### 第一章 机器学习概念

1、机器学习的本质是 **对函数的拟合**

2、假定数据已经预处理好，机器学习训练阶段的三个基本步骤是**： 建立模型、确定损失函数、计算最优函数。**

3、回归问题和分类问题的区别？

**分类问题的输出维度有限（3分），而回归问题输出维度原则上无限（3分）**

1. 简述神经网络中梯度下降方法的原理和作用。

**梯度下降方法是利用数值优化中对待优化函数求导能够获取函数变化最快方向的信息从而逐步更新函数参数的方法。（4分）在神经网络中，参数是待定的，要找到最恰当拟合输入输出数据的函数，就需要构建损失函数，（4分）损失函数中有神经网络的参数作为待优化参数，（4分）随着损失函数相对于神经网络参数的梯度逐步减小直至近似为0，就能够找到最优的目标函数。（4分）**

5、机器学习中模型的训练误差和测试误差是一致的。 **错** X

6、机器学习中选择的模型越复杂越好。 **错** X

7、模型在训练阶段效果比较好，但在测试阶段效果不好的现象称之为\_**过度拟合**。

8、解决模型过拟合问题的一个方法是正则化。 **√**

9、模型过拟合的本质原因是什么？为什么正则化能够解决过拟合问题？

**过拟合的本质是模型过于复杂（3分），过好的拟合了训练数据，却对测试数据泛化能力较差（4分）。正则化本质上是通过对过拟合的模型做约束（3分），使得其拟合训练数据的能力没有那么强，从而提升其泛化能力。（4分）**

10、机器学习中，模型误差的两个主要来源是\_\_**偏差**\_\_\_和\_\_**方差**\_\_\_。

11、模型在训练阶段的效果不太好称之为欠拟合。  **√**

12、随着模型的复杂度增加，模型的偏差会\_\_**减少**\_\_\_，方差会\_\_\_**增加**\_\_\_。

13、如果模型的误差来自于偏差较大，可以采用以下措施解决 **A B**

A．给数据增加更多的特征 B. 设计更复杂的模型

C. 增加更多的数据 D. 使用正则化

14、用解决回归问题的模型来解决分类问题可能存在的问题是\_\_**模型可能会惩罚那些过于正确的实例 回归模型输出的是连续型变量，分类问题中的较远距离的离散变量可能会影响分类问题的边界（4分)**

15、分类问题可以用\_\_**贝叶斯公式**\_\_\_\_\_进行刻画。

### 第二章：神经网络基本概念

1. 一个逻辑回归模型的可看做神经网络中的一个采用sigmoid激活函数的神经元。（ **√** ）
2. 隐藏层神经元的激活函数为sigmoid的神经网络可以看成由多个逻辑回归模型叠加而成。（ **√**  ）
3. 神经网络的待学习参数由所有神经元连接的权重和偏差组成。（ **√**  ）
4. 一个结构确定的神经网络对应一组函数集合，而该神经网络的参数确定后就只对应一个函数。（ **√** ）
5. 请定义全连接前馈神经网络。

一个这样的神经网络，两个网络层之间，n层的任意一个神经元和n+1层所有神经元都相连。并且整个网络中无反馈，信号从输入层向输出层单向传播

**神经网络一般由多层组成（5分），每层有多个神经元（5分），从输入到输出各层之间的神经元连接都是沿着从输入层到输出层的方向（5分），且两相邻层的神经元之间两两互相有连接（5分）。**

6、深度神经网络的深度一般是网络隐藏层的数量。（ **√** ）

7、神经网络从计算上可以看成矩阵运算和非线性运算的多次叠加而组成的复合函数，且网络叠加的层次可看成复合函数的嵌套深度。（ **√** ）

8、神经网络的层次和每层神经元的数量可以通过以下哪些方法确定？（ **BCD** ）

A、可随意设定 B、可人为进行设计

C、可通过进化算法学习出来 D、可通过强化算法学习出来

9、\_\_\_\_**GPU**\_\_\_\_的并行运算功能是深度学习兴起的一个重要原因。

10、深度学习兴起的标志性事件包括（ **ABC**  ）

A、Alpha Go在围棋上击败李世石 B、在图像识别领域的准确率超过人类

C、语音识别达到人类的水平 D、深蓝在国际象棋上击败卡斯帕罗夫

1. 数学上已经证明只有一个隐藏层的神经网络可以拟合任意函数，为什么还需要多个隐藏层的神经网络？

需要多个隐藏层的主要原因是效率问题，通过多个隐藏层我们能够更快的训练我们到模型达到极值点，如果采用的是一个隐藏层，那么可能就在一层有大量的神经元，这样我们模型的训练效率就会大量的降低。而且模型的参数也会大大增加，使得我们更难以达到我们的预期目标

