有监督学习和无监督学习的区别  
有监督学习：对具有标记的训练样本进行学习，以尽可能对训练样本集外的数据进行分类预测。（LR,SVM,BP,RF,GBDT）  
无监督学习：对未标记的样本进行训练学习，比发现这些样本中的结构知识。(KMeans,DL)

正则化  
正则化是针对过拟合而提出的，以为在求解模型最优的是一般优化最小的经验风险，现在在该经验风险上加入模型复杂度这一项（正则化项是模型参数向量的范数），并使用一个rate比率来权衡模型复杂度与以往经验风险的权重，如果模型复杂度越高，结构化的经验风险会越大，现在的目标就变为了结构经验风险的最优化，可以防止模型训练过度复杂，有效的降低过拟合的风险。  
奥卡姆剃刀原理，能够很好的解释已知数据并且十分简单才是最好的模型。

过拟合

如果一味的去提高训练数据的预测能力，所选模型的复杂度往往会很高，这种现象称为过拟合。所表现的就是模型训练时候的误差很小，但在[测试](http://lib.csdn.net/base/softwaretest" \o "软件测试知识库)的时候误差很大。  
产生的原因  
过拟合原因  
1. 样本数据的问题。

样本数量太少

抽样方法错误，抽出的样本数据不能有效足够代表业务逻辑或业务场景。比如样本符合正态分布，却按均分分布抽样，或者样本数据不能代表整体数据的分布

样本里的噪音数据干扰过大

2. 模型问题

模型复杂度高 、参数太多

决策树模型没有剪枝

权值学习迭代次数足够多(Overtraining),拟合了训练数据中的噪声和训练样例中没有代表性的特征.

解决方法

1. 样本数据方面。

增加样本数量，对样本进行降维，添加验证数据

抽样方法要符合业务场景

清洗噪声数据

2. 模型或训练问题

控制模型复杂度，优先选择简单的模型，或者用模型融合技术。

利用先验知识，添加正则项。L1正则更加容易产生稀疏解、L2正则倾向于让参数w趋向于0.

交叉验证

不要过度训练，最优化求解时，收敛之前停止迭代。

决策树模型没有剪枝

权值衰减

泛化能力

泛化能力是指模型对未知数据的预测能力

生成模型和判别模型  
1. 生成模型：由数据学习联合概率分布P(X,Y)，然后求出条件概率分布P(Y|X)作为预测的模型，即生成模型：P(Y|X)= P(X,Y)/ P(X)。（朴素贝叶斯、Kmeans）  
生成模型可以还原联合概率分布p(X,Y)，并且有较快的学习收敛速度，还可以用于隐变量的学习  
2. 判别模型：由数据直接学习决策函数Y=f(X)或者条件概率分布P(Y|X)作为预测的模型，即判别模型。（k近邻、决策树、SVM）  
直接面对预测，往往准确率较高，直接对数据在各种程度上的抽象，所以可以简化模型

线性分类器与非线性分类器的区别以及优劣  
如果模型是参数的线性函数，并且存在线性分类面，那么就是线性分类器，否则不是。  
常见的线性分类器有：LR,贝叶斯分类，单层感知机、线性回归  
常见的非线性分类器：决策树、RF、GBDT、多层感知机  
SVM两种都有(看线性核还是高斯核)  
线性分类器速度快、编程方便，但是可能拟合效果不会很好  
非线性分类器编程复杂，但是效果拟合能力强

特征比数据量还大时，选择什么样的分类器？  
线性分类器，因为维度高的时候，数据一般在维度空间里面会比较稀疏，很有可能线性可分  
对于维度很高的特征，你是选择线性还是非线性分类器？  
理由同上  
对于维度极低的特征，你是选择线性还是非线性分类器？  
非线性分类器，因为低维空间可能很多特征都跑到一起了，导致线性不可分

下面是吴恩达的见解：  
1. 如果Feature的数量很大，跟样本数量差不多，这时候选用LR或者是Linear Kernel的SVM  
2. 如果Feature的数量比较小，样本数量一般，不算大也不算小，选用SVM+Gaussian Kernel  
3. 如果Feature的数量比较小，而样本数量很多，需要手工添加一些feature变成第一种情况

ill-condition病态问题  
训练完的模型测试样本稍作修改就会得到差别很大的结果，就是病态问题（这简直是不能用啊）

L1和L2正则的区别，如何选择L1和L2正则  
http://blog.csdn[.NET](http://lib.csdn.net/base/dotnet" \o ".NET知识库)/xbmatrix/article/details/61624196

