我眼中的人工智能

人工智能是个很大的范畴，我就谈谈人工智能中的机器学习。机器学习指程序从过往的经验中学习知识，用学习的知识来解决未知的问题。在计算机系统中，经验通常以数据的形式存在，而学习的知识，在机器学习中称作“模型”，模型在数学上可以理解为从输入数据到输出数据的一个映射，有了模型后，将新的数据提供给程序，程序能够作出预测。下面，我将从机器学习的历史、问题的分类，一些经典算法对机器学习进行展开。

最早在上世纪五十年代，机器学习被AI先驱Arthur Samuel定义为“使计算机能够在没有显式编程的情况下学习的研究领域”。从五十到七十年代初，人工智能研究处于“推理期”，那时人们认为只要赋予机器逻辑推理的能力，机器就能具有智能，这一时期的代表性成果包括证明了《数学原理》中52条定理的“逻辑理论家”程序。但是人们逐渐意识到仅仅有逻辑推理能力是远远实现不了人工智能的，从七十年代中期开始，人工智能进入“知识期”，代表成果是“专家系统”，该系统是由人来把自己的知识总结出来教给机器，但是，人们逐渐意识到由人把知识总结出来是一件费时费力的事情，人们希望机器能自己学习，于是，机器学习进入了“学习期”。学习期分为四个阶段，八十年代时流行的是以决策树为代表的符号主义学习，九十年代初期以神经网络为代表的连接主义出现，但是由于计算机性能瓶颈，在九十年代中期被统计学习热潮盖住，2010年左右，随着计算机性能的提升，深度学习成为了最热门的研究领域，连接主义重新成为了机器学习的主流。

机器学习问题中，很大一部分是监督学习。监督学习中，有回归问题、分类问题和结构学习等；除此之外还有无监督学习，典型的无监督学习是聚类问题；还有半监督问题、迁移学习和强化学习等。为了更好地解释上述的机器学习问题，需要引入一些概念。在机器学习中，过往经验以一个个“样本”的形式存在，而样本以一个个“特征”对事件或对象进行描述，在有监督学习中，对于训练集中的每个“样本”都有标记。例如，在“预测房价”的问题中，每个样本都是采集到的一个房子的数据，房子的位置、面积、楼层、开发时间等等都是特征，而房子的价格就是标记。样本的集合称为“数据集”。有监督学习，就是样本有标记，无监督学习就是没有标记，半监督学习就是一部分样本有标记。回归问题，研究一组独立变量和另一组独立向量之间的关联，例如预测房价；分类问题，可以理解为离散状态下的回归问题，例如手写识别；结构学习学习数据中的结构，例如语音识别和机器翻译。无监督学习中的聚类问题是没有标识的分类问题，例如在没有标识的情况下自动将0-9分成十个不同的类别；迁移学习指将在解决某个问题中学到的知识应用到另一个问题上，例如程序在一个数据集上学习到识别猫图像的模型后，可以利用少量的样本学习识别狗图像的模型。强化学习旨在利用正负反馈使程序能够对环境做出应对。下面，我将就解决这些机器学习问题列举一些经典的算法。

对于最基本的线性回归问题，我们先用形式化的语言描述一下问题：给定数据集，其中，表示每个样本有d个特征，，代表每个实例的标记值。现在试图学得一个线性函数

使得与尽量接近。对于这个问题，可以将其视作一个优化问题，优化对象就是与的接近程度，这里可以使用二范数来描述它们的接近程度，于是可以写作

为了简化该表达式，将b吸入，定义

同时定义一个维的矩阵**X**，使得

同时，将整合成一个列向量**y**，现在可以把线性回归问题写作

这是一个没有约束条件的凸优化问题，对求偏导，令其为0，可以求出解析解：

但是在实际应用中，可能遇到非满秩，或者特征太多(即d的值太大)，导致的维数太高，求逆的时间复杂度太大的问题。此时，可以使用梯度下降法进行求解。梯度下降法的每次迭代如下：

