**学习大数据需要具备的数学基础**

提到大数据，很多人首先联想到的是每天要和一堆数据相处，数学必须要很好，当然，大数据的学习确实需要一定的数学基础，但是也有一定的侧重点。

大数据有很多细分方向，如大数据开发、大数据分析、大数据挖掘、大数据运维等，不同方向对数学的要求是不同的，但是如果想更好的学习大数据，还是需要有一定的数学基础。

**1. 概率论与数理统计**

这部分与大数据技术开发的关系非常密切，条件概率、独立性等基本概念、随机变量及其分布、多维随机变量及其分布、方差分析及回归分析、随机过程（特别是Markov）、参数估计、Bayes理论等在大数据建模、挖掘中就很重要。大数据具有天然的高维特征，在高维空间中进行数据模型的设计分析就需要一定的多维随机变量及其分布方面的基础。Bayes定理更是分类器构建的基础之一。除了这些这些基础知识外，条件随机场CRF、隐Markov模型、n-gram等在大数据分析中可用于对词汇、文本的分析，可以用于构建预测分类模型。

当然以概率论为基础的信息论在大数据分析中也有一定作用，比如信息增益、互信息等用于特征分析的方法都是信息论里面的概念。

**2. 线性代数**

这部分的数学知识与大数据技术开发的关系也很密切，矩阵、转置、秩 分块矩阵、向量、正交矩阵、向量空间、特征值与特征向量等在大数据建模、分析中也是常用的技术手段。

在互联网大数据中，许多应用场景的分析对象都可以抽象成为矩阵表示，大量Web页面及其关系、微博用户及其关系、文本集中文本与词汇的关系等等都可以用矩阵表示。比如对于Web页面及其关系用矩阵表示时，矩阵元素就代表了页面a与另一个页面b的关系，这种关系可以是指向关系，1表示a和b之间有超链接，0表示a,b之间没有超链接。著名的PageRank算法就是基于这种矩阵进行页面重要性的量化，并证明其收敛性。

以矩阵为基础的各种运算，如矩阵分解则是分析对象特征提取的途径，因为矩阵代表了某种变换或映射，因此分解后得到的矩阵就代表了分析对象在新空间中的一些新特征。所以，奇异值分解SVD、PCA、NMF、MF等在大数据分析中的应用是很广泛的。

**3. 最优化方法**

模型学习训练是很多分析挖掘模型用于求解参数的途径，基本问题是：给定一个函数f:A→R，寻找一个元素a0∈A，使得对于所有A中的a，f(a0)≤f(a)（最小化）；或者f(a0)≥f(a)（最大化）。优化方法取决于函数的形式，从目前看，最优化方法通常是基于微分、导数的方法，例如梯度下降、爬山法、最小二乘法、共轭分布法等。

**4. 离散数学**

离散数学的重要性就不言而喻了，它是所有计算机科学分支的基础，自然也是大数据技术的重要基础。

大数据正在悄然改变我们的世界，搜索引擎、交通出行、金融分析、医疗健康、娱乐消费、数据挖掘等领域，无处不在，因此，现在选择选择学习大数据正当时！