**特征选择**是特征工程中的重要问题（另一个重要的问题是特征提取），坊间常说：**数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已**。由此可见，特征工程尤其是特征选择在机器学习中占有相当重要的地位

**1. 计算每一个特征与响应变量的相关性**：工程上常用的手段有计算皮尔逊系数和互信息系数，皮尔逊系数只能衡量线性相关性而互信息系数能够很好地度量各种相关性，但是计算相对复杂一些，好在很多toolkit里边都包含了这个工具（如sklearn的MINE），得到相关性之后就可以排序选择特征了；

2. **构建单个特征的模型，通过模型的准确性为特征排序，借此来选择特征**，另外，记得JMLR'03上有一篇论文介绍了一种基于决策树的特征选择方法，本质上是等价的。当选择到了目标特征之后，再用来训练最终的模型；

3. **通过L1正则项来选择特征：**L1正则方法具有稀疏解的特性，因此天然具备特征选择的特性，但是要注意， L1没有选到的特征不代表不重要，原因是两个具有高相关性的特征可能只保留了一个，如果要确定哪个特征重要, 应再通过L2正则方法交叉检验；

4. **训练能够对特征打分的预选模型**：RandomForest和Logistic Regression等都能对模型的特征打分，通过打分获得相关性后再训练最终模型；

5. **通过特征组合后再来选择特征**： 如对用户id和用户特征最组合来获得较大的特征集再来选择特征，这种做法在推荐系统和广告系统中比较常见，这也是所谓亿级甚至十亿级特征的主要来源，原因是用户数据比较稀疏，组合特征能够同时兼顾全局模型和个性化模型，这个问题有机会可以展开讲。

6. **通过深度学习来进行特征选择**：目前这种手段正在随着深度学习的流行而成为一种手段，尤其是在计算机视觉领域，原因是深度学习具有自动学习特征的能力，这也是深度学习又叫unsupervised feature learning的原因。

从深度学习模型中选择某一神经层的特征后就可以用来进行最终目标模型的训练了。整体上来说，特征选择是一个

既有学术价值又有工程价值的问题，目前在研究领域也比较热，值得所有做机器学习的朋友重视。

机器学习要注意：海量离散特征+简单模型组合

少量连续特征+复杂模型组合

的权衡，一个折腾特征，一个折腾模型，注意权衡，第一个已广泛使用，第二个前景看起来很好。

首先，**数据标准化**是为了将不同数量级的数据变成同一数量级，**消除数量级的影响**，比如：在K近邻算法中，如果不对解释变量进行标准化，那么具有小数量级的解释变量的影响就微乎其微了。再者，使用神经网络这种对数据比较敏感的算法时，数量级的差别会导致模型出现异常等等其次，数据标准化，也是无量纲化的一种方法，便于理解即说明

神经网络/SVM/PCA/k-means必须做

logistic regression和linear regression做了会更好（如果用L1/L2 regularization的话），如果条件限制无法做的话也能凑合用

tree based model不用做

任何有用到距离，或者利用梯度进行优化的算法

**1、归一化**

特点：对不同特征维度的伸缩变换的目的是**使各个特征维度对目标函数的影响权重是一致的**，**即使得那些扁平分布的数据伸缩变换成类圆形。这也就改变了原始数据的一个分布**。

