《认知过程的信息处理》复习思考题

简答题 5 选 3，计算题 4 选 2，论述题 2 选 1

一、填空题

1、认知系统需要足够多的数据来发掘其中所蕴含的模式和独特的价值，充足的数据对于保障认知系统的分析结果的可靠性和一致性至关重要。在研究认知数据之前，我们需要了解大数据的五个基础特性——5V，它们分别是：**数据容量 volume 超大**、**数据类型多样性 variety**、**数据处理速度 velocity 快**、**数据的真实性 veracity** 和**数据值 values 的变化性**。根据是否有定义好的格式和语义，一般将数据分为**结构化数据**和**非结构化数据**。P3

2、在机器学习过程中，用户必须了解输入数据类型以及模型在创建过程中所扮演的角色。根据学习方式分类，机器学习算法可以分为**监督学习**、**无监督学习**和**半监督学习**。K-均值聚类算法属于**无监督学习**，在聚类方法中，K-均值聚类一般只能发现球状簇，而**基于密度的聚类**方法能够构造出任意形状的簇。P55 P82

3、为了提高模型的效果，我们需要增强输入数据的数量，可以在数据发现、数据收集、数据准备预处理阶段完成，从而使数据更加完整、相关和规整。常用的数据预处理手段包括：**处理缺失数据**、**处理不正确的数据**、**规则化数据**和**可视化支持**。P92

4、深度学习是模仿人类感知的一种特征提取和学习算法，深度学习实现的步骤可以分为**定义神经网络架构**、**确定学习目标**、**开始学习**。人工神经网络是模仿人脑结构和功能的一种抽象的数学模型，使用**后向传播算法**学习网络参数，是深度学习的基础。P125 P144

5、卷积神经网络是目前最常见的深度学习的架构，由多层组成，通常包括**输入层**、**卷积层**、**pooling 层**、**全连接层**和**输出层**。卷积神经网络应用广泛，在计算机视觉领域表现突出。在处理序列数据时，更合适的深度学习方法是**递归神经网络**。P145

6、认知计算代表一种全新的计算模式，它包括信息分析、自然语言处理和机器学习领域的大量技术创新，能够助力决策者从大量非结构化数据中揭示非凡的洞察。认识计算的三个核心分别是：**理解**、**推理**和**学习**。P219

7、认知计算为了处理庞大数量的异构数据，必须有高效率的处理工具作为支撑。使用并行式编程模型能够有效提高系统的吞吐量。考虑到分布式计算系统包含一组网络节点，一个典型的并行程序运行包括的系统问题包括**分区**、**映射**、**同步**、**通信**和**调度**。P172

8、DeepQA（深度问答）是 IBM Waston 的核心组件，它是根据**非结构化信息管理架构(UIMA)**标准建立的，核心设计准则有：**大规模并行处理**、**概率问题和内容分析的整合**、**可行度评估**和**浅层和深层知识的整合**。P275

9、深度增强学习过程主要是由**增强学习**和**深度学习**组合而成，其中**增强学习**是用于解决序列决策问题的算法，通过在一个工作环境下连续选择一系列的行为，使这些行为完成后得到最大的收益。AlphaGO 是使用深度增强学习的成功案例，AlphaGO 程序有两个深度神经网络，分别是**策略网络**和**估值网络**，并使用**蒙特卡罗树搜索**。P255 P263

10、在自然语言处理中，词法分析包括**词形分析**和**词汇分析**两部分，**词形分析**是对单词前缀和后缀等的分析，**词汇分析**是对整个词汇系统的分析。P42

11、RFID 技术被用于手机个人信息，并将它们存储在附着在用户身上的廉价芯片中。典型的 RFID 系统包括 **RFID 标签**，**RFID 读取器**，和**后台数据库**等组成部分。P16

12、主成分分析法、奇异值分析法都属于**降维算法**；决策树、SVM 都属于**分类算法**；K-means、DBSCAN 都属于**聚类算法**。P86 P60 P76

13、云提供了一系列共享计算资源，包括应用、计算服务、存储容量、网络、软件发展、不同的部署形式和业务处理过程。**公共云**通常提供一个共享的多租户环境，其中多个用户在一个服务器内共享一个物理容器。**私有云**的数据与资源保存在数据中心中，而这些资源通常不与其他公司共享。**混合云**提供集成或连接到公有云、私有云和管理服务的功能。P164

14、将感知器模拟成人类的神经系统，那么**输入节点**就相当于神经元，**输出节点**相当于决策神经元，**权重系数**相当于神经元之间链接的强弱程度。人类的大脑可以不断刺激神经元，进而学习未知的知识，同样感知器模型通过**激活函数 f(x)**来模拟人类大脑的刺激。P127

15、Apache Hadoop 是分布式存储和分布式处理的开源软件库，其主要由 **Hadoop Common Package**、**MapReduce 引擎**和 **Hadoop 分布式文件系统**组成。与 Hadoop 相比，Spark 的运算速度一般较**快**。P181

16、MapReduce 软件框架提供了数据控制流的抽象。数据流的步骤是**数据分区**、**映射**和**调度**、**同步**、**通信**以及结果的输出。P175

17、1997 年 5 月 IBM 的计算机程序**深蓝**在正常时限的比赛中首次击败了等级分排名世界第一的象棋棋手加里卡斯帕罗夫；2016 年 7 月，DeepMind 开发的 **AlphaGo** 在围棋世界排名中超过柯洁，成为世界第一。其中，后者的程序核心是**深度增强学习(DRL)**算法。P255