其中，是学习率，是人为设计的一个值。但是学习率设计不当可能导致梯度下降法无法收敛，由此，能引出梯度下降法的各种变种，诸如引入动量项，引入惩罚项，学习率自动计算等等，内容过于繁杂，此处不予赘述。回到线性分类问题，当样本很少的时候，或者选择特征太多之时容易引起机器学习中一个常见问题：过拟合。过拟合指训练出来的模型太过于贴近样本数据，以至于泛化能力弱，即预测未来新数据的能力弱，例如使用一个十次多项式来拟合平面上的十个点。为了避免过拟合，在损失函数中引入正则项，使中的值更贴近于零：

其中，是惩罚系数，是人为设计的，表示对的惩罚力度。

如果回归问题的结果只有有限个值，那就变成了分类问题，分类问题有非常多的算法，如今仍然是一个热门的研究领域，目前分类算法中占据主导的无疑是神经网络，在神经网络大热前，比较热门的算法包括SVM及其变种，现在我们就来讨论一下这两个算法。首先我们来讨论一下SVM。先用形式化的语言描述一下二分类问题：对于数据集，其中，。现寻找一个模型，使得与尽量接近。可以看出，这个定义和回归问题的区别仅在于的范围。SVM的思想是，使得这个模型所表示的高维曲面正好居于正样本和负样本点的正中间。为了更好地讲述SVM的思想，我们需要循序渐进地对其进行讨论，首先，我们假设这个数据集是线性可分的，即存在一个超平面，能将正例和负例完全分开，定义该超平面为：

其中为超平面的法向量，令正例使得，负例使得。刚才我们直观的描述“正中间”，可以用几何间隔来加以描述：

其中，表示第i个样本点的几何间隔，等式右边的后一项的绝对值为该点到分离超平面的距离，而乘上前一项之后，容易验证结果必定为正值，右边第一项的作用是使得正例和负例的几何间隔表示统一。所以表示第i个样本点到分离超平面的距离。现在要使得超平面处于正中间，即要最大化最小的几何间隔，即：

为了化简，我们想要消去项，于是，上式等价于：

的Hessian矩阵是单位阵，约束是线性约束，所以这是个凸优化问题。写出其拉格朗日函数：

之后可以使用KKT条件进行求解。在实际使用的过程中，为了减少开销，会使用SMO算法进行计算。由于噪声的存在，使得数据集中存在一些特异点，使得该数据集线性不可分，即不存在一个超平面，能够完全将正例和反例分开来，但是去除掉这些特异点，其余点都是线性可分的。要解决这个问题，需要使用软间隔最大化，即引入松弛变量，优化问题改写为：

回到一开始未经简化的版本，即如果样例点不是线性可分的，那么就需要引入核函数。而现实中大部分的问题是极其复杂的，核函数的选择是一个难题，此时，深度神经网络的用武之地就显现了出来。

讨论深度神经网络，得先讨论其前身——多层感知机。多层感知器是全连接的神经网络，设l层的输入向量为，输出向量为，偏移为，并且定义系数矩阵，其第i行第j个元素为从(l-1)层的神经元j连接到l层的神经元i的系数，与的关系为：

设激活函数为，其作用是使得损失函数变成非线性函数，常见的激活函数有阶跃函数和sigmoid函数，每一层激活前后的关系为：

多层感知机一般使用梯度下降法来求极小值（不一定是最小值），而求梯度有著名的后向传递算法，其核心思想是使用链式法则，从输出层到输入层依次求梯度。除了全连接层，常见的还有卷积层和池化层，卷积层的作用是提取局部特征，卷积层还有一个功能是将m通道的矩阵投射成n通道的矩阵，这需要m\*n个卷积核，比如在图像处理中，将长和宽分别为L和W的二维图像投影成C\*L\*W的立方体，其中C代表通道数，这种投影或许能提取出图像的某种隐藏特征（比如图像的空间特征），但是这是一个的黑箱子，人们据此搭建出非常复杂的神经网络，能达到很好的效果，却无法解释其背后的逻辑。池化层的作用是消除噪音的影响，常见的池化有最大值池化和平均值池化。

除了上述的三种算法，还有其他无数的算法，在这里无法一一展开。机器学习是一个热门的研究领域，我对此很有兴趣，未来希望能从事机器学习方面的科研。