### 选择题

1. 梯度下降算法的正确步骤是什么？（D）

a.计算预测值和真实值之间的误差

b.重复迭代，直至得到网络权重的最佳值

c.把输入传入网络，得到输出值

d.用随机值初始化权重和偏差

e.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

A.abcde B.edcba C.cbaed **D.dcaeb**

1. 什么情况下神经网络模型被称为深度学习模型？(A)

**A.**加入更多层，使神经网络的深度增加 B.有维度更高的数据

C.当这是一个图形识别的问题时 D.以上都不正确

1. 下面哪项操作能实现跟神经网络中Dropout的类似效果？(B)

A.Boosting **B.Bagging**  C.Stacking D.Mapping

1. 下列哪一项在神经网络中引入了非线性？(B)

A.随机梯度下降 B.修正线性单元（**ReLU**） C.卷积函数 D.以上都不正确

1. 深度学习是当前很热门的机器学习算法，在深度学习中，涉及到大量的矩阵相乘，现在需要计算三个稠密矩阵 A, B, C 的乘积ABC,假设三个矩阵的尺寸分别为m∗n，n∗p，p∗q，且m < n < p < q，以下计算顺序效率最高的是（A）

**A、 (AB)C**  B、 AC(B) C、 A(BC) D、 所有效率都相同

1. 输入图片大小为200×200，依次经过一层卷积（kernel size 5×5，padding 1，stride 2），pooling（kernel size 3×3，padding 0，stride 1），又一层卷积（kernel size 3×3，padding 1，stride 1）之后，输出特征图大小为(C)

A、 95 B、 96 **C、 97** D、 98

1. 神经网络模型（Neural Network）因受人类大脑的启发而得名,神经网络由许多神经元（Neuron）组成，每个神经元接受一个输入，对输入进行处理后给出一个输出，如下图所示。请问下列关于神经元的描述中，哪一项是正确的？(E)

A、 每个神经元可以有一个输入和一个输出

B、 每个神经元可以有多个输入和一个输出

C、 每个神经元可以有一个输入和多个输出

D、 每个神经元可以有多个输入和多个输出

**E、 上述都正确**

1. 如果我们用了一个过大的学习速率会发生什么？(D)

A、神经网络会收敛 B、不好说 C、都不对 **D、**神经网络不会收敛

1. 在一个神经网络中，下面哪种方法可以用来处理过拟合？(D)

A、Dropout B、分批归一化(Batch Normalization) C、正则化(regularization) D、都可以

1. 批规范化(Batch Normalization)的好处（A）

**A、**让每一层的输入的范围都大致固定 B、它将权重的归一化平均值和标准差  
C、它是一种非常有效的反向传播(BP)方法 D、这些均不是

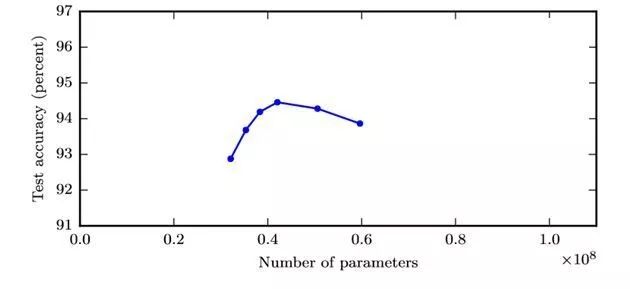
1. 下列哪个神经网络结构会发生权重共享？(D)

A、卷积神经网络 B、循环神经网络 C、全连接神经网络 **D、选项A和B**

1. 下列哪个函数不可以做激活函数？(D)

A、y = tanh(x) B、y = sin(x) C、y = max(x,0) **D、**y = 2x

1. 下图显示了训练过的3层卷积神经网络准确度，与参数数量(特征核的数量)的关系。(C)



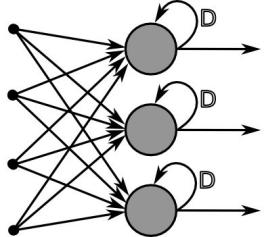
从图中趋势可见，如果增加神经网络的宽度，精确度会增加到一个特定阈值后，便开始降低。造成这一现象的可能原因是什么？

A、即使增加卷积核的数量，只有少部分的核会被用作预测  
B、当卷积核数量增加时，神经网络的预测能力会降低  
C、当卷积核数量增加时，导致过拟合  
D、以上都不正确

1. 假设你需要**调整超参数**来最小化代价函数（cost function），会使用下列哪项技术？(D)

A、穷举搜索 B、随机搜索 C、Bayesian优化 **D、**都可以

1. 构建一个神经网络，将前一层的输出和它自身作为输入。下列哪一种架构有**反馈连接**？(A)

