中国石油大学（华东）
CHINA UNIVERSITY OF PETROLEUM

《计算科学导论》课程总结报告

姓　　名　张文清

学　　号　1915010228

专业班级　计科 1902

学　　院　计算机科学与技术学院

课程认识 30%	问题思考 30%	格式规范 20%	IT 工具 20%	总分	评阅教师

2021 年 12 月 31 日

1 引言

作为一名计算机科学与技术的学生,我深刻意识到计算科学在专业学习中不可或缺的地位。作为一名已经大三的学生，虽然有些内容已经学习过，但是学习《计算科学导论》还是很有必要的，因为《计算科学导论》给我们提供了一个系统学习专业内容的机会，也将学习的内容串联起来，让我们在学习专业课时的路线更明确。就比如在学习离散数学，高数，线性代数和概率论与数理统计时会疑惑这些数学课对我们写代码有什么用呢，但其实计算机的研发最初就是为了能够解决数学问题而产生的，随着科技的不断进步，当下计算科技的发展同样需要依靠数学科学的支持来完成,计算机的出现是为了能够解决人力无法进行的数学科学问题，而数学科学又是计算机发展不可或缺的重要基础组成部分。而且很多计算机大牛的数学功底都很好，如果想更清楚透彻的了解算法，良好的数学基础是必不可少的。

2 对计算科学导论这门课程的认识、体会

《计算科学导论》这门课带我重新认识了计算机的世界，通过老师上课的讲解，我重新认识了计算机，对计算机的起源与发展、计算机体系结构、程序设计、算法、软件工程、操作系统、人工智能以及计算机专业的培养目标都有了更深入更全面的认识，同时在上课学习的过程中，我对计算科学的兴趣也进一步得到了培养。

《计算科学导论》的目的在于从科学哲学的角度用高级科普的形式为初学者提供一个了解和学习计算机科学与技术领域的专门的，具体的，系统的专业技术知识。而且导论并不系统地阐述科学哲学与学科方法论的内容，而是将科学哲学的观点 与学科方法论中大量成熟的内容融入到各章节之中，自始至终贯穿在各个章节的字里行间中。导论中很少涉及具体的，系统的专业知识特别是操作使用计算机的技术知识，不必感到困惑， 因为导论仅是为我们以后学习基础课程和后续计算机科学与技术的一个导引。其中没有解释的一些名词和术语,以后会在一些具体的分支学科课程中学到,其中的许多观点和思想方法，对整个大学生涯是大有裨益的。

2.1 计算科学与数学的关系

在这个信息的时代，在这个几乎所有事都要求量化的时代，在这个任何时候都离不开资源整理分析的时代,我们离不开计算机这个科学进步的代表,更离不开科学的基础数学。从计算机和数学家的关系中可以看出计算机和数学的关系,而计算机在数学中的应用更进一步体现了数学和计算机密不可分。

数学与电子科学构成了我们今天计算机系统的基础,也构成了计算科学的基础。但是，与数学相比，电子技术基础地位的重要性不及数学，原因是数学提供了计算科学最重要的学科思想和学科的方法论基础，而电子技术主要是提供了今日计算机的实现技术，它仅仅

是对计算科学许多数学思想和方法的一种当前最现实、最有效的实现技术。

就目前非常火热的人工智能，其中的算法全部都要以数学为基础。其实计算机专业真正学的是数学应用，通过编程将数学应用起来。像这学期我们还学了机器学习的课程，在完成大作业时发现了这样一句话，“只有了解算法的人才配调参”。毕竟对于一些机器学习模型,很多时候只需要调参数就能得到很好的训练效果,但如果不了解底层的数学逻辑，那也只能停留在“用”的阶段，而不是“会”。如果仔细去看他们的理论分析就会发现，其实这些东西都是数学，看他们的论文就会发现，根本没人在乎你是怎么实现的。数学才是机器学习的理论基础，而计算机不过是工具而已。深度学习的过程中，可以深刻的体会到，人工智能上，每一次技术的突破，一定是数学在应用上的突破，计算机突破导致人工智能突破已经很少了。所以，数学，是相当重要，尤其概率统计，微积分模型。

