复习课 112.2.5

▼ old

矩估计、贝叶斯估计：了解即可

简单3层神经网络MLP，知道如何计算，知道反向传播是怎么回事；复杂的网络不作要求（如带隐藏层、非线性MLP）

112.2.5 复习课

题型：简答题和计算题为主，考试范围是课上讲的范围。

公式的话，会有一些比较简单的计算，但不会涉及复杂的计算。

流数据、图数据等没讲的不在考试范围内。

一、概论

大数据和传统的数据有什么不同：4V是什么意思，能够举出来现实生活中大数据的哪些应用（比如搜索、天猫购物商品），包括数据挖掘等等。

二、数据存储

会写基础的SQL语句。

知道传统的关系型数据库是针对结构化的特点，而大数据的形态是多样化的，种类多，需要关系型和非关系型数据库辅助存储。

知道两种数据库的区别。比如，关系型数据库语言统一，存储效率高；非关系型数据库能存储非关系的数据，如文本、图像。

知道面临的数据有哪些形态：结构化，非结构，半结构化数据。哪些是结构化非结构化？

数据库系统有哪些组成？

常见的数据库：postgresql、mysql、neo4j等等。

SQL语言的基本操作，如：如何创建表？

关系型数据库的优点（3点）：管理效率高、统一界面。

对于文本的存储是关系型数据库难以存储的，因此介绍了NoSQL。知道为什么会出现NoSQL：伴随大数据的发展、关系型数据库的范式约束、事务特性等等，不适合数据源源不断到来的大数据时代。

NoSQL和关系型数据库的区别？如NoSQL没有固定的查询格式、没有ACID的严格约束。非关系数据库是面向大数据的数据存储软件系统，要求的是高性能、高可用、可伸缩性，能够及时对大数据存储、没有压力。二者的关注点不一样。

了解常见的非关系数据库：MongoDB、neo4j、redis、Hbase。

知道MongoDB各种名词和MySQL名词的映射关系，知道neo4j的用法。

三、数据预处理

这一部分了解数据预处理的基本手段和方法，是进行大数据应用开发的必备步骤。因为大数据本身是脏的。因此要了解数据清洗、集成、变换、归约的基本方法。

数据清洗的重要性：与大数据特点相关。因为大数据的各种特性，很难发现其中的模式。—不完整不一致不准确。

缺失是一个常见现象，遇到缺失值时有哪些基本方法处理缺失值？例如数据删除（改变原始数据分布）和数据填充（保留原始数据分布）等。

数据集成：大数据有可能来自多个数据源，整合可以丰富样本信息，为了获得更完整的数据和更全面的用户画像。最重要的是检测冗余的样本和冗余的属性，防止浪费存储空间和重复计算，防止改变数据特性使分析结果不准确。

检测冗余属性：常见的有Pearson相关系数和卡方检验。

检测冗余样本：距离度量、相似度计算。（注意马氏距离的公式PPT上有问题）

数据相似度计算：SMC和Jaccard、余弦相似度（并不是距离，但经常用到）。

练习：知道怎么计算。在应用时，常用的是Jaccard、余弦相似度、欧氏距离、卡方检验。

有序数据的度量：Spearman Rank

数据变换：目的是将数据转化为适合分析建模的形式，前提条件是尽量不改变原始数据规律。

数据规范化和数据离散化。知道各有什么主要的方法。

最小最大规范化、zscore规范化是常用的。要知道为什么要进行数据规范化，是如何变换的，知道不同的数据规范化方法它们的优势和不足。

数据离散化：为什么要进行离散化。主要方法：非监督离散化（分箱和聚类）和监督离散化（基于熵的离散化）。要知道如何计算一个系统内的熵，知道熵的意义是什么，要理解熵的公式是什么意思。例子：如何计算数据集中的熵、如何根据数据集中的熵来划分分割点（根据信息增益确定分割点，常用）。

数据归约：什么叫数据归约？缩小数据挖掘所需的数据及规模。通过归约算法，缩减变量个数，提高挖掘效率。主要方法有维度归约和数据归约（两个分别是什么意思？）。

维度归约：主成分分析、矩阵分解、奇异值分解（有兴趣的话可以看一下，可以编程试一下效果）

特征工程：实际上和数据归约是耦合的，特征工程是数据归约后的一个比较重要的步骤，从数据中提取最有效的特征。数据归约是从很多个属性、变量中寻找有用的，而特征工程是提出最有效的特征、真正对预测产生作用的特征，使后期建立的数据模型能够达到更好的效果。