他们都是可以防止过拟合，降低模型复杂度  
L1是在loss function后面加上 模型参数的1范数（也就是|xi|）  
L2是在loss function后面加上 模型参数的2范数（也就是sigma(xi^2)），注意L2范数的定义是sqrt(sigma(xi^2))，在正则项上没有添加sqrt根号是为了更加容易优化  
L1 会产生稀疏的特征  
L2 会产生更多地特征但是都会接近于0  
L1会趋向于产生少量的特征，而其他的特征都是0，而L2会选择更多的特征，这些特征都会接近于0。L1在特征选择时候非常有用，而L2就只是一种规则化而已。

L1求解

最小角回归[算法](http://lib.csdn.net/base/datastructure" \o "算法与数据结构知识库)：LARS算法

越小的参数说明模型越简单  
过拟合的，拟合会经过曲面的每个点，也就是说在较小的区间里面可能会有较大的曲率，这里的导数就是很大，线性模型里面的权值就是导数，所以越小的参数说明模型越简单。

为什么一些[机器学习](http://lib.csdn.net/base/machinelearning" \o "机器学习知识库)模型需要对数据进行归一化？

http://blog.csdn[.net](http://lib.csdn.net/base/dotnet" \o ".NET知识库)/xbmatrix/article/details/56695825

归一化化就是要把你需要处理的数据经过处理后（通过某种算法）限制在你需要的一定范围内。  
1）归一化后加快了梯度下降求最优解的速度。等高线变得显得圆滑，在梯度下降进行求解时能较快的收敛。如果不做归一化，梯度下降过程容易走之字，很难收敛甚至不能收敛  
2）把有量纲表达式变为无量纲表达式, 有可能提高精度。一些分类器需要计算样本之间的距离（如欧氏距离），例如KNN。如果一个特征值域范围非常大，那么距离计算就主要取决于这个特征，从而与实际情况相悖（比如这时实际情况是值域范围小的特征更重要）

3) 逻辑回归等模型先验假设数据服从正态分布。

哪些机器学习算法不需要做归一化处理？

概率模型不需要归一化，因为它们不关心变量的值，而是关心变量的分布和变量之间的条件概率，如决策树、rf。而像adaboost、gbdt、xgboost、svm、lr、KNN、KMeans之类的最优化问题就需要归一化。

特征向量的归一化方法

线性函数转换，表达式如下：y=(x-MinValue)/(MaxValue-MinValue)

对数函数转换，表达式如下：y=log10 (x)

反余切函数转换 ，表达式如下：y=arctan(x)\*2/PI

减去均值，乘以方差：y=(x-means)/ variance

标准化与归一化的区别  
简单来说，标准化是依照特征矩阵的列处理数据，其通过求z-score的方法，将样本的特征值转换到同一量纲下。归一化是依照特征矩阵的行处理数据，其目的在于样本向量在点乘运算或其他核函数计算相似性时，拥有统一的标准，也就是说都转化为“单位向量”。规则为l2的归一化公式如下：

特征向量的缺失值处理  
1. 缺失值较多.直接将该特征舍弃掉，否则可能反倒会带入较大的noise，对结果造成不良影响。  
2. 缺失值较少,其余的特征缺失值都在10%以内，我们可以采取很多的方式来处理:

1) 把NaN直接作为一个特征，假设用0表示；

2) 用均值填充；

3) 用随机森林等算法预测填充

随机森林如何处理缺失值（http://charleshm.github.io/2016/03/Random-Forest-Tricks/）

方法一（na.roughfix）简单粗暴，对于训练集,同一个class下的数据，如果是分类变量缺失，用众数补上，如果是连续型变量缺失，用中位数补。  
方法二（rfImpute）这个方法计算量大，至于比方法一好坏？不好判断。先用na.roughfix补上缺失值，然后构建森林并计算proximity matrix，再回头看缺失值，如果是分类变量，则用没有缺失的观测实例的proximity中的权重进行投票。如果是连续型变量，则用proximity矩阵进行加权平均的方法补缺失值。然后迭代4-6次，这个补缺失值的思想和KNN有些类似12。

随机森林如何评估特征重要性（http://charleshm.github.io/2016/03/Random-Forest-Tricks/）

衡量变量重要性的方法有两种，Decrease GINI 和 Decrease Accuracy：  
1) Decrease GINI： 对于回归问题，直接使用argmax(Var−VarLeft−VarRight)作为评判标准，即当前节点训练集的方差Var减去左节点的方差VarLeft和右节点的方差VarRight。  
2) Decrease Accuracy：对于一棵树Tb(x)，我们用OOB样本可以得到测试误差1；然后随机改变OOB样本的第j列：保持其他列不变，对第j列进行随机的上下置换，得到误差2。至此，我们可以用误差1-误差2来刻画变量j的重要性。基本思想就是，如果一个变量j足够重要，那么改变它会极大的增加测试误差；反之，如果改变它测试误差没有增大，则说明该变量不是那么的重要。

