

## Risposte – Data Mining B (SD\_IABA23\_1727a\_01-2)

### Lezione 002

01. Che cosa è il Data Mining?

Risposta: C) Il processo che estrae nuova conoscenza o identifica pattern/modelli nei dati mediante l'applicazione di diversi algoritmi.

02. Che cosa è un pattern?

Risposta: A) Una espressione in un determinato linguaggio che descrive i fatti di un sottoinsieme di registrazioni presenti in un data base

03. Quale fra le seguenti attività non fa parte del processo di KDD

Risposta: A) Nessuna delle altre alternative

04. Quale fra le seguenti attività non fa parte del processo di KDD?

Risposta: D) Intervista

05. Quali sono le fasi caratteristiche del data mining?

Risposta (max 5 righe): Tipicamente: (1) esplorazione/understanding dei dati; (2) modellazione (scelta algoritmo, training); (3) valutazione/validazione dei risultati; (4) interpretazione e messa in produzione/uso delle conoscenze.

06. Cosa è una transazione?

Risposta (max 5 righe): Una transazione è un record che rappresenta un singolo evento/operazione (es. acquisto) composto da un insieme di item. Nel market basket, ogni transazione è un carrello con gli articoli comprati insieme in un certo istante.

07. Quali sono le fasi del KDD?

Risposta: A) Set Domande : DATA MINING B

### Lezione 003

01. Cosa si intende per knowledge elicitation?

Risposta: A) Il confronto fra gli esperti di uno specifico dominio e l'ingegnere della conoscenza al fine di aggiungere in fase di modellazione elementi che potrebbero non essere

02. A cosa servono i modelli predittivi?

Risposta: A) Utilizzano i loro parametri e meccanismi caratteristici per effettuare delle previsioni future che possono essere utili nei processi decisionali.

03. A cosa servono i modelli descrittivi?

Risposta: B) Vengono utilizzati per descrivere la struttura organizzativa e la distribuzione dei dati. L'obiettivo di questi modelli è quello di approfondire la conoscenza nascosta dietro ai

04. Quali sono i pilastri operativi su cui si basa ogni algoritmo di data mining?

Risposta: B) Esplorazione, Modellazione, Valutazione

05. Discutere la differenza fra modelli predittivi e descrittivi.

Risposta (max 5 righe): I modelli predittivi stimano un'uscita ignota (classe o valore) su nuovi dati, quindi puntano alla generalizzazione (es. classificazione/regressione). I modelli descrittivi cercano strutture interne ai dati senza target (es. clustering, associazioni) per capire come sono organizzati.

06. Perché il processo di data mining è da considerarsi interattivo?

Risposta (max 5 righe): È interattivo perché le scelte (attributi, pulizia, trasformazioni, algoritmo, parametri) dipendono dai risultati parziali. Spesso si iterano più cicli: si prova, si valuta, si correggono dati/parametri finché il modello è soddisfacente.

07. Quali sono i pilastri operativi su cui poggia il data mining?

Risposta: -)

### Lezione 004

01. Cosa succede durante un addestramento supervisionato?

Risposta: C) Vengono utilizzati dei dati etichettati, ovvero dove si conosce il valore dell'uscita desiderata, sia essa una classe o il valore di una funzione, a fronte di un ingresso specifico

02. A cosa serve la cross validation?

Risposta: C) A migliorare la capacità di generalizzazione dei modelli

03. Che cosa è l'overfitting?

Risposta: A) E' il fenomeno per cui in fase di addestramento un modello si adatta al campione osservato e perde la capacità di generalizzazione

04. Cosa succede durante un addestramento non supervisionato?

Risposta: C) I dati che si utilizzano non sono etichettati e gli algoritmi di apprendimento mirano ad estrarre la conoscenza (i parametri del modello) senza avere una specifica

05. Quale è il dominio degli attributi categorici?

Risposta: B) Nessuna delle altre alternative

06. Discutere il processo della cross validation.

Risposta (max 5 righe): Si divide il dataset in k fold. Per ogni fold: k-1 parti per training e 1 per test; si ripete k volte ruotando il fold di test. Le performance si mediano per stimare la capacità di generalizzazione e ridurre la dipendenza da un singolo split.

07. Elenicare le differenze fra boosting e bagging.

Risposta (max 5 righe): Bagging: addestra più modelli in parallelo su campioni bootstrap e combina (vote/media) per ridurre varianza. Boosting: addestra modelli in sequenza, enfatizzando gli errori dei precedenti, per ridurre bias e migliorare l'accuratezza; più sensibile al rumore.

08. Discutere il problema dell'overfitting.

Risposta: -)

Lezione 005

01. Cosa sono gli istogrammi?

Risposta: D) Sono dei diagrammi che consentono di visualizzare la frequenza con cui si distribuiscono i dati divisi in classi (banalmente in intervalli)

02. Cosa sono gli scatter plot?

Risposta: E) Sono dei grafici che consentono di verificare rapidamente se due variabili sono correlate (dipendenti linearmente) fra di loro

03. Cosa sono i box plot?

Risposta: B) Si tratta di visualizzazioni grafiche che consentono di sintetizzare con 5 valori le caratteristiche di una distribuzione. Tali valori sono: minimo, massimo, mediana, primo

04. Che cosa è la mediana di una variabile casuale X?

Risposta: B) E' un indice di posizione: una volta ordinati tutti i valori assunti da X, la mediana si calcola estraendo il valore del campione che divide la distribuzione in due popolazioni

05. Che cosa è la moda di una variabile casuale X?

Risposta: E) E' l'indicatore che indica quale è il valore di X che si presenta più frequentemente nei dati

06. Fornire un esempio di distribuzione e del suo istogramma

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Cosa si intende per outliers?

Risposta (max 5 righe): Gli outlier sono osservazioni anomale, molto lontane dal comportamento tipico della distribuzione. Possono derivare da errori di misura/inserimento o da eventi rari reali. Influenzano medie e modelli: vanno identificati e gestiti (rimozione, cap, trasformazioni).

08. Quali sono gli indicatori di dispersione?

Risposta (max 5 righe): Esempi: varianza e deviazione standard, range (max–min), interquartile range (IQR), scarto medio assoluto (MAD), coefficiente di variazione. Misurano quanto i dati sono “sparsi” rispetto a un valore centrale.

