1. **试阐述LDA（线性鉴别分析）的分类思想。**
   给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能接近，异类样例的投影点尽可能远离；在对新样本进行分类时，将其投影到同样的这条直线上，再根据投影点的位置来判断新样本的类别。

2. **试分析SVM 对噪声敏感的原因。**
   给定训练集，SVM最优决策边界由支持向量决定。当增加噪声时，那么该噪声有极高的可能是含噪声训练集的一个支持向量，这意味着决策边界需要变。

3. **距离函数的四个基本性质**：非负性、同一性、对称性、直递性

4. **在数据必理时，为什么通常要进行标准化处理。**
   在实标问题中、我们使用的样本通常是多维数据、每一维对应一个特征，这些这些特征的量纲和数量级都是不一样的，这时需要对数据进行标准化处理，是所有特征同样的尺度。

5. **随机变量X的支撑集（也就是非零值域）定义为【a,b】，没有别的限制加在X上，该随机变量的最大熵分布是什么。**均匀分布

6. **随机变量x 的給定均值和方差限制在x上、该随机变量的最大熵分布是什么。**
   根据最大熵模型推导出了概率密度函数是一个高斯分布。

7. **試述将线性函数用作神经元激活函数的缺陷。**
   如果单用线性函数作为激活函数。无论多少层的神经网络都会退化成一个线性回归，不能处理非线性分类任务。

8. **试述学习幸的取值对神经网络训练的影响。**

   如果学习率太低，每次下降的很慢，使得迭代次数非常多。如果学习率太高，在后面迭代时会出现震荡现象，在最小值附近来回波动。

9. **神经网络为什么会产生梯度消失，有什么解决方案。**
   前面层上的梯度是来自于后面层上梯度的乘积。当存在过多的层次时，且激活函数的梯度小于1时，就会使前面层的梯度变得很小，更新速度过慢，导致梯度消失。

   一种解决方案是使用Relu 激活函数替换sigmoid, Relu 函数的梯度不会随着x的增大而变小，sigmoid 在x 取值较大时梯度趋近于0。

10. **.对3个32×32的特征图进行卷积层操作，卷积核10个5X5，Stride是1，po.d为2，输出特征图的尺度是多少？卷积层的参数是多少？写出公式和结果。**
    输出尺度（32+2X2-5）/1+1 = 32
    卷积层的参数（5X5X3+1）X10=760

11. **试析随机森林为何比决策树Bagging集成的训练速度更快。**

    随机森林是Bagging算法的一个扩展变体，以决策树为基学习器构建Bagging集成，Bagging在选择划分属性时需要考察结点的所有属性，而随机森林只需随机地考察一个属性子集，所以随机森林比决策树Bagging训练速度更快。

12. **请指出数据聚类存在哪些挑战性问题。**
    1. 能够处理高维数据：在高维空间聚类更具挑战性，随着维数的增加，具有相同距离的两个样本其相似程度可以相差很远。对于高维稀疏数据，这一点更突出。
    2. 对噪声鲁棒：在实际中，绝大多数样本集都包含噪声、空缺、部分未知属性、孤立点、甚至错误数据。
    3. 具有约束的聚类：在实际应用中，通常需要在某种约束条件下进行聚类，既满足约束条件，以希望有高聚类精度，是一个挑战性问题。对初始输入参数鲁棒：具有自适应的簇数判定能力，对初始聚类中心鲁棒。

4. 能够解决用户的问题：聚类结果能被用户所理解，并能带来经济效益，特别是在数据挖掘领域。

13. **阐述一下对泛化误差的理解。**

    泛化误差=偏差 + 方差+噪声

    偏差：度量了学习算法的期望预测与真实结果的偏离程度，刻画了学习算法本身的拟合能力

    方差：度量了同样大小的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响。

    噪声：表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界，即刻画了学习问题本身的难度。

14. **模型评估过程中，欠拟合和过拟合现象是什么。**

    过拟合是指模型对于训练数据拟合呈过当的情况，反映到评估指标上，就是模型在训练集上的表现很好，但在测试集和新数据上的表现较差。欠拟合是模型在训练和预测时表现都不好的情况

15. **说出几种降低过拟合和欠拟合的方法**

    降低过拟合：

    1. 从数据入手，获得更多的训练数据。使用更多的训练数据是解决过拟合问题最高效的手段，因为更多的样本能够让模型学习到更多更高效的特征。当然，直接增加实验数据一般是很困难的，但是可以通过一定的规则来扩充训练数据。比如在图像分类的问题上，可以通过图像的平移、旋转、缩放等方式扩充数据，更进一步地，可以使用生成式对抗网络来合成大量的新训练数据。
    2. 降低模型复杂度。在数据较少时，模型过于复杂是产生过拟合的主要因素，适当降低模型复杂度可以避免模型拟合过多的采样噪声。例如，在神经网络模型中减少网络层数、神经元个数等；在决策树模型中降低树的深度、进行剪枝等。
    3. 正则化方法。给模型的参数加上一定的正则约束，比如将权值的大小加入到损失函数中。
    4. 集成学习方法。集成学习是把多个模型集成在一起，来降低单一模型的过拟合风险，如Bagging方法

    降低欠拟合：

    1. 添加新特征。当特征不足或者现特征与样本标签的相关性不强时，模型容易出现欠拟合。通过挖掘“上下文特征”“ID 类特征”“组合特征”等新的特征，往往能够取得更好的效果。
    2. 增加模型复杂度。简单模型的学习能力较差，通过增加模型的复杂度可以便模型拥高更强的拟合能力。例如，在线性模型中添加高次项，在神经网络模型中增加网络层数或神经元个数等。
    3. 减小正则化系数。正则化是用来防止过拟合的，但当模型出现欠拟合现象时，则需要针对性地减小正则化系数

16. **常用的决策树算法有ID3，C4.5，CART，它们构建树所使用的启发式函数各是什么**

    ID3：最大信息增益

    C4.5：最大信息增益率

    CART：最小基尼指数

17. **K均值算法的优缺点是什么，如何对其调优。**

    1. k均值算法缺点：例如受初值和离群点的影响每次的结果不稳定、结果通常不是全局最优而是局部最优解、无法很好地解决数据簇分布差别比较大的情况、不太适用于离散分类等。
    2. K均侦聚类的优点：主要体现在对于大数据集，K均值聚类算法相对是高效的，计算复杂度是O（NKt）接近于线性，其中N是数据对象的数目，K是聚类的簇数，t是迭代的轮数。
    3. 调优方法：数据归一化，离群点预处理，采用核函数，合理选择K值。

18. **Relu激活函数的优缺点**

优点：

（1）从计算的角度上，Sigmoid与Tanh激活函数均需要计算指数，复杂度高。而ReLU只需要一个阈值即可得到激活值。

（2）ReLU的非饱和性可以有效地解决梯度消失的问题。

（3） ReLU的单侧仰制提供了网络的稀疏表达能力。

缺点：

在较大学习率设置下Relu可能会出现大量神经元死亡问题。后面神经元方向传播梯度为正，且学习率较大，Relu的梯度为1，梯度下降此时会导致该神经元的参数为负值，可能之后不会再被激活，造成神经元死亡