18、从功能定义的角度出发，物联网的体系架构可以分为 8 层，自底向上分别为协作层、**应用层**、服务层、抽象层、**存储层**、处理层、**网络通信层**、物理层。P18

19、智能移动设备（手机、可穿戴设备、平板电脑）在集成了越来越多的传感器后，拥有了强大的计算和感知能力，在用户群中形成大范围且密集的移动感知网络。这种感知网络通过移动互联网进行协作，实现感知任务分发与感知数据收集。这种感知称为**群智感知**。其通常可分为**机会感知**和**参与式感知**。P24

20、认知活动包括**思维**、**语言**、**定向**和**意识**等四个部分。认知过程指人们获得知识或应用知识的过程，或信息**加工**的过程，这是人的最基本的心理过程。

认知的定义：

1、认知是人们推断和判断客观事物的心理过程，是在过去的经验及有关线索进行分析的基础上形成的对信息的理解、分类、归纳、演绎和计算。

2、认知活动包括思维、语言、定向和意识等四个部分。

3、认知反应个体的思维能力，是制定和执行护理计划的依据。

21、结构化数据：有定义好的**长度**与**格式**，它的元数据、视图和词汇语义是**明确**定义的。P4

二、简答题

1、主成分分析也称主分量分析，旨在利用降维的思想，把多指标转化为少数几个综合指标（即主成分），其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复。请指出主成分与原始变量之间的关系。P86

(1)主成分保留了原始变量绝大多数信息
(2)主成分的个数大大少于原始变量的数目
(3)各主成分间互不相关
(4)每个主成分都是原始变量的线性组合

2、半监督学习是无监督学习和有监督学习的混合。在半监督学习中，实验使用的数据集是不完整的，一部分数据有标签而另一部分数据没有标签。事实上，半监督的学习更接近人类的学习方式。为了尽可能地利用无标签的数据，我们必须给出数据分布的假设。请指出常用的数据分布假设。P91

(1)平滑性假设：越接近彼此的样本数据点越有可能来自同一个类标签。
(2)聚类假设：数据往往形成离散的集群，分在同一个集群的点更有可能共享同一个标签。
(3)流行假设：流行化后的数据空间往往比输入空间的维度低。

3、请解释何为欠拟合现象，为何出现欠拟合现象，如何避免欠拟合现象。P93

定义：一个利用训练数据集训练出来的模型，测试训练数据集时出现很大的偏差。
原因：(1)数据集很少，训练过程和检验过程不能很好的执行
(2)没有选择合适的机器学习算法

避免：(1)改变模型参数（或改用其他模型）
(2)修正损失函数
(3)组合方法或其他修正

(1)使用主成分分析（PCA）算法，将原始数据的多数信息用维度远低于原始数据维度的几个主成分表示出来，大大降低了数据之间的内在联系。
(2)使用组合算法，利用多个同样的模型组合成一个模型来提高准确率，缓解欠拟合现象。

4、请解释何为过拟合现象，为何出现过拟合现象，如何避免过拟合现象。P93

定义：一个模型在训练数据集上能够获得比其他模型更好的效果，但是在训练数据集外的数据集上却不能得到很好的结果。

原因：模型刻意地去记住训练样本的分布状况

避免：(1)增加训练数据（增大样本量，减少噪声影响）
1、使得训练集的分布更加具备普适性、减少偶然性使得噪声的均值趋于零，减少了噪声对数据整体的影响。
2、在不增加样本量的情况下，利用去噪算法修改和完善原始训练样本集，减少噪声的方差，减少噪声对数据整体的影响，如用小波分析法去除噪声。

(2)特征筛选和降维
1、特征与特征之间的联系会影响训练模型，通过对样本特征的分析，发现各特征之间的内在联系，减少那些代表性较差的特征，突出代表性特征。如利用关联分析和相关分析。
2、降低模型的复杂度，而更不容易刻划到噪声数据的分布。如多项式拟合模型中降低多项式次数、神经网络中减少神经网络的层数和每层的节点数、SVM 中增加 RBF-kernel 的带宽。

(3)数据归一化
L1 正则化：让特征获得的权重稀疏化，对结果影响不大的特征，不给其赋予权重。
L2 正则化：尽量打散权重到每个维度上，不让权重集中在某些维度上，出现权重特别高的特征。

5、简述两种有监督学习方法，并概括其原理、应用、优缺点。P57

决策树（课本 P60-63）、基于规则的分类（课本 P63-65）、最近邻分类（课本 P65-67）、支持向量机（SVM）（课本 P67-69）、朴素贝叶斯（课本 P69-72）、随机森林（课本 P72-76）

(1)随机森林

原理：（课本 P72-76）
随机森林是一类专门为决策树分类器设计的组合方法。它组合多棵决策树做出预测,其中每棵树都是基于随机向量的一个独立集合的值产生的。

应用：（网络）

随机森林算法是集成学习中具有代表性的一个算法，它简单高效、应用广泛，在金融学、医学、生物学等众多应用领域均取得了很好的成绩。从市场营销到医疗保健保险，既可以用来做市场营销模拟的建模，统计客户来源，保留和流失，也可用来预测疾病的风险和病患者的易感性。随机森林算法至今在许多方面被优化，学术届引入预处理和新算法，提出了旋转森林算法、霍夫森林算法、拟自适应分类随机森林算法等，在学术届均有广泛的应用。