09. Discutere la differenza fra quartile e quantile

Risposta: A) Set Domande : DATA MINING B

Lezione 006

01. Quali fra i seguenti passaggi non fanno parte della pre-elaborazione dei dati?

Risposta: D) Nessuna delle altre alternative

02. Quali fra i seguenti passaggi non fanno parte della pre-elaborazione dei dati?

Risposta: -)

03. Cosa si intende per riduzione della dimensionalità?

Risposta: F) compressa o ridotta dei dati originali (esempio uso di Trasformate o tecniche di selezione di attributi).

04. Quali sono i principali fattori che caratterizzano la qualità dei dati?

Risposta (max 5 righe): Qualità dei dati: accuratezza (correttezza), completezza (missing), consistenza (assenza di contraddizioni), tempestività/attualità, unicità (no duplicati) e validità (rispetto di domini/vincoli).

05. Perché è necessario il trattamento preliminare dei dati?

Risposta (max 5 righe): Serve perché dati reali sono spesso incompleti, rumorosi, incoerenti e provenienti da fonti diverse. Pre-processing migliora qualità e rappresentazione, riduce dimensioni e rende i modelli più affidabili ed efficienti.

06. Fornire un esempio in cui potrebbe essere necessario effettuare data cleaning o data integration.

Risposta: A) Set Domande : DATA MINING B

Lezione 007

01. Quali fra le seguenti tecniche non si usano in fase di gestione dei missing value?

Risposta: A) Riempimento a mano dei valori mancanti

02. L'analisi degli outlier può essere usata per:

Risposta: A) Ridurre la dimensionanalità

03. L'analisi in regressione può essere usata per:

Risposta: A) Identificare e gestire i dati rumorosi

04. Con quali tecniche si gestiscono i missing value?

Risposta: A) Data cleaning

05. Le fasi del processo di data cleaning si effettuano solo una volta. Vero o falso? Discutere la scelta effettuata.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

06. Cosa si intende per Data Scrubbling e Data Auditing?

Risposta (max 5 righe): Data scrubbing: insieme di attività di pulizia/correzione (standardizzazione, deduplica, fixing errori, gestione missing). Data auditing: analisi/controllo della qualità dei dati con report e metriche per individuare problemi e tracciarne le cause.

07. Descrivere la seconda fase del processo di data cleaning, specificando che tipi di tool commerciali potrebbero essere utilizzati.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

08. Quali sono i passaggi per effettuare l'eliminazione delle discrepanze fra i dati?

Risposta: A) Set Domande : DATA MINING B

#### Lezione 008

01. Quali sono i passi da effettuare nel processo di data integration?

Risposta: D) Identificazione delle entità, Identificazione e risoluzione dei conflitti nei dati, identificazione delle ridondanze

02. Quale è l'obiettivo principale della fase di data integration?

Risposta: D) Ottenere una memorizzazione coerente per i dati derivanti da sorgenti multiple di memorizzazione

03. Per quale fase si può utilizzare l'analisi di correlazione nel processo di data integration?

Risposta: D) Identificazione delle ridondanze

04. Quali sono i conflitti che possono apparire nei dati?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

05. Quando si parla di data integration, cosa sono e come si usano i metadati?

Risposta (max 5 righe): I metadati descrivono i dati: significato, domini, unità, provenienza, mapping tra schemi. In data integration servono per allineare attributi tra fonti, risolvere ambiguità e gestire trasformazioni coerenti.

06. Cosa si intende per identificazione delle ridondanze? Come si possono gestire?

Risposta (max 5 righe): Ridondanza = informazioni duplicate/correlate tra attributi o record (es. duplicati, attributi derivati). Si identifica con analisi di correlazione/chi-quadro, matching di entità e regole di dipendenza; si gestisce rimuovendo duplicati o scegliendo una fonte autorevole.

07. Cosa è il test del chi quadro?

Risposta: A) Come puo essere utilizzato nella fase di data integration?

#### Lezione 009

01. Cosa si intende per riduzione dei dati lossy?

Risposta: A) La tecnica utilizzata per la riduzione consente di ricostruire i dati originali solo tramite una approssimazione degli stessi

02. Per cosa può essere utilizzata la principal component analysis?

Risposta: D) Riduzione della dimensionalità

03. Cosa si intende per riduzione dei dati lossless?

Risposta: A) La tecnica utilizzata consente di ricostruire i dati originali da quelli ridotti senza alcuna perdita di informazione

04. Per cosa può essere utilizzata la trasformata discreta wavelet?

Risposta: B) Riduzione della dimensionalità

05. Discutere la trasformata discreta wavelet e suoi vantaggi

Risposta (max 5 righe): La DWT rappresenta un segnale/dati su basi wavelet, separando componenti a diverse scale (approssimazioni + dettagli). Vantaggi: buona compressione, cattura feature locali, utile per riduzione dimensionalità e denoising mantenendo pattern rilevanti.

06. Discutere il problema della dimensionalità e come la riduzione dei dati aiuta a risolverlo.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Discutere l'idea di base della PCA

Risposta (max 5 righe): La PCA cerca nuove variabili (componenti) come combinazioni lineari degli attributi originali che massimizzano la varianza. Ordinando le componenti per varianza spiegata, si mantengono le prime k riducendo dimensionalità con minima perdita informativa.

08. Quali sono i casi in cui conviene maggiormente usare la PCA oppure la trasformata discreta wavelet?

Risposta: A) Set Domande : DATA MINING B

Lezione 010

01. Quali fra le seguenti sono tecniche parametriche per la riduzione della numerosità?

Risposta: D) Regressione lineare e modelli log-lineari

02. La selezione degli attributi è in sostanza in problema di:

Risposta: A) Ottimizzazione

03. Nella strategia di selezione in avanti, il processo di selezione di attributi in genere comincia con:

Risposta: C) Un insieme iniziale di attributi vuoto

04. Nella strategia di selezione all'indietro, il processo di selezione di attributi in genere comincia con:

Risposta: D) Un insieme iniziale contenente tutti gli attributi

05. Quale fra le seguenti non è una tecnica non parametrica per la riduzione della numerosità?

Risposta: B) Regressione lineare e modelli log-lineari

06. Quali delle seguenti tecniche di classificazione possono essere usate anche per la selezione degli attributi?

Risposta: C) Alberi di decisione

07. Discutere la mutua informazione fra due attributi/variabili e spiegare come può essere utilizzata nel processo di selezione di attributi.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

08. Cosa si intende per attributi ridondanti e irrilevanti?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. Discutere le differenze fra metodi wrapper e metodi filter per la selezione di attributi.

Risposta (max 5 righe): Filter: seleziona attributi usando criteri statistici indipendenti dal modello (veloce, generale). Wrapper: valuta sottoinsiemi addestrando un modello e misurando performance (più accurato ma costoso e rischio overfitting).

