1. **数学在人工智能的应用**
2. **线性代数——如何将研究对象形式化**
3. **从数学概念的角度来分析机器学习中的线性代数**

机器学习中涉及到的线性代数无非是两方面，一方面是**线性空间理论**，也就是我们说的向量、矩阵、变换这样一些问题。第二个是**矩阵分析或者说矩阵乘法**。比如说主成分分析法、奇异值分解、线性回归等等问题。

1. **.主成分分析法（PCA）**

主成分分析法是一种分析、简化数据集的技术。**常用于降低数据集的维数**，同时保持数据集中对方差贡献最大的特征。其方法主要是通过对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）与它们的权值（即特征值）。PCA 是最简单的以特征量分析多元统计分布的方法。在分析复杂数据时尤为有用，比如人脸识别。

1. **.奇异值分解（SVD）**

奇异值分解（Singular Value Decomposition, SVD）是线性代数中一种重要的矩阵分解。PCA 的实现一般有两种，一种是通过特征值分解实现，另一种就是用奇异值分解来实现。但特征值分解只是针对方阵而言的，当处理 m\*n 的矩阵时，就需要奇异值分解了。

上述两种方法都是属于机器学习算法中的**降维算法**。

1. **.线性回归**

解决线性回归问题的最常用方法是通过**最小二乘法**，这一方法是使用线性回归的矩阵分解方法解决的（例如 LU 分解或奇异值分解）。

1. **高斯消元法**

高斯消元法（Gaussian Elimination），是线性代数中的一个算法，可用来为线性方程组求解，求出矩阵的秩，以及求出可逆方阵的逆矩阵。

1. **.最小二乘法**

最小二乘法是对过度确定系统，即其中存在比未知数更多的方程组，以[回归分析](https://link.zhihu.com/?target=https://zh.wikipedia.org/wiki/%E8%BF%B4%E6%AD%B8%E5%88%86%E6%9E%90" \t "https://zhuanlan.zhihu.com/p/_blank)求得近似解的标准方法。选择未知参数，使得理论值与观测值之差的平方和达到最小，最重要的应用是在[曲线拟合](https://link.zhihu.com/?target=https://zh.wikipedia.org/wiki/%E6%9B%B2%E7%B7%9A%E6%93%AC%E5%90%88" \t "https://zhuanlan.zhihu.com/p/_blank)上。

**B.应用的角度来总结机器学习中涉及到的线性代数**

1. **.数据集和数据文件(Dataset and Data Files )**

在机器学习中，需要在数据集上拟合一个模型时就涉及到线性代数，可以用矩阵来进行表征。

1. **.图像和照片(Images and Photographs )**

图像和照片也是线性代数矩阵的一种。与图像相关的操作，如裁剪、缩放、剪切等，都是使用线性代数的符号和运算来描述的。

1. **.编码(One-Hot Encoding one-hot)**

有时候机器学习中要用到分类数据，可能是用于解决分类问题的类别标签，也可能是分类输入变量。对分类变量进行编码以使它们更易于使用并通过某些技术进行学习是很常见的。one-hot 编码是一种常见的分类变量编码。

1. **.正则化(Regularization)**

在机器学习中，往往寻求最简单可行的模型来解决问题。例如回归方法和人工神经网络，较简单的模型通常具有较小的系数值。一种常用于模型在数据拟合时尽量减小系数值的技术称为**正则化**，常见的实现包括正则化的 L2 和 L1 形式。这两种正则化形式实际上是系数矢量的大小或长度的度量，是直接脱胎于名为矢量范数的线性代数方法。

1. **.潜在语义分析(Latent Semantic Analysis)**

在用于处理文本数据的机器学习子领域（称为自然语言处理），通常将文档表示为词出现的大矩阵。这是文本的稀疏矩阵表示。矩阵分解方法（如奇异值分解）可以应用于此稀疏矩阵，该分解方法可以提炼出矩阵表示中相关性最强的部分。以这种方式处理的文档比较容易用来比较、查询，并作为监督机器学习模型的基础。

**6).推荐系统(Recommender Systems)**

推荐系统时机器学习的子领域，推荐系统的开发主要涉及线性代数方法。

比如，使用欧式距离或点积之类的距离度量来计算稀疏顾客行为向量之间的相似度。或者，用**奇异值分解**这样的矩阵分解方法来提取项目和用户数据的有用部分，以备查询、检索及比较。

1. **.深度学习(Deep Learning)**

人工神经网络的执行涉及线性代数数据结构的相乘和相加。如果扩展到多个维度，深度学习方法可以处理向量、矩阵，甚至输入和系数的张量，此处的张量是一个两维以上的矩阵。线性代数是描述深度学习方法的核心，它通过矩阵表示法来实现深度学习方法。

**8).线性回归(Linear Regression)**

**9).主成分分析(Principal Component Analysis)**

**10).奇异值分解(Singular-Value Decomposition)**

1. **概率论——如何描述统计规律**

在具体地给定了目标函数之后，我们在实际地去评价这个目标函数的时候，我们会用到一些概率论。比方说给定了一个分布，我要求解这个目标函数的期望值。在平均意义上，这个目标函数能达到什么程度呢？这个时候就需要使用到概率论。

