<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。