

Università di Pisa	A.A. 2008-2009
Data Mining	
Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda	

Progetti

Informazioni generali

Di seguito vengono alcuni 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.

I dataset

Nell'ambito dell'Unicoop Tirreno vengono forniti due dataset:

- (a) nel primo, i dati descrivono gli acquisti dei clienti relativi ad un punto vendita della catena Supermercati, per un periodo di circa 6 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;
- (b) nel secondo, i dati descrivono gli acquisti effettuati in un piccolo gruppo di punti vendita tra loro relativamente vicini, per un periodo di circa un anno. Gli acquisti sono descritti in forma aggregata per ogni scontrino, ovvero non viene fornito il contenuto dettagliato di ogni scontrino o carrello, bensì solo alcuni aggregati calcolati su ogni scontrino, tra i quali: spesa totale, spesa totale divisa per settori merceologici, numero di prodotti. Tali dati sono collegabili all'anagrafica dei soci, anche essa fornita.

Progetto 1 (Category Management)

Dataset di riferimento: (a)

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 (Segmentazione per profili di acquisto)

Dataset di riferimento: (a)

Contesto: Un problema classico della business intelligence è la scoperta di segmenti di clientela omogenee rispetto ad alcune caratteristiche interessanti, tra cui il modo in cui gli acquisti vengono effettuati.

Obiettivo: Identificare segmenti di clientela omogenei rispetto al proprio profilo di acquisto (quantità ma anche tipologia di spesa), tenendo eventualmente conto anche del modo in cui i comportamenti di acquisto evolvono nel tempo.

Progetto 3 (Segmentazione per risposta alle promozioni)

Dataset di riferimento: (a)

Contesto: Un problema classico della business intelligence è la scoperta di segmenti di clientela omogenee rispetto ad alcune caratteristiche interessanti, tra cui il modo in cui gli acquisti vengono effettuati.

Obiettivo: Identificare segmenti di clientela omogenei rispetto al modo in cui i propri comportamenti di acquisto variano in presenza di promozioni, ovvero durante e/o dopo il periodo di promozione vero e proprio (di solito di breve durata).

Progetto 4 (Impatto delle promozioni sul singolo prodotto)

Dataset di riferimento: (a)

Contesto: Le promozioni tendono tipicamente ad incentivare l'acquisto di prodotti durante il periodo di promozione, e ci si attende che ciò porti anche ad un incremento di vendite del prodotto promosso nei periodi successivi. Una valutazione più esauriente del fenomeno, però, deve tener conto anche degli effetti che la promozione ha sui prodotti “concorrenti”.

Obiettivo: Analizzare l'effetto di (alcune) promozioni sia sui prodotti promossi, sia sui prodotti concorrenti, durante e/o dopo il periodo promozionale. In tali valutazioni, si consideri il fatto che diverse promozioni tipicamente si susseguono o si sovrappongono.

Progetto 5 (Circolarità)

Dataset di riferimento: (b)

Contesto: Uno dei fenomeni di interesse nella grande distribuzione è la cosiddetta “circolarità”, ovvero il fatto che un socio possa frequentare diversi punti vendita della stessa catena con intensità diverse, finalità diverse e anche variabili nel tempo. Ad esempio, alcuni punti vendita possono essere privilegiati per l'acquisto di alcuni prodotti, o durante certi periodi dell'anno.

Obiettivo: Analizzare il fenomeno della circolarità, cercando di estrarre profili tipici ovvero segmentazioni della clientela in base a tale fenomeno – laddove presente.

Altri progetti

Dataset di riferimento: (a) o (b)

Contesto e obiettivo: Proposti dal gruppo e valutato (eventualmente accettato) dai docenti. Sono benvenute varianti dei progetti proposti sopra, così come problemi completamente diversi, purché facenti riferimento ai soli dataset (a) e (b) descritti sopra.