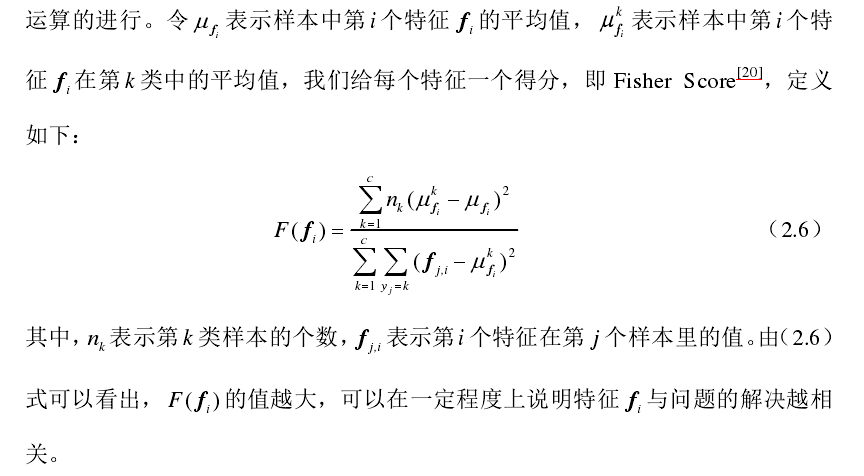
突出 **数据、算法模型、代码**

1. 特征选择：<http://www.doc88.com/p-8919600315838.html>  
   m个特征选最优的d个，有2\*\*m-1种情况，m很大时是个问题。解决这个问题：  
   1.确定搜索所需策略：完全搜索、启发式搜索、随机搜索  
   2.确定对特征的评价准则：过滤和封装。
2. 维数灾难：<https://blog.csdn.net/zbc1090549839/article/details/38929215>

纬度越高，数据越稀疏，超平面越好找，但是这过多的强调了训练集的准确率甚至于对一些错误/异常的数据也进行了学习。导致*训练集上表现良好，但是对*[新数据](https://www.baidu.com/s?wd=%E6%96%B0%E6%95%B0%E6%8D%AE&tn=24004469_oem_dg&rsv_dl=gh_pl_sl_csd)*缺乏泛化能力，得到的分类器对新数据出错，也就是它过拟合了*。我们想要的是泛化能力强的模型，所以特征不是越多越好。过多特征也要求过多样本数据来支撑。也说明维数越高，分类器并不是越优。  
【】给定一组N个特征; 我们如何选择M个特征的最佳子集，使得M <N？一种方法是在图1所示的曲线中搜索最优值。由于为所有特征的所有可能组合训练和测试分类器通常是难以处理的，因此存在几种尝试以不同方式找到该最佳值的方法。这些方法称为**特征选择算法**，并且通常采用启发式（贪婪方法，最佳优先方法等）来定位最佳数量和特征组合。

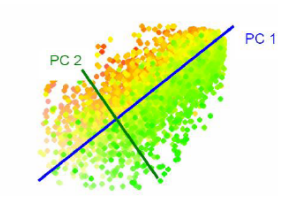
【】另一种方法是用一组M个特征替换N个特征的集合，每个特征是原始特征值的组合。试图找到原始特征的最佳线性或非线性组合以减少最终问题的维度的算法称为**特征提取方法**。一种众所周知的降维技术是主成分分析（PCA），它产生原始N特征的不相关的线性组合。PCA试图找到较低维度的线性子空间，以便保持原始数据的最大方差。但是，请注意，数据的最大差异不一定代表最具辨别力的信息。  
【】在分类器训练期间用于检测和避免过度拟合的宝贵技术是[交叉验证](https://en.wikipedia.org/wiki/Cross-validation_(statistics))。

1. **互信息**：互信息具有对称性，它的值是非负的。当且仅当两个变量相互独立时，互信息的取值为0。一个特征与其类别标签之间的互信息越高，可以在一定程度上认为该特征对后续预测越有用。利用互信息进行特征选择的最简单的做法是，计算出每个特征与其类别标签之间的互信息做法操作简单然后设定一个阂值，选出互信息大于阂值所对应的那些特征。该但由于互信息的值是通过估算得到的，其本身就不精确，因此通过计算互信息的方法选出的特征子集并不是很理想[}29}，并且，阂值该如何设定也是一个问题。概率密度函数是未知的，由于观测样本有限所以也无法准确计算，只能估算。
2. **Fisher Score：**若一个特征是具有鉴别力的，那么该特征与同一类别样本之间的方差应该尽量小，而与不同类别样本之间的方差应该尽量大（类内部样本点的距离尽可能小，类之间的距离尽量大），这样才利于分类、预测等后续运算的进行下面公式是类间方差/类内方差，所以值越大越好。

Fisher score与互信息优势：计算简单得多，计算量小很多，准确性高，可操作性强，节省运算时间。问题：阈值如何确定。仅考阈值来选择太片面，有些低fisher score的特征组合后效果很好；有些高fisher score都被选择了但是是相关的，出现冗余。所以不能一概依据fisher score，而要看整体信息。比如将FS与封装法组合来特征选择。

1. 过滤式：互信息、fisher score、Relief系列、Laplacian Score、Hilbert Schmidt独立准则、追踪率准则等，缺点：阈值如何确定？会有冗余。
2. 封装式：前向选择法、后向消除法、爬山法、分支限界算法、遗传算法。这种运算速度慢。
3. **遗传算法GA**：适用于解决大规模特征选择问题。是美国密歇根大学的Holland教授及其学生于20世纪70年代初提出，思想是根据问题的目标函数构造一个适值函数，通过适值函数评估种群的每个解(一个解对应一条染色体)，选择适应值高的染色体参与形成后代的遗传运算，经多代繁衍，获得适应值最好的个体作为问题的最优解，遗传算法是一种基于群体寻优的方法。