10. Cosa si intende per metodo greedy?

Risposta: A) Set Domande : DATA MINING B

Lezione 011

01. A cosa serve la normalizzazione dei dati?

Risposta: C) Viene spesso utilizzata nella fase di data transformation per fare in modo che tutti gli attributi abbiano lo stesso peso o influenza

02. Se una tecnica di discretizzazione dei dati utilizza le etichette delle classi e le informazioni da esse derivanti di che tipologia è?

Risposta: B) Non supervisionata

03. Cosa si intende per splitting?

Risposta: D) E' fase di un processo di discretizzazione in cui si identificano ricorsivamente uno o più punti in cui suddividere gli intervalli di definizione degli attributi

04. Discutere una tecnica per la generazione di gerarchie di concetti per dati nominali.

Risposta (max 5 righe): Sono strutture che organizzano valori in livelli di astrazione (es. città→regione→stato). Utili per generalizzazione/summarization, OLAP e discretizzazione nominale, e per ridurre granularità migliorando robustezza dei modelli.

05. Cosa sono le gerarchie di concetti? Per cosa possono essere utili?

Risposta (max 5 righe): Sono strutture che organizzano valori in livelli di astrazione (es. città→regione→stato). Utili per generalizzazione/summarization, OLAP e discretizzazione nominale, e per ridurre granularità migliorando robustezza dei modelli.

06. Discutere la discretizzazione dei dati e la tassonomia delle tecniche esistenti.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Cosa si intende per data transformation e quali sono le principali tecniche ad essa collegate?

Risposta: A) Set Domande : DATA MINING B

Lezione 012

01. Quali sono le condizioni che deve soddisfare una misura di similarità fra due punti?

Risposta: D) Riflessività, non negatività, simmetria

02. Quali sono le condizioni che deve soddisfare una misura di distanza fra due punti?

Risposta: C) Riflessività, non negatività, simmetria

03. Quando una disanza si definisce metrica?

Risposta: C) Quando rispetta anche la condizione di diseguaglianza triangolare

04. Quale è la differenza fra distanza e similarità?

Risposta (max 5 righe): La distanza misura “quanto sono diversi” (0 se uguali, cresce con la dissimilarità). La similarità misura “quanto sono simili” (massima quando coincidono).

Spesso sono trasformabili:  $\text{sim} = 1/(1+\text{dist})$  o  $\text{sim} = 1-\text{norm}(\text{dist})$ , a seconda del contesto.

05. Cosa sono la matrice dei dati e la matrice di dissimilarità?

Risposta (max 5 righe): La matrice dei dati è una tabella  $n \times p$  ( $n$  istanze,  $p$  attributi). La matrice di dissimilarità è  $n \times n$  e contiene per ogni coppia di istanze la distanza (o dissimilarità) calcolata con una metrica scelta.

06. Cosa è la diseguaglianza triangolare?

Risposta: A) Fornire un semplice esempio grafico.

Lezione 013

01. Nella formula della distanza di Mahalanobis che tipo di matrice è presente?

Risposta: C) Di covarianze fra i due punti

02. La distanza Manhattan è un caso particolare di quale distanza?

Risposta: D) Minkowski

03. Quali sono le proprietà della distanza Minkowski?

Risposta: B) Definita positiva, simmetrica e gode di diseguaglianza triangolare

04. Definire la distanza di Minkowski e discutere le sue proprietà.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

05. Cosa si intende per ordine della distanza Minkowski?

Risposta (max 5 righe): L'ordine  $p$  è il parametro che controlla come si aggregano le differenze coordinate-per-coordinate. All'aumentare di  $p$ , pesano di più le differenze grandi; casi noti:  $p=1$  Manhattan,  $p=2$  Euclidea,  $p \rightarrow \infty$  Chebyshev.

06. Quali sono le altre distanze che si ottengono al variare dell'ordine della distanza Minkowski?

Risposta: B) INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

Lezione 014

01. Che tipo di distanza si può utilizzare fra due punti descritti da variabili binarie asimmetriche?

Risposta: B) Jaccard

02. Che tipo di distanza si può utilizzare fra due punti descritti da variabili binarie simmetriche?

Risposta: B) SMC

03. Quando una variabile binaria si dice simmetrica?

Risposta: B) Se la probabilità che ciascuno dei suoi stati si verifichi (che assuma valore zero o uno) è uguale ed entrambi gli stati assumono lo stesso peso

04. Discutere i passaggi per calcolare la distanza fra due vettori che contengono variabili ordinali

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

05. Quale è la differenza fra la distanza e la similarità di Jaccard?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

06. Cosa si intende per ranking di una variabile ordinale.

Risposta: A) Set Domande : DATA MINING B

Lezione 015

01. La formula della similarità coseno fra due vettori utilizza al suo interno?

Risposta: C) La norma dei due vettori e il prodotto scalare fra i due vettori

02. E' possibile calcolare la distanza fra istanze descritte da attributi nominali?

Risposta: D) Si utilizzando una distanza basata sul matching

03. Che tipo di distanza si può usare fra due istanze descritte con attributi di tipologia mista?

Risposta: B) Nessuna delle altre alternative

04. Discutere come si calcola la distanza fra oggetti descritti con attributi di tipologia mista.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

05. Come si calcola la similarità coseno e in che contesto si utilizza?

Risposta (max 5 righe):  $\text{SimCos}(x,y) = (x \cdot y) / (\|x\| \cdot \|y\|)$ . Misura l'angolo tra vettori (indipendente dalla norma) ed è molto usata in text mining/IR con rappresentazioni bag-of-words o TF-IDF.

06. La similarità coseno è anche una distanza. Vero o falso?

Risposta: A) Giustificare la Risposta.

## Lezione 016

01. Che tipo di apprendimento si utilizza quando si creano modelli per il clustering?  
Risposta: A) Non supervisionato
02. Gli algoritmi di clustering si utilizzano per generare:  
Risposta: B) Modelli descrittivi
03. Con riferimento agli algoritmi di clustering, quali delle seguenti affermazioni è falsa?  
Risposta: B) E' sempre noto a priori il numero di cluster da cercare
04. Un algoritmo di clustering produce cluster di alta qualità se assicura:  
Risposta: C) Alta similarità intra-cluster e bassa similarità inter-cluster
05. Discutere la principale differenza fra clustering e classificazione.  
Risposta (max 5 righe): Classificazione: supervisionata, classi note e si impara a predire l'etichetta. Clustering: non supervisionato, nessuna etichetta; si cercano gruppi naturali massimizzando similarità intra-cluster e minimizzando inter-cluster.
06. Quali sono i principali requisiti che devono assicurare gli algoritmi di clustering?  
Risposta (max 5 righe): Tipici requisiti: scalabilità, gestione rumore/outlier, capacità di trovare cluster di forme diverse, robustezza a dimensionalità alta, interpretabilità e minimo bisogno di parametri (o metodi per stimarli).
07. Definire il concetto di cluster.  
Risposta (max 5 righe): Un cluster è un insieme di oggetti tra loro simili (secondo una misura scelta) e dissimili rispetto agli oggetti in altri cluster. È quindi un gruppo "compatto" internamente e separato dagli altri.
08. Discutere la principale differenza fra apprendimento supervisionato e non supervisionato.

