AREA SCIENCE PARK

PROFILO A

Traccia n. 2

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

A) Il candidato descriva il metodo di Analisi delle Componenti Principali (PCA) per la riduzione della dimensionalità. Descriva come determinerebbe il numero minimo di componenti necessarie per spiegare almeno il 90% della varianza del dataset. Si assume che la normalizzazione dei dati non sia necessaria, poiché le variabili sono omogenee e su scale comparabili.

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

- Quali sono le principali differenze architetturali tra i Transformer e le Reti Neurali Ricorrenti (RNN) nel processamento di sequenze?
- Cosa si intende per "residual stream" e quali elementi architetturali consentono lo scambio di informazioni tra i residual stream di diversi token?

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 correlati 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, il candidato affronti brevemente almeno due dei seguenti argomenti:

- Gestione dei dati secondo i principi FAIR
- Utilizzo di strumenti open-source per l'analisi dei dati con metodi di machine learning
- Data provenance

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 describe the Principal Component Analysis (PCA) method for dimensionality reduction. The Candidate should explain how they would determine the minimum number of components needed to account at least the 90% of the dataset's variance. 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:

- What are the main architectural differences between Transformers and Recurrent Neural Networks (RNNs) for sequence processing?
- What is meant by "residual stream", and what architectural elements enable the exchange of information between the residual streams of different tokens?

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 should briefly address at least two of the following topics:

- FAIR Data Management
- Use of open-source tools for data analysis with machine learning methods
- Data provenance