BANCA D'ITALIA
EUROSISTEMA

10 ASSUNZIONI DI PERSONE LAUREATE CON COMPENTENZE IN *DATA SCIENCE*

(Bando del 9 maggio 2024)

Testo n. 1

***DATA SCIENCE* E INTELLIGENZA ARTIFICIALE**

Un quesito a scelta tra due proposti dalla Commissione:

QUESITO N. 1

La candidata/il candidato:

1. Illustri le proprietà dei modelli di transazione dei database distinguendo tra ACID *(Atomicity, Consistency, Isolation, Durability)* e BASE (*Basically Available, Soft state, Eventually consistent*).
2. Con riferimento ai database di tipo NoSQL *(Not only SQL)*:
   a. ne identifichi le caratteristiche, evidenziando i principali vantaggi e limiti rispetto ai database di tipo relazionale (RDBMS);
   b. scelga uno tra i seguenti tipi di database NoSQL: *key-value, column-family, document-oriented e graph-based.* Ne descriva il modello dati e le caratteristiche principali.
3. Definisca lo schema di un database di *tipo column-family* per esaminare un *dataset* relativo agli autobus che circolano a Roma, specificando la struttura della chiave di riga, le colonne e il contenuto di ciascuna cella.
   Si consideri che il *dataset* include eventi (*record*) contenenti le seguenti informazioni: istante temporale di registrazione dell'evento, identificativo del veicolo, identificativo della linea (percorso), nome della località di partenza, posizione del punto di partenza (espresso in termini di latitudine e longitudine), nome della località di arrivo, posizione del punto di arrivo (espresso in termini di latitudine e longitudine), posizione del veicolo al momento della registrazione dell'evento (espressa in termini di latitudine e longitudine), istante temporale di arrivo programmato, istante temporale di arrivo effettivo o previsto. Esempio: *1724917721, GC-5846, A60-CTR, PZZ VENEZIA 00186, 41.895894, 12.482528, TRR ARGENTINA 00186, 41.896282, 12.476942, 41.896250, 12.479853, 1724918321, 1724918471.* Gli eventi sono acquisiti ogni 10 minuti a partire da gennaio 2017.

   Si noti che: **i)** le informazioni temporali contenute in ogni *record* vengono memorizzate sotto forma di *timestamp* secondo il modello Unix, ovvero come numero di secondi trascorsi a partire dal primo gennaio 1970 alle ore 00:00 UTC. Ad esempio, il *timestamp* 1724917721 fa riferimento a giovedì 29 agosto 2024 alle ore 07:48:41 UTC; **ii)** l'identificativo di ciascuna linea è un codice costituito da una lettera - che indica la circoscrizione di Roma attraversata (ad esempio, "A" per il Municipio Roma I, "E" per il Municipio Roma V) - e da una stringa alfanumerica. Si supponga che una linea di autobus attraversi un solo Municipio.
   La base dati progettata deve consentire di recuperare:
   - le posizioni (latitudine e longitudine) occupate da uno specifico veicolo che percorre una specifica linea all'interno di un'ora *H* (dalle *H:00:00 alle H:59:59*) di un determinato giorno, utilizzando il *timestamp* della registrazione dell'evento;
   - le posizioni occupate dai veicoli che percorrono una specifica linea all'interno di un'ora *H* di un determinato giorno;

- le posizioni occupate dai veicoli che percorrono le linee del Municipio Roma V ("E") all'interno di un'ora *H* di un determinato giorno.

Si osservi che la granularità delle interrogazioni è su base oraria e pertanto, per rispondere alle interrogazioni, è sufficiente arrotondare il *timestamp* per difetto all'ora (ovvero ignorare minuti e secondi).