**该数学结论只是说明可以拟合（5分），但没有说明拟合的效率（5分），采用多个隐藏层进行拟合的神经网络具有更好的计算效率（5分）**

12、上世纪80年代就已经提出的\_**多层感知机**\_模型已经与现在的深度神经网络模型本质上没有太大的区别。

### 第三章：梯度下降与模型训练

1. Vanilla梯度下降方法的核心思想是？

**在进行梯度下降的开始阶段，因为远离优化目标，可以采用较大的学习律来快速逼近优化目标，经过若干轮迭代后，由于已经离优化目标比较接近，可以采用较小的学习律，以防止由于步长过大而远离优化目标或在优化目标附近震荡。Vanilla梯度下降方法采用的步长是，其中表示迭代次数，表示初始的学习律，表示第次迭代的学习律。**

1. Adagrad梯度下降方法的核心思想是？

**最佳梯度步长不仅取决于一阶梯度而且也取决于二阶梯度（实际上取决于一阶梯度和二阶梯度的比值），因此用两者的比值来计算学习律能够较为有效的反映梯度的变化与损失函数的曲率之间的关系，实现动态的学习律调整，而且通过蒙特卡洛采样的原理用当前迭代步骤之前的一阶梯度的均方根来近似二阶梯度能否进一步减少计算复杂度。**

1. 随机梯度下降运行速度要比梯度下降慢。（ **×**）
2. 特征缩放（feature scaling）目的是使得不同的特征有相同的比例。（ **√** ）
3. 反向传播算法的提出是为了解决深度神经网络中梯度计算的效率问题。（**√** ）
4. 反向传播算法由\_\_\_**前向传递**\_和\_\_**反向传递**\_\_两个阶段构成。
5. 在反向传播算法中，反向传递过程可以看成一个没有激活函数的反向的神经网络。（ **√** ）
6. 在反向传播算法中，在前向传递过程中神经元j的输入信号相对于连接*l*层神经元i和*l*+1层神经元j的边的权重的梯度实际上等价于神经元i的\_\_\_**输出值**。
7. 在反向传播算法中，在反向传递过程中损失函数相对于第*l*层神经元i的输入信号的梯度可以通过从输出层开始通过\_\_**递归**\_\_\_\_\_的方式逐步计算出来。
8. 在深度学习中采用梯度下降方法往往达不到最优解，而是在以下哪些情况下停止了。 **ABCD**

A. 鞍点 B.局部极值点

C.高原（**plateau**）点 D.早停点

1. 随机梯度下降相对于梯度下降，损失函数会呈现较强的抖动。（ **√** ）

### 第六章 CNN

1. CNN图像识别设计所考虑的三个准则是（**A B C** ）
2. 图像中存在远小于整个图像大小的模式
3. 一个模式在不同图像中可能位于不同的位置
4. 改变图像的分辨率不会影响对图像中物体的识别
5. 不采用全连接可以减少网络参数
6. 由CNN构成的分类模型中一般包含如下网络层（**A B C D** ）

A、全连接层 B、拉直层 C、池化层 D、卷积层

1. CNN中卷积层所起的作用包括（ **A B C D** ）

A、识别图像中的子模式

B、识别图像中位于不同位置的模式

C、相比于全连接网络减少网络参数

D、获取图像的恒定表征

1. CNN中池化层所起的作用包括（**C D** ）

A、识别图像中的子模式

B、识别图像中位于不同位置的模式

C、降低图像采样率

D、获取图像的恒定表征

1. 假定CNN中某池化层的输入张量形状为（15，26，26）且该池化层的池化窗口大小为（3，3），在不考虑边缘补零的前提下，其输出张量的大小为\_(15,8,8)\_\_。 计算方法是什么呢？
2. 假定CNN中某卷积层有10个卷积核且每个卷积核的大小为（4，4），该卷积层的每个卷积核有\_\_**16\_**\_个参数（假定上一层的信道数目是1），若该卷积层的输出又接入另一个卷积层，且这个卷积层有15个卷积核且每个卷积核的大小为（2，2），则这个下游的卷积层的每个卷积核有\_\_**40**\_\_个参数。
3. 以下属于CNN可视化可解释性方法的是（**A B D** ）
4. 可视化卷积神经网络的中间输出
5. 可视化卷积神经网络的过滤器
6. 可视化网络权重
7. 可视化图像中类激活的热力图
8. 请概述可视化卷积神经网络的中间输出的基本思想。

