Domande IDS

Intro

1- Cos'è un dato? E un'informazione? Cos'è la Data Science?

Un dato è un elemento che, per fornire informazione, deve essere interpretato in un contesto. Un informazione è tutto ciò che produce variazioni nel patrimonio cognitivo di un soggetto. La DS è un campo multi-disciplinare che usa metodo scientifico, processi algoritmi e sistemi per estrarre conoscenza da dati strutturati e non strutturati.

2- Che nome do alle righe di una tabella? Alle colonne?

R: osservazioni / dati, C: caratteristiche / feature

3- Quali sono le 5 fasi del processo? Come sono organizzate?

Preparazione (*) → Rappresentazione → Esplorazione (*) → Predizione (*)
(*) → Visualizzazione

Nella preparazione troviamo "Integrazione, pulizia, esplorazione, trasformazione", nello storage "file, db?", nell'esplorazione "OLAP, analisi statistica", nella predizione "ML".

4- Differenza tra dato strutturato e non strutturato? Fai esempi.

Strutturato: dati tratti da osservazioni di caratteristiche, normalmente organizzati in formato tabulare (righe e colonne) → ES: oss. scientifiche

Non strutturato: dati che esistono come entità libere e che non seguono alcuna organizzazione standard o gerarchia → ES: dati testuali (tweet)

NB: I dati strutturati (10%) sono considerati più facili da elaborare e analizzare

5- Differenza tra dato quantitativo e qualitativo? Fai esempi.

Quantitativo: dati che possono essere descritti tramite numeri e su cui è possibile eseguire & ha senso operazioni matematiche → ES: temperatura, fatturato

Qualitativo: 1 - Quantitativo → ES: nome, CAP

6- In cosa si dividono i dati quantitativi?

Dati discreti (possono essere contati e possono assumere solo determinati valori)

Dati continui: devono essere misurati e possono assumere una gamma infinita di valori

7- In cosa consiste il livello nominale? Quali operazioni comprende?

Dati qualitativi descritti per nome o categoria. OP: = (anche semantica), appartenenza ad un insieme, moda.

8- In cosa consiste il livello ordinale? Quali operazioni comprende?

Dati qualititativi ordinabili (ma non ancora op. matematiche). OP: tutte quelle del nominale + ordinamento, confronto, calcolo del centro dei dati con la mediana

9- In cosa consiste il livello degli intervalli? Quali operazioni comprende?

Dati quantitativi su cui è possibile eseguire op. matematiche. OP: tutte quelle dell'ordinale + somma, sottrazione, calcolo del centro dei dati con la media

10- In cosa consiste il livello dei rapporti? Quali operazioni comprende?

Tutti i liv. precedenti + moltiplicazioni + punto iniziale naturale o uno zero naturale MA i valori devono essere non negativi

11- Cos'è l'obiettivo analitico?

Un obiettivo che si vuole raggiungere, definito all'inizio della pipeline (preparazione), che guidi le fasi successive

Preparazione

12- In cosa consiste la preparazione del dato?

Pulizia, integrazione, esplorazione, trasformazione

13- Riassumi la tassonomia degli errori possibili.

Divisione in single & integrated data sources, e in schema (semantica schema) & entity level (tuple inconsistenti)

14- Quali sono le 4 caratterstiche da valutare nel Data Cleaning?
Consistenza: violazione regole semantiche (ES: age = 70 e age = 33 per la stessa persona)
Precisione: vicinanza al valore che ci si aspetta (età studenti <= 40 oppure 15 (preciso))
Completezza: possibilità di determinare domande dalle info che ho (ES: presenza di null)
Timeliness: presenza di dati sul periodo a cui voglio rispondere (ES: dati mancanti in anno x)
15- Quali sono 3 motivi per cui un dato è sporco? Da cosa sono introdotti?
incompleto (ci sono null), rumoroso (typo, compatibilità di tipo), inconsistente (dati incoerenti)
Possono essere introdotti da fasi della pipeline o dalla fonte (errori umani, …)
16- Quali trasformazioni posso effettuare per integrare tra loro tabelle?
Conversioni (u di misura), traduzioni, creazione nuove colonne (separando o raggruppando), eliminazione di colonne inutili guidati dall'obiettivo analitico.
17- Fai considerazione sull'efficienza, in base alla dimensione del problema.
Può essere comodo avere un'unica tabella se la dimensione del problema lo consente (denormalizzazone), dato che fare join & trasmettere info è costoso.