Risposta: A) Set Domande : DATA MINING B

## Lezione 017

01. L'algoritmo di clustering SOM (self organizing map) a quale categoria di algoritmi appartiene?  
Risposta: D) Nessuna delle altre alternative
02. Quale tipologia di algoritmi di clustering è più efficiente per l'identificazione di outlier?  
Risposta: C) Density-based
03. L'algoritmo di clustering AGNES a quale categoria di algoritmi appartiene?  
Risposta: A) Gerarchici
04. L'algoritmo di clustering k-means a quale categoria di algoritmi appartiene?  
Risposta: C) Partizionali
05. Discutere le principali caratteristiche degli algoritmi di clustering density-based.  
Risposta (max 5 righe): Gli algoritmi density-based (es. DBSCAN) definiscono cluster come regioni di alta densità separate da regioni di bassa densità. Identificano core/border/noise, trovano forme arbitrarie e gestiscono bene outlier, ma richiedono parametri ( $\epsilon$ , MinPts).
06. Come funzionano gli algoritmi di clustering gerarchici?  
Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.
07. Discutere la tassonomia degli algoritmi di clustering: come si raggruppare tali algoritmi?  
Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.
08. Quale è la principale differenza fra algoritmi di clustering partizionali e density-based?

Risposta: A) Set Domande : DATA MINING B

#### Lezione 018

01. Quale fra i seguenti criteri non rappresenta un criterio di convergenza dell'algoritmo di clustering k-means?

Risposta: A) Riduzione eccessiva delle partizioni iniziali

02. Quali sono i punti di forza dell'algoritmo di clustering K-means?

Risposta: C) 1) Efficienza, in quanto è lineare rispetto al numero di oggetti, 2) Termina in un ottimo locale

03. Quali sono i principali parametri che occorre fissare per effettuare il clustering tramite l'algoritmo K-means?

Risposta: A) Il numero K di cluster e il criterio di stop

04. Quale è la differenza fra cluster e centroide nell'algoritmo di clustering K-means?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

05. Discutere l'idea di base nell'algoritmo di clustering K-means?

Risposta (max 5 righe): K-means sceglie K centroidi iniziali, assegna ogni punto al centroide più vicino, ricalcola i centroidi come media dei punti assegnati e ripete finché converge. Minimizza la somma delle distanze quadrate intra-cluster (SSE).

06. Scrivere in pseudo codice o in un linguaggio di programmazione conosciuto l'algoritmo di clustering K-means?

Risposta (max 5 righe): Pseudo: init centroidi c1..cK; repeat: (1) assegna ogni xi al cluster argmin\_j d(xi,cj); (2) aggiorna cj=media dei punti in cluster j; until assegnazioni stabili o ΔSSE < soglia o maxIter.

07. Quale è la misura che si ottimizza nell'algoritmo di clustering K-means?

Risposta: A) Set Domande : DATA MINING B

#### Lezione 019

01. L'algoritmo gerarchico di clustering DIANA quale approccio segue?

Risposta: D) Divisivo

02. L'algoritmo gerarchico di clustering AGNES quale approccio segue?

Risposta: B) Agglomerativo

03. Che cosa è un dendogramma?

Risposta: C) E' un diagramma ad albero che mostra le sequenze di fusioni fra cluster generati man mano da un algoritmo di clustering gerarchico

04. Negli algoritmi di clustering gerarchici con approccio agglomerativo come avviene l'inizializzazione?

Risposta: B) Si parte con tanti cluster quanti sono gli oggetti

05. Discutere i passi principali dell'algoritmo DIANA.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

06. Discutere lo schema base di un algoritmo di clustering gerarchico agglomerativo.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Come si calcola la distanza fra due cluster nati dalla fusione di due o più cluster?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

08. Che cosa è il dendogramma?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. Quali sono i pro e contro degli algoritmi di clustering gerarchico?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

10. Quale è la principale differenza fra gli algoritmi AGNES e DIANA?

Risposta: A) Set Domande : DATA MINING B

Lezione 020

01. Quali sono i principali parametri che occorre fissare per effettuare il clustering tramite l'algoritmo DBSCAN?

Risposta: C) Nessuna delle altre alternative

02. Oltre al valore del massimo raggio del vicinato, quale altro parametro deve essere specificato per l'algoritmo di clustering DBSCAN?

Risposta: D) Il minimo numero di punti che devono essere contenuti all'interno di un vicinato

03. Quando si parla di algoritmo di clustering DBSCAN, cosa sono i core point?

Risposta: C) Sono i punti la cui densità (numero di punti) è superiore a MinPts

04. Cosa rappresenta il parametro Epsilon dell'algoritmo di clustering DBSCAN?

Risposta: C) Il massimo raggio del vicinato

05. Discutere i pro e i contro dell'algoritmo di clustering DBSCAN.

Risposta (max 5 righe): Pro: trova cluster di forma arbitraria, gestisce rumore/outlier, non richiede K. Contro: scelta di  $\epsilon$  e MinPts non banale, difficoltà con densità variabile e in alta dimensionalità (distanze meno informative).

06. Mostrare una implementazione del DBSCAN usando uno pseudocodice o un linguaggio di programmazione.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Discutere l'idea di base dell'algoritmo di clustering DBSCAN

Risposta (max 5 righe): DBSCAN costruisce cluster come insiemi di punti densamente connessi: un punto è core se ha  $\geq$ MinPts nel suo  $\epsilon$ -vicinato. Dai core si espandono i cluster includendo border; i punti non raggiungibili diventano noise.

08. Definire e mostrare graficamente cosa sono i Core point, Border Point e Noise Point nel contesto dell'algoritmo DBSCAN.

