1、常见的生成式模型和判别式模型有哪些？   
生成式模型：HMM、朴素贝叶斯   
判别式模型：SVM、最大熵模型、神经网络、决策树、条件随机场

2、EM算法的基本概念和应用场景？   
EM算法是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐藏变量（Latent Variable）。假设我们估计知道A和B两个参数，在开始状态下两者都是未知的，并且知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。可以考虑赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。   
应用场景：聚类

3、数据存储和数据流都属于数据，仅仅是所处的状态不同

4、计算机图形学中三维形体表示模型有哪些？   
线框、表面、实体

5、C语言的基本单位是什么？   
函数

6、n个记录使用快速排序，需要的辅助空间是？   
O（log2n）。快速排序对待排序序列的划分大约为log2n次，而快速排序是通过递归算法来完成的，递归深度大约为log2n，因此所需的辅助存储空间为O(log2n)。

7、(7,34,55,64,46,20,10)线性表的散列存储，H(k)=k%9,地址为1的共有多少个元素？   
4个。H（K）= K%9,表示除以9的余数.由于地址重叠造成冲突,所以散列存储时,通常还要有解决冲突的办法,如线性探查法等等.本题的答案分别是：55,64,46,10.

8、二叉树的中序遍历是BCFDEA后续遍历是BFCEAD,前序遍历是？   
DCBFAE

9、栈和队列的共同点是什么？   
只允许在端点处插入和删除元素

10、常见的分类算法有哪些？   
SVM、神经网络、逻辑回归、KNN、随机森林、贝叶斯

11、计算机图形学中多边形的表示方法有哪些？   
顶点表示和点阵表示

12、常见的监督学习算法有哪些？   
感知机、SVM、ANN、决策树、逻辑回归

13、简述SVM过程， 举几个常见的核函数，说说其用途，举例说明什么是支持向量   
线性核、多项式核、高斯核、拉普拉斯核、sigmoid核

14、说说K-means的过程，k的选择方法   
1 初始化聚类中心   
2 计算其他所有点到聚类中心的距离,将点归类到最近的中心   
3 更新聚类中心   
4 重复2-3步,直到聚类中心收敛

k的选择方法：   
1 思想：初始的聚类中心之间相互距离尽可能远   
kmeans++   
2 用层次聚类（从底向上，最近的簇合并）进行初始聚类

15、举个你熟悉的分类算法，简述其原理

16、信息增益、信息增益比、基尼系数分别是什么？信息增益和信息增益比的区别？   
信息增益 = 熵 - 条件熵（偏向选择特征值比较多的属性）   
信息增益比 = 信息增益/所选特征的熵   
基尼系数 = 1 - 全体数据集中每种类别概率的平方之和（构造二叉决策树）   
这三个也是特征选择方法（ID3、C4.5、CART）

17、常见的排序算法中，时间复杂度是nlogn的有哪些？   
快速排序、归并排序、堆排序

18、快速排序过程原理，稳定性及其原因。   
采取分而治之的思想，把大的拆分为小的，每一趟排序，把比选定值小的数字放在它的左边，比它大的值放在右边；重复以上步骤，直到每个区间只有一个数。此时数组已经排序完成。

1．i =L; j = R; 将基准数挖出形成第一个坑a[i]。   
2．j–由后向前找比它小的数，找到后挖出此数填前一个坑a[i]中。   
3．i++由前向后找比它大的数，找到后也挖出此数填到前一个坑a[j]中。   
4．再重复执行2，3二步，直到i==j，将基准数填入a[i]中。

选择排序、快速排序、希尔排序、堆排序不是稳定的排序算法，而冒泡排序、插入排序、归并排序和基数排序是稳定的排序算法

不稳定例子 2 4 1 3 1   
会把最后面的1换到最前面去，打破了1的稳定性

19、基于核的算法？   
SVM RBF LDA

20、数据清理中，处理缺失值的方法是?   
数据清理中，处理缺失值的方法有两种：   
删除法：   
1）删除观察样本   
2）删除变量：当某个变量缺失值较多且对研究目标影响不大时，可以将整个变量整体删除   
3）使用完整原始数据分析：当数据存在较多缺失而其原始数据完整时，可以使用原始数据替代现有数据进行分析   
4）改变权重：当删除缺失数据会改变数据结构时，通过对完整数据按照不同的权重进行加权，可以降低删除缺失数据带来的偏差