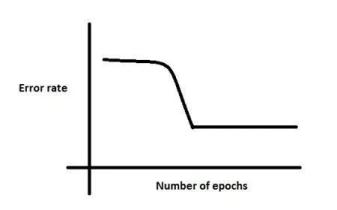


A、循环神经网络 B、卷积神经网络 C、限制玻尔兹曼机 D、都不是

1. 下列哪项关于模型能力（model capacity）的描述是正确的？（指神经网络模型能拟合复杂函数的能力）(A)

**A、**隐藏层层数增加，模型能力增加 B、Dropout的比例增加，模型能力增加  
C、学习率增加，模型能力增加 D、都不正确

1. 在训练神经网络时，损失函数(loss)在最初的几个epochs时没有下降，可能的原因是？(D)



A、学习率(learning rate)太低 B、正则参数太高 C、陷入局部最小值 **D、以上都有可能**

1. 下列哪一项属于特征学习算法（representation learning algorithm）？(C)

A、K近邻算法 B、随机森林 C、神经网络 D、都不属于

1. 假设我们拥有一个已完成训练的、用来解决车辆检测问题的深度神经网络模型，训练所用的数据集由汽车和卡车的照片构成，而训练目标是检测出每种车辆的名称（车辆共有10种类型）。现在想要使用这个模型来解决另外一个问题，问题数据集中仅包含一种车（福特野马）而目标变为**定位**车辆在照片中的位置。(B)

A、除去神经网络中的最后一层，冻结所有层然后重新训练

**B、**对神经网络中的最后几层进行微调，同时将最后一层（分类层）更改为回归层

C、使用新的数据集重新训练模型

D、所有答案均不对

1. 假设你有5个大小为7x7、边界值为0的卷积核，同时卷积神经网络第一层的深度为1。此时如果你向这一层传入一个维度为224x224x3的数据，那么神经网络下一层所接收到的数据维度是多少？(A)

**A、**218x218x5 B、217x217x8 C、217x217x3 D、220x220x5

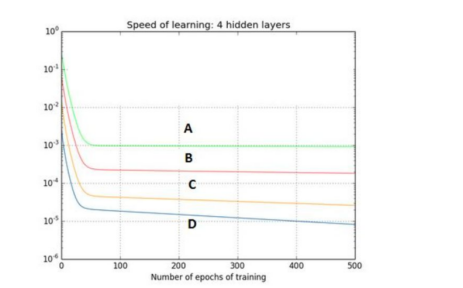
1. 假设我们有一个使用ReLU激活函数(ReLU activation function)的神经网络，假如我们把ReLU激活替换为线性激活，那么这个神经网络能够模拟出同或函数(XNOR function)吗？(D)

A、可以 B、不好说 C、不一定 D、不能

1. 下列的哪种方法可以用来降低深度学习模型的过拟合问题？(D)

1增加更多的数据 2使用数据扩增技术 3使用归纳性更好的架构 4正规化数据 5降低架构的复杂度

A、1 4 5 B、1 2 3 C、1 3 4 5 D、都有用

1. 下图是一个利用sigmoid函数作为激活函数的含四个隐藏层的神经网络训练的梯度下降图。这个神经网络遇到了**梯度消失**的问题。下面哪个叙述是正确的？(A)

**A、**第一隐藏层对应D，第二隐藏层对应C，第三隐藏层对应B，第四隐藏层对应A

B、第一隐藏层对应A，第二隐藏层对应C，第三隐藏层对应B，第四隐藏层对应D

C、第一隐藏层对应A，第二隐藏层对应B，第三隐藏层对应C，第四隐藏层对应D

D、第一隐藏层对应B，第二隐藏层对应D，第三隐藏层对应C，第四隐藏层对应A

1. 考虑某个具体问题时，你可能只有**少量数据**来解决这个问题。不过幸运的是你有一个类似问题已经预先训练好的神经网络。可以用下面哪种方法来利用这个预先训练好的网络？(C)

A、把除了最后一层外所有的层都冻结，重新训练最后一层

B、对新数据重新训练整个模型

**C、**只对最后几层进行调参 (fine tune)

D、对每一层模型进行评估，选择其中的少数来用

1. 在选择神经网络的**深度**时，下面哪些参数需要考虑？ (C)

1 神经网络的类型(如MLP,CNN)

2 输入数据

3 计算能力(硬件和软件能力决定)

4 学习速率

5 映射的输出函数

A、1,2,4,5 B、2,3,4,5 C、都需要考虑 D、1,3,4,5

1. 当数据过大以至于无法在RAM中同时处理时，哪种梯度下降方法更加有效？(A)

A、随机梯度下降法 B、不知道 C、整批梯度下降法 D、都不是

1. 基于二次准则函数的H-K算法较之于感知器算法的优点是(B)？

A、 计算量小 B、 可以判别问题是否线性可分 C、 其解完全适用于非线性可分的情况

1. 在一个神经网络中，知道每一个神经元的权重和偏差是最重要的一步。如果知道了神经元准确的权重和偏差，便可以近似任何函数，但怎么获知每个神经的权重和偏移呢？(B)