Si supponga inoltre:

- di utilizzare un'unica famiglia di colonne;
- che una cella può memorizzare più versioni di un dato. Ad esempio, se due eventi hanno la stessa chiave di riga, il database li memorizza entrambi e fornirà, se non richiesto diversamente, solo la versione del dato memorizzata più di recente. Le diverse versioni dei dati sono ordinate sulla base di un *timestamp* specificato;
- di disporre di una API che consenta di:
  - eseguire le operazioni CRUD (*Create, Read, Update e Delete*) sui *record* registrati nel database. Per quanto riguarda l'operazione di *read*, si supponga che tale API consenta di eseguire operazioni di lettura che rientrano in due categorie: **i)** lettura di una singola riga del database; **ii)** scansione, cioè lettura di un intervallo di righe che condividono un prefisso della chiave. Per eseguire una scansione deve essere specificato un prefisso di chiave di riga o un intervallo di chiavi di riga (ossia la chiave di riga iniziale e finale). Per le suddette operazioni di lettura è possibile inoltre specificare le colonne e il numero di versioni dei dati da restituire;
  - arrotondare un *timestamp* per difetto all'ora. Attraverso la funzione fornita, ad esempio, il *timestamp* 1724917721 (29 agosto 2024, ore 07:48:41 UTC) viene arrotondato a 1724914800 (29 agosto 2024, ore 07:00:00 UTC).

QUESITO N. 2

Bag-of-words (BOW) e *word2vec* sono approcci per la rappresentazione dei dati usati in *Natural Language Processing* (NLP).

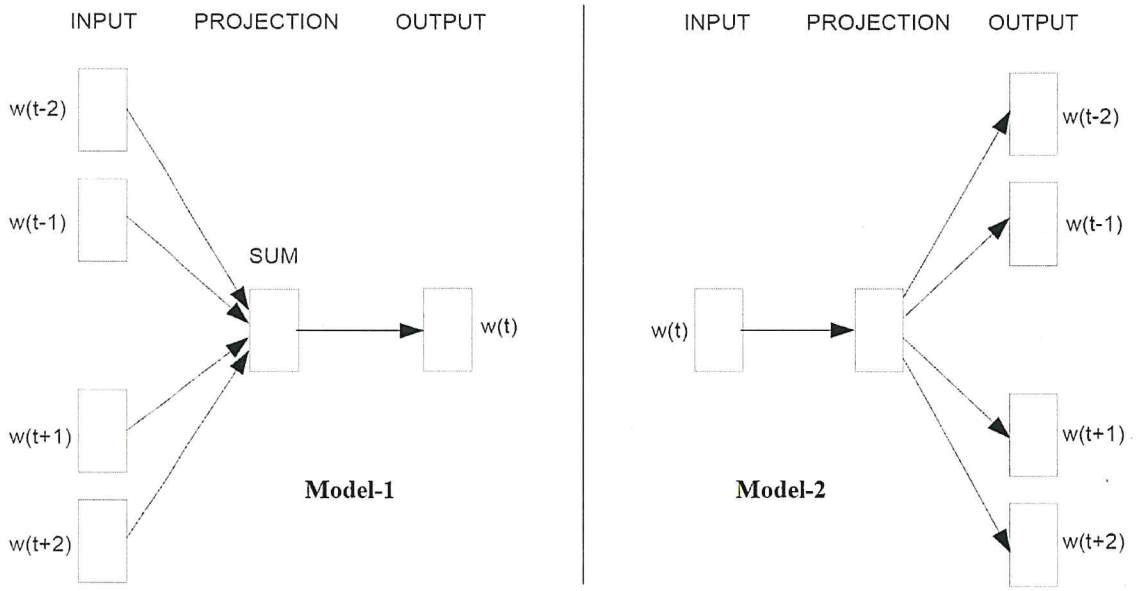
La candidata/il candidato:

1. Illustri, descrivendone vantaggi e limitazioni, l'approccio BOW e le relative tecniche di *pre-processing*. Applichi l'approccio BOW, comprensivo delle tecniche di *pre-processing*, al seguente *corpus:* ["Data Science Is Great.", "Data science and artificial intelligence are amazing in processing data!", "Not all science is based on data"].
2. Descriva il concetto di *word embedding*, fornendone un esempio di possibile uso.
3. Assumendo che King, Man, Woman, Queen, Paris, France, Italy, Rome rappresentino dei vettori prodotti con *word2vec*, spieghi il significato delle seguenti operazioni, motivando la risposta:

   $$\text{King} - \text{Man} + \text{Woman} \approx \text{Queen}$$
   $$\text{Paris} - \text{France} + \text{Italy} \approx \text{Rome}$$

4. Date le architetture *Model-1 e Model-2* in Figura 1, indichi, motivando la risposta, quale modello rappresenti un *Continuous Skip-gram model* e quale un *Continuous bag-of-words model* (CBOW). Descriva inoltre l'obiettivo di previsione di ciascuna architettura.