优点：（网络、第九题）

1、准确率极高

2、能够有效地在大数据集上运行

3、引入了随机性，不容易过拟合

4、有很好的抗噪声能力

5、能处理很高维度的数据，而且不用降维

6、不仅能处理离散型数据，还能处理连续型数据，而且不需要将数据集规范化

7、训练速度快，能够得到变量重要性排序

8、容易实现并行化

9、即使对于缺省值问题也能够获得很好得结果

10、超参数的数量不是很多，并且都能很直观地了解它们多代表的含义

缺点：（网络、第九题）

1、在数据噪音比较大的情况下会过拟合

2、决策树个数很多时，训练时所需要的空间和时间会很大

3、不适用于实时性要求很高的场景

(2)支持向量机

原理：（课本 P67-69）

支持向量机 SVM 是一种对线性和非线性数据进行分类的方法。支持向量是指那些在间隔区边缘的训练样本点，在空间中可以用直线或（超）平面分隔空间内的点，将不同区域的点分为一类。对于在低维空间不可分的点，可以通过核函数将其映射到高维空间，利用高维空间的超平面分隔这些点，从而使分类简单化。

应用：（网络）

SVM 在各领域的模式识别问题中有应用，包括人像识别、文本分类、手写字符识别、生物信息学等。例如由台湾大学开发的 LIBSVM 是使用最广的 SVM 工具。

优点：（网络）

1、计算的复杂性取决于支持向量的数目，而不是样本空间的维数，避免了“维数灾难”。

2、对异常值不敏感，具有较好的“鲁棒性”

3、SVM 学习问题可以表示为凸优化问题，因此可以利用已知的有效算法发现目标函数的全局最小值，

4、有优秀的泛化能力。

缺点：（网络）

1、对大规模训练样本难以实施

2、解决多分类问题困难

3、对参数和核函数选择敏感

6、简述两种无监督学习方法，并概括其原理、应用、优缺点。P57

聚类（课本 P76-84）（K-均值聚类、凝聚层次聚类、基于密度的聚类）、降维（课本 P85-89）（主成分分析法）

降维的两种方法参考（练习题 16）

(1)K 均值聚类

原理：（网络）

K 均值聚类算法是一种迭代求解的聚类分析算法。给定一个数据点集合和需要的聚类数目 K，K 由用户指定，K 均值算法根据某个距离函数反复把数据分入 K 个聚类中。

应用：（网络）

K 均值聚类算法是一个很好的分类算法，在用户分类、图像分类甚至目标识别、目标检测、网络入侵中都有很好的应用。学术界中常常使用该算法结合其他算法训练人工智能模型。

优点：（网络）

1、原理很简单，收敛速度非常快，可解释性好

2、只有 K 一个参数在调参的时候需要被调整

3、对处理大数据集，该算法保持可伸缩性和高效性

4、当簇接近高斯分布时，它的效果较好

缺点：（网络）

1、对于离群点和噪音点非常敏感

2、K 值的选择非常难以估计

3、初始聚类中心对训练结果的影响非常大

4、一般只能发现球状簇，不能发现环状等任意形状的簇

7、请分析有监督机器学习、无监督机器学习、半监督机器学习的异同。P55

有监督：给定训练样本，每个样本的输入 x 都对应一个确定的结果 y，需要训练出一个模型，在位置的样本 x^’给定后，能对结果 y^’做出预测。

无监督：样本没有给出标签或标准答案，需要在样本中抽取出通用的规则。

半监督：给出的数据有一部分有标签，有一部分没有标签。需要在探寻数据组织结构的同时，也能做相应的预测。

异：

1、输入的数据的标签情况不同：有监督学习输入的是一对{输入，输出}数据集。无监督学习的数据集是没有标签的。半监督数据集一部分有标签，一部分没有（大部分情况是没有标签的更多）。

2、学习目标不同：有监督学习的主要目标是构建一个合适的模型，该模型对于新的输入，能够给出正确的输出。无监督学习不给出数据的类标签，只是探索数据的内在联系和潜在关系。半监督学习可以是用有标签数据来训练网络，用无标签得到的深度特征增强有标签的效果；也可以是用无标签训练网络，用有标签来增强效果。

同：都属于机器学习、都需要利用大量的数据来训练出一个模型。

8、什么是结构化数据？非结构化数据、半结构化数据及结构化数据有什么不同？P4

结构化数据有定义好的长度和格式，而且它们的元数据、视图和词汇语义是明确定义的。与结构化数据不同，非结构化数据和半结构化数据没有特定的格式，而且语义也没有明确的定义。语义必须通过自然语言处理、文本分析和机器学习的技术来进行发掘和提取。

9、随机森林是一类专门为决策树分类器设计的组合方法，是组合分类方法的一种。比较随机森林相对于决策树分类器的优缺点。

随机森林是一类专门为决策树分类器设计的组合方法。它组合多棵决策树做出预测,其中每棵树都是基于随机向量的一个独立集合的值产生的。

优点：

降低异常值所带来的影响

降低了过拟合的可能性

缺点：

计算量相对于决策树很大，性能开销很大。

可能会导致有些数据集没有训练到，但这种几率很小。

10、深度学习针对传统的机器学习或者浅层学习相比有哪些优势和不足？

深度学习定义（课本 P121）：深度学习使用多层神经网络结构进行数据的特征提取和学习。

对比浅层学习（课本 X、P121-122）：

1、对比浅层的神经网络，深度神经网络能为更复杂的非线性关系建模。

2、对于简单的模式识别问题，浅层学习的分类工具已经足够。当模式变得非常复杂时，需要深度学习来实现。

3、深度学习的输入将经过更多层的转换。

对比其他机器学习（课本 P122）： 深度学习与其他机器学习方法最大的不同是

具有特征学习能力，可以理解为深度模型是手段,特征学习是目的。深度学习区别于传统的机器学习表现在以下四个方面。
1、强调了 ANN 模型结构的深度，与通常的浅层学习相比，深度学习使用更多隐藏层。
2、突出特征学习的重要性。通过逐层特征变换,将数据在原始空间的特征表示变换到一个新特征空间,使分类或预测变得容易而且精确度得到提高。
3、深度学习来源于人工神经网络的发展,但是训练的方式与传统的人工神经网络不同,采用逐层训练的方式,然后再对网络参数进行微调。
4、深度学习利用大量数据来学习特征，而浅层学习不需要使用。