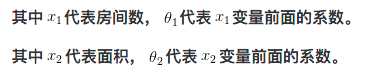
好处：

**1 提高迭代求解的收敛速度**

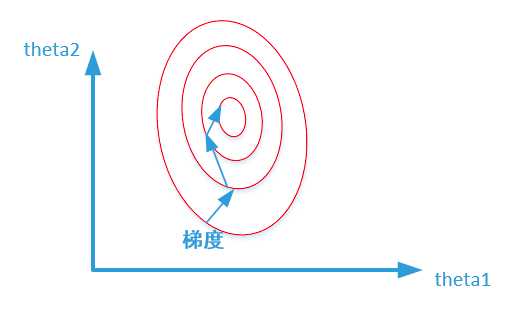
**2 提高迭代求解的精度**

假定为预测房价的例子，自变量为面积，房间数两个，因变量为房价。那么可以得到的公式为：

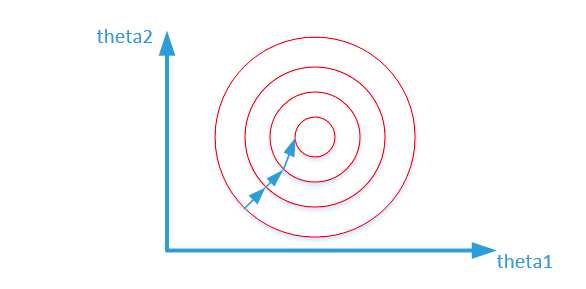
Y=a1 \* x1 + a2 \* x2



未归一化：



归一化



寻找最优解的过程也是使得损失函数值最小的参数theta1,theta2

上述两幅图是随时函数的等高线

当没有归一化时，面积的范围可以从1~1000，房间数的范围为1~10，可以看出面积远大于房间数。

影响：画损失函数时，数据没归一化的表达式可以为

J = ( 3\* theta1 + 6000 \* theta2 - y) \*\* 2

这样等高线为椭圆形状。

归一化之后损失函数

J = (0.5 \* theta1 + 0.55 \* theta2 - y) \*\* 2

变量前面的系数几乎一样，图像的等高线为类似圆形形状

**经过归一化后，最优解的寻优过程会变的平缓，容易收敛到最优解。**

**一、无量纲化：**例如房子数量和收入，因为从业务层知道，这两者的重要性一样，所以把它们全部归一化，这是从业务层面上作的处理。

不同的数据在不同列数据的数量级相差过大的话，计算起来大数的变化会掩盖掉小数的变化。  
**二.避免数值问题：** 太大的数会引发数值问题。  
**三.一些模型求解的需要**：例如梯度下降法。

所说的，就是一种情况-----不归一化，容易产生陕谷，而学习率较大时，就会以之字形下降。学习率较小，则会产生直角形路线，不管怎么样，都不会是好路线。

**2、标准化**

特点：对不同特征维度的伸缩变换的目的是**使得不同度量之间的特征具有可比性**。**同时不改变原始数据的分布**。

好处

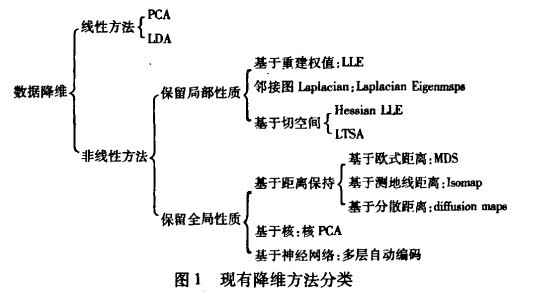
1 使得不同度量之间的特征具有可比性，对目标函数的影响体现在几何分布上，而不是数值上

2 不改变原始数据的分布

标准化和归一化一样吗？

协同过滤是一种

**数据降维**



**主成分分析（PCA）**

　　将n维样本X通过投影矩阵W，转换为K维矩阵Z

　　输入：样本集D，低维空间d

　　输出：投影矩阵W

　　算法步骤：

　　　　1）对所有样本进行中心化操作

　　　　2）计算样本的协方差矩阵

　　　　3）对协方差矩阵做特征值分解

　　　　4）取最大的d个特征值对应的特征向量，构造投影矩阵W

　　注：通常低维空间维数d的选取有两种方法：1）通过交叉验证法选取较好的d  2）从算法原理的角度设置一个阈值，比如t=0.95，然后选取似的下式成立的最小的d值：

　　　　Σ（i->d）λi/Σ（i->n）λi>=t，其中λi从大到小排列

　　PCA降维的准则有以下两个：

　　　　最近重构性：重构后的点距离原来的点的误差之和最小

　　　　最大可分性：样本点在低维空间的投影尽可能分开

===========================

1，特征值分解 和 奇异值分解

矩阵可以看做运动，包括旋转，拉伸，投影三个操作。

首先，矩阵可以认为是一种线性变换，而且这种线性变换的作用效果与基的选择有关。以Ax = b为例，x是m维向量， b是n维向量，m,n可以相等也可以不相等，表示矩阵可以将一个向量线性变换到另一个向量，这样一个线性变换的作用可以包含旋转、缩放和投影三种类型的效应。

奇异值分解正是对线性变换这三种效应的一个析构。A=U∑V，u和v是两组正交单位向量，是对角阵，表示奇异值，它表示我们找到了和这样两组基，A矩阵的作用是将一个向量从u这组正交基向量的空间旋转到v这组正交基向量空间，并对每个方向进行了一定