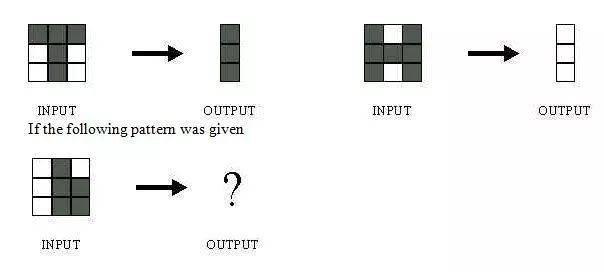
A、搜索每个可能的权重和偏差组合，直到得到最佳值

**B、**赋予一个初始值，然后检查跟最佳值的差值，不断迭代调整权重（梯度下降）

C、随机赋值，听天由命

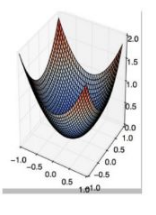
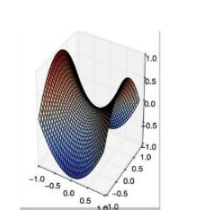
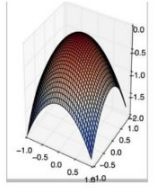
D、以上都不正确的

1. 下图所示的网络用于训练识别字符H和T，如下所示(D)



A、  B、 C、 **D、** 可能是A或B，取决于神经网络的权重设置

1. 在下面哪种情况下，一阶梯度下降不一定正确工作（可能会卡住）？(B)

A、 **B**、 C、

1. 混沌度(Perplexity)是一种常见的应用在使用深度学习处理NLP问题过程中的评估技术，关于混沌度，哪种说法是正确的？(B)

A、混沌度没什么影响 B、混沌度越低越好 C、混沌度越高越好 D、混沌度对于结果的影响不一定

1. 训练神经网络过程中，损失函数在一些时期（Epoch）不再减小, 原因可能是：(D)

1.学习率太低 2.正则参数太大 3.卡在了局部最小值

A. 1 and 2 B. 2 and 3 C. 1 and 3 D. 都是

1. 我们不是想要绝对零误差，而是设置一个称为贝叶斯（bayes）误差（我们希望实现的误差）的度量。使用贝叶斯（bayes）误差的原因是什么？(D)

A.输入变量可能不包含有关输出变量的完整信息 B.系统（创建输入-输出映射）可以是随机的

C.有限的训练数据 D.所有

答案：（D）想在现实中实现准确的预测，是一个神话，所以希望应该放在实现一个“可实现的结果”上。

1. 神经网络中的死神经元（dead unit）是什么？(A)

A.在训练任何其它相邻单元时，不会更新的单元 B.没有完全响应任何训练模式的单元

C.产生最大平方误差的单元 D.以上均不符合

1. 对于分类任务，我们不是将神经网络中的随机权重初始化，而是将所有权重设为零。下列哪项正确？(B)

A.没有任何问题，神经网络模型将正常训练

B.神经网络模型可以训练，但所有的神经元最终将识别同样的事情

C.神经网络模型不会进行训练，因为没有净梯度变化

D.这些均不会发生

1. 开始时有一个停滞期，这是因为神经网络在进入全局最小值之前陷入局部最小值。为了避免这种情况，下面的哪个策略有效？(C)

A.增加参数的数量，因为网络不会卡在局部最小值处

B.在开始时把学习率降低10倍，然后使用梯度下降加速算子(momentum)

C.抖动学习速率，即改变几个时期的学习速率

D.以上均不是

1. 假设在训练时，你遇到这个问题。在几次迭代后，错误突然增加。您确定数据一定有问题。您将数据描绘出来，找到了原始数据有点偏离，这可能是导致出现问题的地方。你将如何应对这个挑战？(B)

A.归一化 B. 应用PCA然后归一化 C.对数据进行对数变换 D.以上这些都不符合

答案：（B）首先将相关的数据去掉，然后将其置零。

1. 下面哪个描述可以最好的描述early stopping? (A)

A在每一个训练epoch后在测试集上模拟网络，当归一化误差(generalization error)减小时停止训练

B 一种更快的后向传播方式，像是"Quickprop"算法

C 在权重值更新时上加一个动量项(momentum term)，使训练能够更快地收敛

D 训练网络直到误差函数出现一个局部最小值

1. 我们建立一个5000个特征, 100万数据的机器学习模型. 我们怎么有效地应对这样的大数据训练？(ABC)