11、在人工神经网络中，学习率反应了感知器模型学习的速度，一般取值在[0,1]中，以便控制循环过程中的调整量。试回答在学习率取值不同时，对应感知器模型学习会出现的情况。并简述学习率过大或过小会导致的问题。P128

λ→0 时,新的权值主要受旧的权值影响，学习速率慢，但是更容易找到合适的权值；λ→1 时，新的权值主要受当前的调整量影响，学习速率快，但是可能出现跳过最佳权值的现象。因此,在某些情况下，往往前几次循环让 λ 值大一些，在后面的循环里逐渐减小。

12、神经网络中，激活函数的作用是给网络中加入一些非线性元素，使得网络的拟合效果更好，从而达到更好的识别率或者准确率。那么神经网络中的激活函数有哪些可以选择呢？至少列举出三个，并画出其函数曲线。P128

Sigmoid 函数 双曲正切函数 ReLU 函数：f(x)=max{x,0}

Sigmoid 函数可以将实数映射到(0,1)区间内。平滑、易于求导。

tanh 激活函数是 0 均值的，tanh 激活函数相比 sigmoid 函数更陡峭，对于有差异的特征区分得更开了。

ReLU 函数优点：1.计算量小 2.激活函数导数维持在 1，可以有效缓解梯度消失和梯度爆炸问题 3.使用 ReLU 会使部分神经元为 0,这样就造成了网络的稀疏性，并且减少了参数之间的相互依赖关系，缓解了过拟合问题的发生。

13、每个算法都有各自的适用范围，虽然很难准确评判哪一种算法更优秀，但是我们可以利用一些常见的指标来了解算法。请列举几个常见评价算法的性能指标。选择合适的机器学习算法需考虑哪些性能指标？P92

(1)准确度：反映算法在测试数据集上的表现，即是否出现过拟合或欠拟合现象。显然，对训练集的测试效果越理想，算法越优秀，这是最重要的一个指标。
(2)训练时间：反映算法收敛的速度以及建立一个模型所需要的时间。显然，训练时间越短，算法越理想。

(3)线性度：反映算法的复杂度，是算法设置本身的要求，尽可能使用低复杂度的算法求解问题。

14、词向量的 one-hot representation 和 distributed representation 有什么区别，请分别阐述并举例说明。P45

One-hot Representation 方法建立的词向量中向量的维度是总词数,每个词对应的向量只有一个维度值为 1,剩下的所有维度值为 0。优点是非常简单直接,建立方便;缺点是任意两个词之间都是孤立的，即使两个词具有相同的语义,其向量之间也没有任何关系,存在“词汇鸿沟”现象。且因为词向量词表维数等于总词数,所以在某些任务中会因为计算负担过大带来维数灾难。

Distributed Representation 方法建立的词向量词表中,每一个词用实数向量表示，类似(0.792,-0.177,..)。向量的维度通常使用 50 维或 100 维,远远小于总词数,避免了维度灾难问题的出现。优点是两个词的语义越近,两个词的向量在向量空间中的距离越近，避免了“词汇鸿沟”的问题。但需要使用大量真实的文本语料进行训练和学习，在学习词向量的过程中确定需要的向量维度。

15、请简述梯度下降法的作用和算法流程。P133

神经网络由连接各层的权重表示模型，机器是使用梯度下降法或随机梯度下降法来拟合参数（各层的连接权重）的。梯度下降法是一种迭代方法，用于找到函数的最小值。训练神经网络时根据给定的一组网络参数，利用梯度下降法计算预测损失或者分类损失，然后调整这些参数以减少损失函数值。

算法流程：1、用随机值初始化权重和偏差
2、把输入传入神经网络，得到输出值
3、计算预测值和真实值之间的误差
4、对每一个产生误差的神经元，调整相应的（权重）值以减小误差
5、重复迭代，直至得到网络权重的最佳值

16、请简述降维算法的作用，并选其中两种降维算法描述其算法流程，并比较其优劣。P85

降维算法是将高维的点通过映射函数转换到低维空间，以此来缓解“维数灾难”，降维不仅可以减少数据之间的相关性，而且由于数据量减少加快了算法的运行速度。

主成分分析法利用降维的思想，把多指标（回归中称为变量）转化为少数几个综合指标（即主成分），其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复，它既能大大减少参与数据建模的变量个数，同时也不会造成信息的大量丢失。

(1)计算初始样本均值和方差执行标准化
(2)计算标准化数据得到标准化矩阵
(3)利用标准化矩阵计算相关性矩阵 C
(4)计算相关性矩阵的特征值和特征向量
(5)根据相关性矩阵的特征值计算主成分的贡献率与累计贡献率
(6)根据用户指定的贡献率确定主成分的个数得到评价矩阵的主成分

局部线性嵌入算法(LLE)是一个非线性降维方法，它能够使降维后的数据保持原有拓扑结构不变。它的主要思想是利用数据的局部线性来逼近全局线性：即假设任意样本点都可表示为其临近样本点的线性组合，在寻找数据的低维嵌入同时，保持这种邻域线性组合关系不变。

