**Domande Data Mining M**

10/01/2017

Sartori:

Classificazione e classificatore bayesiano nel dettaglio, tutto e anche complessità; cosa è l'overfitting e come lo si evita nel classificatore bayesiano

Ravaldi:

Datawarehouse e BI, architetture e operatori OLAP  
Map Reduce e Spark

Sartori si concentra principalmente su un argomento e chiede di tutto su quest'ultimo dalle cose scritte sulle slide per passare poi a cose più particolari tipo la complessità. Ad altri ha chiesto il clustering ed in particolare DBSCAN con ragionamenti sul cambiamento del valore di epsilon e del valore di minPoints, ad altri ha chiesto regole associative o ancora il perché della normalizzazione.  
Ravaldi invece fa domande principalmente su DWH, BI, Hadoop e Open Data ma nulla al di fuori di quanto scritto sulle slide.

31/01/2017

Ravaldi:  
- OLTP e OLAP: paragone, con particolare attenzione a OLAP.  
- Open Data: discorso generale, classificazione a 5 stelle.  
- HDFS.  
- Location Intelligence: quale dev'essere la struttura del DWH per gestirli? (Lo schema con il bridge).  
Sartori:   
- DBSCAN: discorso generale, definizioni relative, in cosa consiste l'algoritmo.  
- Quali sono gli indicatori di performance per il Clustering? Riguardo l'SSE possiamo, e se si come, avere informazioni sul numero di Cluster ottimale?  
Visto che la risposta a questa domanda NON è scritta esplicitamente nelle slide la metto io qui: l'SSE non si può usare direttamente, ma si può avere un indizio se si considera che al limite, quando K = N, è uguale a 0 (perché si ha un clustering dove ogni oggetto è un cluster, cioè ogni oggetto è il centroide di sé stesso). Tuttavia, considerando il grafico a sinistra a pagina 65 di "DataMining-03-clustering-a", troviamo che all'inizio l'SSE cala molto e poi diventa più o meno stabile. Il K ottimale è dalle parti di questo "ginocchio".  
Ora, non è per criticare, ma personalmente avrei gradito un discorso di questo genere nelle slide invece di due figure in una slide non commentata. A questo punto consiglierei a chi deve dare l'esame di elucubrare un po' anche sull'indice di silhouette.

31/01/2017

1) A me hanno chiesto  
Ravaldi:   
-Big Data in generale (cosa sono -definizione standard e definizione 5 V ). Perché sono generati. Hadoop generale, HDFS in particolare caratteristiche, MR caratteristiche implicite.  
Per cosa, rispetto a quanto abbiamo studiato, la tecnologia dihadoop non sarebbe efficiente e perché .  
Sartori:  
-Reti neurali tutto, (si, sono 3 slide ma voleva sapere tutto di quelle 3 slide.)  
-Come valutare un classificatore: accuratezza, specificità, test set.  
  
In generale Ravaldi si studia bene sulle slide e non chiede niente di più di quanto scritto. E' molto tranquillo e se non sai la risposta immediatamente ti ci fa arrivare.  
Sartori in realtà, rispetto a quando detto/letto in giro, l'ho trovato un po' più puntiglioso e "molesto". Fino a quando sono rimasto io gli argomenti classici e generali (regole associative-cluster-classificatori) sono stati chiesti molto meno rispetto a domande specifiche e che magari si tralasciano tranquillamente studiando.  
Per il materiale bastano le slide, di sartori quelle in italiano dell'anno scorso hanno cose in piùnon richieste, ma sono più chiare mi sembra. Materiale in più, personalmente parlando solo documentazioni del mondo hadoop e ecosistema (hive, spark, pig...) ma più a livello informativo personale

31/01/2017

- Ravaldi quando mi sono seduto ha detto che era la terza volta che ricominciava la lista dall'inizio. Mi ha chiesto i Data Warehouse chiedendomi da dove volessi partire, gli ho fatto un discorso partendo dal generico per poi parlare di Data Mart collegando le due lienee di pensiero con le possibili architetture.  
Poi Hadoop e HDFS molto in generale, parallelizzazione di MapReduce, come mi comporto se vengono aggiunti altri nodi ad Hadoop (fa tutto lui, si aggiorna da solo), come gestire elaborazioni non parallelizzabili. In generale è molto tranquillo, chiede un po' di tutto, magari anche poco e gli piace se fai collegamenti tra argomenti differenti.  
  
