什么是自监督学习

机器学习里常见的方法是监督学习：对数据进行人为标注，再把这些数据喂给机器，机器就会象人一样对未知数据进行分类和预测。

监督学习的问题在于成本高，标注花时间长。人们想到无监督学习方法，这种方法不需要标注，通过观察数据样本来获知某些规律，再利用这些信息来帮助监督学习。一个例子是当我们看到很多有毒的蛇都具有鲜艳的花纹时，下次再遇到有花纹的蛇，就算没有人告诉我们这条蛇有毒，我们也会远远躲开。这是因为我们利用了无监督学习里的聚类能力，把带花纹的蛇自然而然地归为一类了，从而实现了知识扩展。

无监督学习虽然很有价值，但传统方法还是以学习数据的分布规律为主。最近，一种称为“自监督学习”的无监督学习方法引起了广泛关注。这种方法事实上是用监督学习的方法来进行无监督学习。更具体地说，这一方法完成的是分类或预测任务，但类别或预测的目标不是人为标注的，而是从数据中自动生成的。

例如，我们知道一幅图片是只猫，对这幅图片换个角度，它依然是这只猫。因此我们就自动生成了一个标注：这两张图片是同一只猫。再比如，视频里人的嘴唇运动和声音是天然同步的，因此可以设计一个预测模型，用嘴唇的运动来预测声音。这里的预测目标同样是自动生成的。在对语言建模时，一段话后接哪些词是有一定规律的，这些规律天然存在于文本数据中，我们可以通过一个预测模型，基于过去的文本来预测未来文本。

类似的方法还有很多[1]，这些方法统称为自监督学习。本质上，自监督学习是利用先验知识来确定数据中的相关性(如旋转操作前后图片的内容相关性，音视频两个模型之间的相关性，过去样本和未来样本的相关性)，再利用深度学习方法来学习这种相关性。学习这些相关性本身不是目的，真正的目的是通过这种学习发现数据中的规律，获得抽象的、有效的数据表示，从而起到预训练的作用。有了这些预训练模型，稍加调整即可显著提升下游任务的性能。

目前，自监督学习已经广泛应用于众多领域，极大提高了对无标注数据的利用效率。

[1]Liu X, Zhang F, Hou Z, et al. Self-supervised learning: Generative or contrastive[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.