2.2 区块链的应用

区块链是以比特币为代表的数字加密货币体系的核心支撑技术。区块链技术的核心优势是去中心化，能够通过运用数据加密、时间戳、分布式共识和经济激励等手段，在节点无需互相信任的分布式系统中实现基于去中心化信用的点对点交易、协调与协作，从而为解决中心化机构普遍存在的高成本、低效率和数据存储不安全等问题提供了解决方案。随着比特币近年来的快速发展与普及，区块链技术的研究与应用也呈现出爆发式增长态势，被认为是继大型机、个人电脑、互联网、移动/社交网络之后计算范式的第五次颠覆式创新，是人类信用进化史上继血亲信用、贵金属信用、央行纸币信用之后的第四个里程碑[1]。区块链技术是下一代云计算的雏形，有望像互联网一样彻底重塑人类社会活动形态，并实现从目前的信息互联网向价值互联网的转变。

区块链独特的技术设计这使得区块链技术不仅可以成功应用于数字加密货币领域,同时在经济、金融和社会系统中也存在广泛的应用场景。根据区块链技术应用的现状，区块链目前的主要应用为数字货币、数据存储、数据鉴证、金融交易、资产管理和选举投票共六个场景[4]。

数据鉴证：区块链数据带有时间戳、由共识节点共同验证和记录、不可篡改和伪造，这些特点使得区块链可广泛应用于各类数据公证和审计场景。例如，区块链可以永久地安全存储由政府机构核发的各类许可证、登记表、执照、证明、认证和记录等，并可在任意时间点方便地证明某项数据的存在性和一定程度上的真实性。包括德勤在内的多家专业审计公司已经部署区块链技术来帮助其审计师实现低成本和高效地实时审计;; Factom 公司则基于区块链设计了一套准确的、可核查的和不可更改的审计公证流程与方法[2] 。

金融交易: 区块链技术与金融市场应用有非常高的契合度。区块链可以在去中心化系统中自发地产生信用， 能够建立无中心机构信用背书的金融市场， 从而在很大程度上实现了 “金融脱媒”， 这对第三方支付、资金托管等存在中介机构的商业模式来 说是颠覆

性的变革; 在互联网金融领域， 区块链特别适合或者已经应用于股权众筹、P2P 网络借贷和互联网保险等商业模式; 证券和银行业务也是区块链的重要应用领域，传统证券交易需要经过中央结算机构、银行、证券公司和交易所等中心机构的多重协调，而利用区块链自动化智能合约和可编程的特点，能够极大地降低成本和提高效率，避免繁琐的中心化清算交割过程，实现方便快捷的金融产品交易；同时，区块链和比特币的即时到帐的特点可使得银行实现比 SWIFT 代码体系更为快捷、经济和安全的跨境转账；这也是目前 R3CEV 和纳斯达克等各大银行、证券商和金融机构相继投入区块链技术研发的重要原因。

资产管理：区块链在资产管理领域的应用具有广泛前景，能够实现有形和无形资产的确权、授权和实时监控。对于无形资产来说，基于时间戳技术和不可篡改等特点，可以将区块链技术应用于知识产权保护、域名管理、积分管理等领域; 而对有形资产来说，通过结合物联网技术为资产设计唯一标识并部署到区块链上，能够形成 “数字智能资产”，实现基于区块链的分布式资产授权和控制。例如， 通过对房屋、车辆等实物资产的区块链密钥授权，可以基于特定权限来发放和回收资产的使用权，有助于 Airbnb 等房屋租赁或车辆租赁等商业模式实现自动化的资产交接; 通过结合物联网的资产标记和识别技术，还可以利用区块链实现灵活的供应链管理和产品溯源等功能。

选举投票：投票是区块链技术在政治事务中的代表性应用。基于区块链的分布式共识验证、不可篡改等特点，可以低成本高效地实现政治选举、企业股东投票等应用； 同时，区块链也支持用户个体对特定议题的投票。例如，通过记录用户对特定事件是否发生的投票，可以将区块链应用于博彩和预测市场等场景[3]; 通过记录用户对特定产品的投票评分与建议，可以实现大规模用户众包设计产品的“社会制造”模式等。