优化Kmeans  
使用kd树或者ball tree(这个树不懂)  
将所有的观测实例构建成一颗kd树，之前每个聚类中心都是需要和每个观测点做依次距离计算，现在这些聚类中心根据kd树只需要计算附近的一个局部区域即可

KMeans初始类簇中心点的选取  
k-means++算法选择初始seeds的基本思想就是：初始的聚类中心之间的相互距离要尽可能的远。  
1. 从输入的数据点集合中随机选择一个点作为第一个聚类中心  
2. 对于数据集中的每一个点x，计算它与最近聚类中心(指已选择的聚类中心)的距离D(x)  
3. 选择一个新的数据点作为新的聚类中心，选择的原则是：D(x)较大的点，被选取作为聚类中心的概率较大  
4. 重复2和3直到k个聚类中心被选出来  
5. 利用这k个初始的聚类中心来运行标准的k-means算法

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。

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：一是减少特征数量、降维，使模型泛化能力更强，减少过拟合;二是增强对特征和特征值之间的理解  
常见的特征选择方式：

1. 去除方差较小的特征

2. 正则化。1正则化能够生成稀疏的模型。L2正则化的表现更加稳定，由于有用的特征往往对应系数非零。

3. 随机森林，对于分类问题，通常采用基尼不纯度或者信息增益，对于回归问题，通常采用的是方差或者最小二乘拟合。一般不需要feature engineering、调参等繁琐的步骤。它的两个主要问题，1是重要的特征有可能得分很低（关联特征问题），2是这种方法对特征变量类别多的特征越有利（偏向问题）。

4. 稳定性选择。是一种基于二次抽样和选择算法相结合较新的方法，选择算法可以是回归、SVM或其他类似的方法。它的主要思想是在不同的数据子集和特征子集上运行特征选择算法，不断的重复，最终汇总特征选择结果，比如可以统计某个特征被认为是重要特征的频率（被选为重要特征的次数除以它所在的子集被测试的次数）。理想情况下，重要特征的得分会接近100%。稍微弱一点的特征得分会是非0的数，而最无用的特征得分将会接近于0。

数据预处理

1. 缺失值，填充缺失值fillna：

i. 离散：None,

ii. 连续：均值。

iii. 缺失值太多，则直接去除该列

2. 连续值：离散化。有的模型（如决策树）需要离散值

3. 对定量特征二值化。核心在于设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0。如图像操作

4. 皮尔逊相关系数，去除高度相关的列

#correlation matrix

corrmat = df\_train.corr()

f, ax = plt.subplots(figsize=(12, 9))

sns.heatmap(corrmat, vmax=.8, square=True);

1. 去除噪声点。通过matplotlib.scatter函数图示某一特征与预测特征的点分布图，明显看出噪声点，去除即可

#bivariate analysis saleprice/grlivareavar = 'GrLivArea'

data = pd.concat([df\_train['SalePrice'], df\_train[var]], axis=1)

data.plot.scatter(x=var, y='SalePrice', ylim=(0,800000));

df\_train.sort\_values(by = 'GrLivArea', ascending = False)[:2]

df\_train = df\_train.drop(df\_train[df\_train['Id'] == 1299].index)

df\_train = df\_train.drop(df\_train[df\_train['Id'] == 524].index)

1. 标签编码，把字符类别特征编码成数值类型，如红绿蓝编码为0、1、2

2. 归一化。将数据按比例缩放，使这些数据落入到一个较小的特定的区间之内。

i. 最小最大缩放

ii. b.Z-score标准化。让数据服从基于 μ=0 和 σ=1的标准正态分布

3. OneHot编码get\_dummies。有的离散特征如颜色需OneHot编码，编码后需特征对齐，因为进行one-hot编码后，会出现一种情况就是：某个特征的某一个取值只出现在训练集中，没有出现在测试集中，或者相反

4. 产生K折交叉数据集：sklearn.cross\_validation.KFold

5. 填充均值前，通过对数运算修正数据基本符合正态分布

i.

查看目标数据统计，count,mean, min, max, std

ii.

#descriptive statistics summary

df\_train['SalePrice'].describe()

iii.

iv.

查看目标数据直方图分布,。如果类似正态分布，计算偏离正态分布程度：偏度和峰度

v.

#histogramsns.distplot(df\_train['SalePrice']);

vi.

#log transform skewed numeric features:

numeric\_feats = all\_data.dtypes[all\_data.dtypes != "object"].index

skewed\_feats = train[numeric\_feats].apply(lambda x: skew(x.dropna())) #compute skewness

skewed\_feats = skewed\_feats[skewed\_feats > 0.75]

skewed\_feats = skewed\_feats.index

all\_data[skewed\_feats] = np.log1p(all\_data[skewed\_feats])

vii.

