Where is Statistics in Artificial Intelligence

李育杰 研究員
中央研究院資訊科技創新研究中心

摘要

In the rapidly advancing field of Artificial Intelligence (AI), the role of statistics is not just relevant; it is foundational. I will begin by introducing the core equation that underpins AI: Machine Learning = Representation + Optimization + Evaluation. This equation captures the essence of the statistical principles that drive AI, and I will dissect each component to uncover the statistical foundations that propel this technology forward.

In the realm of Representation, I will delve into the quest for an optimal latent space, highlighting advanced techniques such as Principal Component Analysis (PCA), Sparse PCA, Kernel PCA, Sliced Inverse Regression (SIR), and Kernel SIR. These methods are essential for capturing the underlying structure of data, addressing one of AI's most significant challenges.

Optimization will be illustrated through the lens of statistical tools like Maximum Likelihood Estimation (MLE) and Least Squares Estimation (LSE). I will connect these principles with the broader concept of Structural Risk Minimization, demonstrating how statistical methods ensure that AI models are not only unbiased and of minimum variance but also robust and reliable.

When it comes to Evaluation, I will emphasize the importance of stratified cross-validation as a rigorous tool for assessing model generalization. This discussion will highlight how statistical techniques play a critical role in ensuring that AI models are not just accurate but dependable across diverse data distributions.

As we move into the era of Generative AI, I will share insights on this transformative technology and introduce TAIDE, the Trustworthy AI Dialogue Engine—a Taiwanese-style large language model based on Meta's Llama family. The talk will conclude with a reflection on AI's evolution from a search engine to a prediction engine, with statistics at the core of this ongoing transformation.