Data Mining

Lab 5:

Reguli de asociere
Modele de regresie

---

1. Reguli de asociere/

Exemplu (problema cosului de cumparaturi). Se consideră un set de tranzacţii (T_1, T_2,…, T_n), fiecare tranzacţie conţinând un set de produse achiziţionate. De exemplu:

T_1: {pâine, lapte, apă}
T_2: {pâine,carne, apă}
T_3: {pâine, unt, carne, apă}
T_4: {fructe, apă}

Se pune problema identificării produselor care sunt frecvent cumpărate împreună şi a unor reguli de asociere de forma IF pâine AND apă THEN carne.

Regulile de asociere sunt de forma IF A THEN B. Termenul A joacă rol de antecedent, iar B rol de consecinţă (totuşi nu exprimă relaţii de cauzalitate ci doar de corelare).

Dintr-un set de tranzacţii pot fi extrase numeroase reguli – e necesar să poata fi evaluată relevanţa lor pentru a fi ierarhizate.

Pentru evaluarea relevanţei unei reguli se folosesc cel puţin două mărimi:

- Suport (support): supp(A->B) = numărul de tranzacţii ce conţin pe A şi B/ numarul total de tranzacţii
- Incredere (confidence): conf(A->B)=numărul de tranzacţii ce conţin pe A şi B/ numărul de tranzacţii ce conţin pe A

Exemplu: IF pâine şi apă THEN carne

A={pâine, apă}, B={carne}
Supp(A->B)=2/4=0.5
Conf(A->B)=2/3=0.6

Obs: pe lângă aceste măsuri se folosesc şi indicatori ai noutăţii regulii (cât este de interesantă sau neobişnuită regula); un exemplu de astfel de indicator este cel denumit lift:

Lift(A->B)=prob(A,B)/(prob(A)prob(B))

(probabilitatea se estimează prin frecvenţa relativă)

Regula e considerată interesantă dacă valoarea măsurii lift este mare. Daca valoarea e apropiată de 1 aceasta sugerează că A şi B nu sunt corelate deci nu poate fi extrasă o regulă utilă de forma A->B

Exemplu: R=IF pâine AND carne THEN apă
Conf(R)=2/2=1
Lift(R)=0.5/(0.5*1)=1

Algoritm de extragere a regulilor de asociere din date (algoritmul APRIORI)

Date de intrare: set de tranzacţii (fiecare tranzacţie conţine o listă de entităţi)
Parametri de control:

- Prag pentru suport minim (ex: 0.2)
- Nivel minim de încredere (ex: 0.9)

Structura general a algoritmului Apriori:

Pas 1: identificare subseturi cu support semnificativ (mai mare decât pragul minim) – "frequent itemsets"; identificarea acestor subseturi se bazează pe:

- Identificarea subseturilor frecvente cu un singur element (lista L_1)
- FOR k=1,K DO construirea listei L_k cu subseturi frecvente având k elemente pornind de la lista L_{k-1} (subseturi frecvente cu k-1 elemente)

Pas 2: construirea regulilor prin partiţionarea subseturilor identificate la pasul 1 în două părţi (o parte pentru antecedentul regulii şi o parte pentru consecinţă); se reţin doar regulile care au nivelul de încredere mai mare decat pragul

Exercitiul 1.

a) Deschideţi în Weka setul de date supermarket.arff
b) Determinaţi reguli de asociere folosind Associate->Apriori şi valorile implicite ale parameterilor
c) Aplicaţi acelasi algoritm pentru alte valori ale pragului pentru support (lowerBoundMinSupport=0.2) si pentru incredere (minMetric=0.75).

2. Modele de regresie

2.1. Regresie liniara

In modelele liniare dependenţa dintre variabila(variabilele) prezise şi cele predictor este descrisă printr-o funcţie liniară de forma Y=WX. Parametrii modelului (elementele vectorului/matricii W) se determină pornind de la date folosind o tehnica de minimizare a sumei pătratelor erorilor

Exercitiul 2.

a) Deschideţi în Weka fisierul autoPrice.arff
b) Utilizaţi Class->Functions->SimpleLinearRegression pentru a determina o dependenţă liniară simplă între atributul de ieşire (preţ) si cel mai relevant dintre atributele de intrare. Analizaţi valorile corespunzătoare coeficientului de corelaţie şi erorii Correlation Coeficient si Mean Absolute Error.
c) Utilizaţi Class->Functions->LinearRegression pentru a determina o dependenţa liniară între atributul de ieşire (pret) şi cele mai relevante dintre atributele de intrare. Analizati valorile

corespunzătoare coeficientului de corelaţie şi erorii Correlation Coeficient şi Mean Absolute Error. Analizaţi influenţa metodei de selecţie a atributelor (attributeSelectionMethod)

2.2. Regresie neliniară

Exercitiul 3. (tot pentru fisierul autoPrice.arff)

d) Utilizati Class->Functions->MultilayerPerceptron cu valorile implicite ale parametrilor. Analizaţi valorile corespunăatoare coeficientului de corelaţie şi erorii Correlation Coeficient si Mean Absolute Error.

e) Identificati în categoria Class->Trees varianta care permite construirea unui arbore de regresie

Obs. Versiunile de Weka mai mici de 3.8 include si implementare a retelelor RBF (RBF Network).

Exercitiul 4. Aplicati prelucrarile de la Exercitiile 2 si 3 in cazul setului de date autoMPG.arff si analizati diferentele (in particular in ceea ce priveste arborii de regresie).