最小最大缩放：  
当使用基于距离的算法时，我们必须尝试将数据缩放，这样较不重要的特征不会因为自身较大的范围而主导目标函数。如KNN。  
在逻辑回归中，每个特征都被分配了权重或系数(Wi)。如果某个特征有相对来说比较大的范围，而且其在目标函数中无关紧要，那么逻辑回归模型自己就会分配一个非常小的值给它的系数，从而中和该特定特征的影响优势，而基于距离的方法，如KNN，没有这样的内置策略，因此需要缩放。  
Z-score标准化:  
很多的机器学习技巧/模型（例如L1，L2正则项，向量空间模型-Vector Space Model，欧几里得距离测量的KNN、k-均值、SVM、感知器、神经网络、线性判别分析、主成分分析）都基于这样的假设：所有的属性取值都差不多是以0为均值且取值范围相近的，标准化数据帮助我们提高预测模型的精度。  
备注：在缩放和标准化中二选一是个令人困惑的选择，你必须对数据和要使用的学习模型有更深入的理解，才能做出决定。对于初学者，你可以两种方法都尝试下并通过交叉验证精度来做出选择。

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。

偏差：

方差：

解决bias和Variance问题的方法：

交叉验证  
High bias解决方案:Boosting、复杂模型（非线性模型、增加神经网络中的层）、更多特征  
High Variance解决方案：agging、简化模型、降维

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。

SVM、LR、决策树的对比？  
模型复杂度：SVM支持核函数，可处理线性非线性问题;LR模型简单，训练速度快，适合处理线性问题;决策树容易过拟合，需要进行剪枝  
损失函数：SVM hinge loss; LR L2正则化; adaboost 指数损失  
数据敏感度：SVM添加容忍度对outlier不敏感，只关心支持向量，且需要先做归一化; LR对远点敏感  
数据量：数据量大就用LR，数据量小且特征少就用SVM非线性核  
  
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。

xgboost怎么给特征评分？

在训练的过程中，通过Gini指数选择分离点的特征，一个特征被选中的次数越多，那么该特征评分越高。

[python] [view plain](http://blog.csdn.net/xbmatrix/article/details/62056589" \o "view plain) [copy](http://blog.csdn.net/xbmatrix/article/details/62056589" \o "copy) [print](http://blog.csdn.net/xbmatrix/article/details/62056589" \o "print)[?](http://blog.csdn.net/xbmatrix/article/details/62056589" \o "?)

1. # feature importance

2. print(model.feature\_importances\_)

3. # plot

4. pyplot.bar(range(len(model.feature\_importances\_)), model.feature\_importances\_)

5. pyplot.show()

6. ==========

7. # plot feature importance

8. plot\_importance(model)

9. pyplot.show()

# feature importance

print(model.feature\_importances\_)

# plot

pyplot.bar(range(len(model.feature\_importances\_)), model.feature\_importances\_)

pyplot.show()

==========

# plot feature importance

plot\_importance(model)

pyplot.show()

什么是OOB？随机森林中OOB是如何计算的，它有什么优缺点？

bagging方法中Bootstrap每次约有1/3的样本不会出现在Bootstrap所采集的样本集合中，当然也就没有参加决策树的建立，把这1/3的数据称为袋外数据oob（out of bag）,它可以用于取代测试集误差估计方法。  
袋外数据(oob)误差的计算方法如下：  
对于已经生成的随机森林,用袋外数据测试其性能,假设袋外数据总数为O,用这O个袋外数据作为输入,带进之前已经生成的随机森林分类器,分类器会给出O个数据相应的分类,因为这O条数据的类型是已知的,则用正确的分类与随机森林分类器的结果进行比较,统计随机森林分类器分类错误的数目,设为X,则袋外数据误差大小=X/O;这已经经过证明是无偏估计的,所以在随机森林算法中不需要再进行交叉验证或者单独的测试集来获取测试集误差的无偏估计。

解释置信区间

置信区间不能用贝叶斯学派的概率来描述，它属于频率学派的范畴。真值要么在，要么不在。由于在频率学派当中，真值是一个常数，而非随机变量（后者是贝叶斯学派），所以我们不对真值做概率描述。比如，95%置信区间，并不是真值在这个区间内的概率是95%，而应该为100次随机抽样中构造的100个区间如果95次包含了参数真值，那么置信度为95%。

监督学习一般使用两种类型的目标变量：标称型和数值型  
标称型：标称型目标变量的结果只在有限目标集中取值，如真与假(标称型目标变量主要用于分类)  
数值型：数值型目标变量则可以从无限的数值集合中取值，如0.100，42.001等 (数值型目标变量主要用于回归分析)