- Sartori mi ha chiesto la classificazione, in generale e nel dettaglio gli Alberi Decisonali.  
Poi mi ha chiesto la Cross Validation  
Anche lui abbastanza tranquillo ma meno di ravaldi, fa sempre una domanda molto generale e una su un argomento preciso.  
  
Come appunti mi son basato sulle slide e ho fatto un riassunto che comprende anche il modulo 2, quando mi torna la voglia lo sistemo dato che ci son alcuni errorini e lo metto sul drive

31/01/2017

Ravaldi:  
-Hadoop in generale, mi ha lasciato parlare di hdfs e mapreduce  
-Data Warehouse  
-DFM e design logico (ROLAP MOLAP e HOLAP), stella e snowflake  
Sartori:  
-Clustering in generale  
-Clustering gerarchico  
-Domanda sul numero di iterazioni minimo nel Single-Linkage, dovrebbe essere proporzionale a log2(N) era una domanda di ragionamento, non c'è nelle slides.  
-Preprocessing dei dati in generale, ho parlato di aggregazione e riduzione della dimensionalità tramite PCA  
Confermo che Sartori cerca di non rifare mai le stesse domande (a meno della introduzione generale), quindi verso la fine si va su argomenti meno scontati

31/01/2017

Ravaldi tra le cose non citate sul questo post, mi ha chiesto la differenza (a grandi linee) tra OLTP e OLAP (slide 57 , lesson 3).

Sartori mi ha chiesto Clustering in generale, Clustering Model Based con algoritmo EM (si, può chiederlo). Poi valutazione di un modello di clustering con dataset supervisionato, indici di misura di Jaccard e Rand e la differenza tra questi ultimi (cioè che Jaccard non considera i punti etichettati come DD) (si, può chiederli).

31/01/2017

Sartori:   
- Percettrone e SVM + una domanda di ragionamento sul percettrone (se ho piu di due classi, come posso distinguerle usando un percettrone? risposta: non posso! devo usarne almeno 2. Se per esempio avessi 4 classi linearmente separabili potreiusare un percettrone per individuare un piano e un secondo percettrone per individuarne un secondo. Il risultato combinato dei due mi distingue una delle quattro classi)

- normalizzazione e standardizzazione: cosa sono, quali algoritmi di classificazione visti le usano (bayes) e quali invece non ne hanno bisogno (alberi)

Ravaldi  
OLAP vs OLTP

operatori OLAP

map reduce (come funziona)

YARN ( che cos'è e perchè ha permesso un sostanziale miglioramento in hadoop)

linked open data (i 4 principi per poter essere un linked open data, fondamentale il 3 punto su rdf)

20/02/2018

Sartori:  
Molto tranquillo, ti mette a tuo agio, però le cose vanno sapute bene e con un buon livello di approfondimento.

A me ha chiesto:  
1) Association Rules con Apriori  
2) Come possiamo sfruttare la gerarchia tra oggetti (esempio: Water fa parte del gruppo/gerarchia beverage, altri item saranno sotto altre gerarchie) nel contesto delle association rules  
3) Come individuare regole ridondanti

Ravaldi:  
Anche lui molto tranquillo, ancora più di Sartori, però secondo me non è una "semplice chiacchierata" come descritto da molti. Livello di approfondimento: medio/basso

Domande:  
1) CAP Theorem  
2) Esempio di booking.com con applicazione CAP Theorem  
3) NOSQL  
4) Hadoop tutto, ma senza entrare nel dettaglio di Spark

Mio punto di vista:  
Il Modulo 1 si prepara con le slides, ma bisogna dedicargli molto tempo e molta concentrazione (15/20 giorni direi)  
Il Modulo 2 si prepara con i riassunti (attenzione agli argomenti inseriti quest'anno e non presenti nei riassunti) in 2/3 di giorni

20/06/2018

Sartori:

1. Dimmi delle Discovery Association Rules, a cosa servono, definizioni di Support, Frequent Itemsets, Confidence, Apriori Algorithm (tutto);
2. Tipi di dato esistenti (Categorici e Nominali, con relative Sottocategorie, tipi di trasformazioni ammissibili, operazioni statistiche che si possono fare per ognuno, cos’è la Mediana e cos’è la Media);
3. Concetti del Clustering, cos’è, misure per stabilire la “bontà” del clustering, metodo empirico per scegliere il “K” più adatto in un algoritmo come K-means (dato un grafico con SSE nell’asse y e numero di “K” nell’asse X, scelgo il punto in cui SSE cala maggiormente e mi fermo se decresce piano piano);
4. Cosa facciamo con i dati “raw” prima di sottoporli agli algoritmi di Mining (tecniche di pulizia dati, campionamento, riduzione della dimensionalità, ecc...);
5. Definizioni di Standardizzazione e Normalizzazione, a cosa servono (per portare i dati ad avere range comuni e dunque annullare le diverse scale di valori su attributi diversi);
6. Classificazione Supervised, principali algoritmi e soluzioni (Neural Network, Perceptron, Decisional Tree, SVM). Di alcuni, non ricordo quali, ha chiesto come funzionano e le principali problematiche;
7. Parlami degli alberi decisionali, over-underfitting, MDL per scegliere il DT “migliore”, come si costruisce il DT e come si stabilisce la lunghezza del DT con MDL (formula annessa);
8. Ensemble Method, che cos’è e come funziona;
9. Density Based Clustering, che cos’è, come funziona, algoritmi specifici funzionamento, quali sono i parametri di un algoritmo Density Based (DBSCAN mi pare ha chiesto) e le proprietà per stabilire i punti appartenenti al cluster data la densità (Directly Density Reachable, Density Reachable, Density Connected);
10. Classificazione in generale e in particolare Naive Bayes Classifier (tutto sul Naive Bayes, anche LAplace Smoothing);
11. Tecniche principali per fare Outlier Detection;
12. Algoritmo Expectation Maximization e come funziona;
13. Association Rules, nel caso delle regole “multidimensional” (item inseriti in colonne separate, non tutti insieme in ogni transazione), come scriviamo gli item nelle colonne, quante sono le colonne, come segnaliamo che in una transazione ci sono solo certi elementi e non altri (il numero di colonne è l’unione di tutti gli item, tolti i duplicati e ne segnaliamo la presenza o meno in ogni transazione con valori binari, 0/1 e quindi le transazioni diventano vettori binari);
14. Da cosa è caratterizzato il k-means, che lo differenzia dagli algoritmi basati su densità (k-means tende a non essere “locale”, quindi non è adatto a regioni non convesse e in più tende intrinsecamente a fare ipersfere mentre DBSCAN e DENCLUE sono più locali, essendo basati su densità entro un certo raggio e numero di punti).

Ravaldi:

1. Definizione di Big Data e loro caratteristiche (5 V dei Big Data), chi genera i Big Data (Organizzazioni, Macchine, Persone);
2. Differenze fra sistemi SQL, No-SQL;
3. Hadoop, HDFS, loro descrizione e funzionamento (fault-tolerance, come sono distribuiti i dati in HDFS, che dati supporta [tutti], parallelismo in Hadoop con Alta Latenza e Alta velocità di trasferimento).. Disk Usage con Map-Reduce tramite HDFS (quindi tante scritture sui dischi locali di dati);
4. Differenza fra visione Schema-on-Read e Schema-on-Write;
5. Concetti OLAP (Data Warehouse, Interrogazioni Online di dati Multidimensionali). Query OLAP e descrizione di ognuna;
6. Map-Reduce, operazioni, a cosa serve, che problemi risolve, principali proprietà, principali limitazioni;
7. Differenza fra computazione Batch e Interattiva;
8. Teorema CAP, esempi di sistemi AP, piuttosto che CP, definizione di Partition Tolerance, che succede quando un sistema è PT e definizione di Eventual Consistency ed un esempio della stessa;
9. Data Warehouse, che cosa sono i processi ETL, a cosa servono, come lavorano;
10. Concetti principali di Spark, cosa migliore di Map-Reduce e di Hadoop, RDD cosa sono e come funzionano, lazy evaluation che cos’è e come funziona, comandi di Action, Control e Stage;
11. Schema-Less database, che significa, le loro peculiarità, tipologie di database Schema-Less.

Entrambi molto tranquilli, le domande sono state all’incirca queste, con dettaglio più o meno elevato a seconda del candidato, Sartori chiede un pò più precisione nei termini mentre Ravaldi sembrava più a la mano e leggermente più alto con i voti.

20/12/2018

Sartori: Association Rules, Lift Chart.

Ravaldi: Data warehouse,Map Reduce.