* 数据库中的主键、索引和外键（数据分析岗经常问）
* 决策树ID3和C4.5的差别？各自优点？
* Boost算法
* CART（回归树用平方误差最小化准则，分类树用基尼指数最小化准则）
* GBDT与随机森林算法的原理以及区别。
* 优化算法中常遇到的KKT条件？作用是？
* 最近邻算法KNN（分类与回归）
* L1和L2函数？L1和L2正则项的比较，如何解决 L1 求导困难？
* L1正则为何可把系数压缩成0，说明坐标下降法的具体实现细节
* LR为什么用sigmoid函数。这个函数有什么优点和缺点？为什么不用其他函数？
* LR和SVM的区别？libsvm和liblinear的区别？
* Logistics与随机森林比较
* Logistics（理论推导）；Logistic回归的推导，如何得到objective function？
* SVM与随机森林的差别？
* SVM为何要引入拉格朗日的优化方法？
* SVM原问题和对偶问题关系？
* SVM在哪个地方引入的核函数, 如果用高斯核可以升到多少维。
* SVM如何防止过拟合？
* VM的目标函数？常用的核函数有哪些？
* SVM的过程，理论推导
* bagging、adaboost、boosting区别在哪
* EM 与 k-means 的关系？
* k-means算法中的k如何选取？
* k-means算法初始点怎么选择？
* k-means的原理，优缺点以及改进。
* k折交叉验证中k取值多少有什么关系？
* L2惩罚项是怎么减小Overfitting的？L1，L2等范数的通式是？差别是？在什么场景下用什么范数？L1在0处不可导，怎么处理？
* 随机森林和GBDT差别？
* RF, GBDT, xgboost的区别？
* 为什么要做数据归一化？
* 梯度下降法的原理以及各个变种（批量梯度下降，随机梯度下降法，mini 梯度下降法），以及这几个方法会不会有局部最优问题。
* 牛顿法原理和适用场景，有什么缺点，如何改进（拟牛顿法）
* 什么情况下一定会发生过拟合？
* 贝叶斯估计？
* 介绍LR、RF、GBDT ，分析它们的优缺点
* 会哪些机器学习算法？信息熵公式？
* 决策树原理；决策树处理连续值的方法；决策树如何防止过拟合；决策树过拟合哪些方法，前后剪枝。
* 分类模型可做回归分析吗？反过来可以吗？
* 分类模型和回归模型的区别？
* 判别模型和生成模型？差别
* 各个模型的Loss function，牛顿学习法、SGD如何训练。
* 在模型的训练迭代中，怎么评估效果？
* 如何防止过拟合（增加数据，减少模型复杂度->正则化）
* 对于同分布的弱分类器，求分类器均值化之后的分布的均值跟方差。
* 对于机器学习你都学了哪些？讲一个印象深的。
* 常见分类模型（ svm，决策树，贝叶斯等）的优缺点，适用场景以及如何选型
* 数据归一化的方式
* 手写k-means的伪代码。
* 手写svm硬软间隔对偶的推导
* 手写逻辑回归（损失函数及更新方式推导）
* BP算法原理
* 改变随机森林的训练样本数据量，是否会影响到随机森林学习到的模型的复杂度？
* 数据挖掘各种算法，以及各种场景下的解决方案
* 是否了解mutual infomation、chi-square、LR前后向、树模型等特征选择方式。
* 是否了解线性加权、bagging、boosting、cascade等模型融合方式
* 有哪些常见的分类器，简单介绍下原理
* 机器学习与深度学习的区别
* 线性回归与逻辑回归区别？
* 机器学习：几种树模型的原理和对比，朴素贝叶斯分类器原理以及公式，出现估计概率值为 0 怎么处理（拉普拉斯平滑），缺点；
* DBSCAN聚类算法原理
* 主成分分析法原理、MapReduce原理、Spark等（数据岗题）
* 梯度下降、牛顿、拟牛顿算法的原理
* 梯度下降的优缺点。
* 深度学习有很大部分是CNN，给他用通俗的语言解释下卷积的概念，解释下CNN中的优势及原因
* 牛顿法、随机梯度下降算法和直接梯度下降算法的区别？
* 牛顿法推导
* 特征选择方法有哪些
* 由数据引申到数据不平衡怎么处理（10W正例，1W负例，牛客上有原题）
* 聊聊SVM，这段说了好久，从基本的线性可分到不可分，相关升维，各种核函数，每个是如何实现升。以及出现了XX问题，分析是样本的原因还是其他原因。针对不同情况，采取什么解决方案较好。
* 自己实现过什么机器学习算法
* 解决过拟合的方法有哪些？
* 解释一下过拟合和欠拟合，有哪些方法防止过拟合。
* 如何构造决策树、计算信息熵和信息增益、说明C4.5 、ID3、 CART的区别及优缺点
* 详细讨论了样本采样和bagging的问题
* 说一下Adaboost，权值更新公式。当弱分类器是LR时，每个样本的的权重是w1，w2...,写出最终的决策公式。
* 说了一下bagging跟boosting。
* 说明L1和L2正则的效果与为什么形成这种情况（L1正则稀疏，L2正则平滑，之后说明就是画图说明正则化）
* 选个你熟悉的机器学习方法 ，着重介绍一下产生原因，推导公式，背后统计意义什么等等
* 逻辑回归估计参数时的目标函数
* 逻辑回归的值表示概率吗？
* 数据挖掘的基础知识，包括SVM,逻辑回归、EM、K-means等，然后给出很多场景问你遇到这些情况你如何处理数据，怎么进行建模等
* 随机梯度下降，标准梯度
* 随机森林和GBDT的区别？LR的参数怎么求解？有没有最优解？