**线性特征提取http://lanbing510.info/2014/10/22/Feature-Extraction-Selection.html**

1. **PCA**：思想：寻找表示数据分布的最优子空间（降维，可以去相关）。就是取协方差矩阵前s个最大特征值对应的特征向量构成映射矩阵，对数据进行降维。

**去除噪音**和冗余，将原有的**复杂数据降维**，优点是简单，而且无参数限制

PCA的目标就是使用另一组基去重新描述得到的数据空间。而新的基要能尽量揭示原有的数据间的关系。==如何寻找到另一组正交基，它们是标准正交基的线性组合，而且能够最好的表示数据集？  
怎样区分什么是信号，什么是噪音呢？这里假设，变化较大的信息被认为是信号，变化较小的则是噪音。而变化的大小则是由方差来描述的，方差较大表示了采样点的主要分布趋势，是主信号或主要分量；而方差较小的分布则被认为是噪音或次要分量。

(a)图所示的情况是低冗余的，从统计学上说，这两个观测变量是相互独立的，它们之间的信息没有冗余。而相反的极端情况如(c)，和高度相关，完全可以用表示，应当去除，只用一个变量就可以表示了。这也就是PCA中“降维”思想的本源。对于上面的简单情况，可以通过简单的线性拟合的方法来判断各观测变量之间是否出现冗余的情况，而对于复杂的情况，需要借助协方差（为0表示独立）来进行衡量和判断。

协方差矩阵性质如下：

* 是一个的平方对称矩阵。（m表示特征数量）
* 对角线上的元素是对应的观测变量的方差。
* 非对角线上的元素是对应的观测变量之间的协方差。
* 在对角线上的元素越大，表明信号越强，变量的重要性越高；元素越小则表明可能是存在的噪音或是次要变量。（这个结论是假设来的，去噪看对角线，降维去冗余看非对角线）
* 在非对角线上的元素大小则对应于相关观测变量对之间冗余程度的大小。

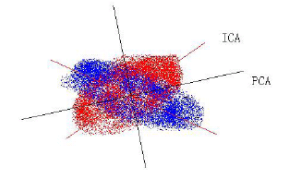
一般，初始矩阵的协方差矩阵总不太好，表现为信噪比不高且变量间相关度大。PCA的目标就是通过基变换对协方差矩阵进行优化，找到相关“主元”。优化方法：对冗余要让非对角元素变小；对信噪比要让对角元素变大。所以优化的目标矩阵应该是一个对角阵，而变换基假设必须是标准正交的。接下来怎么对角化？

**PCA的假设：【1】线性假设**1）数据被限制在一个向量空间中，能被一组基表示；2）隐含的假设了数据之间的连续性关系。【2】使用中值和方差进行充分的概率分布描述的模型只限于指数型概率分布模型。若数据的概率分布不满足高斯或指数型，则PCA会失效，因为方差协方差不能很好的描述噪音和冗余。（根据中央极限定理，现实生活中所遇到的大部分采样数据的概率分布都是遵从高斯分布的。）【3】大方差向量具有较大重要性，最高方差的一维向量就可以被看作是主元，而方差较小的变化则被认为是噪音【4】PCA方法假设主元向量之间都是正交的，从而可以利用线形代数求解，提高了效率和应用的范围。

PCA方法和线形代数中的奇异值分解(SVD)方法有内在的联系，一定意义上来说，PCA的解法是SVD的一种变形和弱化。

1. LDA-线性判别分析：思想：寻找可分性判据最大的子空间。

用到了Fisher的思想，即寻找一个向量，使得降维后类内散度最小，类间散度最大；其实就是取S−1wSbSw−1Sb前s个特征值对应的特征向量构成映射矩阵，对数据进行处理。

1. ICA-独立成分分析：思想：PCA是将原始数据降维，并提取不相关的部分；ICA是将原始数据降维并提取出相互独立的属性；寻找一个线性变换z=Wx,使得z的各个分量间的独立性最大（PCA的是使Z的各分量正交且方差最大。）。

#### PCA VS ICA

PCA的问题其实是一个基的变换，使得变换后的数据有着最大的方差。方差的大小描述的是一个变量的信息量，我们在讲一个东西的稳定性的时候，往往说要减小方差，如果一个模型的方差很大，那就说明模型不稳定了。但是对于我们用于机器学习的数据（主要是训练数据），方差大才有意义，不然输入的数据都是同一个点，那方差就为0了，这样输入的多个数据就等同于一个数据了。

ICA是找出构成信号的相互独立部分(不需要正交)，对应高阶统计量分析。ICA理论认为用来观测的混合数据阵X是由独立元S经过A线性加权获得。ICA理论的目标就是通过X求得一个分离矩阵W，使得W作用在X上所获得的信号Y是独立源S的最优逼近，该关系可以通过下式表示：Y=WX=WAS,A=W−1

ICA相比与PCA更能刻画变量的随机统计特性，且能抑制高斯噪声。

1. 二维PCA
2. CCA-典型对应分析（描述两个高纬变量之间的线性关系）

**非线性特征提取**

1. Kernel PCA
2. Kernel FDA
3. Manifold Learning 流形学习