机器学习深度学习保研常见面试问题：

1. 什么是[梯度爆炸](https://zhida.zhihu.com/search?content_id=203995965&content_type=Article&match_order=1&q=%E6%A2%AF%E5%BA%A6%E7%88%86%E7%82%B8&zd_token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJ6aGlkYV9zZXJ2ZXIiLCJleHAiOjE3NTUyNjM4NDgsInEiOiLmoq_luqbniIbngrgiLCJ6aGlkYV9zb3VyY2UiOiJlbnRpdHkiLCJjb250ZW50X2lkIjoyMDM5OTU5NjUsImNvbnRlbnRfdHlwZSI6IkFydGljbGUiLCJtYXRjaF9vcmRlciI6MSwiemRfdG9rZW4iOm51bGx9._zGxBM68BrKP3eVF1hZwPUkHqLKJR6jB6sJE_PVaJLg&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)和[梯度消失](https://zhida.zhihu.com/search?content_id=203995965&content_type=Article&match_order=1&q=%E6%A2%AF%E5%BA%A6%E6%B6%88%E5%A4%B1&zd_token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJ6aGlkYV9zZXJ2ZXIiLCJleHAiOjE3NTUyNjM4NDgsInEiOiLmoq_luqbmtojlpLEiLCJ6aGlkYV9zb3VyY2UiOiJlbnRpdHkiLCJjb250ZW50X2lkIjoyMDM5OTU5NjUsImNvbnRlbnRfdHlwZSI6IkFydGljbGUiLCJtYXRjaF9vcmRlciI6MSwiemRfdG9rZW4iOm51bGx9.vRJMHcOxeu4SuWxCQ8itpx0u7X666Z7sc6E8-xn_l2I&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)？如何解决梯度消失、梯度爆炸？

在[反向传播](https://zhida.zhihu.com/search?content_id=203995965&content_type=Article&match_order=1&q=%E5%8F%8D%E5%90%91%E4%BC%A0%E6%92%AD&zd_token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJ6aGlkYV9zZXJ2ZXIiLCJleHAiOjE3NTUyNjM4NDgsInEiOiLlj43lkJHkvKDmkq0iLCJ6aGlkYV9zb3VyY2UiOiJlbnRpdHkiLCJjb250ZW50X2lkIjoyMDM5OTU5NjUsImNvbnRlbnRfdHlwZSI6IkFydGljbGUiLCJtYXRjaF9vcmRlciI6MSwiemRfdG9rZW4iOm51bGx9.kRVvYb3C_-gpdvbwCiXp2FYFrxYmJF7AFnXZxafW4vg&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)过程中需要对激活函数进行求导，如果导数大于1，那么随着网络层数的增加梯度更新将会朝着指数爆炸的方式增加这就是梯度爆炸。同样如果导数小于1，那么随着网络层数的增加梯度更新信息会朝着指数衰减的方式减少这就是梯度消失。因此，梯度消失、爆炸，其根本原因在于反向传播训练法则，属于先天不足。

解决方法：

对于[RNN](https://zhida.zhihu.com/search?content_id=203995965&content_type=Article&match_order=1&q=RNN&zd_token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJ6aGlkYV9zZXJ2ZXIiLCJleHAiOjE3NTUyNjM4NDgsInEiOiJSTk4iLCJ6aGlkYV9zb3VyY2UiOiJlbnRpdHkiLCJjb250ZW50X2lkIjoyMDM5OTU5NjUsImNvbnRlbnRfdHlwZSI6IkFydGljbGUiLCJtYXRjaF9vcmRlciI6MSwiemRfdG9rZW4iOm51bGx9.ACT5R2NbW5enq8AqgdtKqYSXxJU3T9nZEnZyEQOLQ3I&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)，可以通过梯度截断，避免梯度爆炸。