Risposta: -)

Lezione 021

01. Che cosa è il supporto di una regola associativa?

Risposta: D) È la frequenza relativa delle transazioni nel database che verificano la regola

02. Che cosa è la confidenza di una regola associativa?

Risposta: A) È la frequenza delle transazioni nel database che verificano la regola rispetto a quelle che ne verificano l'antecedente

03. Quando una regola associativa si definisce forte?

Risposta: A) Si definisce forte, una regola associativa che soddisfa un supporto minimo prefissato ed una confidenza minima prefissata

04. La generazione di regole associative avviene utilizzando che tipologia di apprendimento?

Risposta: D) Non supervisionato

05. Che cosa è una regola associativa?

Risposta: A) E' un'implicazione della forma X->Y, dove X è un itemset ed Y è un item

06. Cosa è la Market Basket Analysis?

Risposta: B) Una analisi ha l'obiettivo di studiare la regolarità, all'interno delle transazioni registrate, nelle vendite dei supermercati

07. Che cosa è un itemset?

Risposta: A) Un insieme di articoli

08. Quando una regola associativa si dice forte?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. Quale è la differenza fra supporto di un itemset e supporto di una regola associativa?

Risposta: -)

11. Definire i concetti di Item, itemset e transazioni.

Risposta: A) Set Domande : DATA MINING B

Lezione 022

01. A cosa serve l'algoritmo APRIORI?

Risposta: C) Per generare regole associative

02. Quale è l'idea di base dell'algoritmo APRIORI?

Risposta: A) L'algoritmo Apriori affronta la fase di generazione degli itemset frequenti per approssimazioni successive, a partire dagli itemset con un solo elemento

03. Che cosa è un itemset frequente?

Risposta: -)

04. Per la generazione di regole associative, in genere è necessario generare come prima cosa:

Risposta: C) Gli itemset frequenti

05. Su quali basi teoriche si fonda l'algoritmo Apriori?

Risposta (max 5 righe): Si basa sulla proprietà di antimonotonicità del supporto: se un itemset è infrequente, tutti i suoi sovrainsiemi sono infrequent. Questo permette di potare i candidati e ridurre drasticamente lo spazio di ricerca.

06. Cosa si intende per association rule mining?

Risposta (max 5 righe): È il processo di estrazione di regole  $X \rightarrow Y$  che descrivono co-occorrenze frequenti tra item in un database transazionale. Si usa tipicamente supporto e confidenza (e altre misure come lift) per selezionare regole interessanti.

07. Discutere brevemente i passi dell'algoritmo Apriori per la generazione di itemset frequenti.

Risposta: B) INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

Lezione 023

01. Quali sono i parametri più importanti da fissare per l'algoritmo APRIORI?

Risposta: A) Valore del supporto minimo, valore della confidenza minima

02. Che approccio usa APRIORI per la generazione di regole associative?

Risposta: -)

03. E' possibile aumentare l'efficienza dell'algoritmo APRIORI?

Risposta: C) Si, ma solo riducendo il numero di candidati da considerare, usando tecniche di indirizzamento e partizionamento

04. Mostrare un esempio di traliccio delle regole associative contenente regole potate.

Risposta (max 5 righe): Esempio: con item {A,B,C}, dal traliccio degli itemset si potano tutti i supersets di un itemset infrequente. Se {A,B} è infrequente, allora {A,B,C} viene potato e non considerato come candidato.

05. Cosa è la proprietà di antimonotinicità? Vale per il supporto, per la confidenza o per entrambi?

Risposta (max 5 righe): Antimonotonicità: se  $X \subseteq Y$  allora  $\text{support}(Y) \leq \text{support}(X)$ . Vale per il supporto (non per la confidenza). È la base del pruning in Apriori; la confidenza non è antimonotona.

06. Riportare e discutere lo pseudocodice dell'algoritmo Apriori per la generazione di regole associative forti (si consideri di avere già gli itemset frequenti).

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Come si può migliorare l'efficienza dell'algoritmo Apriori?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

08. Quale è l'idea di base per la generazione di regole associative forti in Apriori?

Risposta: A) Set Domande : DATA MINING B

## Lezione 024

01. Con riferimento all'algoritmo FP-growth, quante scansioni del data base vengono effettuate per la generazione dell'FP-tree?

Risposta: A) Due

02. L'algoritmo FP-growth consente di evitare cosa?

Risposta: D) La generazione degli itemset frequenti candidati

03. Con riferimento all'algoritmo FP-growth, cosa viene fatto durante la seconda scansione nella fase di generazione dell'FP-tree?

Risposta: C) 1), Si identificano tutti gli 1-itemset frequenti dal database, 2) Si crea una lista L degli 1-itemset frequenti ordinati secondo il loro supporto o frequenza

04. Con riferimento all'algoritmo FP-growth, cosa viene fatto nei primi due passi della generazione dell'FP-tree?

Risposta: A) 1), Si identificano tutti gli 1-itemset frequenti dal database, 2) Si crea una lista L degli 1-itemset frequenti ordinati secondo il loro supporto o frequenza

05. A serve l'FP-tree?

Risposta: C) A comprimere il database considerando solo gli item frequenti

06. Quale è l'idea di base di FP-growth?

Risposta (max 5 righe): FP-growth evita la generazione di candidati comprimendo il DB in un FP-tree. Poi estrae ricorsivamente i pattern frequenti tramite conditional pattern base e conditional FP-tree, esplorando l'albero dal basso verso l'alto.

07. Quali sono i due passi principali dell'algoritmo FP-growth?

Risposta (max 5 righe): 1) Costruzione dell'FP-tree (2 scansioni: conta 1-itemset frequenti e inserisci transazioni filtrate/ordinate). 2) Mining ricorsivo: per ogni item nella header table crea conditional pattern base/FP-tree e genera i frequent pattern.

08. Discutere le principali differenze fra Apriori ed FP-growth

Risposta (max 5 righe): Apriori: genera molti candidati e richiede molte scansioni del DB (costoso su grandi dati). FP-growth: comprime in FP-tree con 2 scansioni e fa mining senza candidati (più efficiente), ma può avere problemi di memoria se l'albero è grande.

