AREA SCIENCE PARK

PROFILO B

Traccia 3

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 inoltre alle seguenti domande: Qual è la funzione di "loss" più utilizzata, come viene definita e qual è la sua interpretazione statistica/probabilistica?

Si supponga ora che i vettori di input abbiano una dimensionalità molto elevata (M) e che si decida di rendere il modello più sparso per facilitarne l'interpretazione.
Il candidato discuta come modificherebbe la funzione di "loss" per rendere il modello più sparso. Infine, il candidato spieghi come il modello ottenuto potrebbe essere utilizzato per interpretare il suo processo decisionale. Si assuma che il modello sparso ottenuto fornisca buone prestazioni, rendendo possibile tentare un'interpretazione.

B) Il candidato descriva brevemente cosa sia una Rete Neurale Convoluzionale (CNN), spiegando per quali tipi di problemi può essere utilizzata e delineandone 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?
- Quali sono i potenziali problemi nell'analizzare dati di tipo immagine utilizzando architetture basate esclusivamente su architetture Fully-Connected?

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 automaticamente avviare 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
- Strumenti open-source per la condivisione dei dati
- Gestione della versione per dati, modelli e codice

Traccia n. 3
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 questions: What is the most used "loss" function, how is it defined, and what is its statistical/probabilistic interpretation? Now suppose that the input vectors have a very high dimensionality M, and it is decided to sparsify the model to make it easier to interpret. The Candidate should discuss how they would modify the loss function to make the model sparser. Finally, the Candidate should explain how the resulting model could be used to interpret its decision-making process.

Assume that the resulting sparsified model performs well, and therefore it is possible to attempt at an interpretation.

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?
- What are the potential problems in analyzing image-type data using architectures based solely on Fully-Connected Layers?

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
- Open-source tools for data sharing
- Versioning management for data, models, and code