人工智能学习札记

（四）

变换器与多元分析及函数分析

在遗传分析中，无论是在会议的报告中，期刊发表的文章中言必称全基因组关联分析，《GWAS》已经成为一个英文单词了。人们不再过问是怎样进行全基因组关联分析的。现在所进行的关联分析大概可分为单点，多点多元分析和基因水平的函数分析。这些分析的共同缺点是(1)忽略DNA之间的联系（2）忽略因果分析。因果分析比较复杂，不在此讨论而留在下一杂记去讨论。

2017年所提出的变换器（transformer)是比现在统计学中用于关联分析的多元和函数分析更为强有力的分析方法。变换器具有以下四个方面显著的优点。（1）它使用了每点或更广元地说token（包括数值，单词，SNP，DNA，氨基酸，一小快图象等)的embedding(欧氏空间的一个向量）。因此有能力刻划基因组信息在人群中的变异，其表示能力強。而在遗传的分析中，一个SNP编码为一个整数，表示能力有限。

（2）变换器对每一点的变换是上下文的变换，它的embedding是由上下文决定的，不同的上下文，它的embedding的值是完全不同的。它可虑各点之间的依赖关系，它可以从一点生成另一点。但多元和函数分析不是上下文的。各点之间的值在各种上下文之间是不变的。DNA 中AG中的G与GG 中的G的编码都是一样的。它不可能由周围的值来生成。由泰勒公式可知，一点邻域的函数不完全是由该点的函数值所决定，而是由该点的函数值和它的许多阶的导数值决定。

（3）变换器可以近似任一连续函数，任一连续子空间。函数分析精确近似往往需要许多基函数，这样会增加检验统计量的自由度。（4）变换器可以实现一序列连续函数到另一序列连续函数的变换。另一函数可以是自身也可以是其它函数。变换器的编码器的输出就是解码器的输入。变换器可以把一个函数变换为自身。其编码器输出形成的隐空间类似于VAE的隐空间，但和VAE的隐空间不同，隐空间的变量之间不独立。任一基因变换器的编码器输出的embedding 来源于该基因DNA 的变化，可以代表，表示该基因的整个DNA，因而可用以检验该基因是否与某一疾病或表象有关联或有因果关系。函数分析只

能将一个函数分解为一组基函数的线索组合，它不如映照到自身。因此离开了一组基函数，它无法生成与自己类似的函数。它不能用于生成语言，生成图象，生成蛋白质，它用在假设检验中，它分解为基函数的线性组合时，其自由度也往往会大于基于变换器编码的检测统计量的自由度。虽然在计算机和人工智能领域变换器得到了大量而深入的研究，但在统计学界和遗传，分子生物学领域研究甚少。现在是我们尝试用变换器代替多元和函数分析，或者保守一点说作为它们的另一选择的时候了。