1. **.最大似然估计**

最大似然估计其实就是最大可能性估计。最大似然估计是一种[统计方法](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%96%B9%E6%B3%95" \t "https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E4%BC%BC%E7%84%B6%E4%BC%B0%E8%AE%A1/_blank)，它用来求一个样本集的相关[概率密度函数](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0" \t "https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E4%BC%BC%E7%84%B6%E4%BC%B0%E8%AE%A1/_blank)的参数。通俗理解来说，就是利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值！换句话说，极大似然估计提供了一种给定观察数据来评估模型参数的方法，即：“模型已定，参数未知”。**线性回归模型**使用的误差函数（也叫代价函数）采用的是平方和，正是最大似然估计来确定的。

1. **.贝叶斯模型**

**综上，极大似然思想、贝叶斯模型** 是理论基础，朴素贝叶斯(Naïve Bayes )、语言模型(N-gram)、隐马尔科夫（HMM）、隐变量混合概率模型是他们的高级形态。常见分布如高斯分布是混合高斯模型(GMM)等的基础

1. **.高斯分布**
2. **.EM算法**

**最大期望算法的思想**是假设我们想估计知道A和B两个参数，在开始状态下二者都是未知的，但如果知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。可以考虑首先赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。

**举个例子**说明一下，假设我们想估计知道A和B两个参数，在开始状态下二者都是未知的，但如果知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。可以考虑首先赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。

最大期望算法经常用在机器学习和计算机视觉的数据聚类（Data Clustering）领域。

1. **.蒙特卡洛**

蒙特卡罗是一类随机方法的统称。蒙特卡洛是依靠足够多次数的随机模拟，来得到近似结果的算法，说白了就是通过频率来估计概率。

1. **.时间序列**

时间序列（time series）是一组按照时间发生先后顺序进行排列的数据点[序列](https://link.zhihu.com/?target=https://zh.wikipedia.org/wiki/%E5%BA%8F%E5%88%97" \t "https://zhuanlan.zhihu.com/p/_blank)。通常一组时间序列的时间间隔为一恒定值（如1秒，5分钟，12小时，7天，1年），因此时间序列可以作为离散时间数据进行分析处理。

1. **数理统计——如何以小见大**

线性回归、逻辑回归都是数理统计的内容，实际上是统计学的内容。

实际上对于数理统计来说，我们在评价模型的时候，不只关注的是一个目标函数，我们可能还关注一些它的统计特性。比如说它的置信度，或者是其他的一些指标。你这个模型建立起来，它的可信性程度到底有多大，这些在早期的机器学习算法当中也是需要考虑的。当然随着神经网络，随着深度学习的兴起，这部分内容实际上渐渐地衰落，或者渐渐地被忽略。你在神经网络当中可能只需要达到一个这个好的目标函数，好的指标就行，至于说它的置信度，这些我们不去考虑。所以说，这也是深度学习不太受学数学，或者说学统计学的人待见的一个原因。因为统计学强调什么呢？强调可解释性，你这个模型能够达到什么样的指标，我们能把它清清楚楚地讲明白，为什么能够达到这样的指标，它的原理在哪？它背后的根据在哪？我给定一个分布，假如说高斯分布，那么再给定一个模型，我就可以通过严谨而简洁的这个数学推导，把这个结果以公式的形式给它呈现出来，这个看起来就很高大上，或者说很清楚。但神经网络和深度学习，现在还达不到这样可解释的程度。

1. **最优化理论——如何找到最优解**

在李航博士的《统计学习方法》中，其将机器学习总结为如下表达式：机器学习 = 模型 + 策略 + 算法，其中算法指的就是优化算法，其中梯度下降法和牛顿法用的比较多。

**梯度**就是函数的导数方向。**梯度法**是求解**无约束多元函数极值**最早的数值方法，很多机器学习的常用算法都是以它作为算法框架，进行改进而导出更为复杂的优化方法。

**牛顿法**也是求解**无约束最优化问题**常用的方法，最大的优点是收敛速度快。从本质上去看，**牛顿法是二阶收敛，梯度下降是一阶收敛，**所以牛顿法就更快。通俗地说，比如你想找一条最短的路径走到一个盆地的最底部，梯度下降法 每次只从你当前所处位置选一个坡度最大的方向走一步，牛顿法在选择方向时，不仅会考虑坡度是否够大，还会考虑你走了一步之后，坡度是否会变得更大。所以，**可以说牛顿法比梯度下降法看得更远一点，能更快地走到最底部。**

梯度下降是用平面来逼近局部，牛顿法是用曲面逼近局部

1. **信息论——如何定量度量不确定性**

**逻辑回归中广泛应用了交叉熵，逻辑回归的损失函数就是交叉熵。**

**相对熵又称为KL散度、信息散度。相对熵应用在生成模型中，比如GAN、EM、贝叶斯学习和变分推导中。信息增益应用在决策树ID3算法中比较多。**

**在最优化理论中，很多算法用熵来作为优化目标。**

1. **数学在人工智能的作用**

**线性代数：**

**1.将具体事物抽象为数学对象**

**2.提升大规模运算的效率**

要处理的数据越来越多，如果只是简单地说，用最传统的方法，用一个一个的 for 循环去处理高维的矩阵，它的效率肯定是相当低下。有了线性代数之后，我们可以把矩阵的运算引入到机器学习的算法当中，通过一些额外的库，或者一些额外的软件包，提升大规模运算的效率。这里面最直观的一个例子就是 MATLAB 软件。

**概率统计：**

**1.利用数据来学习模型**

**2.利用模型推断数据**

**最优化理论：**

**1.**