3 进一步的思考

我在本学期的选题为数据清洗和特征工程。

首先先来解释下什么是特征。中文字典里是这么解释特征的：一事物异于其他事物的特点。英文字典里是这么解释 feature 的：A feature of something is an interesting or important part or characteristic of it.把这两个综合一下，特征就是，于己而言，特征是某些突出性质的表现，于他而言，特征是区分事物的关键，所以，当我们要对事物进行分类或者识别，我们实际上就是提取‘特征’，通过特征的表现进行判断。在机器学习中，特征是被观测对象的一个独立可观测的属性或者特点。比如识别水果的种类，需要考虑的特征（属性）有：大小、形状、颜色等。

其次对于数据清洗其实是属于特征工程的一部分的，但由于数据清洗在特折工程中十分重要，所以在这里单独提出。其次特种工程目前有手动和自动两大类。

3.1 什么是特征工程

图 3-1：特征工程在机器学习中的位置

从上图可以看出，特征工程处在原始数据和特征之间。他的任务就是将原始数据“翻译”成特征的过程。特征：是原始数据的数值表达方式，是机器学习算法模型可以直接使用的表达方式。特征工程是一个过程，这个过程将数据转换为能更好的表示业务逻辑的特征，从而提高机器学习的性能。

图 3-2：特征工程就像烹饪

其实特征工程跟做饭很像：我们将食材购买回来，经过清洗、切菜，然后开始根据自己的喜好进行烹饪，做出美味的饭菜。上面的例子中：食材就好像原始数据清洗、切菜、烹饪的过程就好像特征工程最后做出来的美味饭菜就是特征

人类是需要吃加工过的食物才行，这样更安全也更美味。机器算法模型也是类似，原始数据不能直接喂给模型，也需要对数据进行清洗、组织、转换。最后才能得到模型可以消化的特征。

3.2 特征工程的重要性

美国计算机科学家 Peter Norvig 有句经典名言：更多的数据优于聪明的算法，而好的数据优于多的数据。（More data beats clever algorithms, but better data beats more data.）[5] 这句话说明了特征工程的重要性。所以，如何基于给定数据来发挥更大的数据价值就是特征工程要做的事情。

在 16 年的一项调查中发现，数据科学家的工作中，有 80%的时间都在获取、清洗和组织数据。构造机器学习流水线的时间不到 20%。详情如下[6]：

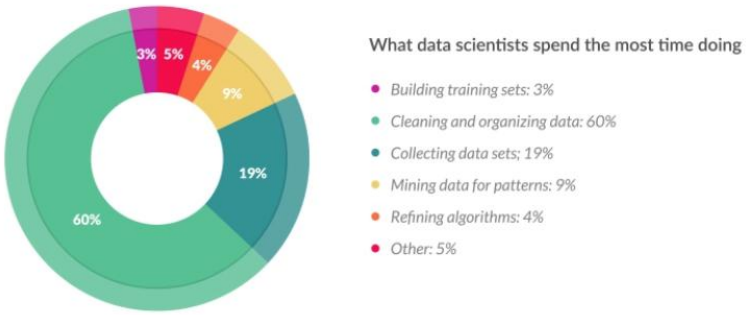

图 3-3：数据科学家各部分工作用时

可以看到特征工程在是实际情况中是十分重要的一部分。其重要性有以下四点：

- 良好的特征意味着更好的灵活性
  即便我们选择了一个相对较差的模型，但是由于良好的特征加持，我们依旧可以获得较好得到结果。因为大部分模型其实都可以很好地利用数据中良好的结构信息。良好的特征可以使我们使用更加简单的的模型，这样更容易训练，也更容易理解和维护。

- 良好的特征意味着更简单的模型
  如果我们提取出了较好的特征，其实它已经具备足够强的表达能力，可以为我们节省大量在模型选择和参数调优上的时间。

- 良好的特征意味着更好的结果
  对于 Kagglers，或者参加其他比赛的同学来说，大家使用的算法都大同小异，大部分时间都花在了特征工程上面。

3.3 手动特征工程

手动特征工程是一种传统的特征工程方法，它主要是利用领域知识来构建特征，一次只能产生一个特征，这是一个繁琐，费时又易出错的过程。此外，每次进行手动特征工程的代码是针对特定的问题，当我们要解决一个新问题、新数据集时，我们需要重写相关代码。