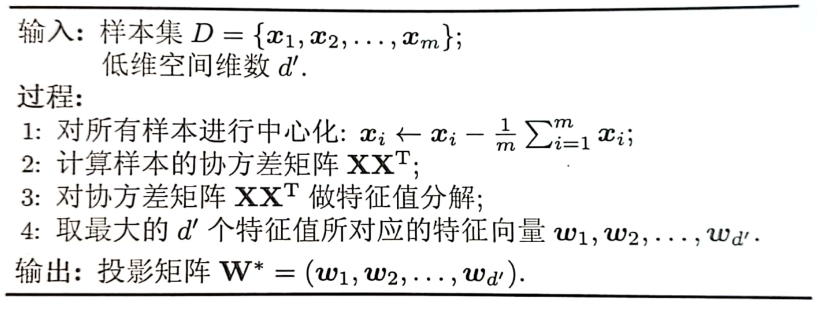
可以通过添加正则项，避免梯度爆炸。

使用[LSTM](https://zhida.zhihu.com/search?content_id=203995965&content_type=Article&match_order=1&q=LSTM&zd_token=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJ6aGlkYV9zZXJ2ZXIiLCJleHAiOjE3NTUyNjM4NDgsInEiOiJMU1RNIiwiemhpZGFfc291cmNlIjoiZW50aXR5IiwiY29udGVudF9pZCI6MjAzOTk1OTY1LCJjb250ZW50X3R5cGUiOiJBcnRpY2xlIiwibWF0Y2hfb3JkZXIiOjEsInpkX3Rva2VuIjpudWxsfQ.DuDw8I6hhjpOBP3SN8Yrqk5EdPmHNd9S2KTNZTqafko&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)等自循环和门控制机制，避免梯度消失。

优化激活函数，譬如将sigmold改为relu，避免梯度消失。

1. 介绍PCA算法

PCA (Principal Component Analysis)是最常用的线性降维方法，它的目标是通过某种线性投影，将高维的数据映射到低维的空间中表示，并期望在所投影的维度上数据的方差最大，以此使用较少的数据维度，同时保留住较多的原数据点的特性。



1. 介绍随机森林算法

随机森林其实算是一种集成算法。它首先随机选取不同的特征(feature)和训练样本(training sample)，生成大量的决策树，然后综合这些决策树的结果来进行最终的分类。

随机森林算法是最常用也是最强大的监督学习算法之一，它兼顾了解决回归问题和分类问题的能力。随机森林是通过集成学习的思想，将多棵决策树进行集成的算法。对于分类问题，其输出的类别是由个别树输出的众数所决定的。在回归问题中，把每一棵决策树的输出进行平均得到最终的回归结果。

集成学习：通过建立几个模型组合的来解决单一预测问题。它的工作原理是生成多个分类器/模型，各自独立地学习和做出预测。这些预测最后结合成单预测，因此优于任何一个单分类的做出预测。

1. 介绍k-means

事先确定常数k，k代表着聚类类别数。首先随机选取k个初始点为质心，并通过计算每一个样本与质心之间的相似度（可以采用欧式距离），将样本点归到最相似的类中，接着重新计算每个类的质心（该类中所有点的平均值），重复这样的过程直到质心不再改变，最终就确定了每个样本所属的类别以及每个类的质心。

优点：原理简单、容易实现。

缺点：收敛太慢、算法复杂度高、需先确定K的个数、结果不一定是全局最优，只能保证局部最优。

由于每次都要计算所有样本与每一个质心之间的相似度，故在大规模的数据集上，K-Means算法的收敛速度比较慢。

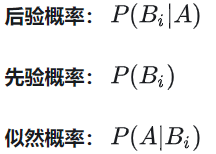
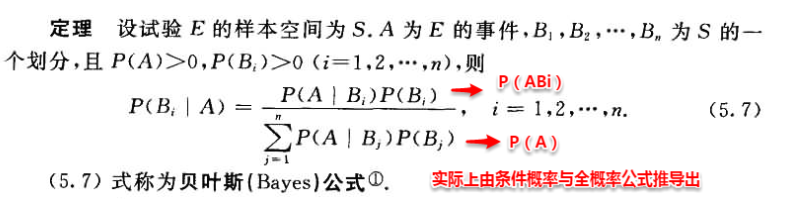
1. 介绍KNN学习（k近邻学习）

k近邻学习是一种常用的监督学习方法，其工作机制非常简单：给定测试样本，基于某种距离度量找出训练集中与其最靠近的k个训练样本，然后基于这k个“邻居”的信息来进行预测。通常，在分类任务中可使用“投票法”，即选择这k个样本中出现最多的类别标记作为预测结果；在回归任务中可使用“平均法”，即将这k个样本的实值输出标记的平均值作为预测结果；还可基于距离远近进行加权平均或加权投票，距离越近的样本权重越大。

1. 介绍朴素贝叶斯（基于概率论的分类方法）

朴素贝叶斯（naive Bayes）法是一种基于贝叶斯定理与特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入输出的联合概率分布；然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。

朴素：一个特征出现的概率与其他特征（条件）独立。



1. 介绍支持向量机

SVM的全称是Support Vector Machine，即支持向量机，主要用于解决模式识别领域中的数据分类问题，属于有监督学习算法的一种。

通俗来讲，SVM是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，其学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

SVM分类，就是找到一个超平面，让两个分类集合的支持向量或者所有的数据（LSSVM）离分类平面最远；SVR回归，就是找到一个回归平面，让一个集合的所有数据到该平面的距离最近。SVR是支持向量回归(support vector regression)的英文缩写，是支持向量机(SVM)的重要的应用分支