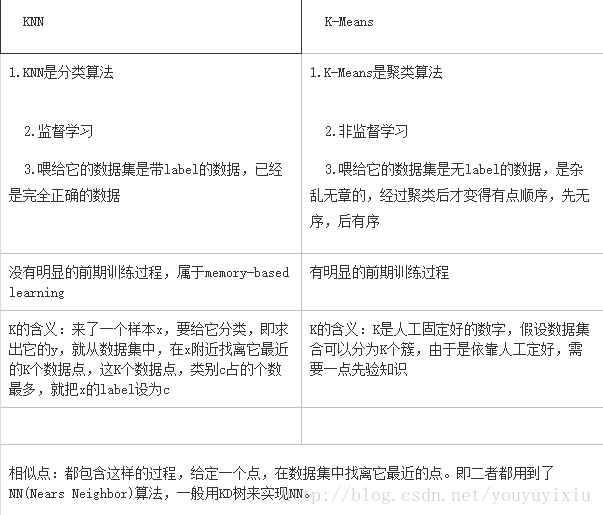
查补法：均值插补、回归插补、抽样填补等

21、下列哪个不属于CRF模型对于HMM和MEMM模型的优势（ B）

A. 特征灵活 B. 速度快 C. 可容纳较多上下文信息 D. 全局最优

首先，CRF，HMM(隐马模型)，MEMM(最大熵隐马模型)都常用来做序列标注的建模。隐马模型一个最大的缺点就是由于其输出独立性假设，导致其不能考虑上下文的特征，限制了特征的选择。最大熵隐马模型则解决了隐马的问题，可以任意选择特征，但由于其在每一节点都要进行归一化，所以只能找到局部的最优值，同时也带来了标记偏见的问题，即凡是训练语料中未出现的情况全都忽略掉。条件随机场则很好的解决了这一问题，他并不在每一个节点进行归一化，而是所有特征进行全局归一化，因此可以求得全局的最优值。

22、KNN和K-means区别



23、以下哪个是常见的时间序列算法模型 B   
A. RSI B. MACD C. ARMA D. KDJ

时间序列模型是指采用某种算法（可以是神经网络、ARMA等）模拟历史数据，找出其中的变化规律。时间序列算法模型主要有：移动平均算法、指数平滑算法及ARMA、ARIMA方法。

24、SVM 的推导，特性？多分类怎么处理？

SVM是最大间隔分类器   
从线性可分情况下，原问题，特征转换后的dual问题，引入kernel(线性kernel,多项式，高斯),最后是soft margin。

线性：简单，速度快，但是需要线性可分   
多项式：比线性核拟合程度更强，知道具体的维度，但是高次容易出现数值不稳定，参数选择比较多。   
高斯：拟合能力最强，但是要注意过拟合问题。不过只有一个参数需要调整。

多分类问题，一般将二分类推广到多分类的方式有三种，一对一，一对多，多对多。

一对一：将N个类别两两配对，产生N(N-1)/2个二分类任务，测试阶段新样本同时交给所有的分类器，最终结果通过投票产生。   
一对多：每一次将一个例作为正例，其他的作为反例，训练N个分类器，测试时如果只有一个分类器预测为正类，则对应类别为最终结果，如果有多个，则一般选择置信度最大的。从分类器角度一对一更多，但是每一次都只用了2个类别，因此当类别数很多的时候一对一开销通常更小(只要训练复杂度高于O(N)即可得到此结果)。   
多对多：若干个类作为正类，若干个类作为反类。注意正反类必须特殊的设计。

25、LR 的推导，特性？

LR的优点在于实现简单，并且计算量非常小，速度很快，存储资源低，缺点就是因为模型简单，对于复杂的情况下会出现欠拟合，并且只能处理2分类问题(可以通过一般的二元转换为多元或者用softmax回归)。

