Università di Pisa	A.A. 2007-2008
Data Mining	
Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda	

Progetti

Informazioni generali

Di seguito vengono proposti 3 progetti rivolti a gruppi di 2-3 persone. Per tutti viene richiesto di:

1. svolgere il lavoro di analisi indicato, seguendo i passi generali del modello CRISP, dalla comprensione del problema fino alla validazione dei risultati ottenuti;
2. produrre una relazione scritta che descrive in modo succinto le fasi svolte nel corso del proprio lavoro, seguendo anche qui (seppur con un certo grado di flessibilità) i passi indicati dal modello CRISP. Indicativamente la relazione dovrà occupare tra le 20 e le 30 pagine;
3. preparare una presentazione di circa 20-30 minuti, da effettuare in sede di esame da tutti i membri del gruppo.

Viene lasciata piena libertà in quanto agli strumenti utilizzati nello svolgimento dei progetti, purché tutte le fasi da essi richieste (esplorazione dei dati, preprocessing, mining, validazione, ecc.) siano adeguatamente coperte.

La persona di riferimento comune ad ogni progetto, cui rivolgersi per chiarimenti e problemi vari è Mirco Nanni (mirco.nanni@isti.cnr.it), affiancato in diversi progetti da un esperto (o quasi) del dominio applicativo.

Progetto 1 (Category Management)

Dati

Nell'ambito dell'Unicoop Tirreno vengono forniti i dati che descrivono gli acquisti dei clienti relativi ad un punto vendita della catena Supermercati, per un periodo di circa 3 mesi. Tali dati contengono il dettaglio dei singoli scontrini emessi dal punto vendita, ovvero ogni singolo prodotto venduto, nonché, per i clienti che sono soci della cooperativa, il codice che identifica univocamente il socio in questione. Vengono forniti inoltre: (a) le tabelle che descrivono la gerarchia di prodotti presenti nel circuito Unicoop, e (b) l'anagrafica dei soci, arricchita da alcuni aggregati precalcolati relativi agli acquisti effettuati durante i 3 mesi.

Contesto

La categorizzazione dei prodotti attualmente adottata da Unicoop, e descritta dalla gerarchia dei prodotti fornita coi dati, segue una divisione razionale che porta a raccogliere in una stessa classe i prodotti dello stesso tipo, formato o marca. Lo stesso tipo di logica viene spesso seguita per disporre i prodotti sugli scaffali. D'altro canto, da studi sociologici rivolti ai clienti della grande distribuzione, emerge la richiesta di organizzare e presentare i prodotti secondo categorie e raggruppamenti più vicini alle loro necessità. Tali raggruppamenti possono rispecchiare tipologie di richieste già note, ad esempio relative ai clienti interessati ai prodotti etnici, oppure possono essere del tutto nuovi.

Obiettivo

Sfruttare la conoscenza puntuale degli acquisti effettuati dai soci Unicoop per farne emergere nuove categorizzazioni o raffinamenti della categorizzazione esistente, ad esempio cercando insiemi (significativi) di prodotti venduti frequentemente insieme.

Progetto 2 (Prevenzione Abbandono)

Dati

Nell'ambito dell'Unicoop Tirreno vengono forniti i dati che descrivono gli acquisti dei clienti relativi ad un punto vendita della catena Ipermercati (i punti vendita più grossi), per un periodo di circa 6 mesi. Tali dati contengono i "riassunti" dei singoli scontrini emessi dal punto vendita, ovvero i pagamenti effettuati e diversi aggregati precalcolati relativi ai prodotti acquistati, oltre che, per i clienti che sono soci della cooperativa, il codice che identifica univocamente il socio in questione. Vengono forniti inoltre: (a) una descrizione delle promozioni cui gli acquisti sono stati soggetto, e (b) l'anagrafica dei soci.

Contesto

Un obiettivo generale delle aziende nella grande distribuzione consiste nel coltivare la fedeltà dei propri clienti, in quanto acquisire nuovi clienti è tipicamente molto più difficile e dispendioso che non offrire sconti e vantaggi per evitare la perdita di clienti già consolidati. Non essendo economicamente sostenibile offrire a tutti i soci vantaggi di questo genere, nasce il problema di individuare i soci che realmente ne hanno bisogno – ovvero, quelli che stanno per lasciare l'azienda, ad esempio per diventare clienti abituali di altri negozi.

Obiettivo

Studiare alcune possibili definizioni di "abbandono", ovvero condizioni che stabiliscano quando un socio è da considerarsi perduto. Quindi, estrarre dei modelli predittivi che consentano di riconoscere con almeno un mese di anticipo i soci in procinto di abbandonare – secondo le definizioni di abbandono studiate.

Suggerimento: dividere il dataset in due segmenti contenenti, rispettivamente, i primi 5 mesi di vendite e l'ultimo mese. I primi costituiscono il *passato*, da cui estrarre variabili predittive, e l'ultimo rappresenta il *futuro*, da cui determinare se ogni socio abbandonerà o meno (= variabile target).

Progetto 3 (Rilevamento Frodi)

Dati

Nell'ambito del controllo delle dichiarazioni dei redditi, limitatamente a dichiarazioni di una determinata tipologia, vengono forniti i dati relativi a circa 200.000 dichiarazioni, di cui un quarto accertate e quindi fornite dei dati che descrivono l'accertamento (frode/non frode, nonché differenza tra ammontare dichiarato e ammontare accertato). Ogni dichiarazione viene descritta attraverso numerose variabili, principalmente numeriche.

Contesto

Le dichiarazioni accertate sono state scelte con criteri a noi non noti – in parte a caso, ma, probabilmente, in parte seguendo regole suggerite dall'esperienza dei revisori. Risulta quindi interessante: (i) capire se le dichiarazioni accertate hanno caratteristiche che le distinguono chiaramente da quelle non accertate (quindi svelando quali “regole” di selezione sono state adottate); (ii) ottimizzare il processo di selezione, suggerendo regole che portano ad un maggior numero di frodi accertate o ad un miglior risultato gobale in termini di ricavato economico degli accertamenti.

Obiettivo

Cercare caratteristiche discrimanti dei due gruppi di accertamenti, quelli accertati e quelli no. Inoltre, definire un criterio per selezionare le dichiarazioni più promettenti, o come probabilitàdi reale frode, o come convenienza economica.