(1)寻找每个样本点的 k 个近邻点
(2)由每个样本点的近邻点计算出该样本点的局部重建权制矩阵 W
(3)由该样本点的局部重建权制矩阵 W 和其近邻点计算出该样本点的输出值

比较优劣：PCA 算法是线性算法，依赖于投影，会破坏原本高维空间样本之间的某些线性关系。LLE 是非线性算法，可以学习任意维的局部线性的低维流形，因此保留了高维空间的局部线性关系。

17、云交付模型中，基础设施即服务、软件即服务、平台即服务分别是什么？P167

基础设施即服务(IaaS)是基础性的云服务，IaaS 通过一个虚拟的图像或直接在计算机系统中定义计算、存储和网络服务，这被称为本地实施，典型的 IaaS 模型依赖于虚拟化。

软件即服务(SaaS)是一个定义的应用程序，允许用户在公共云服务中操作。

平台即服务(PaaS)是一个完整的基础设施包，用于设计、实施和部署在任何公共或私有云的应用和服务。

18、请简述深度学习的实现步骤。P126

1、定义神经网络学习架构（定义函数）
2、确定学习目标（确定什么样的函数最好）
3、开始学习（寻找最优函数）

19、请简述大数据的特点，并简单介绍两种数据采集和预处理方法。P3 P5

特点：数据容量超大、数据类型多样性、数据处理速度快、数据的真实性和数据值的变化性。

日志文件(CAs)是一种广泛使用的数据采集方法，日志文件是由数据源系统自动生成的记录文件，以记录指定的文件格式的活动，以供后续分析。

传感器在日常生活中经常被用于测量物理量，而物理量会被转换成可读的数字

信号用于后续的处理（和存储）

数据清理：根据决定的策略清理和预处理，数据能按要求处理丢失的域和更改数据。数据清理是一个识别不准确、不完整或不合理的数据，然后修改或删除这些数据以提高数据质量的过程。

数据集成：数据集成是现代商业信息学的基石，它涉及不同来源的数据的组合，为用户提供了一个统一的数据视图。

20、什么是强化学习？请简述其原理。

强化学习是用于解决序列决策问题的算法，通过在一个工作环境中连续选择一系列的行为，使这些行为完成后得到最大的收益。在事先不了解任何规则的情况下，一个智能体先观察当前的环境状态，并尝试一些行动，以改善收益。奖励是智能体调整其行动策略的反馈。通过不断调整，算法能够学习到在什么样的情况下选择什么样的行为可以得到最好的结果。（课本 P255）

如果 Agent 的某个行为策略导致环境正的奖赏(强化信号)，那么 Agent 以后产生这个行为策略的趋势便会加强。Agent 的目标是在每个离散状态发现最优策略以使期望的折扣奖赏和最大。

强化学习把学习看作试探评价过程，Agent 选择一个动作用于环境，环境接受该动作后状态发生变化，同时产生一个强化信号(奖或惩)反馈给 Agent，Agent 根据强化信号和环境当前状态再选择下一个动作，选择的原则是使受到正强化(奖)的概率增大。选择的动作不仅影响立即强化值，而且影响环境下一时刻的状态及最终的强化值。（网络）

21、简述“吴方法”实现初等几何定理机器证明的核心思想。

将几何问题代数化，即用多项式来表达几何问题的条件以及结论，通过证明条件所组成的多项式交集的零点是结论对应多项式的子集来完成证明。

主要步骤如下：

1.将条件和结论用代数多项式表达

2.确定自由变元和约束变元，对约束变元排序，确定消元的次序

3.将条件所对应的多项式三角化，设三角化后的多项式为 F1,F2,,,,Fk,..Fn，保证 Fk 只包含前面 k 个约束变元

4.将结论所对应的多项式从三角形底部到顶部依次消除最后的约束变元，如果最后所得到的剩余多项式为 0 则表明命题为真，否则命题为假。

22、AlphaGo 策略网络和估值网络的作用分别是什么？两者有何区别？P263

策略网络的输入是棋局状态，预测每一个合法的下一步的概率，选择概率较大的

位置落子。目标是获得在围棋盘面下的落子棋感，从而将计算力分配到最有希望的选点，节省计算力。

价值网络的结构与策略网络的相似，但是输出单个预测结果而不是概率分布，通过随机梯度下降法最小化预测值和实际值的均方根来训练网络。目标是获得胜负棋感，从而利用直觉来进行对每一个落子点给一个当时的快速的胜负估算。