为什么说朴素贝叶斯是高偏差低方差？  
它简单的假设了各个特征之间是无关的，是一个被严重简化了的模型。所以，对于这样一个简单模型，大部分场合都会bias部分大于variance部分，也就是高偏差，低方差

1.什么是机器学习

机器学习是为了应对系统程序设计，属于计算机科学类的学科，它能根据经验进行自动学习和提高。例如：一个由程序操纵的机器人，它能根据从传感器搜集到的数据，完成一系列的任务和工作。它能根据数据自动地学习应用程序。

2.机器学习与数据挖掘的区别

机器语言是指在没有明确的程序指令的情况下，给予计算机学习能力，使它能自主的学习、设计和扩展相关算法。数据挖掘则是一种从非结构化数据里面提取知识或者未知的、人们感兴趣的图片。在这个过程中应用了机器学习算法。

3.什么是机器学习的过度拟合现象

在机器学习中，当一个统计模型首先描述随机误差或噪声，而不是自身的基本关系时，过度拟合就会出现。当一个模型是过于复杂，过拟合通常容易被发现，因为相对于训练数据类型的数量，参数的数量过于五花八门。那么这个模型由于过度拟合而效果不佳。

4.过度拟合产生的原因

由于用于训练模型的标准并不等同于判断模型效率的标准，这导致了产生过度拟合的可能性。

5.如何避免过度拟合

当你使用较小的数据集进行机器学习时，容易产生过度拟合，因此使用较大的数据量能避免过度拟合现象。但是，当你不得不使用小型数据集进行建模时，可以使用被称为交叉验证的技术。在这种方法中数据集被分成两节，测试和训练数据集，测试数据集只测试模型，而在训练数据集中，数据点被用来建模。

在该技术中，一个模型通常是被给定有先验知识的数据集（训练数据集）进行训练，没有先验知识的数据集进行测试。交叉验证的思想是：在训练阶段，定义一个数据集用来测试模型。

6.什么是感应式的机器学习？

感应机器学习涉及由实践进行学习的过程，能从一组可观测到的例子的尝试推导出普遍性规则。

7.什么是机器学习的五个流行的算法？

00001. 决策树2. 神经网络（反向传播）3. 概率网络4.最邻近法5. 支持向量机

8.机器学习有哪些不同的算法技术？

在机器学习不同类型的算法技术是：

00001. 监督学习2.非监督学习3. 半监督学习4. 转导推理（Transduction）5.学习推理（Learning to Learn）。

9.在机器学习中，建立假设或者模型的三个阶段指的是什么？

1.建模2.模型测试3.模型应用。

10.什么是监督学习的标准方法？

监督学习的标准方法是将一组示例数据的分成训练数据集和测试数据集。

11.什么是训练数据集和测试数据集？

在类似于机器学习的各个信息科学相关领域中，一组数据被用来发现潜在的预测关系，称为“训练数据集”。训练数据集是提供给学习者的案例，而试验数据集是用于测试由学习者提出的假设关系的准确度。

12.下面列出机器学习的各种方法？

机器学习的各种方法如下“

1.概念与分类学习（Concept Vs Classification Learning）。

2.符号与统计学习（Symbolic Vs Statistical Learning）。

3.归纳与分析学习（Inductive Vs Analytical Learning）。

13.非机器学习有哪些类型？

人工智能、规则推理。

14.什么是非监督学习的功能？  
1.求数据的集群2. 求出数据的低维表达3. 查找数据有趣的方向4. 有趣的坐标和相关性5.发现显著的观测值和数据集清理

15.什么是监督学习的功能？

1.分类、2.语音识别3.回归4.时间序列预测5. 注释字符串

16.什么是算法独立的机器学习？

机器学习在基础数学领域独立于任何特定分类器或者学习算法，被称为算法独立的机器学习。

17.人工智能与机器学习的区别？

基于经验数据的特性而设计和开发的算法被称为机器学习。而人工智能不但包括机器学习，还包括诸如知识表示，自然语言处理，规划，机器人技术等其它方法。

18.在机器学习中分类器指的是什么？

在机器学习中，分类器是指输入离散或连续特征值的向量，并输出单个离散值或者类型的系统。

19.朴素贝叶斯方法的优势是什么？

朴素贝叶斯分类器将会比判别模型，譬如逻辑回归收敛得更快，因此你只需要更少的训练数据。其主要缺点是它学习不了特征间的交互关系。

20.在哪些领域使用模式识别技术？

模式识别被应用在：

00001. 计算机视觉2.语言识别3.统计4.数据挖掘5. 非正式检索6. 生物信息学。

21.什么是遗传编程？

遗传编程的机器学习中两种常用的方法之一。该模型是基于测试，并在一系列的结果当中，获取最佳选择。

22.在机器学习中归纳逻辑程序设计是指什么？

归纳逻辑程序设计（ILP）是利用逻辑程序设计表达的背景知识和实例，它是机器学习的一个分支。

23.在机器学习中，模型的选择是指？

在不同的数学模型中，选择用于描述相同的数据集的模型的过程被称为模型选择。模型选择吧被应用于统计，机器学习和数据挖掘的等相关领域。  
24.用于监督学习校准两种方法是什么？