**A**. 我们随机抽取一些样本, 在这些少量样本之上训练

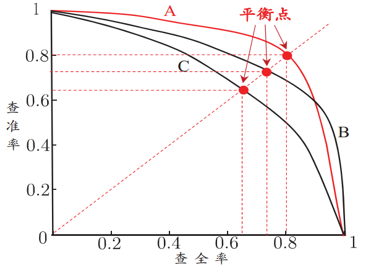
**B.** 我们可以试用在线机器学习算法

**C**. 我们应用PCA算法降维, 减少特征数

D. 可以采用SVM分类器

1. 假设一个二分类问题，现在采用一个大于0.5的阈值，那么关于模型的说法正确的是（AC）

**A.**召回率降低或不变 B.召回率升高 **C.**分类准确率升高或不变 D.分类准确率降低



1. 可以用于特征降维包括（）

A.PCA B.线性判断分析 C.稀疏自动编码器 D.SVD

1. 影响聚类算法效果的主要原因有（ABC）

**A.**特征选取  **B.**模式相似性测度 **C**.分类准则 D.已知类别的样本质量

1. 其他条件不变，以下哪种做法不会引起过拟合问题（AB）

A.增加训练集数量 B. 减少神经网络隐藏节点数

C.删除稀疏的特征 D. SVM算法使用高斯核函数/RBF代替线性核

1. 一阶优化方法的是（B）

A.牛顿法 **B**. **RMSProp** C. L-BFGS D.共轭梯度下降

1. 以下属于多层前馈神经网络的是（B）

A.线性回归 **B**.逻辑回归 C.SVM D.都不是

1. 机器学习中L1正则化和L2正则化的区别是？(AD)

**A**、使用L1可以得到稀疏的权值 B、使用L1可以得到平滑的权值  
C、使用L2可以得到稀疏的权值  **D**、使用L2可以得到平滑的权值

1. 机器学习中做**特征选择**时，可能用到的方法有？(ABCD)

A、卡方 B、信息增益 C、平均互信息 D、期望交叉熵

1. 我们想要减少数据集中的特征数, 即降维。选择以下适合的方案 :(D)

1. 使用前向特征选择方法 2. 使用后向特征排除方法

3. 我们先把所有特征都使用, 去训练一个模型, 得到测试集上的表现. 然后我们去掉一个特征, 再去训练, 用交叉验证看看测试集上的表现. 如果表现比原来还要好, 我们可以去除这个特征.

4.查看相关性表, 去除相关性最高的一些特征

A. 1 和 2 B. 2, 3和4 C. 1, 2和4 **D.** All

1. 数据科学家可能会同时使用多个算法（模型）进行预测， 并且最后把这些算法的结果集成起来进行最后的预测（集成学习），以下对集成学习说法正确的是 :(B)

A. 单个模型之间有高相关性

**B.** 单个模型之间有低相关性

C. 在集成学习中使用“平均权重”而不是“投票”会比较好

D. 单个模型都是用的一个算法

1. 以下哪个概念的**外延**最小（C）

A.机器学习 B.特征学习 C.深度学习 D.人工智能

1. 以下属于多层前馈神经网络的是（B）

A．线性回归 **B.逻辑回归** C.支持向量机 D都不是

1. 以下哪种方法属于判别式模型（D）

A．隐马尔可夫模型HMM B 朴素贝叶斯 C K近邻 **D**支持向量机

1. 哪种方法可以较好地处理RNN中梯度爆炸的问题（B）

A Dropout B 梯度裁剪 C. LSTM（梯度消失） D都不行

1. Dropout技术在哪种神经层中无法发挥优势作用？（AD）

A RNN层 B 卷积层 C仿射层 D输出层

1. 五层神经网络单个数据测试需要2s，新架构分别在第二层和第四层添加Dropout，则测试时间为（C）

A 大于2s B 小于2s C等于2s D不确定

1. CNN的缺点是（D）

A 共享卷积核，减少网络自由参数的个数，对高维数据的处理无压力

B 无需手动提取特征，训练好权重，既得特征

C 降低神经网络的复杂性

**D** 需要调参，需要大量样本

1. 在其他条件不变的情况下，以下哪种做法不会引起过拟合问题（AB）

A 增加训练样本 B减少隐藏层数量 C 删除稀疏的特征 D SVM高斯核/RBF代替线性核

1. 如果要设计一个有2000个特征，100万数据的机器学习模型，如何有效的处理这样大数据训练（ABC）

**A随机抽样，在小样本上训练 B 在线机器学习算法 C. PCA降维，减少特征数** D. SVM分类器

1. 哪种技术不能解决过拟合（D）

A dropout B 样本增强 C 权值衰减 D SGD

1. 一般建议将卷积生成对抗网络中生成部分的池化层替换为什么？(C)

A 跨距卷积层 B ReLU层 C局部跨距卷积层 D仿射层

1. 在CNN中使用1x1的卷积，下列哪项是正确的（D）

A可以帮助降低维数 B 可以用于特征池 C由于小的内核大小会减少过拟合 D上述所有

1. 深度学习可以用在下列哪些NLP任务中？(D)