1. 关于深度学习

深度学习的优点在于它在很多问题上都变现出更好的性能，并且简化了问题的解决步骤，因为它将特征工程完全自动化。

利用深度学习，你可以一次性学习所有特征，而无须自己手动设计。这极大地简化了机器学习工作流程，通常将复杂的多阶段流程替换为一个简单的、端到端的深度学习模型。

1. 介绍SVD奇异值分解

[奇异值](https://link.zhihu.com/?target=https://baike.baidu.com/item/%E5%A5%87%E5%BC%82%E5%80%BC" \t "https://zhuanlan.zhihu.com/p/_blank)分解（Singular Value Decomposition）是[线性代数](https://link.zhihu.com/?target=https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7%E4%BB%A3%E6%95%B0/800%22%20%5Ct%20%22_blank" \t "https://zhuanlan.zhihu.com/p/_blank)中一种重要的[矩阵分解](https://link.zhihu.com/?target=https://baike.baidu.com/item/%E7%9F%A9%E9%98%B5%E5%88%86%E8%A7%A3/4035386" \t "https://zhuanlan.zhihu.com/p/_blank)，是[特征分解](https://link.zhihu.com/?target=https://baike.baidu.com/item/%E7%89%B9%E5%BE%81%E5%88%86%E8%A7%A3/12522621" \t "https://zhuanlan.zhihu.com/p/_blank)（矩阵必须为方阵）在任意矩阵上的推广。奇异值分解是一个适用于任意矩阵的一种分解的方法。

奇异值分解在统计中的主要应用为主成分分析（PCA），一种数据分析方法，用来找出大量数据中所隐含的“模式”，它可以用在模式识别，数据压缩等方面。

1. 介绍梯度

在微积分里面，对多元函数的参数求∂偏导数，**把求得的各个参数的偏导数以向量的形式写出来，就是梯度**。比如函数f(x,y), 分别对x,y求偏导数，求得的梯度向量就是(∂f/∂x, ∂f/∂y)T,简称grad f(x,y)或者▽f(x,y)。对于在点(x0,y0)的具体梯度向量就是(∂f/∂x0, ∂f/∂y0)T.或者▽f(x0,y0)，如果是3个参数的向量梯度，就是(∂f/∂x, ∂f/∂y，∂f/∂z)T,以此类推。

那么这个梯度向量求出来有什么意义呢？它的意义从几何意义上讲，**就是函数变化增加最快的地方**。具体来说，对于函数f(x,y)，在点(x0,y0)，沿着梯度向量的方向就是(∂f/∂x0, ∂f/∂y0)T的方向是f(x,y)增加最快的地方。或者说，沿着梯度向量的方向，更加容易找到函数的最大值。反过来说，沿着梯度向量相反的方向，**也就是 -(∂f/∂x0, ∂f/∂y0)T的方向，梯度减少最快，也就是更加容易找到函数的最小值**。

梯度下降算法（Gradient Descent Optimization）是神经网络模型训练最常用的优化算法。

a. 批量梯度下降法（Batch Gradient Descent，BGD）

批量梯度下降法是梯度下降法最原始的形式，它的具体思路是在更新每一参数时都使用所有的样本来进行更新。

优点：全局最优解；易于并行实现； 缺点：当样本数目很多时，训练过程会很慢。

b. 随机梯度下降（Stochastic Gradient Descent，SGD）

随机梯度下降的思路是在每次迭代时，只使用一个样本，当样本个数很大的时候，随机梯度下降迭代一次的速度要远高于批量梯度下降方法。两者的关系可以这样理解：随机梯度下降方法以损失一部分精确度和增加一定数量的迭代次数为代价，换取了总体的优化效率的提升。增加的迭代次数远远小于样本的数量。如果样本量很大的情况（例如几十万），那么可能只用其中几万条或者几千条的样本，就已经迭代到最优解了。

1. 过拟合

[过拟合](https://zhida.zhihu.com/search?content_id=222803661&content_type=Article&match_order=1&q=%E8%BF%87%E6%8B%9F%E5%90%88&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)：学习能力过强。表现为训练误差较小，测试误差较大

[欠拟合](https://zhida.zhihu.com/search?content_id=222803661&content_type=Article&match_order=1&q=%E6%AC%A0%E6%8B%9F%E5%90%88&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)：学习能力太差。表现为训练误差和测试误差都较大。

