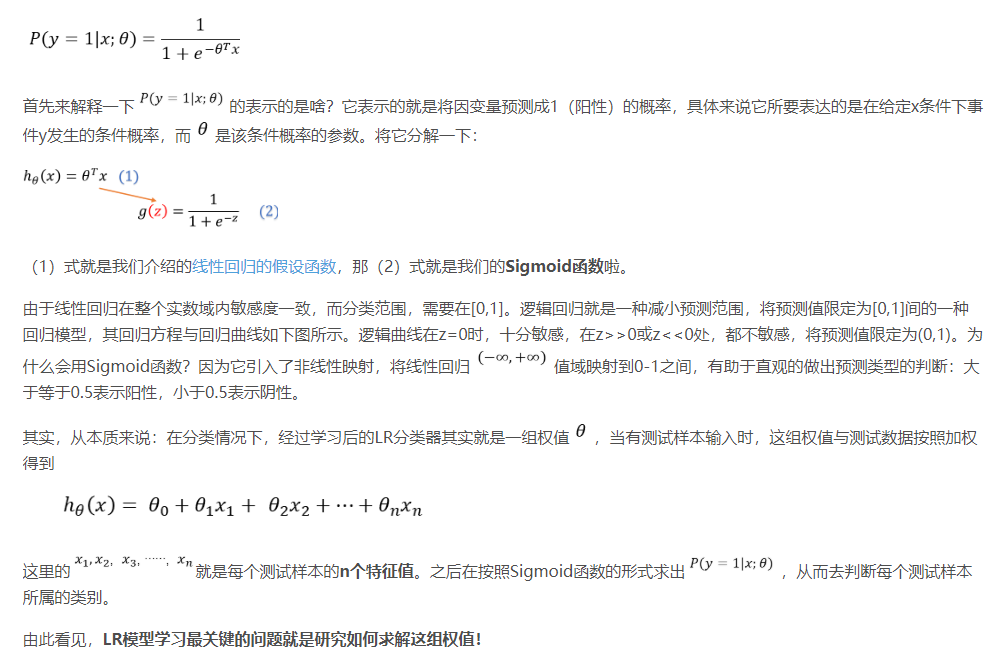
**分类：**

LR（Logistic Regression）虽然是回归模型，但却是经典的分类方法。



LR 回归是当前业界比较常用的机器学习方法，用于估计某种事物的可能性。它与多元线性回归同属一个家族，即广义线性模型。简单来说多元线性回归是直接将特征值和其对应的概率进行相乘得到一个结果，逻辑回归则是在这样的结果上加上一个逻辑函数。在此选择LR 作为回归分析模型的代表进行介绍。