Esplorazione

18- Parla di frequenza, moda, percentile, media, mediana, var, stdev.
freq: #dati / N, moda: dato + freq, percentile: dato attributo ordinale o continuo x, il p-esimo percentile xp è un valore di x t.c il p% dei valori è < xp, media: media, mediana: valore al centro (media se pari), var: 1/(m-1)*SUM(xi - media)^2, stdev: sqrt(var)
19- Cosa sono gli outliers? Come influenzano media, moda, mediana e varianza?
Sono valori fuori dalla distribuzione che influenzano la media e la varianza.
20- Cosa sono gli indici di dispersione? E AAD, MAD, IRQ?
Ci dicono se i valori di un certo attributo sono "sparpagliati" tra il minimo ed il massimo oppure concentrate intorno ad un valore (range, var, AAD: deviazione media abs, MAD: deviazione mediana abs, IRQ: intervallo interquartile x75% - x25%).
21- Cos'è l'analisi multivariata? Cosa sono covar e correlazione?
Analisi su + variabili. covar: quanto due variabili variano insieme e dipende dalla loro magnitude = cov(xi, xj) = 1/(m-1)*SUM[(xki - media(i))(xkj - media(j))]
correlazione = cov(xi, xj) / si * sj (aka normalizzazione per stdev) → [-1, 1]

Visualizzazione

22- Parla dei principi ACCENT
Apprehension: Capacità di percepire correttamente le relazioni tra variabili
Clarity: Capacità di distinguere visivamente tutti gli elementi di un grafico
Consistency: capacità di interpretare un grafico per confronto (similarità) con grafici precedenti
Efficiency: capacità di rappresentare una relazione anche complessa in modo semplice
Necessity: la necessità che si ha di usare il grafico (ci sono modi migliori per rappresentare la stessa informazione?)
Truthfulness: capacità di determinare il valore rappresentato da ogni elemento del grafico osservando la sua magnitude relativamente ad una scala implicita o esplicita
23- Parla dei plot: Scatter, Line, Bar, Hist, Violin, Pie, varie matrici, Grafi.
V.SLIDE
24- Parla dettagliatamente del Box Plot.
Mette in evidenza il 1 quartile = 25-es percentile e il 3 quartile = 75-es percentile (ai due bordi del quadrato), la mediana (linea al centro del quadrato), i baffi (che coincidiono con il più vicino tra val_max/min e IRQ * 1.5). In caso di presenza di valori oltre al baffo: outliers.