在监督学习中，用于预测良好概率的两种方法是：

00001. 普拉特校准，2. 保序回归。

这些方法被设计为二元分类，而且有意义的。  
25. 什么方法通常用于防止过拟合？

当有足够的数据进行等渗回归时，这通常被用来防止过拟合问题。

26.规则学习的启发式方法和决策树的启发式方法之间的区别是什么？

决策树的启发式方法评价的是一系列不相交的集合的平均质量；然而规则学习的启发式方法仅仅评价在候选规则覆盖下的实例集。

27.什么是感知机器学习？

在机器学习，感知器是一种输入到几个可能的非二进制输出的监督分类算法。

28.贝叶斯逻辑程序的两个组成部分是什么？

贝叶斯逻辑程序由两部分组成。第一成分由一组贝叶斯条款组成，能捕捉特定域的定性结构。第二组分是定量的，它能对域的量化信息进行编码。

29.什么是贝叶斯网络？

贝叶斯网络是用来表示一组变量之间为概率关系的图像模型。

30.为什么基于实例的学习算法有时也被称为懒惰学习算法？

基于实例的学习算法也被称为懒惰学习算法，因为它们延缓诱导或泛化过程，直到分类完成。

31.支持向量机能处理哪两种分类方法？

1.结合二分类法2. 修改二进制纳入多类学习法。

32.什么是集成学习？

为了解决特定的计算程序，如分类器或专家知识等多种模式，进行战略性生产和组合。这个过程被称为集成学习。

33.为什么集成学习被应用？

集成学习能提高模型的分类，预测，函数逼近等方面的精度。

34.什么使用集成学习？

当你构建一个更准确，相互独立的分类器时，使用集成学习。

35.什么是集成方法的两种范式？

集成方法的两种范式是：

00001. 连续集成方法2. 并行集成方法。

36.什么是集成方法的一般原则，在集成方法中套袋（bagging）和爆发（boosting）指的是什么？

集成方法的一般原则是要结合定的学习算法多种预测模型，相对于单一模型，其有更强的健壮性。套袋是一种能提高易变的预测或分类方案集成方法。爆发方法被依次用来减少组合模型的偏差。爆发和装袋都可以通过降低方差减少误差。

37.什么是集成方法分类错误的偏置方差分解？

学习算法的期望误差可以分解为偏差和方差。偏置项衡量由学习方法产生的平均分类器与目标函数是否匹配。

38.在集成方法中什么是增量合成方法？

增量学习方法是一种从新数据进行学习，并能应用于后续由现有的数据集生成的分类器的算法。

39.PCA，KPCA和ICE如何使用？

PCA（主成分分析），KPCA（基于内核主成分分析）和ICA（独立成分分析）是用于降维的重要特征提取技术。

40.在机器学习中降维是什么意思？

在机器学习和统计应用中，降维是指在计算时减少随机变量数目的处理过程，并且可以分为特征选择和特征提取。

41.什么是支持向量机？

支持向量机是一种监督学习算法，适用于分类和回归分析。

42.关系评价技术的组成部分是什么？

关系评价技术的重要组成部分如下：

1.数据采集2. 地面实况采集3. 交叉验证技术4. 查询类型5. 评分标准6. 显着性检验。

43.连续监督学习有什么不同方法？

连续监督学习问题的不同解决办法如下：

00001. 滑动窗口方法2. 复发性推拉窗3. 隐藏马尔科夫模型4. 最大熵马尔科夫模型5. 条件随机域6. 图变换网络。

44.在机器人技术和信息处理技术的哪些方面会相继出现预测问题？

在机器人技术和信息处理技术中，相继出现预测问题的是：

00001. 模仿学习2. 结构预测3. 基于模型的强化学习。

45.什么是批量统计学习？

统计学习技术允许根据一组观察到的数据进行学习功能和预测，这可以对无法观察和未知的数据进行预测。这些技术提供的学习预测器对未来未知数据的预测提供性能保证。

46什么是PAC学习？

可能近似正确模型 (PAC) 学习是一个已经被引入到分析学习算法和统计效率的学习框架。

47有哪些不同的类别可以分为序列学习过程？

序列预测2. 序列生成3. 序列识别4. 顺序决定.