一般手动进行特征工程时，主要有以下几个步骤。首先进行数据清洗，然后是数据预处理，最后是特征选择。

数据清洗是对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。一般来说按以下几种清洗方法对数据进行清洗：

- 格式内容清洗：时间日期格式不一致，数值格式不一致，数据类型不符等。

- 逻辑错误清洗：数据重复清洗如存在各个特征值完全相同的两条/多条数据，数据不完全相同，但从业务角度看待数据是同一个数据。不合理值清洗，根据业务常

识，或者使用但不限于箱型图（Box-plot）发现数据中不合理的特征值。

- 异常值清洗：异常值是数据分布的常态，处于特定分布区域或范围之外的数据通常被定义为异常或噪声。异常分为两种：“伪异常”，由于特定的业务运营动作产生，是正常反应业务的状态，而不是数据本身的异常；“真异常”，不是由于特定的业务运营动作产生，而是数据本身分布异常，即离群点。

- 缺失值清洗：在处理缺失值时，可以直接将缺失太多的特征直接删除，也可以使用填充的方法将缺失值填充完整。在填充时可以用基于统计值的填充如：平均数，中位数，众数等。也可用模型填充，使用待填充字段作为 Label，没有缺失的数据作为训练数据，建立分类/回归模型，对待填充的缺失字段进行预测并进行填充。

接下来是数据预处理，因为未经处理的特征常会有下问题：

- 不属于同一量纲：即特征的规格不一样，不能够放在一起比较。无量纲化可以解决这一问题。

- 信息冗余：对于某些定量特征，其包含的有效信息为区间划分，例如学习成绩，假若只关心“及格”或不“及格”，那么需要将定量的考分，转换成“1”和“0”表示及格和未及格。二值化可以解决这一问题。

- 定性特征不能直接使用：某些机器学习算法和模型只能接受定量特征的输入，那么需要将定性特征转换为定量特征。最简单的方式是为每一种定性值指定一个定量值，但是这种方式过于灵活，增加了调参的工作。通常使用哑编码的方式将定性特征转换为定量特征：假设有 N 种定性值，则将这一个特征扩展为 N 种特征，当原始特征值为第 i 种定性值时，第 i 个扩展特征赋值为 1，其他扩展特征赋值为 0。哑编码的方式相比直接指定的方式，不用增加调参的工作，对于线性模型来说，使用哑编码后的特征可达到非线性的效果。

- 信息利用率低：不同的机器学习算法和模型对数据中信息的利用是不同的，之前提到在线性模型中，使用对定性特征哑编码可以达到非线性的效果。类似地，对定量变量多项式化，或者进行其他的转换，都能达到非线性的效果。

最后是特征选择，就是从多个特征中，挑选出一些对结果预测最有用的特征。因为原始的特征中可能会有冗余和噪声。 我们需要选择有意义的特征输入机器学习的算法和模型进行训练。通常来说，从两个方面考虑来选择特征：

- 特征是否发散：如果一个特征不发散，例如方差接近于 0，也就是说样本在这个特征上基本上没有差异，这个特征对于样本的区分并没有什么用。

- 特征与目标的相关性：这点比较显见，与目标相关性高的特征，应当优选选择。除方差法外，本文介绍的其他方法均从相关性考虑。

根据特征选择的形式又可以将特征选择方法分为 3 种：

- Filter：过滤法，按照发散性或者相关性对各个特征进行评分，设定阈值或者待选择阈值的个数，选择特征。
- Wrapper：包装法，根据目标函数（通常是预测效果评分），每次选择若干特征，或者排除若干特征。
- Embedded：嵌入法，先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据系数从大到小选择特征。类似于 Filter 方法，但是是通过训练来确定特征的优劣。

3.4 自动特征工程

自动化特征工程是一种相对较新的技术,用于解决真实世界数据集所面临的一系列科学问题。自动化特征工程是通过从一组相关的数据表中自动提取有用且有意义的特征，这种方法能够改变标准的工作流程，并适用于任务数据集的有关问题。此外，它不仅减少了特征工程所需的时间，还创建了可解释性的特征，并通过过滤与时间相关的数据来防止数据泄漏。