LR算法的优点

1.对数据中小噪声的鲁棒性好；

2.LR 算法已被广泛应用于工业问题中；

3.多重共线性并不是问题，它可结合正则化来解决。

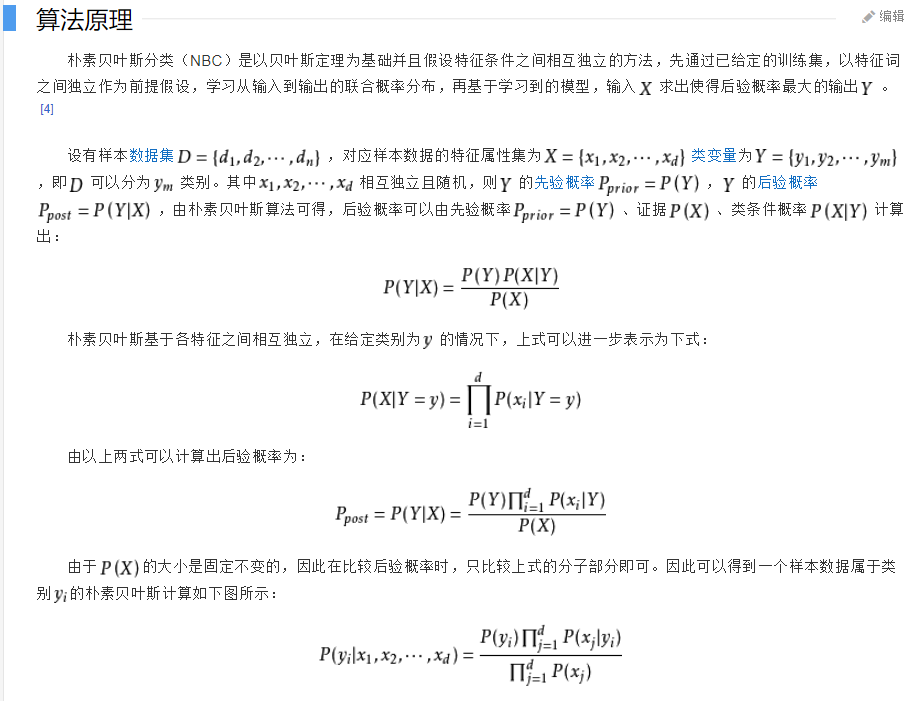
LR算法的缺点

1.对于非线性特征，需要转换

2.当特征空间很大时，LR的性能并不是太好

NBC（Naive Bayesian Classifier，朴素贝叶斯分类）算法。

朴素贝叶斯方法是在[贝叶斯](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF/1405899" \t "https://baike.baidu.com/item/%E6%9C%B4%E7%B4%A0%E8%B4%9D%E5%8F%B6%E6%96%AF/_blank)算法的基础上进行了相应的简化，即假定给定目标值时属性之间相互条件独立。也就是说没有哪个属性变量对于决策结果来说占有着较大的比重，也没有哪个属性变量对于决策结果占有着较小的比重。虽然这个简化方式在一定程度上降低了贝叶斯分类算法的分类效果，但是在实际的应用场景中，极大地简化了贝叶斯方法的复杂性。



NBC 模型发源于古典数学理论，有着坚实的数学基础。该算法是基于条件独立性假设的一种算法，当条件独立性假设成立时，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。

NBC算法的优点

1. 算法逻辑简单，易于实现；
2. 算法所需估计的参数很少；
3. 算法对缺失数据不太敏感；
4. 算法具有较小的误差分类率；
5. 算法性能稳定，健壮性比较好；

NBC算法的缺点

1.在属性个数比较多或者属性之间相关性较大时，NBC 模型的分类效果相对较差；

2.算法是基于条件独立性假设的，在实际应用中很难成立，故会影响分类效果

SVM（支持向量机）、

SVM 算法是建立在统计学习理论基础上的机器学习方法，为十大数据挖掘算法之一。通过学习算法，SVM 可以自动寻找出对分类有较好区分能力的支持向量，由此构造出的分类器可以最大化类与类的间隔，因而有较好的适应能力和较高的分准率。SVM 算法的目的在于寻找一个超平面H，该超平面可以将训练集中的数据分开，且与类域边界的沿垂直于该超平面方向的距离最大，故SVM 法亦被称为最大边缘算法。

SVM算法的优点

1.SVM 模型有很高的分准率；

2. SVM 模型有很高的泛化性能；

3. SVM 模型能很好地解决高维问题；

4. SVM 模型对小样本情况下的机器学习问题效果好。

SVM算法的缺点

1.SVM 模型对缺失数据敏感；

2.对非线性问题没有通用解决方案，得谨慎选择核函数来处理。

ID3算法

ID3 算法是一种基于决策树的分类算法，该算法是以信息论为基础，以信息熵和信息增益为衡量标准，从而实现对数据的归纳分类。信息增益用于度量某个属性对样本集合分类的好坏程度。ID3 算法的时间复杂度为O(n\*|D|\*log|D|)。