的缩放，缩放因子就是各个奇异值。如果维度比大，则表示还进行了投影。可以说奇异值分解将一个矩阵原本混合在一起的三种作用效果，分解出来了。

而特征值分解其实是对旋转缩放两种效应的归并。（有投影效应的矩阵不是方阵，没有特征值）特征值，特征向量由Ax=x得到，它表示如果一个向量v处于A的特征向量方向，那么Av对v的线性变换作用只是一个缩放。也就是说，求特征向量和特征值的过程，我们找到了这样一组基，在这组基下，矩阵的作用效果仅仅是存粹的缩放。对于实对称矩阵，特征向量正交，我们可以将特征向量式子写成，这样就和奇异值分解类似了，就是A矩阵将一个向量从x这组基的空间旋转到x这组基的空间，并在每个方向进行了缩放，由于前后都是x，就是没有旋转或者理解为旋转了0度。

总结一下，特征值分解和奇异值分解都是给一个矩阵(线性变换)找一组特殊的基，特征值分解找到了特征向量这组基，在这组基下该线性变换只有缩放效果。而奇异值分解则是找到另一组基，这组基下线性变换的旋转、缩放、投影三种功能独立地展示出来

了。我感觉特征值分解其实是一种找特殊角度，让旋转效果不显露出来，所以并不是所有矩阵都能找到这样巧妙的角度。仅有缩放效果，表示、计算的时候都更方便，这样的基很多时候不再正交了，又限制了一些应用。

QR分解:如果实（复）非奇异矩阵A能够化成正交（酉）矩阵Q与实（复）非奇异上三角矩阵R的乘积，即A=QR，则称其为A的QR分解。

QR（正交三角）分解法是目前求一般矩阵全部特征值的最有效并广泛应用的方法，一般矩阵先经过正交相似变化成为Hessenberg矩阵，然后再应用QR方法求特征值和特征向量。它是将矩阵分解成一个正规正交矩阵Q与上三角形矩阵R，所以称为QR分解法，与此正规正交矩阵的通用符号Q有关。

奇异矩阵:首先，看这个矩阵是不是方阵（即行数和列数相等的矩阵。若行数和列数不相等，那就谈不上奇异矩阵和非奇异矩阵）。 然后，再看此矩阵的行列式|A|是否等于0，若等于0，称矩阵A为奇异矩阵； 若不等于0，称矩阵A为非奇异矩阵。

三角分解法是将原正方 (square) 矩阵分解成一个上三角形矩阵或是排列(permuted) 的上三角形矩阵和一个 下三角形矩阵， 这样的分解法又称为LU分解法。它的用途主要在简化一个大矩阵的行列式值的计算过程，求逆矩阵，和求解联立方程组。 不过要注意这种分解法所得到的上下三角形矩阵并非唯一，还可找到数个不同 的一对上下三角形矩阵，此两三角形矩阵相乘也会得到原矩阵。

QR分解法是将矩阵分解成一个正规正交矩阵与上三角形矩阵,所以称为QR分解法,与此正规正交矩阵的通用符号Q有关。

奇异值分解 (singular value decomposition,SVD) 是另一种正交矩阵分解法；SVD是最可靠的分解法，但是它比QR分解法要花上近十倍的计算时间。[U,S,V]=svd(A)，其中U和V分别代表两个正交矩阵，而S代表一对角矩阵。和QR分解法相同， 原矩阵A不必为正方矩阵。使用SVD分解法的用途是解最小平方误差法和数据压缩。

==========================================

2，范数

0范数，向量中非零元素的个数。

一定条件下1范数和0范数可以混淆。1因具有比L0更好的优化求解特性而被广泛应用。

向量范数:

1-范数：即向量元素绝对值之和，matlab调用函数norm(x, 1) 。

2-范数：Euclid范数（欧几里得范数，常用计算向量长度），即向量元素绝对值的平方和再开方，matlab调用函数norm(x, 2)。

\infty-范数,无穷大范数:即所有向量元素绝对值中的最大值，matlab调用函数norm(x, inf)。

-\infty-范数，负无穷大范数：即所有向量元素绝对值中的最小值，matlab调用函数norm(x, -inf)。

p-范数：即向量元素绝对值的p次方和的1/p次幂，matlab调用函数norm(x, p)。

矩阵范数：

1-范数：列和范数，即所有矩阵列向量绝对值之和的最大值，matlab调用函数norm(A, 1)。

2-范数：谱范数，即A'A矩阵的最大特征值的开平方。matlab调用函数norm(x, 2)。

\infty-范数：行和范数，即所有矩阵行向量绝对值之和的最大值，matlab调用函数norm(A, inf)。

F-范数：Frobenius范数，即矩阵元素绝对值的平方和再开平方，matlab调用函数norm(A, ’fro‘)。

核范数：即奇异值之和

=========================================

3，协方差

当两个变量相关时，用于评估它们因相关而产生的对应变量的影响。

当多个变量独立时，用方差来评估这种影响的差异。

当多个变量相关时，用协方差来评估这种影响的差异。

==========================================

4,大数定理，中心极限定理

中心极限定理是说：和的分布收敛于正态分布

样本的平均值约等于总体的平均值。

不管总体是什么分布，任意一个总体的样本平均值都会围绕在总体的整体平均值周围，并且呈正态分布。

小数定律：

如果统计数据很少，那么事件就表现为各种极端情况，而这些情况都是偶然事件，跟它的期望值一点关系都没有。

如果数据少，随机现象可以看上去很不随机。甚至非常整齐，感觉好像真有规律一样。

大数定律：频率收敛于概率

如果统计数据足够大，那么事物出现的频率就能无限接近他的期望值。

===========================================

离散数据--概率函数

连续数据--概率密度函数

概率分布：http://blog.csdn.net/dymodi/article/details/54231728

正态曲线呈钟型，两头低，中间高，左右对称因其曲线呈钟形，因此人们又经常称之为钟形曲线。 若随机变量X服从一个数学期望为μ、方差为σ^2的正态分布，记为N(μ，σ^2)。其概率密度函数为 正态分布的期望值μ决定了其位置，其标准差σ决定了分布的幅度。当μ = 0,σ = 1时的正态分布是标准正态分布。

Thin-tail distribution:

Heavy-tailed distribution:

指数分布在x->infinity时，以指数的速度趋近于0，以指数分布为分界线，下降速度更快的 称为Thin-tail distribution如正态分布，在远离峰值的尾部区域，时间发生的概率更低一些。

所以正态分布用来对那些主流事件发生较多，非主流事件发生较少的情况进行建模更为合适。 x->infinity时下降速度慢于指数分布的称为重尾分布Heavy-tailed distribution重尾分布 更适用于对那些离峰值较远的稀有事件也会有相当的概率发生的情况。重尾分布作为一个大的类别， 还包含三个重要的子类别，分别是肥尾分布（Fat-tailed distribution），长尾分布（Long-tailed distribution） 和次指数分布（Subexponential distribution）。

长尾分布，或者说长尾理论是一个与互联网发展分不开的概念。说到这里就不得不先提一下传统商业中的 帕累托法则（Pareto principle），又称为二八定律。比如80%的财富集中在20%的人手里，图书馆里20%的书 可以满足80%的顾客。于是大家往往只关注在PDF图中最左面的20%的顾客，以期满足80%，如下图绿色的部分， 来实现效益的最大化。

但在一些网上零售业中，如Amazon和Netflix，数据表明右端黄色的尾巴虽然平均需求小但是由于数量巨大， 导致其总的营销收益甚至超过主流的商品。这一发现似乎对商业界的触动极大，也说明了正确建模的重要性。 如果用指数分布进行建模，这些远端的需求也许就会被忽视；而用长尾分布进行建模就可以发现这些新的需求 从而带来效益的提高。

==========================================

插值和拟合：

插值：是指已知某函数的在若干离散点上的函数值或者导数信息，通过求解该函数中待定形式的插值函数以及 待定系数，使得该函数在给定离散点上满足约束。

拟合：是指已知某函数的若干离散函数值{f1,f2,…,fn}，通过调整该函数中若干待定系数f(λ1, λ2,…,λn), 使得该函数与已知点集的差别(最小二乘意义)最小。如果待定函数是线性，就叫线性拟合或者线性回归(主要在统计中)， 否则叫作非线性拟合或者非线性回归。表达式也可以是分段函数，这种情况下叫作样条拟合。

从几何意义上将，拟合是给定了空间中的一些点，找到一个已知形式未知参数的连续曲面来最大限度地逼近这些点； 而插值是找到一个( 或几个分片光滑的)连续曲面来穿过这些点。

==========================================

正定矩阵，半正定矩阵：

半正定矩阵定义为: X^T M X >= 0

其中X 是任意向量，M 是变换矩阵

考虑矩阵的特征值。

若所有特征值均不小于零，则称为半正定。

若所有特征值均大于零，则称为正定。

==========================================

向量内积和外积：

外积：右手法则 ab sinc

内积: ab cosc

==========================================

==========================================

LDA是一种监督学习的降维技术，也就是说它的数据集的每个样本是有类别输出的。这点和PCA不同。

PCA是不考虑样本类别输出的无监督降维技术。LDA的思想可以用一句话概括，就是“投影后类内方差最小，

类间方差最大”。什么意思呢？ 我们要将数据在低维度上进行投影，投影后希望每一种类别数据的投影点

尽可能的接近，而不同类别的数据的类别中心之间的距离尽可能的大。