A情感分析 B 问答系统 C机器翻译 **D所有选项**

1. 随着句子的长度越来越多，神经翻译机器将句意表征为固定维度向量的过程愈加困难，可采用（B）

A 使用递归单元代替循环单元 B使用注意力机制

C 使用字符级别翻译 D均不对

1. 在输出层不能使用以下哪种激活函数来分类图像？（C）

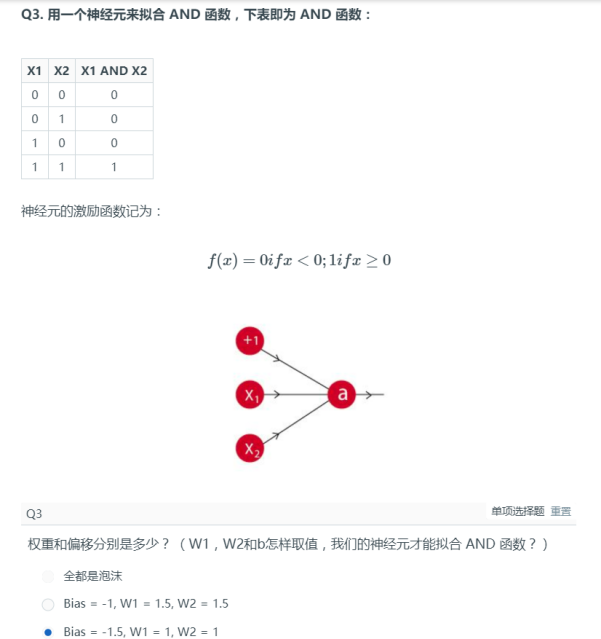
A sigmoid函数 B Tanh C ReLU D. if（x>5,1,0）

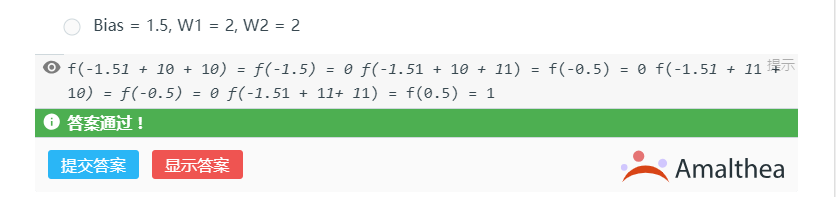
1. 多义现象可以被定义为在文本对象中一个单词或短语的多种含义共存，下列哪种方法最好解决此问题（B）

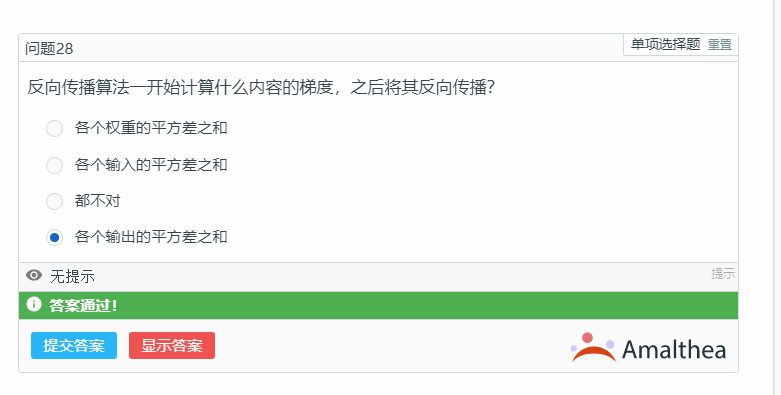
A 随机森林分类器 B 卷积神经网络 C梯度爆炸 D所有

66、在特定神经元给定任意输入，会得到输出-0.0001，可能是以下哪个激活函数（B）

A ReLU B tanh C sigmoid D都不是

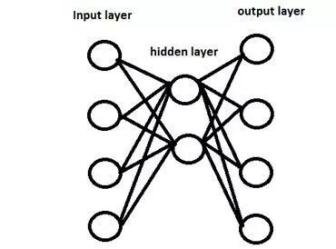






### 判断题

1. 假设我们有一个如下图所示的隐藏层。隐藏层在这个网络中起到了一定的降维作用。假如现在我们用另一种维度下降的方法，比如说主成分分析法(PCA)来替代这个隐藏层。两者的输出效果是一样的吗？×
2. 使用降维技术的网络与隐层网络总是有相同的输出吗？×
3. 主成分分析是一种有监督学习算法 ×
4. PCA 提取的是数据分布方差比较大的方向，隐藏层可以提取有预测能力的特征 √