ID3算法的优点

ID3 算法建立的决策树规模比较小；

查询速度快。

ID3算法的缺点

1.不适合处理连续数据；

2.难以处理海量数据集；

3.建树时偏选属性值较大的进行分离，而有时属性值较大的不一定能反应更多的数据信息。

C4.5 算法

C4.5 算法是ID3 算法的修订版，采用信息增益率来加以改进，选取有最大增益率的分割变量作为准则，避免ID3 算法过度的适配问题。

C4.5算法优点

1.C4.5 继承了ID3 优点；

2.在树构造过程中进行剪枝；

3.能对不完整数据进行处理；

4.能够完成对连续属性的离散化处理；

5.产生的分类规则易于理解，准确率较高；

6.用增益率来选择属性，克服了用增益选择属性时偏向选择取值多的属性。

C4.5 算法缺点

1.构造树时，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效；

2.只适合于能驻留于内存的数据集，当训练集达到内存无法容纳时程序无法运行。

C4.5 用于遥感分类过程中，首先依据通常的方式建立第一个模型。随后建立的第二个模型聚焦于被第一个模型错误分类的记录。以此类推，最后应用整个模型集对样本进行分类，使用加权投票过程把分散的预测合并成综合预测。Boosting 技术对于噪声不大的数据，通常通过建立的多模型来减少错误分类的影响，提高分类精度。

C5.0算法

C5.0 算法是 Quinlan 在C4.5 算法的基础上改进而来的产生决策树的一种更新的算法，它除了包括C4.5 的全部功能外，还引入许多新的技术，其中最重要的技术是提升（Boosting）技术，目的是为了进一步提高决策树对样本的识别率。同时C5.0 的算法复杂度要更低，使用更简单，适应性更强，因此具有更高的使用价值。

C5.0算法的优点

1.C5.0 模型能同时处理连续和离散的数据

2.C5.0 模型估计

模型通常不需要很长的训练时间；

3.C5.0 引入Boosting 技术以提高分类的效率和精度；

4.C5.0 模型易于理解，模型推出的规则有非常直观的解释；

5.C5.0 模型在面对数据遗漏和特征很多的问题时非常稳健。

C5.0算法的缺点

目标字段必须为分类字段。

美国地质调查局(USGS)在进行土地覆盖分类项目过程中研发了支持决策树分类的软件。软件分类模块主要是针对庞大数据量的数据集进行数据挖掘，找出特征，然后建立规则集进行决策分类。在分类模块中采用C5.0 模型来完成决策树分类、形成分类文件，实现遥感影像的分类。

RF（随机森林）、

随机森林（Random Forest）是属于集成学习的一种组合分类算法（确切说是属于bagging），集成学习的核心思想就是将若干个弱（基）分类器组合起来，得到一个分类性能显著优越的强分类器。如果各弱分类器之前没有强依赖关系、可并行生成，就可以使用随机森林算法。

  随机森林利用自主抽样法（bootstrap）从原数据集中有放回地抽取多个样本，对抽取的样本先用弱分类器—决策树进行训练，然后把这些决策树组合在一起，通过投票得出最终的分类或预测结果。

随机森林的生成方法

1.从样本集中通过重采样的方式产生n个样本。

2.建设样本特征数目为a，对n个样本选择a中的k个特征，用建立决策树的方式获得最佳分割点。

3.重复m次，产生m棵决策树。

4.多数投票机制进行预测。

优点：

a）随机森林算法能解决分类与回归两种类型的问题，表现良好，由于是集成学习，方差和偏差都比较低，泛化性能优越；

b）随机森林对于高维数据集的处理能力很好，它可以处理成千上万的输入变量，并确定最重要的变量，因此被认为是一个不错的降维方法。此外，该模型能够输出特征的重要性程度，这是一个非常实用的功能。

c) 可以应对缺失数据；

d）当存在分类不平衡的情况时，随机森林能够提供平衡数据集误差的有效方法；

e ) 高度并行化，易于分布式实现

f) 由于是树模型 ，不需要归一化即可之间使用

缺点：

a）随机森林在解决回归问题时并没有像它在分类中表现的那么好，这是因为它并不能给出一个连续型的输出。当进行回归时，随机森林不能够作出超越训练集数据范围的预测，这可能导致在对某些还有特定噪声的数据进行建模时出现过度拟合。