策略网络关注的是每一步怎么下，是局部的，而估值网络侧重于对全局形势的判断。

23、请简述 Google Duplex 的三个主要模块，并分别说明其功能。

Google Duplex 三大主要模块为自动语言识别系统(ASR)、循环神经网络(RNN)、文本到语音系统(TTS)。

自动语言识别系统：该模块功能为将听到的声音转换为文本信息。

循环神经网络：该模块功能为理解输入的文本信息，并产生对话内容(回答)。

文本到语音系统：将文本信息转换为语音，并决定语音在语调、语气以及一些语言习惯上的特征，使之更加自然。

三、分析与计算题

1、给定下图所示神经网络特征图，（a）池化步长为 2，计算其最大值池化和平均值池化结果；（b）池化步长为 3，计算其最大值池化和平均值池化结果。P148

1	2	3	3	4	4
5	6	6	8	5	7
3	5	3	2	9	6
6	8	5	6	5	8
3	6	9	2	11	3
8	5	10	7	6	6

步长为 2：

最大池化			平均池化		
6	8	7	3.5	5	5
8	6	9	5.5	4	7
8	10	11	5.5	7	6.5

步长为 3：

最大池化	平均池化

6	10	34/9	16/3
9	11	20/3	6

2、下表是 10 位病人是否患有糖尿病体检数据集，请利用规则分类的方式构建其规则集。并验证新 id“血糖含量高，体重偏旁，血脂含量=1.23”是否患有糖尿病。

id	血糖	体重	血脂含量	是否患病
1	高	正常	1.21	否
2	高	正常	1.18	否
3	低	偏胖	2.46	否
4	高	偏胖	3.02	是
5	高	偏胖	2.46	是
6	低	正常	1.98	否
7	低	偏胖	2.58	否
8	高	偏胖	1.03	否
9	高	偏胖	8.59	是
10	高	正常	0.65	否

P65

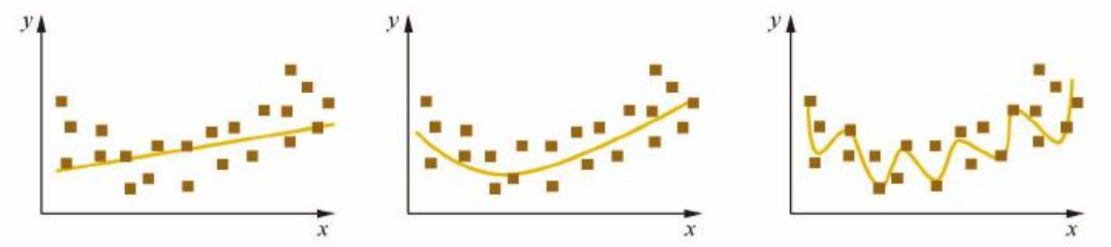
3、请构建一个包含 10 个单词的单词集合，并写出每个单词的 one-hot representation。

P45

4、(1) 请分别指出图中(a)(b)(c)分别对应哪种拟合。

(2) 为什么会出现欠拟合和过拟合的情况？

(3) 有哪些方法可以解决欠拟合和过拟合？



(a) (b) (c)

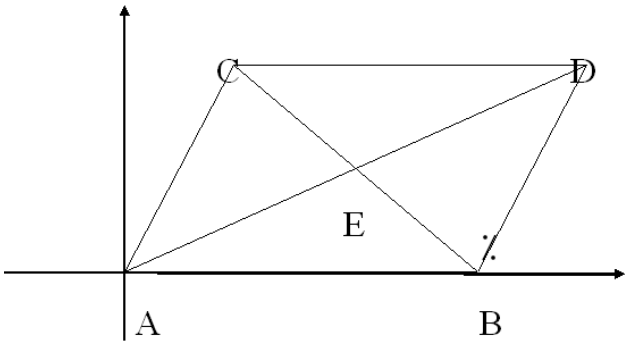
(1)(a)欠拟合(b)正确拟合(c)过拟合

(2)(3)P93

5、“The math exam will take place on next Monday. The English exam will be held this Friday.”用 MapReduce 方法对上述句子进行单词计数，假设一个数据块可以存储 6 个单词，请简述整个算法流程。P178

Map Reduce 库按照 6 个单词为一块将输入数据文件分为 3 块：The math exam will take place、on next Monday The English exam、will be held this Friday，每个 Map 机器会读入各自的输入数据，Map 函数接受输入数据块并产生中间数值对，中间数值键为单词本身，中间数值键为 1，共产生了 17 个中间数值对：(The,1) (math,1) (exam,1) (will,1) (take,1) (place,1) (on,1) (next,1) (Monday,1) (The,1) (English,1) (exam,1) (will,1) (be,1) (held,1) (this,1) (Friday,1)。接着，MapReduce 库收集所有产生的中间数值对，将它们排序并分组，每个键唯一：(English,1) (Friday,1) (Monday,1) (The,1,1) (be,1) (exam,1,1) (held,1) (math,1) (next,1) (on,1) (place,1) (take,1) (this,1) (will,1,1)。分组完成后会将组并行地发送至 Reduce 函数，Reduce 函数可以将“1”值相加并且得出实际上每个句子中每个单词所出现的次数，最终得到(English,1) (Friday,1) (Monday,1) (The,2) (be,1) (exam,2) (held,1) (math,1) (next,1) (on,1) (place,1) (take,1) (this,1) (will,2)，即每个单词的计数。

6、假设 A,B,C,D,E 的坐标为(0,0)， (x_1,0),C(x_2, y_2),D(x_3, y_3),E(x_4, y_4)。AB 平行 CD， AC 平行 BD。证明 AE=DE



试将上述几何问题的条件与结论转化为代数描述。

由条件 AB//CD 有 $y_2=y_3$
由条件 AC//BD 有 $(x_3-x_1)y_2=x_2y_3$
由条件 AE//DE 有 $x_4y_3=x_3y_4$
由条件 CB//EB 有 $y_2x_4+y_4x_1=y_2x_1+y_4x_2$
求证目标 $4(x_4^2+y_4^2)=x_3^2+y_3^2$