特征工程的意义？特征越好，模型越出色…等等。

特征工程的过程：测试、筛选…

基本特征模式：独热特征等。了解什么是维度灾难和语义鸿沟现象。

TF-IDF表示文档，有兴趣可以看一下计算方法。

如何从原始数据设计特征：固有特征，如统计值；或者引入专家知识等方式构建型的特征。

介绍了生成特征子集的三种方式。

传统特征工程和深度学习特征工程的区别。

简单介绍了几种神经网络的构造，有兴趣的话可以看一下。

提到一句：backward words，文本的词频表示。

深度学习遇到的困境和一些解决方案。

四、探索性数据分析

知道为什么要对大数据进行探索性数据分析：4V特性，处理代价，先利用统计手段了解基本信息，获得一个对数据的基本认识，然后再对数据进行处理。

了解总体和样本的概念。

数据分布的特征：集中趋势、离散程度、…

平均数、中位数、众数..

方差、标准差…

在参数估计中，要知道什么是参数估计：用样本数据估计总体的参数。

主要讲了点估计和区间估计，主要讲的是点估计的基本方法。知道矩估计是怎么计算的。知道极大似然估计是怎么推导的，能够做一些简单的计算。最大后验估计和贝叶斯估计，知道他们和极大似然估计有什么不同，有兴趣的同学可以推导一下如何用最大后验估计来推导扔硬币的参数概率。

（最大后验估计、贝叶斯估计的PPT快快的划过）

知道最大后验估计、贝叶斯估计是怎么回事，它们的优点缺点。

PPT上的例子：从应用角度理解这三种估计有什么不同。

总结这三种估计在估计的时候，估计和预测的目标、特点。

有兴趣的话可以推导用不同估计方式估计线性方程的例子。

假设检验是什么？有什么作用？与参数估计的相同和不同。出发点不一样，结果不一样。

知道假设检验的基本概念，如备择假设、第一第二类错误等。

知道假设检验的过程：提出原假设备择假设构造统计量进行检验…

分布：T检验、标准正态分布…

抽样方法：常见的抽样方法。知道各种抽样方法分别是什么。

五、数据分析方法

知道什么叫机器学习，机器学习的概念、流程。

文本数字化的过程：如何将文本转化为计算机计算的形式，通过什么方式（独热编码、词频、…）

知道机器学习的分类：有监督、无监督、半监督、强化学习，分别是什么意思，有什么区别。

统计分析和机器学习的对比，知道这两个东西方法有什么区别，有什么优势。

例子：用梯度下降方法解决线性回归的思路，有兴趣的话可以推导一下。

给出了一个SVM基本的一些原理。如果有兴趣的话可以在课后研究一下这些基本原理。

知道逻辑斯蒂回归是什么形式。

其他机器学习方法：决策树，介绍了如何构造一个决策树。

知道决策树实际是贪心策略，不修改前面，只修改后面，每一次都是局部最优解。

知道决策树…??连续型数据

六、神经网络

知道向下传导和反向传播的概念，要知道什么是一个前向传播，是一个什么过程。反向传导是怎么回事（…进行梯度传导）。

例子：给定…向量和神经网络，输出值，误差，有兴趣的话可以看一下计算过程。

深度学习的特点和一些优势。知道深度学习的局限性。知道可以寻找一些方法来客服它的局限性。

七、数据可视化

数据可视化是什么。

数据可视化的作用。

从数据科学上来说，数据可视化的最终作用是用感知代替认知，协助思考。

知道各种可视化技术的一些方法，如何对各种类型数据进行可视化，有哪些类型，有哪些特点，有哪些优缺点。

八、文本分析

知道文本分析的意义，为什么要进行文本分析。

知道以文本为核心的应用包含的主要内容：包括数据获取、数据处理、数据检索、文本可视化等过程。

给出一个用爬虫爬取网络数据的典型过程。

回答同学提问：异步IO模型无法减少单个页面抓取速度，但对于全局来说平均每个页面时间降下来了。

介绍了爬虫的一些机制。

网页解析：可以用DOM解析。知道DOM的概念。

了解文本分析的概念，词法分析，为什么要引入分词。会产生分词的一些歧义。

知道基于各种东西的分词方法，在应用时知道他们的优缺点，根据实际需求选择不同的算法。

介绍了一些文本分析的技术模型，如hmm算法、维特比算法。大家有兴趣可以课后调研一下。

实际上维特比算法就和最短路径的寻找、动态规划的算法。

知道现有那些中文分词工具：jieba、北大中文分词、深度学习中文分词等等。

回顾文本的编码方式：什么是独热编码，怎么把文本转化为独热编码。什么是绝对词频，如何转化为绝对词频。掌握TF、IDF、TF-IDF的概念。

要了解分布式表示：把文本向量化来得到各文本的表示。

??自然语言处理的发展，文本分析的主要工具。