**可视化中间激活，是指对于给定输入，展示网络中各个卷积层和池化层输出的特征图（层的输出通常被称为该层的激活 ，即激活函数的输出）。这让我们可以看到输入如何被分解为网络学到的不同过滤器。我们希望在三个维度对特征图进行可视化：宽度、高度和深度（通道）。每个通道都对应相对独立的特征，所以将这些特征图可视化的正确方法是将每个通道的内容分别绘制成二维图像。**

1. 请概述可视化卷积神经网络的过滤器的基本思想。

**为了观察卷积神经网络学到的过滤器，可显示每个过滤器所响应的视觉模式。这可以通过在输入空间中进行梯度上升来实现：从空白输入图像开始，将梯度上升应用于卷积神经网络输入图像的值，其目的是让某个过滤器的响应最大化 。得到的输入图像是选定过滤器具有最大响应的图像。这个过程很简单：需要构建一个损失函数，其目的是让某个卷积层的某个过滤器的值最大化；然后使用随机梯度上升来调节输入图像的值，以便让这个激活值最大化。**

1. 请概述可视化图像中类激活的热力图的基本思想。

**类激活图 （class activation map ）可视化方法是指对输入图像生成类激活的热力图。所谓类激活热力图是与特定输出类别相关的二维分数网格，对任何输入图像的每个位置都要进行计算，它表示每个位置对该类别的重要程度。举例来说，对于输入到猫狗分类卷积神经网络的一张图像，类激活图可视化可以生成类别“猫”的热力图，表示图像的各个部分与“猫”的相似程度，类激活图可视化也会生成类别“狗”的热力图，表示图像的各个部分与“狗”的相似程度。**

**该方法有助于了解一张图像的哪一部分让卷积神经网络做出了最终的分类决策，从而有助于对卷积神经网络的决策过程进行调试，特别是出现分类错误的情况下。这种方法还可以定位图像中的特定目标。**

**类激活图方法具体思路如下：给定一张输入图像，对于一个卷积层的输出特征图，用类别相对于通道的梯度对这个特征图中的每个通道进行加权。直观上来看，理解这个技巧的一种方法是，你是用“每个通道对类别的重要程度”对“输入图像对不同通道的激活强度”的空间图进行加权，从而得到了“输入图像对类别的激活强度”的空间图。**

1. AlphaGo中用到了池化层。（ × ）
2. 在文本分析和语音识别中都可以用到卷积神经网络。（ **√** ）

### 第七章：WhyDeep

1. 深层神经网络和浅层神经网络从对函数的表示能力角度来看是差不多的。（ **√** ）
2. 对于同样的机器学习任务，浅层神经网络和深层神经网络所需要的数据量是一样的。（ **×** ）
3. 对于同样的机器学习任务，深层神经网络相对于浅层神经网络效率上更高。（ **√** ）
4. 对于同样参数规模的浅层神经网络和深层神经网络，其预测的准确率是差不多的。（ **×** ）
5. 相比于浅层神经网络，深层神经网络实际上利用了模块化的思想。（ **√** ）
6. 请概述一下为什么浅层神经网络中如果各个分类模块具有相关性，而其中某些分类模块具有数据量较少的问题，如何通过深层神经网络来解决这一问题。

可以利用编程中模块化的思想，把某些具有相关性的模块，采用分层的结构来搭建我们的模型。提取一些共同的特征分类靠近输入层，另外一些分类放到下一层，这样使得前面的分类的力度变大，相应的数据也会变多。而对于下一层的分类来说，要处理的分类已经经过前面一层处理，变得更加简单。相应需要的数据量也就变少了。这样就能解决这个问题。

**考虑靠这些分类模块具有相关性，我们可以把浅层网络中基础的分类模型分解成更为基础的分类模块，使得这些基础的分类模块每个都有相对较为充分的数据，从而能够训练出较好的基础分类模块，在此基础上，再叠加浅层网络中的分类模块，从而较好地训练处这些分类模块。**

1. 在图像识别和语音识别中，深层神经网络模型训练后都发现越是靠近输入层的神经元学出的特征越是\_\_\_**基础**\_\_。
2. 请概述普遍性定理（Universality Theorem）的主要思想及为什么该定理不能说明深层神经网络不如浅层网络。