25- Cos'è OLAP? Come si procede?
Per alcune tipologie di analisi ci conviene usare una rappresentazione alternativa e multidimensionale, su cui è più "facile" operare. Si identificano gli attributi dimensionali (in genere discreti) e le misure da analizzare → si calcola il valore di ogni entry dell'array aggregando i valori degli oggetti corrispondenti.
26- Quali 4 operazioni posso fare con OLAP?
Slicing: selezionare un gruppo di celle con uno specifico valore per una dimensione
Dicing: selezionare un sottoinsieme di celle specificando una combinazione di condizioni per le diverse dimensioni
Roll-up: aumentare il livello di aggregazione dei dati (secondo la struttura gerarchica)
ES: film usciti nei vari mesi indipendentemente dagli anni
Drill-down: ridurre il livello di aggregazione dei dati
ES: selezioniamo l'anno x, e visualizziamo il numero di film usciti per mese
Focus on: statistica
27- Parla del teorema di Bayes.
P(A|B) = [P(A)P(B|A)]/P(B) con P(B) = prob_tot = [P(A)P(B|A)] +[P(!A)P(B|!A)]
28- Cos'è la stima dei punti?
è una stima di un parametro della popolazione sulla base dei dati di un campione
29- E la distribuzione del campionamento? Esempi? (bimodale, normale, ..)
Una distribuzione delle stime di più campioni di uguale dimensione, campionando N volte da un insieme di M campioni e calcolando un istogramma delle N stime (es: media). Ad es. possiamo passare da studiare una distribuzione bimodale ad una distribuzione normale (tramite la ripetizione del campionamento, e grazie al TcdL), su cui ora possiamo applicare test statistici.
30- Cosa sono gli intervalli di confidenza? Cosa ci dice sulle ipotesi?
In casi in cui è difficile ottenere stime precise dai campionamenti, può essere opportuno usare un intervallo di valori basato su una stima che sappiamo contenere il vero parametro della popolazione con un certo grado di confidenza → rappresenta quindi la frequenza con cui la risposta ottenuta è accurata. Possiamo dunque verificare le ipotesi tramite test statistici, decidendo se accettare l'ipotesi, rigettarla o rigettarla in favore dell'alternativa.
31- 3 livelli di confidenza notevoli, con relativi livelli di significatività?
0.95 int_conf con 0.05 liv_sign
0.98 int_conf con 0.01 liv_sign
0.99 int_conf con 0.001 liv_sign
32- Cosa ci serve per queste stime?
1. Una stima del punto
2. Una stima della deviazione standard della popolazione [ad es. deviazione standard del campione / radice quadrata della dimensione del campione]
33- Cos'è H0? E Ha? Cos'è il p-value?
H0: ipotesi nulla (HP da verificare), Ha: ip. alternativa, p-value: frequenza con cui il risultato ottenuto si otterrebbe per caso
Quando i dati presentano prove molto forti contro l'ipotesi nulla, la statistica del test tende a crescere (in positivo o negativo) ed il valore di p diventa molto piccolo → significa che il test mostra risultati netti e quello che dimostra è dovuto al caso
34- Spiega i 5 passi della verifica delle ipotesi.
1- specificare le ipotesi → 2- determinare le dimensioni del campione → 3- scegliere liv_sign → 4- raccogliere i dati → 5- decidere se accettare l'ipotesi

35- Spiega il T-Test
Il t-test per un campione è un test statistico per determinare se un campione di dati numerici (quantitativi) differisce in modo significativo da un altro dataset.
Abbiamo due condizioni:
- La distribuzione della popolazione deve essere normale e il campione deve essere >=30
- La dimensione della popolazione deve essere almeno 10 volte superiore a quella del campione (10n<N) → questo garantisce che il campione sia tratto in modo indipendente
→ Se p-value < livello di significatività rigettiamo ipotesi nulla
Se p-value << livello di significatività rigettiamo ipotesi nulla in favore di quella alternativa (a due code se !=, a una coda se solo < o >).
Se p-value > livello di significatività non possiamo rigettare ipotesi nulla
36- Spiega il Chi-Quadrato dell'idoneità
è un test statistico che lavora su dati QUALITATIVI ragionando in termini di conteggi. Si usa quando vogliamo analizzare una variabile categorica da una popolazione o vogliamo determinare se una variabile segue una certa distribuzione.
Abbiamo due condizioni:
- Tutti i conteggi previsti devono essere almeno 5
- Le singole osservazioni devono essere indipendenti e le dimensioni della popolazione devono essere almeno 10 volte quelle del campione
→ gradi di libertà = #colonne -1 & x = chi-q = (osservato - atteso)^2 / atteso
Per trarre una conclusione confrontiamo il valore ottenuto dal test ed il valore della distribuzione del chi-q corrispondente al livello di confidenza e gradi di libertà del nostro problema (da tabella). Se la statistica di test è inferiore al valore del chi-quadrato non possiamo rifiutare l'ipotesi nulla (uguale distribuzione), e viceversa.
37- Spiega il Chi-Quadrato per associazione/indipendenza
Ci aiuta a determinare se due variabili categoriche sono indipendenti fra loro. Le condizioni necessarie e le conclusioni sono le stesse del test chi-q. I gradi di libertà in questo caso si calcolano come (righe-1)x(colonne-1)
Predizione (Intro e ML supervisionato)
38- Spiega il problema della regressione
Vogliamo trovare una funzione che descriva approssimativamente il modello dei dati a disposizione, dato Training Set Sn = {(x1, y1), …} e Y contenuto in R.
39- Spiega il problema della classificazione
Vogliamo trovare una classe di appartenenza che descriva approssimativamente il modello dei dati a disposizione, dato ad esempio Training Set Sn = {(x1, y1), …}, Y = {-1, 1}, X contenuto in R^2 e xi = [xi1, xi2]
40- Che accade se prendo punti affetti da rumore? Come evitarlo?
Se assegnamo, ad esempio, un nuovo punto ad una classe, potremmo erroneamente assegnarlo ad una classe sbagliata se il punto più vicino appartiene a tale classe errata (per via di rumore o altri problemi). Possiamo evitarlo usando metodi globali invece che locali.
41- Esponi la differenza tra Training Set e Validation Set
Il Training Set può essere inizialmente diviso in Training Set e Test Set
Il Training Set (ad es. 70-80%) è il set di dati che usiamo per il training, mentre il Validation Set lo preserviamo per le fasi successive in cui vorremo validare il modello da noi trovato.
42- Definisci f cappuccio
è una funzione da X a Y (ad esempio, che associa il colore a seconda della codifica), data in output dal Machine Learning supervisionato, che a sua volta prende in input i dati.