26、决策树的特性？

决策树基于树结构进行决策，与人类在面临问题的时候处理机制十分类似。其特点在于需要选择一个属性进行分支，在分支的过程中选择信息增益最大的属性。   
在划分中我们希望决策树的分支节点所包含的样本属于同一类别，即节点的纯度越来越高。决策树计算量简单，可解释性强，比较适合处理有缺失属性值的样本，能够处理不相关的特征，但是容易过拟合，需要使用剪枝或者随机森林。   
信息增益是熵减去条件熵，代表信息不确定性较少的程度，信息增益越大，说明不确定性降低的越大，因此说明该特征对分类来说很重要。由于信息增益准则会对数目较多的属性有所偏好，因此一般用信息增益率。其中分母可以看作为属性自身的熵。取值可能性越多，属性的熵越大。

Cart决策树使用基尼指数来选择划分属性，直观的来说，Gini(D)反映了从数据集D中随机抽取两个样本，其类别标记不一致的概率，因此基尼指数越小数据集D的纯度越高，一般为了防止过拟合要进行剪枝，有预剪枝和后剪枝，一般用cross validation集进行剪枝。

连续值和缺失值的处理，对于连续属性a,将a在D上出现的不同的取值进行排序，基于划分点t将D分为两个子集。一般对每一个连续的两个取值的中点作为划分点，然后根据信息增益选择最大的。与离散属性不同，若当前节点划分属性为连续属性，该属性还可以作为其后代的划分属性。

27、 SVM、LR、决策树的对比？   
SVM既可以用于分类问题，也可以用于回归问题，并且可以通过核函数快速的计算，LR实现简单，训练速度非常快，但是模型较为简单，决策树容易过拟合，需要进行剪枝等。   
从优化函数上看，soft margin的SVM用的是hinge loss,而带L2正则化的LR对应的是cross entropy loss，另外adaboost对应的是exponential loss。所以LR对远点敏感，但是SVM对outlier不太敏感，因为只关心support vector，SVM可以将特征映射到无穷维空间，但是LR不可以，一般小数据中SVM比LR更优一点，但是LR可以预测概率，而SVM不可以，SVM依赖于数据测度，需要先做归一化，LR一般不需要，对于大量的数据LR使用更加广泛，LR向多分类的扩展更加直接，对于类别不平衡SVM一般用权重解决，即目标函数中对正负样本代价函数不同，LR可以用一般的方法，也可以直接对最后结果调整(通过阈值)，一般小数据下样本维度比较高的时候SVM效果要更优一些。

28、GBDT 和随机森林的区别？   
随机森林采用的是bagging的思想，bagging又称为bootstrap aggreagation，通过在训练样本集中进行有放回的采样得到多个采样集，基于每个采样集训练出一个基学习器，再将基学习器结合。随机森林在对决策树进行bagging的基础上，在决策树的训练过程中引入了随机属性选择。传统决策树在选择划分属性的时候是在当前节点属性集合中选择最优属性，而随机森林则是对结点先随机选择包含k个属性的子集，再选择最有属性，k作为一个参数控制了随机性的引入程度。

另外，GBDT训练是基于Boosting思想，每一迭代中根据错误更新样本权重，因此是串行生成的序列化方法，而随机森林是bagging的思想，因此是并行化方法。

29、如何解决类别不平衡问题？

有些情况下训练集中的样本分布很不平衡，例如在肿瘤检测等问题中，正样本的个数往往非常的少。从线性分类器的角度，在用 y = wx +b 对新样本进行分类的时候，事实上在用预测出的y值和一个y值进行比较，例如常常在y>0.5的时候判为正例，否则判为反例。几率y/1-y反映了正例可能性和反例可能性的比值，阈值0.5恰好表明分类器认为正反的可能性相同。在样本不均衡的情况下，应该是分类器的预测几率高于观测几率就判断为正例，因此应该是 y/1-y > m+/m- 时预测为正例，这种策略称为rebalancing。但是训练集并不一定是真实样本总体的无偏采样，通常有三种做法，一种是对训练集的负样本进行欠采样，第二种是对正例进行升采样，第三种是直接基于原始训练集进行学习，在预测的时候再改变阈值，称为阈值移动。注意过采样一般通过对训练集的正例进行插值产生额外的正例，而欠采样将反例划分为不同的集合供不同的学习器使用。

