

**PROGRAMMA DEFINITIVO DEL CORSO DI  
BASI DI CONOSCENZA E DATA MINING  
A.A. 2010/2011**

*Docente: Donato Malerba*

**Obiettivi.** La crescente disponibilità di dati nella attuale società dell'informazione ha evidenziato la necessità di disporre di strumenti adeguati per la loro analisi al fine di scoprire delle regolarità (o modelli) che descrivano adeguatamente i fenomeni di interesse o che possano essere utilizzati in processi decisionali. L'obiettivo del corso è quello di fornire un'introduzione ai concetti di base del processo di estrazione di conoscenza dai dati, alle principali tecniche di data mining ed ai relativi algoritmi. Nel corso ci si propone anche di approfondire l'impiego di tecniche di data mining in diverse applicazioni.

**Organizzazione:** Il corso si articola in due moduli:

Modulo: **Business Intelligence**

Modulo: **Data Mining**

Il primo modulo introduce le tecnologie di Business Intelligence (Data Warehousing, OLAP, il processo di scoperta di conoscenza e il passo di Data mining nello specifico, applicazioni).

Il secondo modulo, introduce i metodi di data mining per i seguenti task: classificazione, regressione, associazione, e clustering. Si considerano inoltre sia rappresentazioni a singola tabella (o proposizionali) che a tabella multipla (multi-relazionali).

**Prerequisiti:** basi di dati II, intelligenza artificiale e rappresentazione della conoscenza.

**Modalità d'esame:** prova orale e discussione di uno studio di caso di Data Mining.

**Crediti Formativi Universitari<sup>1</sup>:** 12 CFU, equamente ripartiti fra i due moduli in 8 di tipo T1 e 4 di tipo T2.

**Programma del corso**

**MODULO A: Business Intelligence (6 CFU)**

1. Tecnologie di Business Intelligence.

Dati operazionali e dati decisionali. Tecnologie di Business Intelligence. Sistemi di supporto alle decisioni (DSS), Executive Information Systems (EIS) e Management Information Systems (MIS). Caratteristiche di un data warehouse. Architettura di un data warehouse. Il modello multidimensionale. Schema del data warehouse: a stella, a fiocco di neve, a costellazione. OLAP e operazioni per l'analisi dei dati: drill down e roll up. ROLAP e MOLAP. Uno studio di caso: il modello relazionale di un DW per il settore agro/alimentare.

2. Scoperta di conoscenza nelle basi di dati: il processo

La scoperta di conoscenza nelle basi di dati: definizione. Il processo della scoperta di conoscenza nelle basi di dati. Il processo CRISP-DM: business understanding, data understanding, data preparation, modelling, evaluation, deployment.

Lab-A. Strumenti di Data Warehousing, OLAP e data mining

Oracle Warehouse Builder, Strumenti di ETL, Data Object Editor, Progettazione logica, definizione delle dimensioni di analisi, definizione dei fatti.

**MODULO B: Data Mining (6 CFU)**

3. La classificazione

*Basata su alberi di decisione:* Il modello ad albero. Test ai nodi interni e loro criteri di scelta. Il trattamento di attributi continui e mancanti. Complessità del problema di costruzione di alberi di decisione. La strategia "greedy" per la costruzione automatica di alberi di decisione. Limitazioni del modello ad albero. Trasformare alberi di decisione in insiemi di regole. Il rasoio di Occam. Costruire alberi di decisione della giusta dimensione: la semplificazione (pruning).

---

<sup>1</sup> La tipologia di CFU è la seguente:

- T1: 8 h di lezione in aula e 17 di studio individuale
- T2: 15 h di laboratorio ed esercitazioni guidate e 10 di rielaborazione personale

*Basata su insiemi di regole:* Costruzione automatica di descrizioni di un concetto a partire da esempi. Strutturare lo spazio delle ipotesi con ordini di generalità. L'algoritmo Find-S per la ricerca di ipotesi massimamente specifiche. Limiti di Find-S. Lo spazio delle versioni e l'algoritmo di eliminazione dei candidati. Classificare con spazi delle versioni. Trattare dati rumorosi. Il ruolo della polarizzazione (bias) induttiva nella costruzione automatica di ipotesi. Apprendere definizioni disgiuntive di concetti: la strategia di di copertura sequenziale (sequential covering o separate-and-conquer). Il passo della 'conquista' nella copertura sequenziale: la ricerca a raggio (beam search). Complessità computazionale di algoritmi fondamentali per l'apprendimento di classificatori a regole. Costruzione automatica di descrizioni di più concetti a partire da esempi.

*Basato su teorema di Bayes:* Il teorema di Bayes e nozioni di base correlate. Apprendimento Maximum-A-Posteriori forza bruta. Ipotesi MAP e sistemi di apprendimento consistenti. Il principio MDL (minimum description length). Classificatori ottimali di Bayes. Algoritmo di Gibbs. Il classificatore di naive Bayes e la stima delle probabilità. La classificazione di testi basata su classificatori naive Bayes.

#### 4. La regressione parametrica e non parametrica

Modelli di regressione con errore additivo. Modelli di regressione (lineare) semplice: stima dei parametri con il metodo dei minimi quadrati, aspetti inferenziali del modello di regressione semplice, diagnostiche grafiche del comportamento dei residui. Modelli di regressione lineare semplice con regressore qualitativo: stima dei parametri con il metodo dei minimi quadrati, aspetti inferenziali del modello di regressione semplice, diagnostiche grafiche del comportamento dei residui. Estensioni: regressione polinomiale e regressione lineare multipla. La notazione matriciale. Funzioni costanti a tratti per la regressione semplice. I modelli non parametrici: alberi di regressione. Combinare modelli parametrici e non: gli alberi dei modelli. L'algoritmo SMOTI per l'apprendimento di alberi di modelli.

#### 5. Le associazioni di variabili

Il caso di due variabili - Correlazioni di variabili quantitative: coefficiente di correlazione (parziale) di Pearson. Coefficienti di correlazione e modelli di regressione. Associazioni di variabili qualitative: lambda, coefficiente di Spearman.. Il caso di più variabili - I limiti dei modelli log-lineari. Le regole di associazione: le configurazioni (pattern) frequenti, supporto di una configurazione, confidenza di una regola di associazione, le regole di associazione forti, il metodo Apriori per la generazione di regole di associazione forti.

#### 6. Tecniche avanzate di data mining

Data Mining Multi-Relazionale: assunzione di tabella singola, dati distribuiti su più tabelle, pattern relazionali, come promuovere gli algoritmi proposizionali all'analisi di dati relazionali. Uno studio di caso: SPADA. La classificazione relazionale con FOIL, Progol e TILDE.

#### Lab-B.Strumenti di Data Mining

Weka: introduzione, preprocessing e trasformazione dati; la classificazione mediante regole di classificazione, alberi di decisione, naive Bayes e k-NN; modelli di regressione; scoperta di regole di associazione.

### ***Principali testi e articoli di riferimento***

T. Mitchell  
*Machine Learning*  
Morgan Kaufmann, 1997  
Capitoli: 2-3-6

Richard J. Roiger, Michael W. Geatz.  
*Introduzione al Data Mining.*  
McGraw-Hill, 2003  
Capitoli: 1-2-3-4-5-6-8-9

A. Azzalini, B. Scarpa  
*Analisi dei dati e data mining*  
Sprinter, 2004  
Capitoli: 1-4-5-6

Articoli scientifici selezionati e copia delle trasparenze proiettate durante le lezioni e durante le esercitazioni in laboratorio sono disponibili sul sito:

<http://www.di.uniba.it/~malerba/courses/bcdm.htm>