防止过拟合（掌握）

增加数据量（最直接，最暴力，最有用的方法）

[数据增强](https://zhida.zhihu.com/search?content_id=222803661&content_type=Article&match_order=1&q=%E6%95%B0%E6%8D%AE%E5%A2%9E%E5%BC%BA&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)：如图像中的平移、旋转、缩放。原理：通常使用CNN实现图像分类，而CNN只关注提取到的特征是什么，而不会在意空间位置，特征之间的相互关系，所以平移、旋转、缩放后的图像对于CNN来说是一种新的物体，就相当于间接扩充了数据量。

合理的数据划分（训练集和验证集的划分比例）

[正则化](https://zhida.zhihu.com/search?content_id=222803661&content_type=Article&match_order=1&q=%E6%AD%A3%E5%88%99%E5%8C%96&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)：常见的正则化参数有**L1和[L2范数](https://zhida.zhihu.com/search?content_id=222803661&content_type=Article&match_order=1&q=L2%E8%8C%83%E6%95%B0&zhida_source=entity" \t "https://zhuanlan.zhihu.com/p/_blank)**

L1范数惩罚项的目的是将参数的绝对值最小化。在固定的网络中，使用L1范数会趋向于使用更少的参数，而其他的参数都是0，从而增加网络的稀疏性，防止过拟合，也叫做Lasso回归。服从拉普拉斯分布L2范数惩罚项的目的是将参数的平方和最小化。利用L2范数会使用更多的参数，但这些参数都会接近于0，防止模型过拟合，也叫做Ridge回归。服从高斯分布。

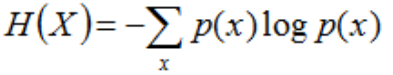
Dropout：以概率p使部分神经元失活。（比较常用）

提前终止训练（需要一定的炼丹经验才推荐）

还有BatchNormalization也可以

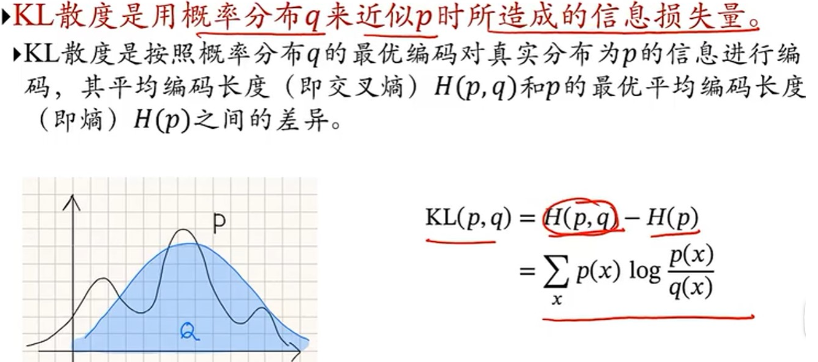
1. 熵与交叉熵

信息（熵）:离散随机事件的出现概率

 交叉熵损失用的就是这个

1. KL散度

样本越相似 KL散度越低



14.[如何确定CNN的卷积核通道数和卷积输出层的通道数？](https://link.zhihu.com/?target=https://www.julyedu.com/question/big/kp_id/26/ques_id/932" \t "https://zhuanlan.zhihu.com/p/_blank)

CNN的卷积核通道数 = 卷积输入层的通道数；CNN的卷积输出层通道数 = 卷积核的个数

（卷积核是一个可学习的权重张量，表示为 [kernel\_height, kernel\_width, input\_channels, num\_kernels]，它的深度必须与输入层的通道数匹配）