b）对于许多统计建模者来说，随机森林给人的感觉像是一个黑盒子——你几乎无法控制模型内部的运行，只能在不同的参数和随机种子之间进行尝试。

c) 忽略属性之间的相关性

PS:最后几个重要的点

RF采用多个决策树的投票机制来改善决策树。

为什么不能用全样本去训练m棵决策树？

       答：全样本训练忽视了局部样本的规律，对于模型的泛化能力是有害的（如果有m个决策树，那就需要m个一定数量的样本集来训练每一棵树）

     3.产生n个样本的方法，采用Bootstraping法，这是一种又放回的抽样方法，产生n个样本。

     4.最终采用Bagging的策略来获得，即多数投票机制。

决策树如何分裂节点，三种决策树的差别

**聚类：**

k-means、

经典K-means算法流程：   
1. 随机地选择k个对象，每个对象初始地代表了一个簇的中心；   
2. 对剩余的每个对象，根据其与各簇中心的距离，将它赋给最近的簇；   
3. 重新计算每个簇的平均值，更新为新的簇中心；   
4. 不断重复2、3，直到准则函数收敛。

优点：对于大型数据集也是简单高效、时间复杂度、空间复杂度低。   
缺点：

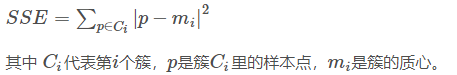
最重要是数据集大时结果容易局部最优；

需要预先设定K值，对最先的K个点选取很敏感；

对噪声和离群值非常敏感；

确定最优聚类个数：

手肘法的评价K值好坏的标准是SSE（sum of the squared errors,误差平方和）



手肘法的核心思想是：随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于最佳聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达最佳聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的最佳聚类数。这也是该方法被称为手肘法的原因。

DBSCAN

两个全局参数：Eps:领域半径，MinPts:在领域中点的最少个数

核心点：领域半径对象个数大于密度阈值MinPts

边界点：领域半径对象个数小于密度阈值MinPts

1.DBSCAN以一个从未访问过的任意起始数据点开始。这个点的邻域是用距离ε（所有在ε距离的点都是邻点）来提取的。

2.如果在这个邻域中有足够数量的点（根据 minPoints），那么聚类过程就开始了，并且当前的数据点成为新聚类中的第一个点。否则，该点将被标记为噪声（稍后这个噪声点可能会成为聚类的一部分）。在这两种情况下，这一点都被标记为“访问（visited）”。

3.对于新聚类中的第一个点，其ε距离附近的点也会成为同一聚类的一部分。这一过程使在ε邻近的所有点都属于同一个聚类，然后重复所有刚刚添加到聚类组的新点。

4.步骤2和步骤3的过程将重复，直到聚类中的所有点都被确定，就是说在聚类附近的所有点都已被访问和标记。

5.一旦我们完成了当前的聚类，就会检索并处理一个新的未访问点，这将导致进一步的聚类或噪声的发现。这个过程不断地重复，直到所有的点被标记为访问。因为在所有的点都被访问过之后，每一个点都被标记为属于一个聚类或者是噪音。

DBSCAN比其他聚类算法有一些优势。首先，它不需要一个预设定的聚类数量。它还将异常值识别为噪声，而不像均值偏移聚类算法，即使数据点非常不同，它也会将它们放入一个聚类中。此外，它还能很好地找到任意大小和任意形状的聚类。

DBSCAN的主要缺点是，当聚类具有不同的密度时，它的性能不像其他聚类算法那样好。这是因为当密度变化时，距离阈值ε和识别邻近点的minPoints的设置会随着聚类的不同而变化。这种缺点也会出现在非常高维的数据中，因为距离阈值ε变得难以估计。