1. K均值聚类算法是一种无监督学习算法 √
2. 深度学习与机器学习算法之间的区别在于，后者过程中无需进行特征提取工作，也就是说，我们建议在进行深度学习过程之前要首先完成特征提取的工作。×
3. 提升卷积核(convolutional kernel)的大小会显著提升卷积神经网络的性能。×
4. 当在卷积神经网络中加入RAM (pooling layer)时，变换的不变性会被保留。√
5. CNN采用最大池化能够产生一定程度上的平移不变性 √
6. 线性方程 (y = mx + c) 可以说是一种神经元吗？√
7. 在监督学习任务中，输出层中的神经元的数量应该与类的数量（其中类的数量大于2）匹配。×
8. Y = ax^2 + bx + c（二次多项式）这个方程可以用具有线性阈值的单个隐层的神经网络表示。×
9. 卷积神经网络可以对一个输入完成不同种类的变换（旋转或缩放）。×
10. 可以用神经网络对函数(y=1/x)建模。√
11. 增加卷积核的大小对于改进卷积神经网络的效果是必要。 ×（要考虑数据集）
12. 假设在ImageNet数据集（对象识别数据集）上训练卷积神经网络。然后给这个训练模型一个完全白色的图像作为输入。这个输入的输出概率对于所有类都是相等的。×
13. 训练CNN时，可以对输入进行旋转、平移、缩放（增强数据）等预处理提高模型泛化能力。 √
14. 决策树是一种面向分类问题的有监督学习算法 √
15. 通过增加模型复杂度，那么模型的测试错误率总是会降低 ×
16. BP算法仅适用于多层前馈神经网络 ×（RNN、CNN）
17. 前馈神经网络的隐藏单元必须在所有的输入点都可微 ×
18. 如果在一个高度非线性且复杂的一些变量中，一个树模型可能比一般回归模型效果更好√
19. 过拟合只在监督学习中出现，非监督学习没有过拟合 ×
20. 数据维数很高时，很多机器学习问题会变得相当困难，这种现象叫做维数灾难 √
21. 训练好的模型在测试集准确率100%，则在新的数据集上也会达到 100% ×
22. 欠拟合是模型不能在训练集上获得足够低的误差，过拟合是指训练误差和测试误差差距小×
23. 一般而言，L2正则化可产生比L1正则化更稀疏的解 ×
24. 感知机学习一定收敛 ×（条件样本线性可分）
25. 梯度为0的点不一定是局部极值 √
26. 深度学习是一种特殊的表示学习方法 √
27. 多层神经网络本质是是一个复合函数 √
28. 贝叶斯派统计方法比频率派统计方法估计模型参数更优？ √
29. RNN适用于序列数据的处理 √
30. LSTM网络结构有利于解决RNN训练出现的梯度消失或者爆炸的问题 √
31. GAN只能用于深度神经网络 √
32. 所有深度学习模型都是神经网络 √？
33. 支持向量机仅能处理二分类问题，产生线性平面 ×
34. 一般来说，L1正则化可产生比L2正则化更稀疏的解√
35. 梯度为0的点只能是局部极小点或局部极大点 ×
36. 如果增加多层感知机（Multilayer Perceptron）的隐藏层层数，分类误差便会减小。×
37. 神经网络可以生成任意的决策边界 √
38. 神经网络被称为通用函数拟合器，所以它理论上可以表示任何决策边界。√
39. 单计算层感知器只能解决线性可分问题 √
40. 一个不包含非线性的神经元可以看作是线性回归函数（Linear Regression Function） √
41. SVM算法中高斯核/RBF核代替限行和容易引起过拟合问题 √
42. 在数据很稀疏和神经网络参数量较少的场景下，偏向于使用L-BFGS而不是SGD √
43. 负对数似然度损失函数和欧式距离损失函数都是用于计算像素区域的修复 √
44. RNN可以被展开为一个完全连接的，具有无限长度的普通网络 √
45. 循环神经网络可以被认为是一个具有无限时间长度的神经元序列 √
46. 理论上神经网络可以解决任何问题，因为神经网络可以逼近任何函数 √
47. 神经网络进行图片语义分割时，先用CNN处理输入，再用反CNN得到输出 √
48. 在神经网络中ReLU永远不会饱和 ×
49. Dropout率（神经元的激活率）越高，正则化程度越高 ×
50. 与普通反向传播不同的是，随时间的反向传播算法会在每个时间步长内叠加所有对应权重的梯度√
51. 在神经网络中，每个参数可以有不同的学习率 √
52. 使用深度学习的情感分析是对多一的预测任务 √
53. Gated Recurrent Units的出现可以帮助防止在RNN中梯度消失的问题 √
54. 将Sigmoid激活函数改为ReLU，将有助于克服梯度消失的问题 √
55. 激活函数为有限值时，基于梯度的优化方法更加稳定 √
56. 多层网络全部初始化为0，会使梯度为0或各层相同，相当于浅层网络
57. **浅层学习：人工神经网络、SVM、Boosting、最大熵方法、LR，Logistic Regression**