: ML: dati → (X → Y) = dati → f^ / f(xk) = yk per ogni (xk, yk) in S con xk in R^d e yk dipende (nel nostro es. y è in R nella regressione e in {-1, 1} nella classificazione)

43- Differenza tra ML supervisionato e non

Nel supervisionato gli algoritmi imparano dai dati tramite un insegnante (Y nota), nel non supervisionato viceversa (abbiamo solo input).

44- Cosa sono fitting e stabilità? Come non innamorarsi dei dati?

La f^ che cerchiamo deve avere due importanti proprietà: - Capacità di rappresentare i dati di training (fitting) - La capacità di generalizzare ai dati che avremo a disposizione in futuro (stability/generalization). Ogni algoritmo di ML deve trovare un compromesso tra le prop.

45- Esponi la regressione lineare

Cerchiamo una relazione lineare tra input e output, cioè (dato x in R^d e y in R)
$y = f(x) = a^0 + a^1x^1 + \dots + a^dx^d$

46- Come stimare la bontà di f? Metriche per validazione?

Per la bontà, che ci fornirà una predizione sugli input di S, usiamo la supervisione e calcoliamo i residui R = sommatoria (y predetti - y reali) ^ 2 = $SUM(y^k - yk)^2$
Per verificare la stabilità usiamo come metrica l'errore quadratico medio (MSE) e la sua radice (RMSE).

47- Che formula posso usare come training del modello?

Il processo di identificare, tra tante soluzioni possibili, la soluzione migliore a partire dai dati a disposizione prende il nome di fase di training di un metodi di ML. Nel Training dobbiamo trovare i coefficienti della f migliore, tramite ad es. la minimizzazione dello scarto quadratico ($a^* = \arg\min_{(a \text{ in } R^{d+1})} SUM(y^k - yk)^2$)

48- Esponi la regressione logistica. Per cosa si usa?

Una generalizzazione del modello di regressione lineare ai problemi di classificazione: prevediamo la probabilità che il dato appartenga ad una certa classe.
$Pr(y = 1 \mid x) = e^{(a0 + a1x)} / (1 + e^{(a0 + a1x)}) = e^{f^(x)} / (1 + e^{f^(x)})$

49- Cos'é odds?

L'odds di un evento è il rapporto tra la sua probabilità e la probabilità complementare

50- Cos'é log(odds), perché è importante? Come derivo p?

log(p/(1-p)) è importante perchè ci permette di condensare meglio le informazioni graze alle proprietà del logaritmo. Stimo log(odds) usando un modello di regressione lineare (derivo la regressione logistica dalla lineare): v. slide per formula

Predizione (ML non supervisionato)

51- Come cambia il training set? In quali due casi può essere utile?

S = {X1, … Xn}, cioè non abbiamo l'output.
Può essere utile nel clustering e nella riduzione della dimensionalità dei dati

52- Qual è l'obiettivo degli alg di clustering? Cosa sono cluster e centroide?