评价聚类效果：

purity--正确聚类的文档数占总文档的比例

**特征提取（降维）：**

PCA：

PCA算法本质就是找一些投影方向，使得数据在这些投影方向上的方差最大，而且这些投影方向是相互正交的。这其实就是找新的正交基的过程，计算原始数据在这些正交基上投影的方差，方差越大，就说明在对应正交基上包含了更多的信息量。原始数据协方差矩阵的特征值越大，对应的方差越大，在对应的特征向量上投影的信息量就越大。反之，如果特征值较小，则说明数据在这些特征向量上投影的信息量很小，可以将小特征值对应方向的数据删除，从而达到了降维的目的。

PCA把可能具有相关性的高维变量合成线性无关的低维变量，称为主成分（ principal components）。新的低维数据集会尽可能保留原始数据的变量。

简而言之，PCA本质上是将方差最大的方向作为主要特征，并且在各个正交方向上将数据“离相关”，也就是让它们在不同正交方向上没有相关性。

MDS：

MDS利用的是成对样本间相似性，目的是利用这个信息去构建合适的低维空间，是的样本在此空间的距离和在高维空间中的样本间的相似性尽可能的保持一致。

MDS的目标是在降维的过程中将数据的dissimilarity(差异性)保持下来，也可以理解降维让高维空间中的距离关系与低维空间中距离关系保持不变。

**特征选择：**

（1）特征选择的定义

　　对当前学习任务有价值的属性称为是“相关特征”，没有价值的属性称为是“无关特征”，从给定的特征集中选择出相关特征子集的过程，就称为是“特征选择”。

其中还有一种特征称为是“冗余特征”，这些特征指的是可以从其他特征中推演出来的特征。

特征选择能剔除不相关(irrelevant)或亢余(redundant )的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的。另一方面，选取出真正相关的特征简化了模型，使研究人员易于理解数据产生的过程。

（2）特征选择的重要性

　　特征选择是一个“数据预处理”过程，它的重要性体现在两个方面：

　　1）减轻维度灾难问题。

2）去除无关特征可以降低学习的难度。

 特征选择的一般过程首先从特征全集中产生出一个特征子集，然后用评价函数对该特征子集进行评价，评价的结果与停止准则进行比较，若评价结果比停止准则好就停止，否则就继续产生下一组特征子集，继续进行特征选择。选出来的特征子集一般还要验证其有效性。

       综上所述，特征选择过程一般包括产生过程，评价函数，停止准则，验证过程，这4个部分。

产生过程：产生过程是搜索特征子空间的过程。搜索的算法分为完全搜索(Complete)，启发式搜索(Heuristic)，随机搜索(Random) 3大类

遗传算法( GA,  Genetic Algorithms )属于随机搜索

算法描述：首先随机产生一批特征子集，并用评价函数给这些特征子集评分，然后通过交叉、突变等操作繁殖出下一代的特征子集，并且评分越高的特征子集被选中参加繁殖的概率越高。这样经过N代的繁殖和优胜劣汰后，种群中就可能产生了评价函数值最高的特征子集。随机算法的共同缺点：依赖于随机因素，有实验结果难以重现。

评价函数：

评价函数的作用是评价产生过程所提供的特征子集的好坏。

       评价函数根据其工作原理，主要分为筛选器(Filter)、封装器( Wrapper )两大类。

**筛选器**通过分析特征子集内部的特点来衡量其好坏。筛选器一般用作预处理，与分类器的选择无关。

**封装器**实质上是一个分类器，封装器用选取的特征子集对样本集进行分类，分类的精度作为衡量特征子集好坏的标准。

下面简单介绍常见的评价函数。

(1)相关性( Correlation)

运用相关性来度量特征子集的好坏是基于这样一个假设：好的特征子集所包含的特征应该是与分类的相关度较高（相关度高），而特征之间相关度较低的（亢余度低）。

可以使用线性相关系数(correlation coefficient) 来衡量向量之间线性相关度。

IMG_256

(2)距离(Distance Metrics )