09. Che cosa è un FP-tree e come si crea?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

10. Cosa sono e a cosa servono la header table e i node-links?

Risposta: A) Set Domande : DATA MINING B

Lezione 025

01. Con riferimento all'algoritmo FP-growth, che filosofia si utilizza per la generazione degli itemset frequenti?

Risposta: C) Si segue un approccio bottom-up: si parte ad esplorare l'albero dei pattern frequenti a partire dalle foglie verso la radice

02. Quale potrebbe essere uno svantaggio dell'algoritmo FP-growth?

Risposta: D) L'FP-tree potrebbe essere computazionalmente difficile da generare

03. Quale potrebbe essere uno svantaggio dell'algoritmo FP-growth?

Risposta: C) L'FP-tree potrebbe non entrare in memoria

04. In quale fase dell'algoritmo FP-growth si genera il conditional pattern tree?

Risposta: A) Nella fase di estrazione degli itemset frequenti

05. Come si genera un conditional FP-tree?

Risposta (max 5 righe): Si parte dal conditional pattern base di un item (insieme dei percorsi prefix che portano a quell'item con conteggi). Si filtrano item non frequenti rispetto al minSup e si costruisce un FP-tree “condizionato” su quell'item, su cui si mina ricorsivamente.

06. In che punto dell'algoritmo FP-growth è presente la parte ricorsiva?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Discutere la differenza fra: suffix pattern base, conditional pattern base e conditional FP-tree.

Risposta (max 5 righe): Si parte dal conditional pattern base di un item (insieme dei percorsi prefix che portano a quell'item con conteggi). Si filtrano item non frequenti rispetto al minSup e si costruisce un FP-tree “condizionato” su quell'item, su cui si mina ricorsivamente.

08. Come si costruisce un conditional pattern base?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. Come si identificano i pattern frequenti nell'algoritmo FP-growth?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

10. Quali sono i principali vantaggi e svantaggi dell'FP-growth?

Risposta: B) INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

Lezione 026

01. Cosa si intende per classificazione binaria?

Risposta: C) Si considerano solo due classi

02. A che categoria di classificatori appartiene il KNN?

Risposta: D) Lazy

03. Che tipo di apprendimento si utilizza quando si creano modelli di classificazione?

Risposta: A) Semi supervisionato

04. Come sono i dataset che in genere si utilizzano nel contesto della classificazione?

Risposta: A) Etichettati

05. Esistono differenze fra problemi di clustering e di classificazione?

Risposta: A) Si

06. Discutere la differenza fra classificazione binaria e classificazione multi-classe.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

07. Discutere le principali differenze fra classificazione e clustering.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

08. Cosa sono i multi-classificatori?

Risposta (max 5 righe): Sono strategie per estendere classificatori binari a multi-classe: one-vs-rest, one-vs-one o strutture gerarchiche. Si addestrano più modelli e si combina la decisione (voto o punteggi) per scegliere la classe finale.

09. Riassumere brevemente le fasi del processo di classificazione.

Risposta: A) Set Domande : DATA MINING B

## Lezione 027

01. Su cosa si basa il calcolo dell'information gain?

Risposta: B) Sull'entropia

02. Che tipo di problema risolvono gli alberi di decisione?

Risposta: D) Classificazione

03. Cosa rappresenta ogni nodo interno in un albero di decisione?

Risposta: B) Un possibile valore di un attributo

04. Cosa rappresenta una foglia in un albero di decisione?

Risposta: B) Il valore predetto per la classe

05. Con riferimento agli alberi di decisione, che filosofia si utilizza per la generazione della classe stimata?

Risposta: D) Si segue un approccio top-down: si parte ad esplorare l'albero dalla radice verso le foglie

06. Nella generazione degli alberi di decisione, l'information gain in genere viene usato come misura di cosa?

Risposta: A) Misura di rilevanza

07. Discutere l'approccio top-down per generare alberi di decisione.

Risposta (max 5 righe): Si parte dalla radice con tutti i campioni. A ogni nodo si sceglie l'attributo migliore (es. max information gain/gain ratio/Gini) e si split il dataset. Si ripete ricorsivamente sui sottoinsiemi finché una condizione di stop è soddisfatta.

08. Quale è l'euristica di base per generare alberi di decisione?

Risposta (max 5 righe): L'euristica base è scegliere, a ogni nodo, l'attributo che rende i sottoinsiemi più "puri" (massimizza information gain/gain ratio o minimizza impurità Gini/entropia), perché porta a decisioni più informative.

09. Come avviene il processo di classificazione con alberi di decisione?

Risposta (max 5 righe): Per classificare un'istanza si percorre l'albero dalla radice: a ogni nodo si verifica il test sull'attributo e si segue il ramo corrispondente. Arrivati a una foglia, si assegna la classe prevista (o una distribuzione di probabilità).

10. Quali sono le condizioni di terminazione per la generazione di alberi di decisione, utilizzando l'approccio top-down?

Risposta (max 5 righe): Si parte dalla radice con tutti i campioni. A ogni nodo si sceglie l'attributo migliore (es. max information gain/gain ratio/Gini) e si split il dataset. Si ripete ricorsivamente sui sottoinsiemi finché una condizione di stop è soddisfatta.

11. Discutere i concetti di entropia e information gain.

Risposta: A) Set Domande : DATA MINING B

#### Lezione 028

01. A che categoria di classificatori appartiene il classificatore C45?

Risposta: B) Alberi di decisione

02. A che categoria di classificatori appartiene il classificatore ID3?

Risposta: D) Nessuna delle altre alternative

03. Con che tipo di attributi lavora il classificatore ID3?

Risposta: A) Nominali

04. Che cosa si usa come misura di rilevanza degli attributi nell'algoritmo ID3?

Risposta: -)

05. Quale fra le seguenti condizioni di terminazione della ricorsione dell'algoritmo ID3 non è corretta?

Risposta: B) Nessuna delle altre alternative

06. Che tipo di attributi può gestire senza pre-elaborazione il classificatore C45 rispetto al classificatore ID3?

Risposta: C) Nominali

07. Dopo la generazione di un albero con l'algoritmo C45, che ulteriore passo può essere effettuato per ridurre la possibilità di overtraining?

Risposta: A) Potatura

08. Come si identificano gli split point quando si utilizzano attributi reali?

Risposta (max 5 righe): Si ordinano i valori dell'attributo, si considerano candidati tra valori consecutivi (tipicamente a metà) e si sceglie lo split che massimizza la metrica (gain/gain ratio o riduzione Gini).

09. Perché l'algoritmo C4.5 prevede una fase di potatura dell'albero?

Risposta: C) INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

#### Lezione 029

01. A che categoria di classificatori appartiene il classificatore CART?

Risposta: D) Alberi di decisione

02. Cosa utilizza l'algoritmo CART per misurare l'impurità di un database o di un generico insieme di transazioni

Risposta: C) L'indice di Gini

03. Quale misura di rilevanza è stata introdotta in una versione migliorata dell'algoritmo C45?

Risposta: B) Il gain ratio

04. Quale è la migliore misura di rilevanza degli attributi?

Risposta (max 5 righe): Non esiste una misura "migliore" in assoluto: dipende dai dati e dagli obiettivi. Information gain può favorire attributi con molti valori; gain ratio corregge questo bias; Gini è efficace e veloce (usato in CART). Si sceglie in base a contesto e prestazioni.