### 分析计算题

1. 正则化在深度神经网络训练的作用，并说明L1和L2正则化特点
2. 激活函数的作用，sigmoid和ReLU的优缺点
3. Adagrad算法，如何选择合适的优化算法求解深度模型
4. 如何判断过拟合和欠拟合，解决方案及原理
5. 什么叫做迁移学习，有几种，分别对应的应用场景（目的及方法）
6. 简要说明Bagging 和 Boosting的区别
7. 给出梯度爆炸和梯度消失定义及解决方案
8. 在训练用于图像分类的深度CNN时

1 如何构建训练集、验证集和测试集

2 不同类别数量分布不均匀，会影响网络识别率吗，有影响的话如何处理

3 如何判断深度CNN过拟合，处理方法有哪些？

1. 卷积神经网络的层次结构及各部分的作用
2. 简要说明生成式模型和判别式模型的区别，有哪些？
3. 请给出“梯度下降”的定义，并简述和讨论任一种梯度下降算法
4. 比较最大似然估计ML和最大后验估计MAP的异同点和适用场景
5. 监督学习和无监督学习及哪些具体的算法
6. 梯度消失和梯度爆炸的解决方案
7. 什么叫做泛化能力，如何提高？
8. 什么叫做代价曲线？
9. 什么叫做偏差、方差和协方差？
10. AlexNet、ResNet、VggNet、DenseNet的结构及计算
11. 朴素贝叶斯分类器、决策树、K-近邻、主成分分析、随机森林算法
12. 前馈神经网络的定义和应用范围
13. 全连接网络的定义
14. 什么叫做LSTM网络，其优点是什么
15. 什么叫做GAN网络
16. 为什么对于小图像数据集一般要采用预训练网络
17. 对于小数据集一般采用什么网络？ SVM
18. 在自然语言处理中用到哪些卷积神经网络？
19. 评估模型通常用训练集和测试集，为何要加上验证集？python p77

评估模型通常用训练集和测试集，为何要加上验证集？python p77

因为测试集用来调整模型参数从而选择最优模型，模型本身已经同时知道了输入和输出，所以从中得出的误差（Error)会有偏差（Bias)。如果不需要调整超参数，则可以不用验证集。验证集仅用于对已经训练好的分类器进行性能评估的样本集。为拟合后的最终模型提供无偏评估。

什么叫做LSTM网络？其优点是什么？课件 10.10

优点：LSTM 被明确设计用来避免长期依赖性问题。长时间记住信息实际上是 LSTM 的默认行为，而不是需要努力学习的东西。具有长期记忆功能，主要其采用门的机制，能够一定程度上解决梯度爆炸，梯度消失问题，但其性能还是受其不能并行而有损失。

什么叫做GAN网络？GAN有两个神经网络 。

第一个叫**判别器**(**Discriminator**)，记做D(Y)。它得到输入Y(比如一张图)后输出一个 值，这个值表示了Y看起来是否"真实"。D(Y)可以看作某种能量函数，当Y是真实样本时，函数的值接近0, 反之，当图片Y的噪声很大或者很奇怪时，函数值为正。

另一个网络叫做**生成器**(Generator), 记为*G*(*Z*)。 这里的Z通常是从一个简单分布(例如高斯分布)随机抽样得到的向量，生成器*G*(*Z*)的作用是生成图片，这些 生成的图片会被用来训练判别器*D*(*Y*)(给真实图片较低的值，其他的图片较高的值)。

训练*D*的过程中,给它一张真实的图片，使其调整参数输出较低的值；再给它一张G生成的图片，让它调整参数 输出较大的值*D*(*G*(*Z*))。另一方面，在训练*G*的时候, 它会调整内部的参数使得它生成的图片越来越真实。也就是它一直在优化使得它产生的 图片能够骗过*D*, 想要让*D*认为它生成的图片是真实的。也就是说，对这些生成的图片，*G*想要最小化*D*的输出，而*D*想要最大化*D*的输出。所以这样的训练就叫做 **对抗训练**(adversarial training), 也叫做**GAN**。

什么叫做迁移学习，有几种，分别对应的应用场景

所谓迁移学习，一般就是要将从源领域（Source Domain）学习到的东西应用到目标领域（Target Domain）上去。源领域和目标领域之间往往有gap/domain discrepancy（源领域的数据和目标领域的数据遵循不同的分布）。迁移学习能够将适用于大数据的模型迁移到小数据上，实现个性化迁移。迁移什么，怎么迁移，什么时候能迁移，这是迁移学习要解决的主要问题。迁移学习就是一层层网络中每个节点的权重从一个训练好的网络迁移到一个全新的网络里，而不是从头开始，为每个特定任务训练一个神经网络。另一种迁移学习的方法是对整个网络进行微调