30、 解释对偶的概念。

一个优化问题可以从两个角度进行考察，一个是primal 问题，一个是dual 问题，就是对偶问题，一般情况下对偶问题给出主问题最优值的下界，在强对偶性成立的情况下由对偶问题可以得到主问题的最优下界，对偶问题是凸优化问题，可以进行较好的求解，SVM中就是将primal问题转换为dual问题进行求解，从而进一步引入核函数的思想。

31、如何进行特征选择？

特征选择是一个重要的数据预处理过程，主要有两个原因，首先在现实任务中我们会遇到维数灾难的问题(样本密度非常稀疏)，若能从中选择一部分特征，那么这个问题能大大缓解，另外就是去除不相关特征会降低学习任务的难度，增加模型的泛化能力。冗余特征指该特征包含的信息可以从其他特征中推演出来，但是这并不代表该冗余特征一定没有作用，例如在欠拟合的情况下也可以用过加入冗余特征，增加简单模型的复杂度。

在理论上如果没有任何领域知识作为先验假设那么只能遍历所有可能的子集。但是这显然是不可能的，因为需要遍历的数量是组合爆炸的。一般我们分为子集搜索和子集评价两个过程，子集搜索一般采用贪心算法，每一轮从候选特征中添加或者删除，分别成为前向和后先搜索。或者两者结合的双向搜索。子集评价一般采用信息增益，对于连续数据往往排序之后选择中点作为分割点。

常见的特征选择方式有过滤式，包裹式和嵌入式，filter,wrapper和embedding。Filter类型先对数据集进行特征选择，再训练学习器。Wrapper直接把最终学习器的性能作为特征子集的评价准则，一般通过不断候选子集，然后利用cross-validation过程更新候选特征，通常计算量比较大。嵌入式特征选择将特征选择过程和训练过程融为了一体，在训练过程中自动进行了特征选择，例如L1正则化更易于获得稀疏解，而L2正则化更不容易过拟合。L1正则化可以通过PGD,近端梯度下降进行求解。

32、为什么会产生过拟合，有哪些方法可以预防或克服过拟合？

一般在机器学习中，将学习器在训练集上的误差称为训练误差或者经验误差，在新样本上的误差称为泛化误差。显然我们希望得到泛化误差小的学习器，但是我们事先并不知道新样本，因此实际上往往努力使经验误差最小化。然而，当学习器将训练样本学的太好的时候，往往可能把训练样本自身的特点当做了潜在样本具有的一般性质。这样就会导致泛化性能下降，称之为过拟合，相反，欠拟合一般指对训练样本的一般性质尚未学习好，在训练集上仍然有较大的误差。

欠拟合：一般来说欠拟合更容易解决一些，例如增加模型的复杂度，增加决策树中的分支，增加神经网络中的训练次数等等。

过拟合：一般认为过拟合是无法彻底避免的，因为机器学习面临的问题一般是np-hard,但是一个有效的解一定要在多项式内可以工作，所以会牺牲一些泛化能力。过拟合的解决方案一般有增加样本数量，对样本进行降维，降低模型复杂度，利用先验知识(L1,L2正则化)，利用cross-validation，early stopping等等。

33、什么是偏差与方差？

泛化误差可以分解成偏差的平方加上方差加上噪声。偏差度量了学习算法的期望预测和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，刻画了数据扰动所造成的影响，噪声表达了当前任务上任何学习算法所能达到的期望泛化误差下界，刻画了问题本身的难度。偏差和方差一般称为bias和variance，一般训练程度越强，偏差越小，方差越大，泛化误差一般在中间有一个最小值，如果偏差较大，方差较小，此时一般称为欠拟合，而偏差较小，方差较大称为过拟合。