05. Paragonare brevemente le misure di rilevanza degli attributi.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

06. Discutere come si utilizza l'indice di Gini per la selezione dell'attributo corrente significativo.

Risposta: B) INGEGNERIA INFORMATICA E DELL'AUTOMAZIONE

Lezione 030

01. I modelli bayesiani si usano per risolvere quali tipi di problemi?

Risposta: C) Classificazione

02. Quale è l'assunzione su cui si basano i classificatori Naïve Bayes?

Risposta: A) L'indipendenza fra gli attributi che descrivono un oggetto quando la classe è nota

03. Su cosa si basano i classificatori bayesiani?

Risposta: C) Teoria della probabilità

04. In cosa consiste la creazione del modello di classificatore bayesiano?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

05. Introdurre il teorema di Bayes.

Risposta (max 5 righe): Teorema di Bayes:  $P(C|X) = P(X|C) \cdot P(C) / P(X)$ . Permette di invertire probabilità condizionate. Nei classificatori si stima la classe massimizzando la posterior  $P(C|X)$ .

06. Discutere la differenza fra probabilità marginale, probabilità congiunta e probabilità condizionale.

Risposta (max 5 righe): Marginale:  $P(A)$  (prob. di A). Congiunta:  $P(A,B)$  (A e B insieme). Condizionale:  $P(A|B)=P(A,B)/P(B)$ , cioè prob. di A sapendo che B è vero.

07. Discutere la principale differenza fra information gain e gain ratio.

Risposta (max 5 righe): Information gain misura la riduzione di entropia ma tende a preferire attributi con molti valori. Gain ratio normalizza l'information gain dividendo per l'entropia dello split (split info), riducendo questo bias.

08. Come si possono stimare le probabilità condizionate da utilizzare nell'apprendimento dei parametri dei classificatori Naïve Bayes?

Risposta: A) Set Domande : DATA MINING B

Lezione 031

01. Come avviene la classificazione di un nuovo punto quando si utilizza il classificatore KNN?

Risposta: B) Si considerano tutti i K punti del training set vicini al punto da classificare e si sceglie a caso la classe che ricorre maggiormente fra i vicini

02. Quando si usano i classificatori lazy che tipo di apprendimento si usa per la costruzione del modello?

Risposta: D) Nessuna delle altre alternative

03. Quali fra i seguenti elementi non deve mai mancare quando si utilizza un classificatore KNN?

Risposta: B) Tutti e tre gli elementi specifici nelle altre alternative

04. In quali classificatori si può usare il majoring voting pesato?

Risposta: B) KNN

05. Cosa è il rote classifier?

Risposta: C) Il più semplice classificatore lazy

06. Mostrare e discutere uno pseudocodice per un classificatore KNN.

Risposta (max 5 righe): Dato  $x$ : calcola distanza da tutti i punti di training; seleziona i  $K$  più vicini; raccogli le loro classi; predici la classe con maggioranza (eventualmente pesata con  $1/distanza$ ).

07. Discutere vantaggi e svantaggi dei classificatori KNN.

Risposta (max 5 righe): Pro: semplice, nessun training esplicito, buone prestazioni con dati ben separati. Contro: predizione lenta (serve tutto il training), sensibile a scala/metriche e rumore, soffre in alta dimensionalità; scelta di  $K$  cruciale.

08. Discutere l'importanza della funzione di distanza nei classificatori KNN.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. Che cosa è il majority voting pesato?

Risposta: A) Set Domande : DATA MINING B

Lezione 032

01. In quali tipi di classificatori si può utilizzare il FOIL\_gain?

Risposta: D) Classificatori a regole

02. Con riferimento ai classificatori basati su regole, cosa è la copertura di una regola?

Risposta: D) Frazione delle istanze nel database che soddisfano l'antecedente

03. Con riferimento ai classificatori basati su regole, cosa è l'accuratezza di una regola?

Risposta: A) Frazione delle istanze nel database che soddisfano antecedente e conseguente rispetto alla copertura

04. Con riferimento ai classificatori basati su regole, cosa sono le regole mutuamente esclusive?

Risposta: C) Un insieme di regole  $R$  è detto mutuamente esclusivo se nessuna coppia di regole può essere attivata dalla stessa istanza

05. Con riferimento ai classificatori basati su regole, cosa sono le regole esaustive?

Risposta: D) Un insieme di regole  $R$  ha una copertura esaustiva se esiste una regola per ogni combinazione di valori degli attributi

06. A che categoria di classificatori appartiene l'algoritmo RIPPER?

Risposta: A) Classificatori a regole

07. Quali sono le modalità di ordinamento delle regole? Perché potrebbe essere necessario ordinare le regole?

Risposta (max 5 righe): Si possono ordinare per accuratezza/confidenza, copertura/supporto, FOIL\_gain o complessità (numero condizioni). Ordinare serve perché, in caso più regole attivabili, si applica la prima (o la migliore) e si gestiscono conflitti tra regole.

08. Discutere brevemente l'algoritmo di classificazione RIPPER

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. A cosa serve il FOIL\_gain?

Risposta (max 5 righe): FOIL\_gain misura quanto l'aggiunta di una condizione a una regola aumenta la capacità di discriminare positivi da negativi. Si usa come euristica per scegliere il letterale migliore durante la costruzione incrementale delle regole.

10. Discutere la differenza fra metodi diretti e indiretti per la generazione di regole?

Risposta: A) Set Domande : DATA MINING B

Lezione 033

01. Cosa deve simulare un percettrone?

Risposta: A) Il funzionamento di un neurone

02. Su cosa si basa l'algoritmo di classificazione RIPPER?

Risposta: B) Sequential covering

03. Che cosa è il percettrone?

Risposta: C) L'elemento base delle reti neurali artificiali

04. Cosa significa addestrarre un percettrone?

Risposta: B) Nessuna delle altre alternative

05. Di che tipo è l'uscita del percettrone?

Risposta: B) Binaria

06. Come si chiama l'algoritmo di apprendimento dei parametri di un percettrone?

Risposta: C) Delta rule

07. Quali sono i parametri caratteristici di un rete neurale artificiale?

Risposta: A) L'insieme dei neuroni, l'insieme dei pesi associati ai collegamenti, l'insieme delle soglie o dei livelli di attivazione

08. In cosa consiste l'addestramento di un percettrone? Discutere l'algoritmo delta rule.

Risposta (max 5 righe): Delta rule aggiorna i pesi:  $w \leftarrow w + \eta \cdot (t - y) \cdot x$ , dove t è target e y output. Si ripete su esempi (online o batch) finché l'errore diminuisce; per il percettrone classico l'output è binario e l'aggiornamento avviene quando c'è errore.

