机器学习与深度学习常见面试题（下）

1、为什么随机森林能降低方差？

随机森林的预测输出值是多课决策树的均值，如果有 n 个独立同分布的随机变量 x_i，它们的方差都为 ^{2}，则它们的均值的方差为：

$$D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{^2}{n}$$

2、对于带等式和不等式约束的优化问题，KKT 条件是取得极值的充分条件还是必要条件？对于 SVM 呢？

对于一个一般的问题，KKT 条件是取得极值的必要条件而不是充分条件。对于凸优化问题，则是充分条件，SVM 是凸优化问题

3、解释维数灾难的概念

当特征向量数理很少时，增加特征，可以提高算法的精度，但当特征向量的维数增加到一定数量之后，再增加特征，算法的精度反而会下降

4、Logistic 回归为什么用交叉熵而不用欧氏距离做损失函数？

如果用欧氏距离，不是凸函数，而用交叉熵则是凸函数

5、解释 hinge loss 损失函数

如果样本没有违反不等式约束，则损失为 0；如果违反约束，则有一个正的损失值

6、解释 GBDT 的核心思想
用加法模拟，更准确的说，是多棵决策树树来拟合一个目标函数。每一棵决策树拟合的是之前迭代得到的模型的残差。求解的时候，对目标函数使用了一阶泰勒展开，用梯度下降法来训练决策树

7、解释 XGBoost 的核心思想
在 GBDT 的基础上，目标函数增加了正则化项，并且在求解时做了二阶泰勒展开

8、解释 DQN 中的经验回放机制，为什么需要这种机制？
将执行动作后得到的状态转移构造的样本存储在一个列表中，然后从中随机抽样，来训练 Q 网络。为了解决训练样本之间的相关性，以及训练样本分布变化的问题

9、什么是反卷积？
反卷积也称为转置卷积，如果用矩阵乘法实现卷积操作，将卷积核平铺为矩阵，则转置卷积在正向计算时左乘这个矩阵的转置 $\mathbf{W}^{\mathrm{T}}$，在反向传播时左乘 $\mathbf{W}$，与卷积操作刚好相反，需要注意的是，反卷积不是卷积的逆运算

10、反卷积有哪些用途？
实现上采样；近似重构输入图像，卷积层可视化

11、PCA（主成分分析）优化的目标是什么？
最小化重构误差/最大化投影后的方差

12、LDA（线性判别分析）优化的目标是什么？
最大化类间差异与类内差异的比值

13、解释神经网络的万能逼近定理
只要激活函数选择得当，神经元的数理足够，至少有一个隐含层的神经网络可以逼近闭区间上任意一个连续函数到任意指定的精度

14、softmax 回归训练时的目标函数时凸函数吗？
是，但有不止一个全局最优解

15、SVM 为什么要求解对偶问题？为什么对偶问题与原问题等价？

原问题不容易求解，含有大量的不易处理的不等式约束。原问题满足 Slater 条件，强对偶成立，因此原问题与对偶问题等价

16、神经网络是生成模型还是判别模型？
判别模型，直接输出类别标签，或者输出类后验概率 p(y|x)

17、logistic 回归是生成模型还是判别模型？
判别模型，直接输出类后验概率 p(y|x)，没有对类条件概率 p(x|y)或者联合概率 p(x, y)建模

18、对于支持向量机，高斯核一般比线性核有更好的精度，但实际应用中为什么一般用线性核而不用高斯核？
如果训练样本的量很大，训练得到的模型中支持向量的数量太多，在每次做预测时，高斯核需要计算待预测样本与每个支持向量的内积，然后做核函数变换，这会非常耗；而线性核只需要计算 $\mathrm{w}^{\mathrm{T}}\mathrm{x}+b$

19、高斯混合模型中，为什么各个高斯分量的权重之和要保证为 1？
为了保证这个函数时一个概率密度函数，即积分值为 1

20、介绍 beam search 算法的原理
这是一种解码算法，每次选择概率最大的几个解作为候选解，逐步扩展

21、介绍 seq2seq 的原理
整个系统由两个 RNN 组成，一个充当编码器，一个充当解码器；编码器依次接收输入的序列数据，当最后一个数据点输入之后，将循环层的状态向量作为语义向量，与解码器网络的输入向量一起，送入解码器中进行预测

22、介绍 CTC 的原理
CTC 通过引入空白符号，以及消除连续的相同符号，将 RNN 原始的输出序列映射为最终的目标序列。可以解决对未对齐的序列数据进行预测的问题，如语音识别

23、介绍广义加法模型的原理
广义加法模型用多个基函数的和来拟合目标函数，训练的时候，依次确定每个基函数

24、为什么很多时候用正态分布来对随机变量建模？

现实世界中很多变量都服从或近似服从正态分布。中心极限定理指出，抽样得到的多个独立同分布的随机变量样本，当样本数趋向于正无穷时，它们的和服从正态分布

25、Batch Normalization 和 Group Normalization 有何区别？
BN 是在 batch 这个维度上进行归一化，GN 是计算 channel 方向每个 group 的均值和方差