è di individuare strutture (gruppi di dati «coerenti» rispetto ad una qualche misura) all'interno dei dati. Un cluster è un gruppo di dati che si «comporta in modo analogo» e un centroide è il «centro» del cluster (ad esempio il punto medio).

53- Spiega il K-Means con un esempio. Condizioni d'arresto?

1. Scegliere k centroidi iniziali (k è un input!)
2. Per ogni punto: assegnare il punto al centroide più vicino
3. Per ogni centroide: aggiornare la posizione del centroide
4. Ripetere i passi 2 e 3 fino a raggiungere un criterio di arresto

Le condizioni d'arresto potrebbero essere o #iterazioni fissate a priori oppure tramite una similutidine (entro una certa tolleranza) tra la posizione precedente e quella attuale.

54- Esponi lo pseudocodice di K-Means.

```
def K-Means(X, centers, maxiter):
# X: n x d
# centers : k x d
n, d = X.shape k = centers.shape[0]
for i in range(maxiter):
        # Compute Squared Euclidean distance between each ...cluster centre and each observation
        dist = all_distances(X, centers)
        # Assign data to clusters:
        # for each point, find the closest center in terms of euclidean distance
        c_ass = np.argmin(dist, axis=1)
        # Update cluster center
        for c in range(k):
                centers[c] = np.mean(X[c_ass == c], axis=0)
return c_ass, centers
```

55- Possibili problemi? Quali due soluzioni?

La difficoltà è se i centroidi sono presi troppo vicini (l'init è random)
Prima sol.) ripetere l'alg n volte e prendere la media
Seconda sol.) K-M++ = i centroidi campionati devono appartenere al dataset & preso uno prendo l'altro alla distanza massima.

56- Cos'è il coeff. di Silhouette? Come cambia al variare di k?

è una misura per capire la bontù dei cluster individuati (best_value = 1)
SC = (b - a) / max(a, b) con b: distanza media extra cluster (meglio se alta) e a: distanza media intra cluster (meglio se bassa). Se k è troppo alto (cioè troppi centroidi) ho b basso.
Rispetto alle slide: idealmente dovrei avere un k scelto vicino al numero di cluster che vedo

57- Cos'è e perchè è importante la standardizzazione dei dati?

(-avg & / stdev) è importante per avere i giusti ordini di grandezza / rapporti tra gli assi

58- Parla del problema della riduzione della dimensionalità? Curse of dim?

Quando i dati vengono rappresentati con un numero di caratteristiche troppo alto rispetto al numero di dati potremmo incontrare la cosiddetta curse of dimensionality. Maggiore è il numero di caratteristiche usate per la rappresentazione di un dato, più i dati stessi risultano distanti gli uni dagli altri. Quando le caratteristiche sono troppe rispetto alle reali necessità rischiamo di non migliorare la capacità descrittiva della rappresentazione, e anzi peggiorare i risultati. Quando la dimensionalità dei dati (D) cresce, il volume dello spazio dei dati cresce velocemente ed i dati diventano presto sparsi… Perchè i risultati siano affidabili il numero di dati (N) deve crescere esponenenzialmente con la loro dimensionalità. In questi casi ci affidiamo a tecniche di riduzione della dimensionalità, che ci permettono anche di poter visualizzare e intepretare meglio i dati.

59- Quali 3 buone proprietà sono desiderabil per buone rappresentazioni?

Varianza alta: caratteristiche con varianza alta contengono «molto segnale»
• Non correlazione: caratteristiche correlate sono ridondanti e poco informative
• Non troppe dim: deve essere sempre un buon bilanciamento tra #dati e dimensionalità

60- In cosa consiste PCA? Come usiamo l'SVD? Algoritmo?

è un algoritmo applicabile a matrici di qualsiasi dim NxD e consiste nell'identificare una nuova base le cui componenti catturano quanta più varianza possibile dai dati originali.
Come ingrediente base ci sono le SVD ($X = USV^t$: ort - diag - ort, *v. slide …*), con alg:

1) matrice di covarianza $C = X^tX$ (i, j: covarianza tra i e j / i, i: varianza di quell'obj)
2) calcoliamo SVD di C
3) identifichiamo nuova base, cioè gli assi del nuovo sistema di riferimento (le prime k colonne di V se vogliamo K componenti principali)