09. Discutere le principali funzioni di attivazione che possono essere usate nei percettroni.  
Set Domande : DATA MINING B INGEGNERIA INFORMATICA E  
DELL'AUTOMAZIONE  
Docente: Antonelli Michela . Cosa è il percettrone? Disegnare la sua struttura e discutere i suoi parametri.

Risposta (max 5 righe): Funzioni comuni: soglia/step (percettrone), sigmoide logistica, tanh, ReLU (e varianti). Servono a introdurre non-linearietà e a mappare la somma pesata in un output interpretabile.

11. In cosa consiste la progettazione di una rete neurale?

Risposta (max 5 righe): Consiste nel definire architettura (numero strati e neuroni), funzioni di attivazione, funzione di costo, algoritmo di training (backprop), iperparametri (learning rate, epoche, regolarizzazione) e criteri di stop/validazione.

12. Quali sono le principali componenti di una rete neurale?

Risposta: A) Set Domande : DATA MINING B

Lezione 034

01. Come avviene la generazione dei pesi iniziali nell'algoritmo backpropagation?

Risposta: A) A caso

02. Quale fra le seguenti caratteristiche non è tipica di una rete neurale feedforward?

Risposta: A) Presenza di pesi sulle connessioni

03. Quale fra le seguenti caratteristiche non è tipica di una rete neurale feedforward?

Risposta: A) Presenza di pesi sulle connessioni

04. Quante uscite in genere si usano nelle reti neurali feedforward quando si vuole risolvere un problema di classificazione con C classi?

Risposta: A) Dipende dal numero di campioni disponibili per ciascuna classe

05. Che problemi possono essere risolti con le reti neurali feedforward?

Risposta: C) Classificazione

06. Come si chiama l'algoritmo di apprendimento delle reti feedforward?

Risposta: D) Back propagation

07. Cosa sono le epoche di addestramento di una rete feedforward?

Risposta (max 5 righe): Un'epoca è un passaggio completo dell'algoritmo di addestramento su tutto il training set. Più epocha permettono di aggiornare i pesi più volte; troppe possono portare overfitting se non si usa early stopping/validazione.

08. Come si possono utilizzare le reti feedforward per la classificazione?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

09. Come avviene l'apprendimento di una rete feedforward?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

10. Disegnare e discutere lo schema e i parametri di una rete feedforward.

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

11. Perché sono state introdotte le reti multi-strato? Non bastava il semplice perceptrone?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

12. Cosa è la matrice di confusione?

Risposta: A) Discutere sia il caso di classificazione binaria sia il caso di classificazione multi-classe.

14. Quali sono i problemi e le possibili soluzioni che si possono incontrare nell'addestramento di una rete feedforward?

Risposta (max 5 righe): Risposta sintetica: definisci i concetti chiave richiesti dalla domanda, descrivi il processo/passi principali e cita 1–2 motivazioni o esempi pratici legati al data mining.

15. Come si può effettuare il confronto fra due o più classificatori? Che tipo di test statistici si possono usare?

Risposta (max 5 righe): Si confrontano con cross-validation e metriche (accuracy, F1, AUC). Per verificare differenze significative si usano test statistici: t-test su fold, Wilcoxon, McNemar (su stessi esempi), Friedman/Nemenyi per più modelli.

16. Discutere l'importanza dell'area sotto la curva ROC (AUC).

Risposta (max 5 righe): La ROC mette in grafico TPR (sensibilità) vs FPR (1-specificità) al variare della soglia decisionale. Ogni punto deriva da una soglia diversa sui punteggi del classificatore; l'AUC riassume la capacità di discriminazione.

17. Cosa sono la sensibilità e la specificità?

Risposta (max 5 righe): Sensibilità (TPR/recall): TP/(TP+FN), quota di positivi correttamente trovati. Specificità (TNR): TN/(TN+FP), quota di negativi correttamente riconosciuti.

18. Cosa è e a cosa serve la curva ROC? Come si individuano i punti che la formano?

Risposta (max 5 righe): La ROC mette in grafico TPR (sensibilità) vs FPR (1-specificità) al variare della soglia decisionale. Ogni punto deriva da una soglia diversa sui punteggi del classificatore; l'AUC riassume la capacità di discriminazione.

19. Quali potrebbero essere i criteri di stop per l'addestramento di una rete feedforward?

Risposta: A) Set Domande : DATA MINING B

Lezione 035

01. Quale è la formula per la Recall, utilizzata come metrica per valutare la bontà di un classificatore binari?

Risposta: A)  $R = TP/(TP+FN)$

02. Quale metrica è più consigliata quando si vuole confrontare la bontà di due o più modelli per la classificazione binaria?

Risposta: -)

03. Quale è l'area associata alla curva ROC di un classificatore random?

Risposta: A) 1

04. Cosa viene riportato sugli assi di una curva ROC?

Risposta: A) TNR, FPR

05. Quando è sconsigliato utilizzare l'accuratezza come metrica per valutare la bontà di un classificatore?

Risposta: A) Quando si ha un problema multiclass e tutte le classi sono di interesse

06. Quale strumento può essere utilizzato per valutare la bontà di un classificatore binario?

Risposta: A) Matrice di confusione

07. Quale è la formula per la Precisione, utilizzata come metrica per valutare la bontà di un classificatore binari?

Risposta: A)  $P = TP/(TP+FN)$

## Lezione 051

01. Quale è stata la soluzione proposta per soddisfare i requisiti del progetto?

Risposta (max 5 righe): Nel file non sono presenti i dettagli della soluzione specifica del progetto. In generale una "soluzione proposta" va descritta indicando: dati in input, pre-processing, algoritmo/i scelti e motivazione, schema sperimentale (train/test o CV), metriche e risultati attesi.

02. Quali sono stati i risultati raggiunti con la soluzione proposta per soddisfare i requisiti del progetto? Powered by TCPDF ([www.tcpdf.org](http://www.tcpdf.org))

Risposta (max 5 righe): Nel file non sono presenti i dettagli della soluzione specifica del progetto. In generale una "soluzione proposta" va descritta indicando: dati in input, pre-processing, algoritmo/i scelti e motivazione, schema sperimentale (train/test o CV), metriche e risultati attesi.