运用距离度量进行特征选择是基于这样的假设：好的特征子集应该使得属于同一类的样本距离尽可能小，属于不同类的样本之间的距离尽可能远。

常用的距离度量（相似性度量）包括欧氏距离、标准化欧氏距离、马氏距离等。

(3)信息增益( Information Gain )

　　假设存在离散变量Y，Y中的取值包括{y1，y2，....，ym} ，yi出现的概率为Pi。则Y的信息熵定义为：

IMG_257

(5)分类器错误率 (Classifier error rate )

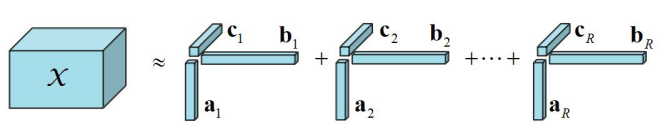
使用特定的分类器，用给定的特征子集对样本集进行分类，用分类的精度来衡量特征子集的好坏。

以上5种度量方法中，相关性、距离、信息增益、一致性属于筛选器，而分类器错误率属于封装器。筛选器由于与具体的分类算法无关，因此其在不同的分类算法之间的推广能力较强，而且计算量也较小。而封装器由于在评价的过程中应用了具体的分类算法进行分类，因此其推广到其他分类算法的效果可能较差，而且计算量也较大。

**模式提取：**

CP分解、

CP分解的张量形式：   
将一个张量表示成有限个秩一张量之和，比如一个三阶张量可以分解为 。



LDA、

LDA技术在document集合中检测潜在的topics,每个document都包含一组word。

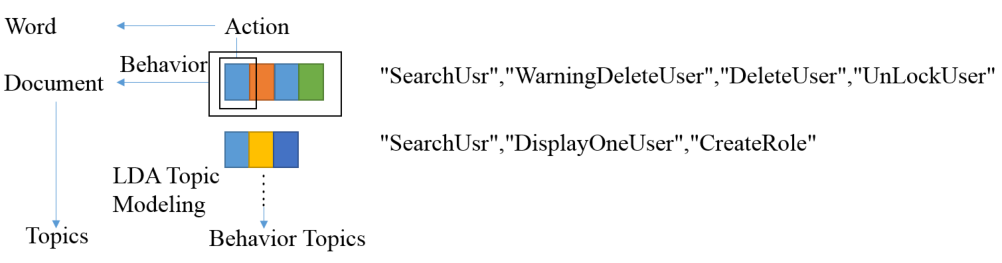
输入： document-word矩阵DW(DWij在document i中word j词的频数)

输出： topic-word矩阵TW（K\*J），document-topic矩阵DT（I\*K）。

(TWkj表示在topic k中word j的概率。

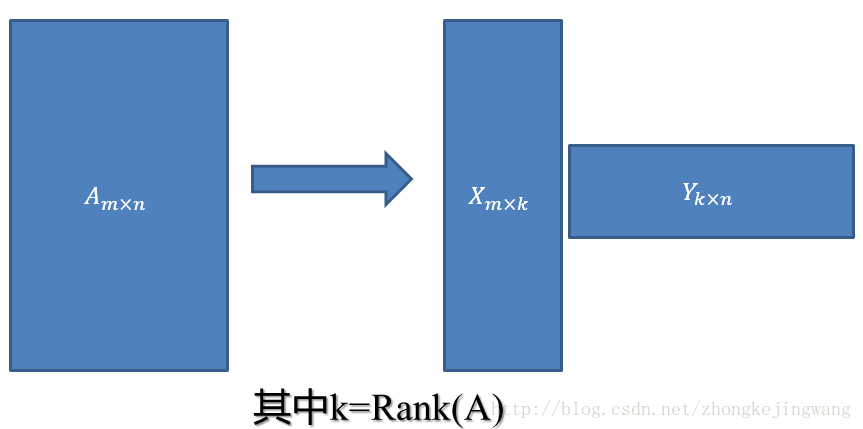
DTik表示在document i中topic k的概率。概率是数字值在0到1之间。)

