PROGRAMMA DEFINITIVO DEL CORSO DI
BASI DI CONOSCENZA E DATA MINING
A.A. 2011/2012

Docente: Donato Malerba

Obiettivi formativi. 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. Pertanto gli obiettivi formativi sono:
Acquisizione di conoscenze su:

- Metodologia industriale CRISP-DM.
- Metodi di pre-elaborazione e trasformazione dei dati, validazione dei pattern e dei modelli estratti.
- Algoritmi di data mining per compiti di classificazione, regressione e analisi di associazione.
- Metodi di data mining per la scoperta di pattern in dati con struttura complessa (relazionale).

Sviluppo di capacità applicative su:

- Utilizzo di strumenti per la selezione, preleaborazione e trasformazione dei dati, e per la validazione dei pattern estratti.
- Utilizzo di strumenti di data mining per l'estrazione di conoscenza finalizzata a scopi predittivi e descrittivi in diversi contesti applicativi (aziendali e scientifici).

Obiettivi professionalizzanti. Acquisizione di competenze nello sviluppo di sistemi di business intelligence con funzionalità di analisi (data mining).
Prerequisiti: basi di dati II, intelligenza artificiale e algoritmi e strutture dati.
Modalità d'esame: prova orale e discussione di uno studio di caso di Data Mining.
Crediti Formativi Universitari[1]: 12 CFU, equamente ripartiti fra i due moduli in 8 di tipo T1 e 4 di tipo T2.

Programma del corso

1. 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.

2. 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).
   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).

[1] 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

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.

3. 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.

4. 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.

5. 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.

Laboratorio:

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