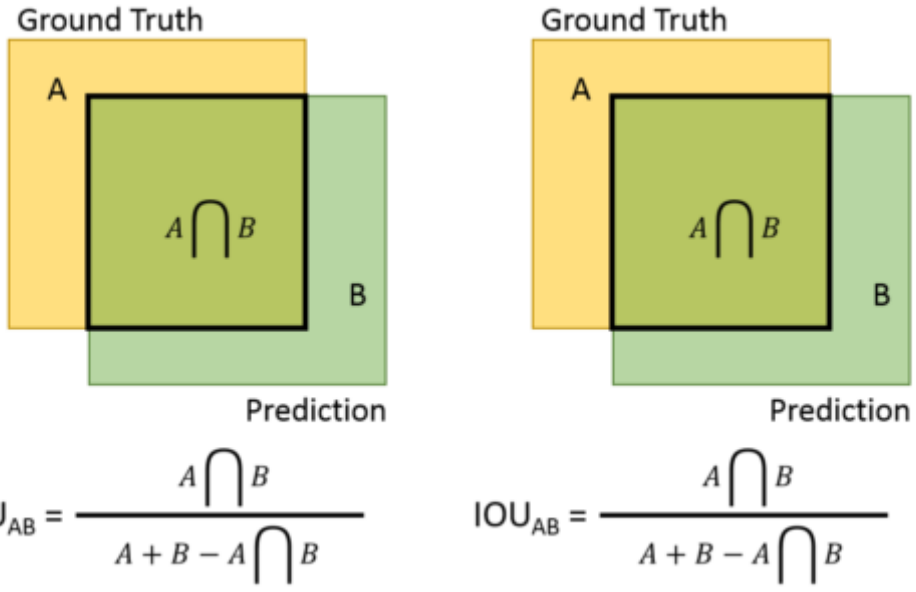
26、GAN 中模型坍塌（model collapse）是指什么？
模型坍塌，即产生的样本单一，没有了多样性。

27、 目前 GAN 训练中存在的主要问题是什么？
（1） 训练不易收敛 （2）模型坍塌

28、Shufflenet 为什么效果会好？
通过引入“通道重排”增加了组与组之间信息交换

29、模型压缩的主要方法有哪些？
（1）从模型结构上优化：模型剪枝、模型蒸馏、automl 直接学习出简单的结构
（2）模型参数量化将 FP32 的数值精度量化到 FP16、INT8、二值网络、三值网络等

30、目标检测中 IOU 是如何计算的？
检测结果与 Ground Truth 的交集比上它们的并集，即为检测的准确率 IoU



$$IOU_{AB} = \frac{A \cap B}{A + B - A \cap B}$$

$$IOU_{AB} = \frac{A \cap B}{A + B - A \cap B}$$

31、给定 0-1 矩阵，如何求连通域？
可采用广度优先搜索

32、OCR 任务中文本序列识别的主流方法是什么？
RNN+CTC

33、在神经网络体系结构中，哪些会有权重共享？？

（1）卷积神经网络
（2）递归神经网络
（3）全连接网络
答案 （1）&（2）

34、 一个典型人脸识别系统的识别流程？
人脸检测--》人脸对齐--》人脸特征提取--》人脸特征比对

35、平面内有两个矩形，如何快速计算它们的 IOU？

36、使用深度卷积网络做图像分类如果训练一个拥有 1000 万个类的模型会碰到什么问题？
提示：内存/显存占用；模型收敛速度等

37、 HMM 和 CRF 的区别？
前者描述的是 P(X,Y)=P(X|Y)*P(Y)，是 generative model; 后者描述的是 P(Y|X)，是 discriminative model. 前者你要加入对状态概率分布的先验知识，而后者完全是 data driven.

38、深度学习中为什么不用二阶导去优化？
Hessian 矩阵是 n*n， 在高维情况下这个矩阵非常大，计算和存储都是问题

39、深度机器学习中的 mini-batch 的大小对学习效果有何影响？
mini-batch 太小会导致收敛变慢，太大容易陷入 sharp minima，泛化性不好

40、线性回归对于数据的假设是怎样的？
http://en.wikipedia.org/wiki/Linear_regression
（1） 线性，y 是多个自变量 x 之间的线性组合
（2） 同方差性，不同的因变量 x 的方差都是相同的
（3） 弱外生性，假设用来预测的自变量 x 是没有测量误差的
（4） 预测变量之中没有多重共线性

41、什么是共线性，跟过拟合有啥关联?
共线性：多变量线性回归中，变量之间由于存在高度相关关系而使回归估计不准确。
共线性会造成冗余，导致过拟合。

解决方法：排除变量的相关性 / 加入权重正则。

42、Bias 和 Variance 的区别？

Bias 量了学习算法的期望预测与真实结果的偏离程度，即刻画了算法本身的拟合能力。Variance 度量了同样大小的训练集的变动所导致的学习性能变化，即刻画了数据扰动所造成的影响。