下面介绍三个自动化特征工程工具包：Featuretools，Boruta，tsfresh。

Featuretools 使用一种称为深度特征合成（Deep Feature Synthesis，DFS）的算法，该算法遍历通过关系数据库的模式描述的关系路径。当 DFS 遍历这些路径时，它通过应用于数据的操作（包括和、平均值和计数）生成综合特征。例如，对来自给定字段 client_id 的事务列表应用 sum 操作，并将这些事务聚合到一个列中。尽管这是一个深度操作，但该算法可以遍历更深层的特征。Featuretools 最大的优点是其可靠性和处理信息泄漏的能力，同时可以用来对时间序列数据进行处理。

Boruta 主要是用来进行特征选择。所以严格意义上，Boruta 并不是我们所需要的自动化特征工程包。Boruta-py 是 brouta 特征约简策略的一种实现，在该策略中，问题以一种完全相关的方式构建，算法保留对模型有显著贡献的所有特征。这与许多特征约简算法所应用的最小最优特征集相反。boruta 方法通过创建由目标特征的随机重排序值组成的合成特征来确定特征的重要性，然后在原始特征集的基础上训练一个简单的基于树的分类器，在这个分类器中,目标特征被合成特征所替代。所有特性的性能差异用于计算相对重要性。Boruta 函数通过循环的方式评价各变量的重要性，在每一轮迭代中，对原始变量和影子变量进行重要性比较。如果原始变量的重要性显著高于影子变量的重要性，则认为该原始变量是重要的；如果原始变量的重要性明显低于影子变量的重要性，则认为该原始变量是不重要的。其中，原始变量就是我们输入的要进行特征选择的变量；影子变量就是根据原始变量生成的变量。

tsfresh 是基于可伸缩假设检验的时间序列特征提取工具。该包包含多种特征提取方法和鲁棒特征选择算法。tsfresh 可以自动地从时间序列中提取 100 多个特征。这些特征描述了时间序列的基本特征，如峰值数量、平均值或最大值，或更复杂的特征，如时间反转对称性统计量等。这组特征可以用来在时间序列上构建统计或机器学习模型，例如在回归或分类任务中使用。时间序列通常包含噪声、冗余或无关信息。因此，大部分提取出来的特征对当前的机器学习任务没有用处。为了避免提取不相关的特性，tsfresh 包有一个内置的过滤过程。这个过滤过程评估每个特征对于手头的回归或分类任务的解释能力和重要性。它建立在完善的假设检验理论的基础上，采用了多种检验方法。

自动化特征工程解决了特征构造的问题，但同时也产生了另一个问题：在数据量一定的前提下，由于产生过多的特征，往往需要进行相应的特征选择以避免模型性能的降低。事实上，要保证模型性能，其所需的数据量级需要随着特征的数量呈指数级增长。

4 总结

通过《计算科学导论》的学习。我对原本学习的专业知识有了更加深刻的理解。除此以外课程还围绕计算机科学与技术学科的定义、特点、基本问题、发展主线、主流方向、学科方法论、历史渊源、发展变化、知识组织结构与分类体系、学科发展的潮流与未来发展方向、学科人才培养目标、教学重点与科学素养等内容进行了系统而又深入浅出的论述，以科学办学思想和内涵发展优先的理念为基础，全面阐述了在培养计算机科学与技术一级学科创新人才与高素质专业技术开发人才的过程中，如何使学生正确地认识和学好计算机科学与技术学科。到现在，我能够整体系统地了解本专业，能够从迷茫与困惑到找到自己应该努力的方向，做出更清楚的自己未来的规划。我认为，这是我从本课程中学到的最重要的知识与得到的最大收获。