卷积操作后的结果，表示为 [new\_height, new\_width, output\_channels]，每个卷积核生成一个独立的输出通道

CNN的本质是完成了特征提取或者说是对原始特征的特征组合工作

调参部分：激活函数推荐使用relu，dropout层数不易设置过大，过大会导致不收敛，调节步长可以是0.05，一般调整到0.4或者0.5就可找到最佳值。

优化函数SGD——随机梯度下降！

1. LSTM与RNN

LSTM有进有出且当前的cell informaton是通过input gate控制之后叠加的，RNN是叠乘，因此LSTM可以防止梯度消失或者爆炸。

防止梯度爆炸还可以通过梯度截断

LSTM通过门控机制解决RNN的梯度消失问题，包含三个关键门：

遗忘门 (Forget Gate)：决定丢弃哪些信息

输入门 (Input Gate)：决定存储哪些新信息

输出门 (Output Gate)：决定输出哪些信息

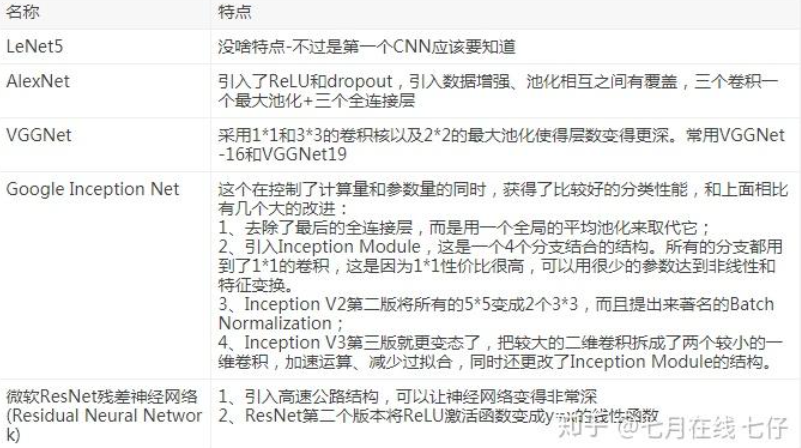
循环神经网络RNN：循环连接：隐藏状态会传递到下一个时间步

参数共享：所有时间步使用相同的权重矩阵

简单计算：h\_t = tanh(W\_h \* h\_{t-1} + W\_x \* x\_t + b)

深度学习的前提是神经网络的隐层加上了非线性激活函数，提升了模型的非线性表达能力，使得神经网络可以逼近任意复杂的函数

1. CNN的几个常用模型



1. 如何判断训练时出现梯度爆炸？

模型无法从训练数据中获得更新（如低损失）。

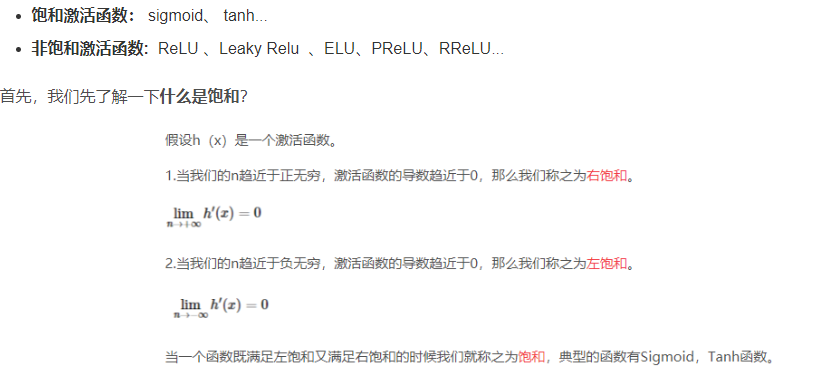
模型不稳定，导致更新过程中的损失出现显著变化。

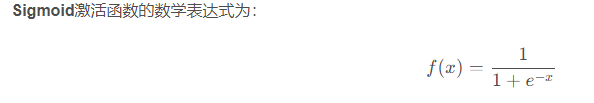
训练过程中，模型损失变成 NaN。

1. 什么是RNN？

RNNs的目的使用来处理序列数据。在传统的神经网络模型中，是从输入层到隐含层再到输出层，层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。例如，你要预测句子的下一个单词是什么，一般需要用到前面的单词，因为一个句子中前后单词并不是独立的。  
RNNs之所以称为循环神经网路，即**一个序列当前的输出与前面的输出也有关**。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即**隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出**。

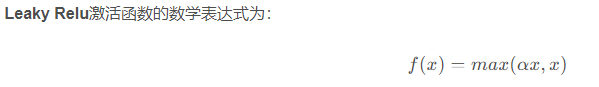
1. 常见的激活函数

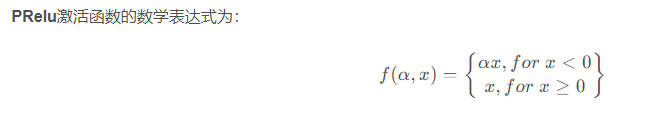














1. 温度对模型输出的影响

 模型每一步预测会输出一个所有可能词汇的原始分数（logits）， 这些 logits 会通过 Softmax 函数转化为概率分布。这个分布决定了每个词被选中的“相对可能性”。引入温度后，在 Softmax 计算之前，用温度参数调整 logits：

adjusted\_logits = logits / temperature

调整后的 logits 再输入 Softmax 函数，得到最终的概率分布。

对于高温情况，temp>1，缩小 logits 之间的差异，使 Softmax 输出的概率分布变得更加“平坦”。所有词的概率趋向于更均匀，罕见词和常见词之间的差距减小

对于低温，放大 logits 之间的差异，使 Softmax 输出的概率分布变得更加“尖锐”（峰度更高）。高分（高概率）的词变得更高，低分（低概率）的词变得更低