普遍性定理表明只要有足够多的神经元，一层隐藏层的神经网络就可以拟合出任意的连续性函数。

为什么不能说明深层网络不如浅层网络的原因是，该定理只是在存在性上证明了可以用一层隐藏层来拟合任意连续函数，但是没有考虑到训练的效率问题，模型的效果。在相同的参数的情况下，深层神经网络的准确率要优于浅层神经网络，并且深层神经网络运用了模块化的编程思想，可以通过更少的数据，更快地速度，训练出我们的模型。

**普遍性定理的主要思想是指只有一个隐藏层的浅层神经网络可以拟合任意的函数。该定理只从数学上说明了存在性，并不意味着浅层神经网络的效率上会好**

1. 请解释一下什么是端到端学习并举例说明深度学习是如何实现端到端的学习的。

端到端的学习指的是只给模型输入和输出，让模型自己学习生产线的每个节点（函数）需要做什么事情。

深度学习通过叠一个很深的神经网络模型来实现端到端的学习，比如深度学习语音识别出文字，我们不再需要自己手动实现其中一个一个的函数，比如DFT等操作，只需要一个很深的神经网络，让它自己训练出每个函数应该做的事情。

**所谓端到端学习就是类似于生产线，将一个复杂的模型分解成多个简单函数，并按序列层次的方式叠加在一起。无论在语音识别还是在图像识别领域，都可以通过深度神经网络来进行训练，每一层的神经元所起到的作用都相当于一个简单的函数，多层神经元一起实现了复杂的模型。**

### 第八章 词嵌入

1. 词嵌入（Word Embedding）是一种无监督的学习方法。（**√** ）
2. 词嵌入不需要损失函数。（ **×** ）
3. 请解释一下自然语言中的独热编码（one-hot encoding）。

建立一个维度和所有字（单词）的数量一样的二进制向量，每一个字（单词）对应着一个向量，这个向量的某一个维度是1.

**所谓独热编码就是设置一个词汇表大小的二进制向量，向量中每一个维度对应词汇表中的一个词汇，这样每一个词汇所对应的独热编码就是只有该词汇所对应的维度为1而其它维度均为0的向量。**

1. 请解释一下词嵌入模型的工作原理。

有两种方法，第一种是基于计数的原理：

如果两个单词经常一起出现，那么他们的词向量的值应该很相近，也就是说两个词汇的词向量做内积应该和两个词汇在一篇相同的文章里面出现的次数成正比。

第二种是基于预测的原理：

将一段文本中前面n个词汇独热编码后的词向量输入到一个神经网络中，使其输出的各个维度中对应于文本中下一个词汇的那个维度的几率最大时，那么隐藏层对应的参数就是我们要的词向量。除此之外也可以使用中间的词汇去预测上下文，或者使用上下文预测中间的词汇来求词汇对于的词向量

**词嵌入模型利用一个词汇通常是通过其上下文词汇来理解的这一思想，通过从单词上下文预测单词或从单词预测其上下文的任务，基于一个浅层的没有激活函数的神经网络学出单词的隐藏层表达作为该词汇的嵌入表达。**

1. 请解释一下在词嵌入模型中为什么语义相近的词汇通常有相似的嵌入表示。

因为通常语义相近的词汇很大可能也有相同或者相似的上下文，也就是在神经网络中有相同或相似的输出，使得神经网络通过其拟合能力修改模型的参数使得这两个词汇有相似的隐藏层。也就是有相似的嵌入表示。

**语义相近的词汇通常有类似的上下文词汇，而词嵌入模型是通过词汇及其上下文的相互约束关系来获得词汇的嵌入表达的，这就使得语义相近的词汇能够通过其有交集的上下文词汇来学出相似的嵌入表达**

1. 通过词汇预测其上下文词汇的词嵌入模型Skip-gram输入的是词汇的\_\_\_**独热编码**\_\_\_\_\_，输出的是\_\_**上下文词汇出现的概率**。
2. 词嵌入模型中输出的即为词汇的嵌入表示。（ **×** ）
3. 通过上下文词汇预测词汇的词嵌入模型CBOW输入的是\_\_**上下文词汇的独热编码**\_，输出的是\_\_**词汇出现的概率**。
4. 以下关于词嵌入模型说法正确的是（ **A B C D** ）
5. 国家和首都直接的关系在词嵌入模型中可以保留
6. 英语词汇的时态关系在词嵌入模型中可以保留
7. 动物类别及其亚种之间的关系在词嵌入模型中可以保留
8. “国王-王后”与“叔叔-阿姨”之间的类别关系在词嵌入模型中可以保留
9. 以下关于词嵌入模型说法不正确的是（ **D**）