48什么是序列学习？

序列学习是一种以合乎逻辑的方式进行教学和学习的方法。

49.机器学习的两种技术是什么？

机器学习的两种技术是：  
1.遗传编程2.归纳学习

50.你在日常工作中看到的机器学习的一个流行应用是什么？

各大电商网站上已部署好的推荐引擎使用的是机器学习。

　　问1：你会在时间序列数据集上使用什么交叉验证技术？是用k倍或LOOCV？

　　答：都不是。对于时间序列问题，k倍可能会很麻烦，因为第4年或第5年的一些模式有可能跟第3年的不同，而对数据集的重复采样会将分离这些趋势，而我们最终可能只是需要对过去几年的进行验证，这就不能用这种方法了。相反，我们可以采用如下所示的5倍正向链接策略：

　　fold 1 : training [1], test [2]

　　fold 2 : training [1 2], test [3]

　　fold 3 : training [1 2 3], test [4]

　　fold 4 : training [1 2 3 4], test [5]

　　fold 5 : training [1 2 3 4 5], test [6]

　　1，2，3，4，5，6代表的是年份。

    问2：你是怎么理解偏差方差的平衡的？

　　答：从数学的角度来看，任何模型出现的误差可以分为三个部分。以下是这三个部分：

　　偏差误差在量化平均水平之上，预测值跟实际值相差多远时有用。高偏差误差意味着我们的模型表现不太好，因为没有抓到重要的趋势。而另一方面，方差量化了在同一个观察上进行的预测是如何彼此不同的。高方差模型会过度拟合你的训练集，而在训练集以外的数据上表现很差。

　　问3：给你一个有1000列和1百万行的训练数据集，这个数据集是基于分类问题的。经理要求你来降低该数据集的维度以减少模型计算时间，但你的机器内存有限。你会怎么做？（你可以自由做各种实际操作假设。）

　　答：你的面试官应该非常了解很难在有限的内存上处理高维的数据。以下是你可以使用的处理方法：

　　1.由于我们的RAM很小，首先要关闭机器上正在运行的其他程序，包括网页浏览器等，以确保大部分内存可以使用。

　　2.我们可以随机采样数据集。这意味着，我们可以创建一个较小的数据集，比如有1000个变量和30万行，然后做计算。

　　3.为了降低维度，我们可以把数值变量和分类变量分开，同时删掉相关联的变量。对于数值变量，我们将使用相关性分析；对于分类变量，我们可以用卡方检验。

　　4.另外，我们还可以使用PCA（主成分分析），并挑选可以解释在数据集中有最大偏差的成分。

　　5.利用在线学习[算法](http://lib.csdn.net/base/datastructure" \o "算法与数据结构知识库)，如VowpalWabbit（在[Python](http://lib.csdn.net/base/python" \o "Python知识库)中可用）是一个不错的选择。

　　6.利用Stochastic GradientDescent（随机梯度下降法）建立线性模型也很有帮助。

　　7.我们也可以用我们对业务的理解来估计各预测变量对响应变量的影响的大小。但是，这是一个主观的方法，如果没有找出有用的预测变量可能会导致信息的显著丢失。

    问4：全球平均温度的上升导致世界各地的海盗数量减少。这是否意味着海盗的数量减少引起气候变化？

　　答：不能够这样说。这是一个“因果关系和相关性”的经典案例。全球平均温度和海盗数量之间有可能有相关性，但基于这些信息，我们不能说因为全球平均气温的上升而导致了海盗的消失。我们不能断定海盗的数量减少是引起气候变化的原因，因为可能有其他因素（潜伏或混杂因素）影响了这一现象。

　　问5：给你一个数据集，这个数据集有缺失值，且这些缺失值分布在离中值有1个标准偏差的范围内。百分之多少的数据不会受到影响？为什么？

　　答：约有32%的数据将不受缺失值的影响。因为，由于数据分布在中位数附近，让我们先假设这是一个正态分布。我们知道，在一个正态分布中，约有68%的数据位于跟平均数（或众数、中位数）1个标准差范围内，那么剩下的约32%的数据是不受影响的。因此，约有32%的数据将不受缺失值的影响。

    问6：你意识到你的模型受到低偏差和高方差问题的困扰。那么，应该使用哪种算法来解决问题呢？为什么？

　　答：可以使用bagging算法（如随机森林）。因为，低偏差意味着模型的预测值接近实际值，换句话说，该模型有足够的灵活性，以模仿训练数据的分布。这样貌似很好，但是别忘了，一个灵活的模型没有泛化能力，意味着当这个模型用在对一个未曾见过的数据集进行[测试](http://lib.csdn.net/base/softwaretest" \o "软件测试知识库)的时候，它会令人很失望。在这种情况下，我们可以使用bagging算法（如随机森林），以解决高方差问题。bagging算法把数据集分成重复随机取样形成的子集。然后，这些样本利用单个学习算法生成一组模型。接着，利用投票（分类）或平均（回归）把模型预测结合在一起。另外，为了应对大方差，我们可以：

