1，CNN中1*1卷积的作用：

实现跨通道的交互和信息整合

只有一个参数，输出核心的数量，所以可能对进行卷积核通道数的降维和升维

在保持feature map 尺寸不变（即不损失分辨率）的前提下大幅增加非线性特性，把网络做得很deep

2，KNN, kmeans

KNN基于实例的学习，是k个最近的邻居，判断新来的节点的分类，如果一个样本在特征空间中的k个最邻近的样本中的大多数属于某一个类别，则该样本也划分为这个类别。距离常用计算方法：欧几里得距离、余弦值（cos），相关度 （correlation）

Kmeans 是一种聚类算法：

- 首先输入k的值，即我们希望将数据集经过聚类得到k个分组。
- 从数据集中随机选择k个数据点作为初始质心，Centroid
- 对集合中每一个点，计算与每一个中心的距离（距离的含义后面会讲），选择最近的一个中心点。
- （其实是通过算法选出新的质心）。
- 表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛，可以认为我们进行的聚类已经达到期望的结果，算法终止。

缺点：有可能收敛到局部最小，收敛结果和初始化中心点的位置相关

3，RELU，sigmod，tanh区别，用法

sigmoid:非零均值，网络中神经元（除了第一层）的所有权重要么为正，要么都为负数，容易饱和停止学习

Relu：整流线性，输入大于1，永不饱和，计算非常简单，收敛速度极快，区间不对称，某些神经元永远不会被激活，参数不能被更新

hanh: 零均值，但是梯度消失

4，Bias和variation的区别

机器学习中的总Error = Bias + Variance，

Bias是模型拟合出来的值和真实值之间的差距；当模型复杂度上升时，Bias减小。当模型复杂度降低时，Bias增加，欠拟合的时候bias通常比较大；

Variance是对一个模型使用不同的数据建模产生不同的结果，在某一个点上的方差就是variance，过拟合的时候容易过大。

5，SDG为什么是一阶求导，而不是二阶求导

6，L1，L2正则化的区别

L2正则化也叫权重衰减，所有权重参数的平方和，迫使所有权重趋向0但大于0；

L1正则化是所有权重的的绝对值和，迫使不需要的权重为0，使得特征变得稀疏

7，Batch Normalization

对网络的下一层的输入进行归一化处理，使得输入量的均值为0，方差为1
能够加速模型训练
8，Dropout
在训练的过程中，让神经元以p的概率被设置为0，目的是为了减少过拟合，使得所有的神经元都能得到充分的训练。
9，梯度消失的问题
使用的不合理的激活函数，比如sigmoid，当层数很深的时候，链式法则求导，计算梯度，损失函数的梯度会梯度相乘，导致乘积越来越接近0，所以无法学习到信息。
解决方法：
预训练
梯度剪切：投影到很小的尺度上，不能超过预设的值
权重正则化，使用不同的激活函数，
使用batchNormalization
使用残差网络（shortcut)
LSTM网络中的结构设计
10， Kmeans/KNN中不同距离的区别
欧氏距离：它将样本的不同属性之间的差别等同看待
曼哈顿距离： 投影距离之和
11，生成模型，判别模型
生成根据数据，来学习一个联合概率分布，HMM，贝叶斯（每个特征都是同样重要的，并且相互独立），高斯混合模型，LDA
判别模型：学习决策函数，LR，SVM，MLP，CRF
区别：生成模型目的是还原模型的联合概率分布，判别模型不能
生成模型能够收敛于真实的模型
判别模型能够直接面对预测结果，准确率会更好，也能够简化学习问题。
12，BN ，能够增加算法收敛的速度、可能提高精度
在深度网络中，之前层的神经网络参数变化，会导致本层的的输入分布发生很大的变化，使用随机梯度下降的时候，每次参数更新都会使得输入输出分布发生变化。
那么会出现一些，参数重新学习的问题，所以为了让每次输入数据的分布保持一致，使用一个简单的归一化操作，就是把前一层的输入减去平均值，除以标准差
13，LR和SVM的区别
都可以处理分类的问题，都能添加正则化项L1,L2正则化
区别：LR是参数模型，SVM是非参数的模型
LR通常使用交叉熵损失函数，SVM是hinge损失函数，增加对分类影响较大的样本点的权重，减少与分类关系影响小的数据点的权重

14，优化器算法：BGD，SGD，MBGD
BGD：使用整个数据集来计算loss funcion，对于凸函数可以收敛到全局最小值，非凸函数收敛到局部最小
SGD：每次对每个样本进行梯度更新，会出现很大的震荡，但是可能会跳到更好的局部最小值
MBGD：minibatch 中和了上述两种的特点
Momentum：SGD 在 鞍点的情况下容易被困住，在梯度方向不变的维度上速度变快，梯度方向有所改变的维度上的更新速度变慢
Adagrad ： adaptive gradient 可以对低频的参数做较大的更新，对高频的做较小的更新
Adam：保存了过去梯度，保存了momentum
15,loss function 和target function的区别
loss function计算一个样本和真实的误差，通常不包括正则化项
target function包括正则化项
16,词向量的作用：
可以表达各个单词之间的关系，维度相比one hot低很多，可以迁移到其他的任务中
对于每个维度，无法解释
17，数据降维的主要方法：
奇异值分解：SVD；主成分分析：PCA
PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征
1) 去平均值，即每一位特征减去各自的平均值。
2) 计算协方差矩阵。
3) 通过SVD计算协方差矩阵的特征值与特征向量。
4) 对特征值从大到小排序，选择其中最大的k个。然后将其对应的k个特征向量分别作为列向量组成特征向量矩阵。
5) 将数据转换到k个特征向量构建的新空间中。
18，precision, recall, accurcy, F1，AUC
precision精确度 实际正样本/预测正样本
recall 被预测到实际正样本/实际正样本
accuracy 预测正确的次数/总数
F1 实际正样本*2/实际正样本*2+预测错误的样本
19，常用词向量：word2vec, fastText, glove, ELMO, BERT
Cbow 通过附近词预测中心词，skipgram中心词预测附近的词
fastText:通过使用句子中的所有Ngram去预测标签，只有一层的隐层以及输出层，并且使用根据类别的频率构造霍夫曼树来代替标准softmax

Glove通过共现矩阵，计算两个词之间共同出现的次数，考虑到了局部和整体的信息
ELMO为了解决词的多义性问题，使用多层的双向LSTM，使用第1~k-1个Token的隐藏层输出预测第k个Token。下游任务在ELMO的基础上进行微调
BERT：使用多层的transformer构建模型，Transformer可以综合的考虑两个方向的信息，并且可以捕捉长距离的依赖，而且有非常好的并行性质；训练过程类似于完形填空，随机覆盖15%的单词，用其他的词预测被覆盖的词。下游任务
20，Bagging和boosting的区别
bagging中是强模型，偏差低，方差高，为了降低方差，所以将很多模型的结果平均
boosting中大多是较弱的模型，偏差高，方差低，训练时对于分类误差小的分类器设置成更大的权重。
21， maxpooling 和average pooling
max-pooling和average-pooling都对数据做了下采样，但是max-pooling感觉更像是做了特征选择，选出了分类辨识度更好的特征，提供了非线性
22，模型loss不下降

- 模型本身不合理，模型容量太小，出现梯度消失，权重初始化方案有问题
- 正则化过度
- 激活函数不合理，sigmoid函数饱和停止学习relu出现失活
- 学习速率太大或太小，导致不收敛
- 未进行归一化