonier). O serie nestaționară poate fi transformată într-una staționară prin diferențiere

Obs: cele 3 componente ale unei serii (tendința, componenta sezonieră și zgomotul) se pot determina direct în R folosind funcția decompose.

- *Predicţie*: estimarea valorilor ulterioare din serie pe baza valorii curente şi a celor anterioare (folosind un model care descrie dependenţa valorii curente din serie de valorile anterioare). Principalele etape:
    - se elimină tendința și component sezonieră
    - se alege un model care explică variabilitatea din componenta de zgomot
        - ARMA(p,q) – pt serii staționare
        - ARIMA(p,d,q) – pt serii nestaționare

Obs: în cazul în care mai multe modele conduc la valori similare ale erorii (suma pătratelor erorilor) se allege modelul mai simplu: cel pentru care valoarea AIC (Akaike Information Criterion) este mai mică.

Obs: Un proces de predicţie este caracterizat prin:

- *Intrare:* datele de intrare sunt valori anterioare din serie
- *Iesire:* rezultatul reprezintă valoarea/valorile următoare din serie
- *Model:* un model de regresie care descrie legătura dintre valoarea curentă a seriei şi valorile anterioare (numărul de valori anterioare despre care se consideră că influenţează valoarea curentă este denumit intârzierea seriei (*time-lag*))

Considerăm seria $X_1, X_2, \ldots, X_n$ şi intârzierea T. Deci valoarea curenta X_i depinde de valorile $X_{i-1}, X_{i-2}, \ldots, X_{i-T}$. Prin urmare secventa de valori din serie poate fi transformată într-un alt set de date în care sunt T atribute predictor si un atribut prezis:

Atribute predictor	*Atribut prezis*
$X_1\ X_2\ \ldots X_i \ldots X_T$	X_{T+1}
$X_2\ X_3\ \ldots X_{i+1} \ldots X_{T+1}$	X_{T+2}
...	
$X_{n-T}\ X_{n-T+1}\ \ldots X_{n-i} \ldots X_{n-1}$	X_n

Folosind acest set de date se poate construi un model de regresie (in aceeaşi manieră ca pentru date care nu sunt temporale). In acest context pot fi folosite modele de regresie, rețele neuronale sau rețele RBF. Una dintre principalele dificultăţi este alegerea adecvată a valorii T.

Exercițiul 1. Analiza în R a unor serii temporale și construirea unor modele de predicție bazate pe netezire (Holt-Winters) și a unor modele de tip ARIMA. Parcurgeți și executați comenzile din fișierul Ex1.r.

Exercițiul 2. Parcurgeți și executați comenzile din fișierul Ex2.r.

Exercitiul 1/Weka.

a) Deschideti fisierul airlines.arff (conţinând nr de pasageri ai unei companii aeriene înregistrat lunar in perioada 1949 – 1960)
b) Construiţi un nou set de date folosind o întârziere T=12. *Indicatie:* utilizaţi Weka pt eliminarea atributului corespunzător datei şi Excel (sau un limbaj de programare) pt construirea noului set de date
c) Aplicaţi un model de regresie pentru noul set de date şi analizaţi rezultatele obţinute

Exercitiul 2/Weka.

(doar pt versiune Weka >=3.7.3)

a) Instalati pachetul Time Series Forecasting utilizand Weka GUI Chooser ->Tools->Package manager si selectand pentru instalare timeSeriesForecasting
b) Deschideti fisierul airlines.arff
c) Preziceti urmatoarele 6 valori utilizand unul dintre urmatoarele modele: (i) linear regression; (ii) multilayer perceptron; (iii) random forests. *Indicatie*: selectia modelului se realizeaza utilizand panelul Advanced Configuration->Based Learner

Obs: detalii privind pachetul TimeSeriesForecasting pot fi gasite la http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka.

(pentru Weka 3.8.1) se utilizează Forecast (Basic configuration) şi se specifică doar Number of time units to forecast (6 dacă se doreşte estimarea următoarelor 6 valori).

Temă.

1. Scrieți funcții R care transformă o serie prin:
    a. Netezire folosind medie mobilă bidirecțională/centrată (moving average smoothing) – dimensiunea ferestrei mobile este parametru al funcției
    b. Netezire exponențială (exponential smoothing) – parametrul regulii de netezire este parametru al funcției
    c. Diferențiere (diferențe între valori vecine) – ordinul diferențierii este parametru al funcției

Obs. valorile de la extremitățile seriei vor fi tratate adecvat (nr de elemente din fereastra mobilă poate fi mai mic la extremități).