第十五讲课程小结

刘子源 2022310709

2022年11月11日

今天听了樊平毅老师的《数据科学与信息论》讲座。

目前人工智能技术非常流行，一个好的AI模型少不了足够的数据的支撑，也就是说AI离不开数据科学的基础。数据科学的影子无处不在，它是一门综合性非常强的应用学科，与大到生物、自然环境，小到我们的日常生活等观测密切相关，是AI技术的物质基础。

老师首先为我们介绍了数据科学的历史。在1962年，Tukey完成了归纳处理，但是没有提出演绎；1977年，Tukey认为应该更重视利用数据做出哪些需要假设，哪些需要验证，这其实就是机器学习的处理逻辑，他还指出探索性的数据分析和论证性的数据分析应该能够并驾齐驱，这个观点放到现在的机器学习研究领域中或许可以成为一个指导性原则；1974年，Peter Naur提出数据是对于事实或想法的规范化表现。

数据科学有三个特点。一是它是在数据形成后，处置数据的科学，二是数据与其所代表的东西间的关系分析，三是数据科学的应用范畴探索与理论延伸。

数据科学的研究内容分为四个层次。基础理论研究：我们做基本理论问题研究时，都要考虑数据测度、时间、数据相似性、数据类型等等；实验和逻辑推理方法研究：在对核心问题进行求解和推断时，需要建立数据科学的实验方法，并用哲学方法揭示数据背后的规律；领域数据学研究：数据科学的理论和方法可以应用在很多领域之中，并形成了许多分支融合学科；数据资源的开发利用方法和技术研究：所谓数据就是资源，尤其在当今大数据时代，掌握了数据就掌握了资源，其重要程度可见一斑。

数据科学与信息论如今面临着一些数据层面的挑战，如随着数据指数级爆炸，我们需要超大规模、分布式的、多种类的数据采集方式；采集到数据之后，我们还面临超大规模的数据存储、传输、处理的问题；拥有如此庞大的数据量之后，我们还需要基于用户个性化需求做出数据挖掘的决策。