Figura 1



STATISTICA

Un quesito a scelta tra due proposti dalla Commissione:

QUESITO N. 3

Sono stati estratti due campioni indipendenti di 100 studenti universitari, uno dalla classe degli studenti del primo anno e un altro da quella del terzo anno. Nel primo campione si sono osservati 40 fumatori, nel secondo 50.

La candidata/il candidato:

1. Descriva la funzione di probabilità di una distribuzione bernoulliana e di una binomiale, i valori attesi e le varianze. Discuta anche il legame tra le due distribuzioni.
2. Considerate le informazioni disponibili, definisca uno stimatore corretto per la proporzione di studenti fumatori. Discuta, inoltre, le condizioni per cui tale stimatore sia distribuito asintoticamente come una normale.
3. Si sta valutando la possibilità di estrarre un nuovo campione di 10 unità dalla popolazione di studenti del primo anno. Utilizzando l'informazione proveniente dal primo campione, stimi la probabilità che nessuno studente del primo anno del nuovo campione sia un fumatore.
4. Descriva la differenza tra test d'ipotesi e intervallo di confidenza e verifichi con un test statistico con un livello di significatività del 5 per cento la seguente ipotesi: "gli studenti del terzo anno fumano di più di quelli del primo anno".
5. Considerando la statistica test ricavata al punto precedente, discuta come varia la regione critica del test all'aumentare del livello di significatività, anche esemplificando graficamente.

QUESITO N. 4

La candidata/il candidato:

1. Data la seguente distribuzione doppia di frequenza di 100 studenti iscritti al primo anno di una facoltà, ripartiti per tipo di diploma conseguito nelle scuole superiori e numero di esami superati nel primo semestre:

Tipo di diploma	Numero di esami superati					Totale
	0	1	2	3	4	
A	3	6	9	30	12	60
B	2	10	12	14	2	40
Totale	5	16	21	44	14	100

   a. costruisca le distribuzioni di frequenza relative cumulate condizionate del numero di esami dato il tipo di diploma di provenienza;
   b. definisca e calcoli la mediana e l'intervallo interquartile delle distribuzioni condizionate del numero di esami dato il tipo di diploma.
2. Illustri gli indici che possono essere utilizzati per valutare la presenza di associazione (o connessione) tra due variabili qualitative nominali.
3. Definisca il concetto di correlazione lineare tra due variabili quantitative continue, indichi quale tipo di grafico può essere utilizzato per rappresentare tale relazione e riporti la formula del coefficiente di correlazione lineare di Bravais-Pearson e le sue proprietà.
4. Fornisca un esempio numerico di distribuzione congiunta di due variabili discrete, X e Z, che presenti una dipendenza in media perfetta di X da Z, ma che non implichi una dipendenza perfetta di Z da X.