A.在多语言嵌入模型中，来自不同语言的具有相似语义信息的词汇可以得到类似的嵌入表达

B.在多领域嵌入模型中，来自不同领域（文本、图像、语音等）的具有相似语义信息的实体可以得到类似的嵌入表达

C.整个文档的嵌入表达可以由词汇的嵌入表达构成

D.词嵌入模型有考虑词汇在句子中的先后关系。

### 第九章：autoEncoder

1. **√**
2. **参考答案：输入、输出**
3. **X**
4. **参考答案：因为自编码器在做各特征维度的信息融合和压缩时保留了非线性关系，而PCA只保留了线性关系，所以自编码器在信息压缩过程中保留了更多的信息，从而在聚类效果上要更好。**
5. **参考答案：去噪自编码器通过人为给数据加入噪声，并通过对比原始未加入噪声的数据和自编码器输出的重构数据的差异来训练模型的抗噪声能力。**
6. **√**
7. **参考答案：自然界中同一类别的高维数据，往往集中在某个低维流形附近。**
8. **参考答案：根据数据科学的流形分布定律，自然数据背后隐藏着流形结构，深度学习方法可提取这些流形结构并用神经网络来表达流形间的映射，给出流形本身的参数化和参数表示。这些流形结构和其上的特定概率分布是整体先验知识的有效表示。正是因为具备这些先验知识，很多视觉和机器学习的问题能够被有效解决。流形能够表达一类数据的整体先验知识，传统方法只能利用局部较少的先验知识。**

### 第十章：RNN

1. 循环神经网络给神经网络提供了时序记忆能力。√
2. 在循环神经网络中，不同时刻相同的输入信号为什么有可能产生不同的输出？

不同时刻，储存在记忆单元里面的值可能不同，记忆单元里面的值会随同输入信号一起输入到循环神经网络中去。所以不同时刻循环神经网络的隐藏层的输出值可能会不一样，那么输出也可能会不一样。

**因为循环神经网络中的神经元有记忆单元能够记录先前的输入信号所蕴含的信息，且在不同时刻，记忆单元的内容也不尽相同，再加上记忆单元所记忆的信息在后续时刻会作为另一种形式的输入信号反馈网络，因此不同时刻网络所接收的输入信息实际上一般不太相同，所以会产生不同的输出。**

1. 在循环神经网络的不同版本中，Elman网络和Jordan网络的差别在于隐藏层输入信号的来源不尽相同。√
2. 循环神经网络在训练过程中为什么容易出现梯度消失或梯度爆炸问题？

**因为循环神经网络一般会迭代很多次，当网络的权重小于1时，在迭代很多次后网络的输出会接趋近于0，因此会导致梯度消失；而当网络的权重大于1时，在迭代很多次后网络的输出会趋近于无穷大，会导致梯度爆炸。**

1. LSTM网络为什么能够较好地解决RNN的梯度消失问题？

因为在LSTM网络中，只要输入能够影响记忆单元（memory），那么这个影响会一直存在于记忆单元中。而不是像rnn每次记忆单元的值被替换。除非遗忘门被关闭，导致记忆单元的值被重置，而大部分时间遗忘门是不会被关闭的。从而解决了梯度消失的问题。

**RNN在每一个时间点，记忆中的值都会被覆盖，而LSTM则是利用输入值以及原来记忆中的值得到新记忆的值。因此，RNN和LSTM处理记忆的方式是完全不同的，LSTM中的权重一旦可以影响到记忆中的值时，这种影响将是永久的，除非遗忘门关闭（在训练时可以给遗忘门较大的偏差，以保证遗忘门在大多数时候都是打开的状态），因此LSTM不会有梯度消失的问题。**

1. 以下属于循环神经网络的是？ **A B C D**
2. Elman网络 B.Jordan网络 C.LSTM D.GRU
3. 循环神经网络的潜在应用包括？**A B C D**

A．机器翻译 B. 聊天机器人 C.视频字幕生成 D.阅读理解

8、 GRU相比于LSTM要复杂一些。**×**

9、 循环神经网络训练时采用的优化算法是\_\_BPTT\_\_\_。