5 附录

Github

申请 Github 账户，给出个人网址和个人网站截图

Github 网址：https://github.com/passerby-bye

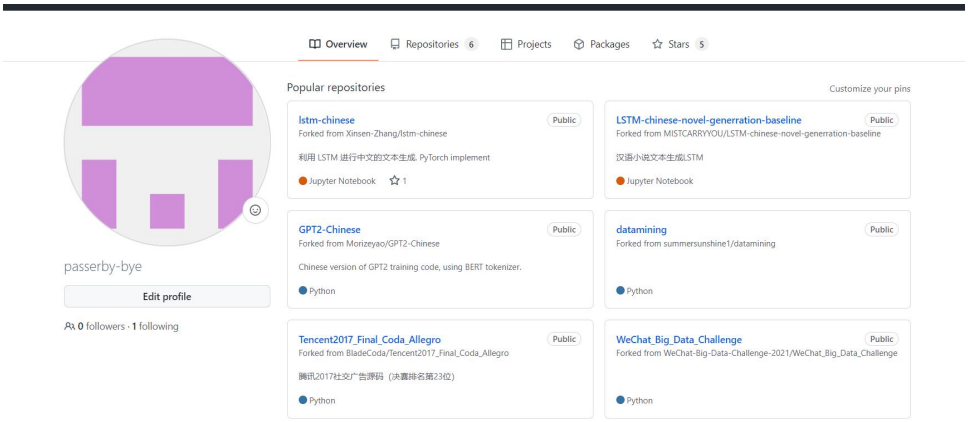

图 5-1：Github 截图

观察者

注册观察者 APP，给出对应的截图

图 5-2：观察者截图

学习强国

注册学习强国 APP，给出对应的截图

图 5-3：学习强国截图

哔哩哔哩

注册哔哩哔哩 APP，给出对应的截图

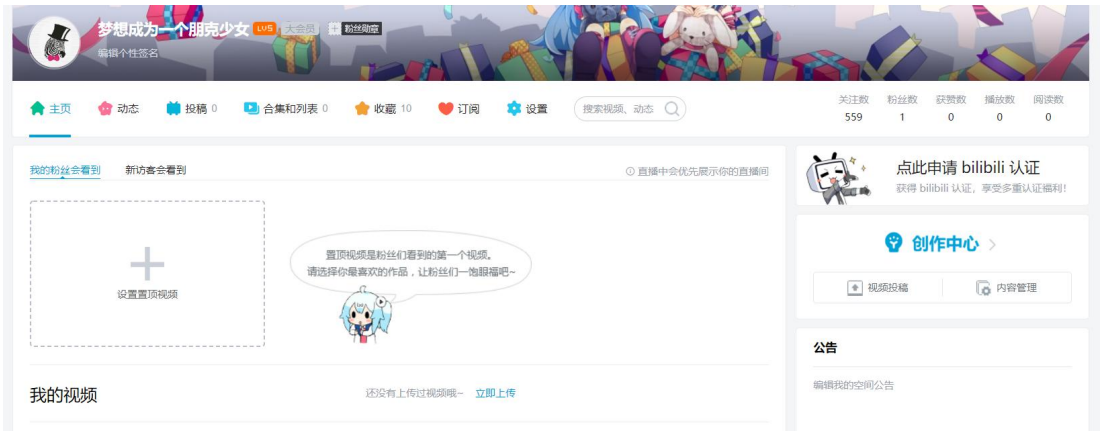

图 5-4：哔哩哔哩截图

CSDN

注册 CSDN 账户，给出个人网址和个人网站截图

CSDN 网址：https://blog.csdn.net/H_1008?spm=1001.2101.3001.5343

图 5-5：CSDN 截图

博客园

注册博客园账户，给出个人网址和个人网站截图

博客园网址：https://home.cnblogs.com/u/2718896

图 5-6：博客园截图

小木虫

注册小木虫账户，给出个人网址和个人网站截图

图 5-7：小木虫截图

参考文献

注意，参考文献至少五篇，其中至少两篇为英文文献，参考文献必须在正文中有引用

[1] Swan M. Blockchain: Blueprint for a New Economy. USA: O0 Reilly Media Inc.,2015.

[2] Factom White Paper [Online], available: http://bite01. com/bit/1421, December 29, 2015.

[3] Brito J, Shadab H, Castillo A. Bitcoin fifinancial regulation:securities, derivatives, prediction markets, and gambling.The Columbia Science & Technology Law Review, 2014, 16:

144−221

[4] 袁勇，王飞跃.区块链技术发展现状与展望自动化学报. 2016,42(04)

[5]https://shaz13.medium.com/rare-feature-engineering-techniques-for-machine-learning-competitions-de36c7bb418f

[6]https://whatsthebigdata.com/2016/05/01/data-scientists-spend-most-of-their-time-cleaning-data/