ECONOMETRIA E *STATISTICAL LEARNING*

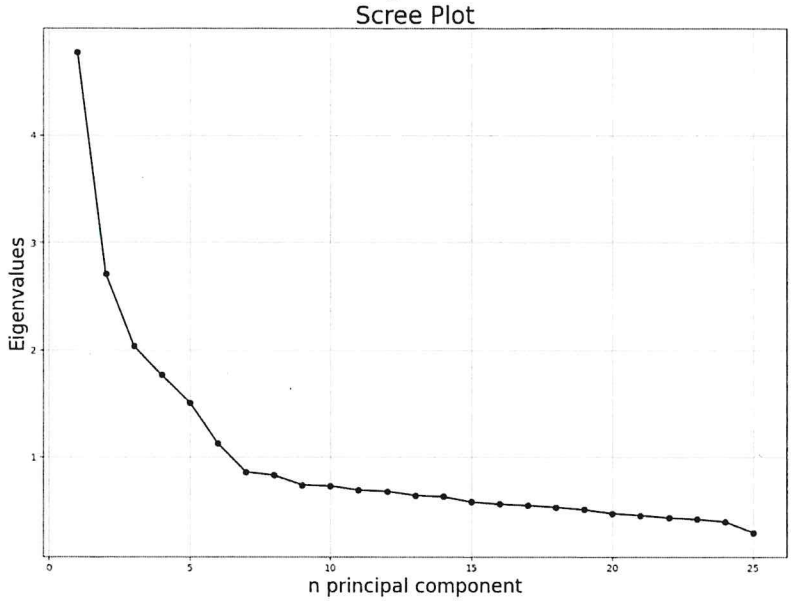
Un quesito a scelta tra due proposti dalla Commissione:

QUESITO N. 5

La candidata/il candidato:

1. Date le osservazioni $x_1, \dots, x_m$ in $\mathbb{R}^n$:
   a. definisca il problema della riduzione della dimensionalità dei dati e descriva un metodo per il calcolo delle prime p componenti principali;
   b. illustri una regola empirica comunemente utilizzata per scegliere il numero di componenti principali e la applichi utilizzando l'informazione riportata nella Figura 2.

Figura 2



2. Descriva la *principal component regression* e ne illustri i contesti in cui il suo utilizzo è preferibile al modello di regressione lineare.
3. Descriva la struttura delle reti di tipo *autoencoder* e illustri quale vantaggio tali reti forniscono rispetto all'analisi per componenti principali (*Principal Component Analysis* - PCA). Si evidenzino inoltre le differenze, in termini di proprietà della soluzione, tra la PCA e una rete *autoencoder* lineare con un solo *hidden-layer* con p neuroni.

QUESITO N. 6

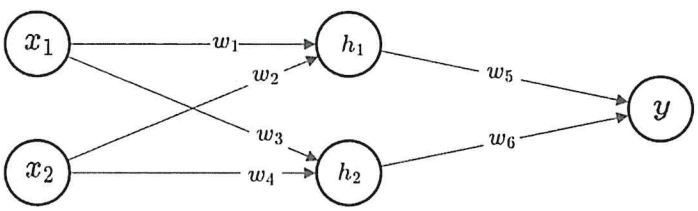
La candidata/il candidato:

1. Dato il modello di regressione multipla

$$y_i = \beta_0 + \sum_{j=1}^{K} x_{ij}\beta_j + \varepsilon_i,$$

dove $(y_1, \dots, y_n)$ rappresenta la variabile risposta osservata su un campione di n individui, K il numero di variabili esplicative e $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$, l'errore, derivi:
   a. lo stimatore $\boldsymbol{b}$ dei $\boldsymbol{\beta}$ con il metodo dei minimi quadrati;
   b. la varianza di previsione con il relativo intervallo.
2. Illustri due differenti test di ipotesi per verificare la presenza di eteroschedasticità (*heteroskedasticity*) negli errori.
3. Considerato il *multilayer perceptron* in Figura 3

Figura 3



dove

- x_1 e x_2 sono le unità di input,
- h_1 e h_2 sono i neuroni nell'*hidden layer* con funzione di attivazione ReLU,
- y è l'unità di output ottenuta con funzione di attivazione uguale all'identità,
- i valori dei pesi sono i seguenti

w_1	w_2	w_3	w_4	w_5	w_6
0.2	-0.6	-0.1	0.7	0.3	0.1

a. descriva il metodo della *backpropagation* nell'ambito dell'addestramento di una rete neurale;
b. calcoli il valore del gradiente della funzione di *loss*

$$L(y_{out}, y_{true}) = \frac{1}{2}(y_{true} - y_{out})^2$$

nel punto di coordinate $(x_1, x_2) = (2, 3)$ e $y = 4$.

PROVA IN LINGUA INGLESE

Many people complain that news shows focus too much on sensational items, such as local crimes and celebrity gossip, and spend too little time on important national and international news. In your opinion, should television news devote more time and coverage to international news and global issues? Why or why not?