34、 神经网络的原理，如何进行训练？

神经网络自发展以来已经是一个非常庞大的学科，一般而言认为神经网络是由单个的神经元和不同神经元之间的连接构成，不同的结构构成不同的神经网络。最常见的神经网络一般称为多层前馈神经网络，除了输入和输出层，中间隐藏层的个数被称为神经网络的层数。BP算法是训练神经网络中最著名的算法，其本质是梯度下降和链式法则。

35、介绍卷积神经网络，和 DBN 有什么区别？

卷积神经网络的特点是卷积核，CNN中使用了权共享，通过不断的卷积和下采样得到不同的特征表示，采样层又称为pooling层，基于局部相关性原理进行下采样，在减少数据量的同时保持有用的信息。DBN是深度信念网络，每一层是一个RBM，整个网络可以视为RBM堆叠得到，通常使用无监督逐层训练，从第一层开始，每一层利用上一层的输入进行训练，等各层训练结束之后再利用BP算法对整个网络进行训练。

36、采用 EM 算法求解的模型有哪些，为什么不用牛顿法或梯度下降法？

用EM算法求解的模型一般有GMM或者协同过滤，k-means其实也属于EM。EM算法一定会收敛，但是可能收敛到局部最优。由于求和的项数将随着隐变量的数目指数上升，会给梯度计算带来麻烦。

37、用 EM 算法推导解释 Kmeans。

k-means算法是高斯混合聚类在混合成分方差相等，且每个样本仅指派一个混合成分时候的特例。注意k-means在运行之前需要进行归一化处理，不然可能会因为样本在某些维度上过大导致距离计算失效。k-means中每个样本所属的类就可以看成是一个隐变量，在E步中，我们固定每个类的中心，通过对每一个样本选择最近的类优化目标函数，在M步，重新更新每个类的中心点，该步骤可以通过对目标函数求导实现，最终可得新的类中心就是类中样本的均值。

38、 用过哪些聚类算法，解释密度聚类算法。

k-means算法，聚类性能的度量一般分为两类，一类是聚类结果与某个参考模型比较(外部指标)，另外是直接考察聚类结果(内部指标)。后者通常有DB指数和DI，DB指数是对每个类，找出类内平均距离/类间中心距离最大的类，然后计算上述值，并对所有的类求和，越小越好。类似k-means的算法仅在类中数据构成簇的情况下表现较好，密度聚类算法从样本密度的角度考察样本之间的可连接性，并基于可连接样本不断扩展聚类蔟得到最终结果。DBSCAN(density-based spatial clustering of applications with noise)是一种著名的密度聚类算法，基于一组邻域参数进行刻画，包括邻域，核心对象(邻域内至少包含个对象)，密度直达(j由i密度直达，表示j在i的邻域内，且i是一个核心对象)，密度可达(j由i密度可达，存在样本序列使得每一对都密度直达)，密度相连(xi,xj存在k,i,j均有k可达)，先找出样本中所有的核心对象，然后以任一核心对象作为出发点，找出由其密度可达的样本生成聚类蔟，直到所有核心对象被访问过为止。

39、 聚类算法中的距离度量有哪些？

聚类算法中的距离度量一般用闽科夫斯基距离，在p取不同的值下对应不同的距离，例如p=1的时候对应曼哈顿距离，p=2的情况下对应欧式距离，p=inf的情况下变为切比雪夫距离，还有jaccard距离，幂距离(闽科夫斯基的更一般形式),余弦相似度，加权的距离，马氏距离(类似加权)作为距离度量需要满足非负性，同一性，对称性和直递性，闽科夫斯基在p>=1的时候满足读来那个性质，对于一些离散属性例如{飞机，火车，轮船}则不能直接在属性值上计算距离，这些称为无序属性，可以用VDM(Value Diffrence Metrix)

40、解释贝叶斯公式和朴素贝叶斯分类   
41、 L1和L2正则化的作用   
42、余弦距离是什么