　　1.使用正则化技术，惩罚更高的模型系数，从而降低了模型的复杂性。

　　2.使用可变重要性图表中的前n个特征。可以用于当一个算法在数据集中的所有变量里很难寻找到有意义信号的时候。

　　问7：协方差和相关性有什么区别？

　　答：相关性是协方差的标准化格式。协方差本身很难做比较。例如：如果我们计算工资（$）和年龄（岁）的协方差，因为这两个变量有不同的度量，所以我们会得到不能做比较的不同的协方差。为了解决这个问题，我们计算相关性来得到一个介于-1和1之间的值，就可以忽略它们各自不同的度量。

　　问8：真阳性率和召回有什么关系？写出方程式。

　　答：真阳性率=召回。它们有相同的公式（TP / TP + FN）。

　　问9：Gradient boosting算法（GBM）和随机森林都是基于树的算法，它们有什么区别？

　　答：最根本的区别是，随机森林算法使用bagging技术做出预测；而GBM是采用boosting技术做预测的。在bagging技术中，数据集用随机采样的方法被划分成n个样本。然后，使用单一的学习算法，在所有样本上建模。接着利用投票或者求平均来组合所得到的预测。bagging是平行进行的，而boosting是在第一轮的预测之后，算法将分类出错的预测加高权重，使得它们可以在后续一轮中得到校正。这种给予分类出错的预测高权重的顺序过程持续进行，一直到达到停止标准为止。随机森林通过减少方差（主要方式）提高模型的精度。生成树之间是不相关的，以把方差的减少最大化。在另一方面，GBM提高了精度，同时减少了模型的偏差和方差。

    问10：你认为把分类变量当成连续型变量会更得到一个更好的预测模型吗？

　　答：为了得到更好的预测，只有在分类变量在本质上是有序的情况下才可以被当做连续型变量来处理。

    问11：“买了这个的客户，也买了......”亚马逊的建议是哪种算法的结果？

　　答：这种推荐引擎的基本想法来自于协同过滤。协同过滤算法考虑用于推荐项目的“用户行为”。它们利用的是其他用户的购买行为和针对商品的交易历史记录、评分、选择和购买信息。针对商品的其他用户的行为和偏好用来推荐项目（商品）给新用户。在这种情况下，项目（商品）的特征是未知的。

　　问12：在k-means或kNN，我们是用欧氏距离来计算最近的邻居之间的距离。为什么不用曼哈顿距离？

　　答：我们不用曼哈顿距离，因为它只计算水平或垂直距离，有维度的限制。另一方面，欧氏距离可用于任何空间的距离计算问题。因为，数据点可以存在于任何空间，欧氏距离是更可行的选择。例如：想象一下国际象棋棋盘，象或车所做的移动是由曼哈顿距离计算的，因为它们是在各自的水平和垂直方向做的运动。

　　问13：我知道校正R2或者F值是用来评估线性回归模型的。那用什么来评估逻辑回归模型？

　　答：我们可以使用下面的方法：

　　1.由于逻辑回归是用来预测概率的，我们可以用AUC-ROC曲线以及混淆矩阵来确定其性能。

　　2.此外，在逻辑回归中类似于校正R2的指标是AIC。AIC是对模型系数数量惩罚模型的拟合度量。因此，我们更偏爱有最小AIC的模型。

　　3.空偏差指的是只有截距项的模型预测的响应。数值越低，模型越好。残余偏差表示由添加自变量的模型预测的响应。数值越低，模型越好。

　　问14：为什么朴素贝叶斯如此“朴素”？

　　答：因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知，这个假设在现实世界中是很不真实的，因此，说朴素贝叶斯真的很“朴素”。

　　问15：花了几个小时后，现在你急于建一个高精度的模型。结果，你建了5 个GBM（Gradient Boosted Models），想着boosting算法会展现“魔力”。不幸的是，没有一个模型比基准模型表现得更好。最后，你决定将这些模型结合到一起。尽管众所周知，结合模型通常精度高，但你就很不幸运。你到底错在哪里？

　　答：据我们所知，组合的学习模型是基于合并弱的学习模型来创造一个强大的学习模型的想法。但是，只有当各模型之间没有相关性的时候组合起来后才比较强大。由于我们已经试了5个GBM也没有提高精度，表明这些模型是相关的。具有相关性的模型的问题是，所有的模型提供相同的信息。例如：如果模型1把User1122归类为1，模型2和模型3很有可能会做同样的分类，