7、已知 $F_1=x_4y_3-x_3y_4=0$， $F_2=y_2x_4+y_4x_1-y_2x_1-y_4x_2=0$

求 F_1 关于 F_2 以及变元 y_4 的带余除法，并将其写成关于 x_4 的标准一元多项式。

由 $F_2 = 0$ 可以得到：

$$y_4 = \frac{y_2x_1 - y_2x_4}{x_1 - x_2}$$

因此：

$$Prem(F_1, F_2, y_4) = -x_3(y_2x_1 - y_2x_4) + x_4y_3(x_1 - x_2)$$

标准化后可得如下关于 x_4 的一元多项式：

$$(x_1y_3 - x_2y_3 + x_3y_2)x_4 - x_3y_2x_1 = 0$$

8、给定函数 $f(x):[0, 1]\rightarrow R$，如果使用简单遗传算法求 $f(x)$ 精度不低于 0.001 的解，那么个体二进制位串编码长度至少应多长？为什么？

10 位。因为串长取决于求解的精度，如果确定求解精度到 0.001，即三位小数，由于区间长度为 1，必须将区间[0, 1]分为 1×10^3 等份。因为 $512=2^9<10^3<2^{10}=1024$，所以编码的二进制串长至少需要 10 位。

9、问题 7：已知 $F_1=(u_2)^2+(x_1)^2-(u_1)^2$，$F_2=x_2-u_1-u_2$，$F_3=x_2(u_2-u_1)+(x_1)^2$。

依次求解：F_2 关于 F_3 以及变元 x_2 的带余除法，记其余项为 F_4；F_1 关于 F_4 以及变元 x_1 的带余除法，记其余项为 F_5。上述过程的非退化条件是什么？

(1)令 $F_3=0$ 得 $x_2=x_1^2/(u_1-u_2)$

$F_4=Prem(F_2, F_3, x_2)=x_1^2/(u_1-u_2)-u_1-u_2=0$

即 $x_1^2-u_1^2+u_2^2=0$

(2)令 $F_4=0$ 得 $x_1^2=u_2^2-u_1^2$

$F5= Prem(F_1, F_4, x_1)=u_2^2+u_2^2-u_1^2-u_1^2=0$

即 $u_2^2-u_1^2=0$

(3)非退化条件：$u_1-u_2\neq0$ 即 $u_1\neq u_2$

解答：$F4=F2*(u_2-u_1)-F3=(u_1-u_2)(u_1+u_2)-x_1^2$

$F5=F4+F1=(u_1)^2-(u_2)^2+(u_2)^2-(u_1)^2=0$

没有非退化条件

10、已知 F1= (u1-u3)x2-x5(x1-x3)=0，F2= x1x4-x3x6=0，F3= x5-x6=0。求 F2 关于 F3 以及变元 x6 的带余除法的余式（记为 F4），F1 关于 F4 以及变元 x5 的带余除法的余式（记为 F5）。上述过程的非退化条件是什么？

$F_4=F_2-x_3*F_3=x_1x_4-x_3x_5$，记 $F_6=F_1+F_4=(u_1-u_3)x_2-x_5x_1+x_1x_4$，则 $F_5=F_6*x_3-F_4*x_1=(u_1-u_3)x_2x_3+x_1x_3x_4-(x_1)^2x_4$。非退化条件为 x_1 不等于 0，x_3 不等于 0。

四、论述题

1、请结合自身经验有监督学习和无监督学习的优劣点，并简述其再实际场景中的应用。

P57-84

2、请选取一种计算机视觉应用场景并对其涉及的技术进行分析论述。

光学字符识别（OCR）是指对文本资料的图像文件进行分析识别处理，获取文字及版面信息的过程。亦即将图像中的文字进行识别，并以文本的形式返回。目前已在文档识别、证照识别、票据识别、车牌识别等领域广泛使用。主要用到了 CNN(卷积神经网络)来提取特征以及 RNN 来对序列进行分析相关性，然后还用 CTC 作为损失函数来解决对齐问题。

传统 OCR 基于数字图像处理和传统机器学习等方法对图像进行处理和特征提取，现在普遍使用基于 CNN 的神经网络作为特征提取手段。感受野是用来表示网络内部不同神经元对图像的感受范围，也就是在 CNN 中表示原图的区域大小。CNN 关注局部像素的相关性比较强，而较远像素的相关性则比较弱，所以神经元的感受野越大，说明它能感受到全图的范围就越大，越小则说明它越关注局部和细节。随着层数的增多，深度神经网络可以提取比较复杂的图像特征。得益于 CNN 强大的学习能力，配合大量的数据可以增强特征提取的鲁棒性，面临模糊、扭曲、畸变、复杂背景和光线不清等图像问题均可以表现良好的鲁棒性。

文本识别在传统技术中采用模板匹配的方式进行分类，通过识别每个单字符以实现全文的识别，这一过程导致了上下文信息的丢失。为了引入上下文的信息，使用了 RNN 和 LSTM 等依赖于时序关系的神经网络。RNN 也叫循环神经网络，它以序列数据为输入，然后在输出的方向上不断地递归往回循环，可以挖掘其中的时序信息以及语义信息。但 RNN 会忘记它在较长序列中看到的内容，因此只有短期记忆。为了解决 RNN 在训练过程中梯度消失的问题，引入了 LSTM(长短期记忆)。LSTM 具有输入门，遗忘门，输出门，以及隐藏的记忆细胞，可以决定哪一些数据、重要的特征需要保存下来，而哪一些数据需要丢弃，这样就可以将相关的信息传递到较远的神经元结点中。

OCR 常见的做法为利用 CRNN 模型，以 CNN 特征作为输入，双向 LSTM 进行序列处理使得文字识别的效率大幅提升，也提升了模型的泛化能力。在实际情况中，输出并不一定能与字符一一对应,标记对齐样本非常困难，为了解决这一问题，CRNN 引入了 CTC （连接主义时间分类）。CTC 本质上使所有路径的概率和最大，根据梯度修改 LSTM 中的权重，对字符序列删除连续重复的字符，然后删除所有的空白字符。因此先由分类方法得到特征图，之后通过转录层(CTC)对结果进行翻译得到输出结果。