I个document，J个word，K个topic



发现行为集合中的潜在topics。

SVD奇异值分解、



应用：推荐系统、数据压缩、降维

* 图像压缩（image compression）：较少的奇异值就可以表达出图像中大部分信息，舍弃掉一部分奇异值来实现压缩。
* 图像降噪（image denoise）：噪声一般存在于图像高频部分，也表现在奇异值小的部分，故可以借助SVD实现去噪。
* 音频滤波（filtering）：Andrew Ng的机器学习课程上有个svd将混杂声音分离的例子，其实和噪声滤波类似。
* 求任意矩阵的伪逆（pseudo-inverse）：由于奇异矩阵或非方阵矩阵不可求逆，在特殊情况下需要广义求逆时可用svd方法。
* 模式识别（pattern recognition）：特征为矩阵，数据量较大时，可以用svd提取主要的成分。
* 潜在语义索引（Latent Semantic Indexing）：NLP中，文本分类的关键是计算相关性，这里关联矩阵A=USV’，分解的三个矩阵有很清楚的物理含义，可以同时得到每类文章和每类关键词的相关性。

SVD将数据分解成三个矩阵U，S，VT，这里得到的S是一个对角阵，其中对角元素为奇异值，它代表着矩阵的重要特征，从左上角到右下角重要程度递减。因为奇异值往往对应着矩阵中隐含的重要信息，而且奇异值大小与重要性正相关。

**异常检测：**

LOF、

缺失值怎么处理？异常值怎么判断？

连续型我答用均值填充，又问离散型怎么弄，想了半天每想到，面试官提示介绍下极大似然估计法，最后说用众数去填充。  
异常值看数据分布，均值、标准差，画图，箱线图判断等

**评价模型：**

模型的评价指标有哪些：

accuracy,precision,recall,F1,ROC,AUC

<https://blog.csdn.net/weixin_45252110/article/details/98957266>

其他方面的评价:模型的稳定性，参数的显著性检验。

过拟合：

自己答了模型层面上的，面试官一直问还有吗，最后他回答：从样本层面上有样本不均衡，维度过多，数据样本太少等问题。

**指标**

成交率，回复率，服务态度，转化率，

怎么评价砍价活动的广告效果？有哪些指标？

一、是提高商品的销售额，增加利润，使企业获得经济效益；

二、是使商品或企业在消费者心目中竖立良好的形象，为企业长远的发展奠定良好的基础。

三、点击率，浏览量。

朴素贝叶斯

数据分析的流程

A/B test，怎么做，什么时候用z test，什么时候用t test。

gbdt和xgboost和rangdomforest是同一种集成学习吗

word2vec

特征值分解和奇异值分解两者有着很紧密的关系，特征值分解和奇异值分解的目的都是一样，就是提取出一个矩阵最重要的特征。

特征值分解是一个提取矩阵特征很不错的方法，但是它只是对方阵而言的，在现实的世界中，我们看到的大部分矩阵都不是方阵，比如说有N个学生，每个学生有M科成绩，这样形成的一个N \* M的矩阵就不可能是方阵，我们怎样才能描述这样普通的矩阵呢的重要特征呢？奇异值分解可以用来干这个事情，奇异值分解是一个能适用于任意的矩阵的一种分解的方法：

这里的σ就是上面说的奇异值，u就是上面说的左奇异向量。奇异值σ跟特征值类似，在矩阵Σ中也是从大到小排列，而且σ的减少特别的快，在很多情况下，前10%甚至1%的奇异值的和就占了全部的奇异值之和的99%以上了。也就是说，我们也可以用前r大的奇异值来近似描述矩阵，这里定义一下部分奇异值分解：

IMG_256

 r是一个远小于m、n的数，这样矩阵的乘法看起来像是下面的样子：

