<https://arxiv.org/> 论文网站

做算法：算法工程师，要看最新的论文，看看可行性

做工程：”调参工程师”，读旧论文，用已知有效的方法解决问题

（链，树，图的构建、遍历、查找、排序）

【1】特征选取：从业务角度区分输入数据包含的特征，并认识到这些特征对结果的贡献。

a) 对数据本身和其对应的业务领域有所了解。

b) 能够根据需要标注数据。

c) 知道如何从全集中通过划分特征子集、加减特征等方法选取有效特征集。

【2】向量空间模型（VSM）构建：了解如何将自然语言、图片等人类日常使用的信息转化成算法可以运算的数据。

a) 能够把文字、语音、图像等输入转化成算法所需输入格式（一般为实数空间的矩阵或向量）。

b) 能够根据信息熵等指标选取有效特征。

【3】数据清洗和处理：对直接的业务数据进行筛选并转换为模型可处理形式。

a) 能够运用统计学方法等ETL手段清洗输入数据。

b) 能够对数据进行归一化（normalization）， 正则化（regularization）等标准化操作。

c) 能够采用bootstrap等采样方法处理有限的训练/测试数据，以达到更好的运算效果。

2.4.3 模型验证

算法+数据就能够得到模型。但是，

这个模型的质量如何？

这个模型和那个模型比较，哪个更适合解决当前问题？

在做了如此这般的优化之后得出了一个新的模型，怎么能够确认它比旧的模型好？

为了解答这些问题，就需要掌握度量模型质量的方法。为此，需要做到：

i） 了解 bias，overfitting 等基本概念，及针对这些情况的基本改进方法。

ii） 了解各种模型度量指标（e.g. Accuracy, Precision,Recall, F1Score……）的计算方法和含义，及其对模型质量的影响。

iii) 能够构建训练集、测试集，并进行交叉验证。

iv) 能够运用多种不同的验证方法（e.g. 2-Fold cross-validation,K-Fold cross-validation, Leave-One-Out cross-validation……）来适应不同的数据集。

现阶段而言数据远比算法重要，对模型的影响远胜算法。有直接作用的都是有监督学习，深度学习也是依赖海量标注好的数据。

目前还没有完全自动化标注技术出现，人工标注仍然是必要和主流。

3.3.1 什么叫做标注

举个很简单的例子说明一下什么是数据标注：

在开发聊天机器人的时候，我们需要训练意图判定和实体识别模型，因此也就需要标注用户问题的意图和出现的实体。

这是用户问题原始数据：“00183号商品快递到伊犁邮费多少？”

这样一句话，很显然问它的用户是想知道某一种商品发往某地的邮费。邮费是商品的一个属性，我们把所有查询商品属性的意图都定义为“商品查询”。

因此，这样一句话的意图是“商品查询”。其中有包含了几个实体，分别是商品Id，目的地和商品属性。

这句话被标注出来以后，就是下面这个样子：

[00183]<-{商品Id}号商品快递到[伊犁]<-{目的地}[邮费]<-{商品属性}多少？||商品查询

具体格式不必纠结。核心一点：标注就是将原始数据内全部或者部分内容，按照业务需求打上定义好的标签。------没有专业要求，属于脏活累活，谁都能干。

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：

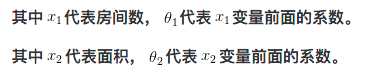
1 提高迭代求解的收敛速度

2 提高迭代求解的精度

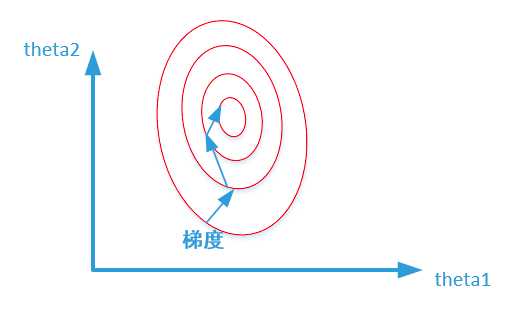
假定为预测房价的例子，自变量为面积，房间数两个，因变量为房价。

那么可以得到的公式为：

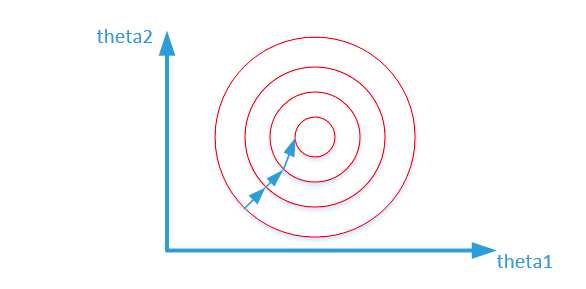
Y=a1 \* x1 + a2 \* x2



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

变量前面的洗漱几乎一样，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。

1. 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

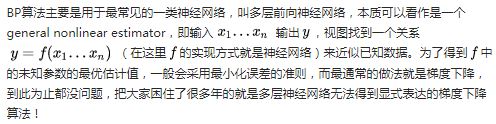
1 使得不同度量之间的特征具有可比性，对目标函数的影响体现在几何分布上，而不是数值上

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

**BP传播算法：**



BP算法实际上是一种近似的最优解决方案，背后的原理仍然是梯度下降，但为了解决上述困难，其方案是将多层转变为一层接一层的优化：只优化一层的参数是可以得到显式梯度下降表达式的；而顺序呢必须反过来才能保证可工作——由输出层开始优化前一层的参数，然后优化再前一层……跑一遍下来，那所有的参数都优化过一次了。但是为什么说是近似最优呢，因为数学上除了很特殊的结构，step-by-step的优化结果并不等于整体优化的结果！不过，好歹现在能工作了，不是吗？至于怎么再改进（已经很多改进成果了），或者采用其他算法（例如智能优化算法等所谓的全局优化算法，就算是没有BP这个近似梯度下降也只是局部最优的优化算法）那就是新的研究课题了。