**样本迁移，特征迁移，模型迁移，关系迁移**

假设已有了一个可以高精确度分辨猫和狗的深度神经网络，之后想训练一个能够分别不同品种的狗的图片模型，需要做的不是从头训练那些用来分辨特征的神经网络的前几层，而是利用训练好的网络，提取初级特征，之后只训练最后几层神经元，让其可以分辨狗的品种。

另一种迁移学习的方法是对整个网络进行**微调**，假设已训练好了识别狗种类的神经网络，已有网络能对50多种狗进行分类。接下来想对网络进行升级，让其能够识别100种狗，不应该只训练网络的最后一层，而应该逐层对网络中每个节点的权重进行微调。如只训练最后几层，是迁移学习的1.0版，而对节点权重进行微调（fine turing），就是更难的2.0版，通过将其他层的权重固定，只训练一层这样的逐层训练，可以更好的完成上述任务。

另一种迁移学习的方式是借用**网络的结构**，即不是使用已训练好的网络的权重，而是使用相同的网络结构，例如多少层，每层多少节点这样的信息，然后使用随机生成的权重作为训练的起点。例如你要训练世界上第一个识别fMRI图像的神经网络，你就可以借鉴识别X光图片的神经网络。

为什么对于小图像数据集一般要采用预训练网络

预训练网络是一个保存好的之前已在大型数据集（大规模图像分类任务）上训练好的卷积神经网络。如果这个原始数据集足够大且足够通用，那么预训练网络学到的特征空间层次结构可以作为有效的提取视觉世界特征的模型。

在这种情况——**数据集小，数据相似度高**下，因为数据与预训练模型的训练数据相似度很高，因此我们不需要重新训练模型。我们只需要将输出层改制成符合问题情境下的结构就好。

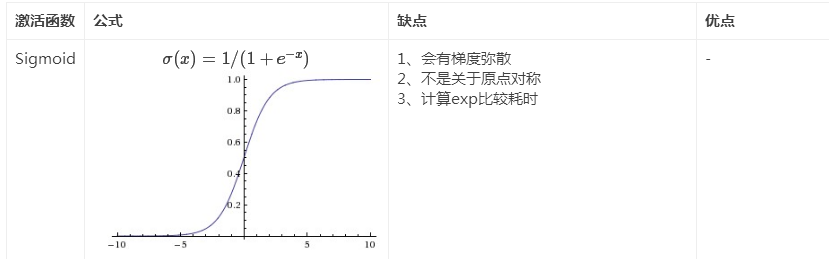
在这种情况——**数据集小，数据相似度不高**下，我们可以冻结预训练模型中的前k个层中的权重，然后重新训练后面的n-k个层，当然最后一层也需要根据相应的输出格式来进行修改。因为数据的相似度不高，重新训练的过程就变得非常关键。而新数据集大小的不足，则是通过冻结预训练模型的前k层进行弥补。

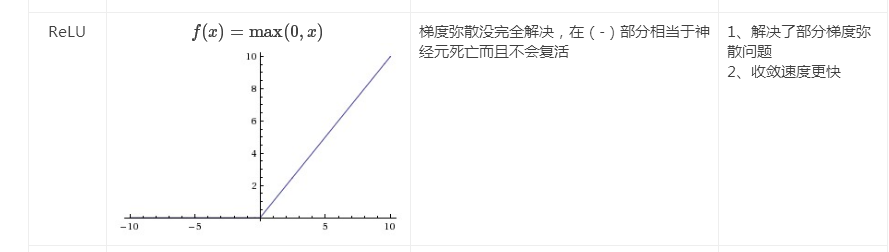
正则化在深度NN训练的作用，并说明L1和L2正则化特点

对学习算法的修改，以减少泛化误差而不是训练误差，提高泛化能力。具体为通过对目标函数添加一个参数范数来惩罚学习到的模型，限制模型的学习能力，避免过拟合，降低原始目标函数关于训练集数据的测试误差并同时减少在某些衡量标准下参数（或参数子集）的规模。L1正则化是在损失函数后面加上参数项的绝对值之和，使得模型在迭代的过程中，将大量参数置0，参数矩阵更加稀疏，而L2正则化是在损失函数后面加上参数矩阵平方和，使参数的值尽可能的小，从而更加平滑。

激活函数的作用，sigmoid和ReLU的优缺点

激活函数是用来**加入非线性因素**的，**提高神经网络对模型的表达能力，解决线性模型所不能解决的问题**。





简要说明Bagging 和 Boosting的区别

Boosting：个体学习器存在强依赖关系；串行生成；每次调整训练数据的样本分布

Bagging：个体学习器不存在强依赖关系；并行化生成；自助采样法

生成式模型：