3、请选取一种自然语言处理技术进行具体分析论述。

语义极性分析和观点抽取是指利用计算机技术自动分析带有观点信息的句子或文档，从而提取出用户感兴趣的主题或特征，并分析其语义极性倾向(褒义、贬义或中性)和强度，广泛用于互联网上消费者的评价分类与整合。运用自然语

言处理技术分析词语的上下文，可以计算计算词语的上下文极性。句法分析还可以分析主题词和极性成分的匹配关系，从而判断句子中每个主题的极性倾向。

Kim 和 Hovy 对观点的定义为：观点是个四元组[Topic, Holder, Claim, Sentiment]。其中，观点持有者(holder)相信有关某个主题(topic)的声明(claim)，在很多情况下，这种信念包含着一定的情感因素。观点抽取的任务就是自动地从自然语言中找出观点中的四元组。由此可见，观点抽取可以分为四个任务：

1、主题和特征抽取：这个子任务主要是识别出观点中的特征词，特征词可以是主题也可以是观点持有者，并且找到主题词之间的关系，包括从属关系和并列关系。有时候还需要指代消解。

2、观点分析：区别主观描述和客观描述，提取出观点描述部分。

3、极性分析：这一步的任务，不仅是要分析观点的极性倾向。还要计算其极性强弱。

4、观点归纳：归纳最终观点和对应的极性。

采用句法分析的方法对句子进行分析，就需要确保句子的结构完整，而且词语没有歧义或者指代不明。因此，需要将句子还原到上下文环境中，进行指代消解和省略恢复的预处理。主题是句子评论的对象，如果一个句子中会同时出现多个或多类特征词，就需要明确它们之间的从属关系。利用 Ontology 定义一个层次式分类体系，以树的结构表示。通过遍历树，利用父子节点的关系，就可以知道特征词之间的从属关系。

对于一个带有观点的句子，仅仅知道观点的极性褒贬是不够的，用户更希望知道观点的持有者是谁，观点讨论的主题或特征是什么。因此，需要对句子进行句法分析。通过对依存关系的分析，发现 SBV 结构(主谓结构)可以提供主语和谓语的修饰关系等信息。WebFountain 的情感模式库主要包含的就是谓语含有极性的模式，根据这一模式，可以把谓语的极性传递给宾语中的特征词。

这种算法在句子比较规范的条件下，可以识别通过依存关系对，找到句子中谓语的极性，然后再传递给主语，从而实现语义极性分析和观点抽取。

4、请简述机器人的支撑技术，分析现有机器人，并展望机器人的未来发展。

P213

5、请简述 IBM Watson 系统的起源与发展历程，并展望其未来发展。

P270

6、请简述 Google AlphaGo 的特点，并展望其未来发展。

AlphaGo 首先使用深度学习训练策略网络，然后利用之前深度学习的结果经过自我对弈训练出增强策略网络，接着在之前的基础上训练出增强价值网络，训练的网络用在蒙特卡洛树上，如此层层递进，实现了准确的预测。

首先，AlphaGo 使用监督学习从当前已有的棋局数据中进行学习，构建基于

深度卷积网络的策略网络。策略网络的输入是棋局状态，预测每一个合法的下一步的概率，选择概率较大的位置落子。虽然策略网络的精度高，但是速度比较慢，在实际对弈中往往有对速度的要求，因此，又训练了一个速度快、精度低的快速走棋网络。因为专家的落子也有可能出现失误，因此通过自我对弈使用增强学习进一步加强训练。

策略网络、快速走棋网络和增强策略网络关注的是每一步怎么下，是局部的，然而围棋是个需要大局观的游戏，常常会有“牵一发而动全身”的情况，因此使用估值网络，它侧重于对全局形势的判断。价值网络的结构与策略网络的相似，但是输出单个预测结果而不是概率分布，通过随机梯度下降法最小化预测值和实际值的均方根来训练网络。

当进行线上对决时，需要采用蒙特卡洛搜索树将策略网络和估值网络结合起来进行选择动作，蒙特卡洛树搜索是一种用于某些决策过程的启发式搜索算法，每个循环包括选择、扩展、模拟、反向传播四个步骤，用以评估赢棋概率选择比较好的落子方案。

2017 年，AlphaGo Zero 抛弃人类经验，从空白状态学起，在无任何人类输入的条件下，经过 3 天的训练便以 100：0 的战绩击败了 AlphaGo Lee，经过 40 天的训练便击败了 AlphaGo Master；2020 年，DeepMind 再推新模型 MuZero，该模型可以在不知道游戏规则的情况下下自学规则，不仅在更灵活、更多变化的 Atari 游戏上代表了 AI 的最强水平，同时在围棋、国际象棋、日本将棋领域也保持了相应的优势地位；2021 年，AlphaFold2 预测了人类 98.5%的蛋白质结构。其创建了一个基于注意力的神经网络系统，经过端到端训练，试图解释图的结构，同时对所构建的隐式图进行推理。并使用进化相关序列，多序列比对和氨基酸残基对表示来完善此图。DeepMind 公司这些接连的成就与突破都让我们感受到了当前人工智能发展速度之快涉猎范围之广，且曾经预测的机器无法超越人类的领域并不是无法解决。未来人工智能所要解决的则是算法的通用性，即构建通用人工智能，并应用到我们的生活中。