AREA SCIENCE PARK

PROFILO A

Traccia n. 1
Il candidato discuta in maniera sintetica due dei seguenti argomenti, indicando esplicitamente la scelta all'inizio del proprio elaborato:

A) Il candidato descriva brevemente il metodo di Analisi delle Componenti Principali (PCA) per la riduzione della dimensionalità. Inoltre, specifichi le condizioni in cui tale metodo può essere utilizzato per determinare che un dataset costituito da un gran numero di vettori di dimensione M è, con buona approssimazione, distribuito su una superficie piana di dimensione K<M. Si assume che la normalizzazione dei dati non sia necessaria, poiché le variabili sono omogenee e su scale comparabili.

B) Il candidato descriva brevemente cos'è un Transformer, evidenziandone gli elementi architetturali caratteristici.
Successivamente, il candidato risponda alle seguenti domande:

1. Quali sono le principali differenze tra i Transformer e le Reti Neurali Ricorrenti (RNN) in termini di efficienza computazionale e capacità di catturare dipendenze tra i "token" nel processamento di sequenze?
2. Perché è necessario aggiungere un "embedding" posizionale?
3. Dove viene utilizzato il mascheramento causale e qual è il suo scopo?

C) Si consideri un laboratorio dotato di strumenti scientifici in grado di generare grandi volumi di dati. I dati grezzi devono essere trasferiti e salvati automaticamente in un centro di calcolo. Una volta salvati con successo, si deve avviare automaticamente una pipeline di analisi dei dati. Inoltre, i dati e gli strumenti computazionali associati devono essere resi accessibili ai ricercatori tramite una piattaforma cloud che supporti l'uso di strumenti open-source per l'analisi, il machine learning e il deep learning. L'obiettivo è garantire che la ricerca sia il più possibile riproducibile.
Il candidato delinei gli elementi essenziali di un'infrastruttura di ricerca di questo tipo. Inoltre, affronti brevemente almeno due dei seguenti argomenti:

- Gestione dei dati secondo i principi FAIR
- Creazione e mantenimento di un piano di gestione dei dati (Data Management Plan)
- Progettazione di pipeline di analisi dei dati

The candidate should discuss, in a concise and effective manner, two of the following topics, explicitly indicating their choice at the beginning of their work:

A) The Candidate should briefly describe the Principal Component Analysis (PCA) method for dimensionality reduction. The Candidate should also specify the conditions under which this method can be used to determine that a dataset consisting of a large number of vectors of dimension M is, to a good approximation, distributed on a flat surface of dimension K < M.

Assume that data normalization is not required, as the variables are homogeneous and on comparable scales.

B) The Candidate should briefly describe what a Transformer is, outlining its characteristic architectural elements.
The Candidate should then answer the following questions:

1. What are the main differences between Transformers and Recurrent Neural Networks (RNNs) in terms of computational efficiency and the ability to capture dependencies between "tokens" in sequence processing?
2. Why is it necessary to add a positional embedding?
3. Where is causal masking used, and what is its purpose?

C) Consider a laboratory equipped with high-throughput scientific instruments capable of generating large volumes of data. The raw data must be automatically transferred and stored in a computing center. Once the data has been successfully saved, a data analysis pipeline should automatically start. Additionally, the data and related computational tools must be made accessible to researchers via a cloud platform that supports the use of open-source tools for analysis, machine learning, and deep learning. The goal is to ensure that the research is as reproducible as possible.
The Candidate should outline the essential elements of such a research infrastructure. Additionally, the Candidate briefly address at least two of the following topics:

- FAIR Data Management
- Creation and maintenance of a Data Management Plan
- Design of data analysis pipelines