Appello 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.

Appello 20/12/2018

Sartori: Association Rules, Lift Chart.

Ravaldi: Data warehouse,Map Reduce.