AREA SCIENCE PARK

PROFILO B

Traccia 2

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 regressione logistica binaria e indichi un possibile ambito di applicazione. Il candidato risponda alla seguente domanda: Qual è la funzione di "loss" più utilizzata, come viene definita e qual è la sua interpretazione statistica/probabilistica?

Si supponga ora che alcune delle M dimensioni dei vettori di input siano fortemente correlate, e che il candidato decida di effettuare un'Analisi delle Componenti Principali (PCA) per proiettare i dati in una dimensione inferiore K, tale che le K dimensioni risultanti spieghino circa il 90% della varianza totale del dataset. Il candidato discuta come procederebbe per costruire il nuovo dataset.

B) Il candidato descriva brevemente cosa sia una Rete Neurale Convoluzionale (CNN), spiegando per quali tipi di problemi può essere utilizzata e ne delinei gli elementi architetturali caratteristici. Il candidato risponda inoltre alle seguenti domande:

- Quali sono le principali proprietà dell'operazione di convoluzione (così come intesa in questo contesto specifico) e quali sono le motivazioni per la sua introduzione?
- Per quale scopo vengono introdotti "Pooling Layers" in un'architettura convoluzionale, e quali sono i tipi più comuni di Pooling?

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, deve avviarsi 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, il candidato affronti brevemente almeno due dei seguenti argomenti:

- Gestione dei dati secondo i principi FAIR
- Gestione del controllo degli accessi basato su modelli come RBAC (Role-Based Access Control) o MAC (Mandatory Access Control)
- Gestione della versione per dati, modelli e codice

Traccia n. 2

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 binary logistic regression method and indicate a possible area of application. The Candidate should also answer the following question: What is the most used "loss" function, how is it defined, and what is its statistical/probabilistic interpretation?

Now suppose that some of the M dimensions of the input vectors are highly correlated, so that the Candidate takes the decision to perform a Principal Component Analysis (PCA) to project the data on a lower dimension K, such that the K resulting dimensions explain around the 90% of the dataset's total variance. The Candidate should discuss how to proceed to construct the new dataset.

B) The Candidate should briefly describe what a Convolutional Neural Network (CNN) is, explaining what types of problems it can be used for and outlining its characteristic architectural elements. The Candidate should then answer the following questions:

- What are the main properties of the convolution operation (as understood in this specific context), and what are the motivations for its introduction?
- For what purpose are Pooling Layers introduced in a convolutional architecture, and what are the most common types of Pooling?

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
- Access control management based on models such as RBAC (Role-Based Access Control) or MAC (Mandatory Access Control)
- Versioning management for data, models, and code