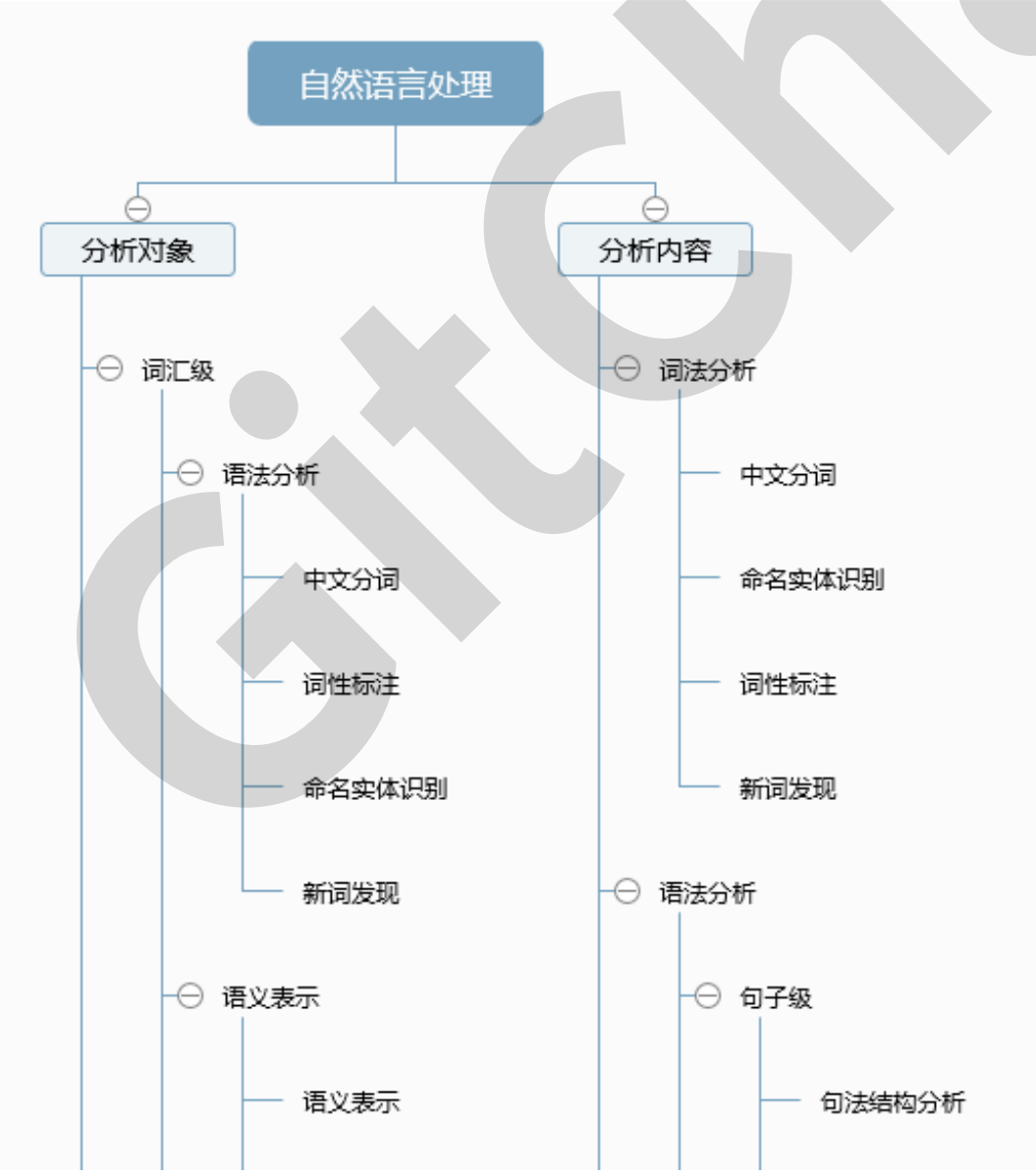
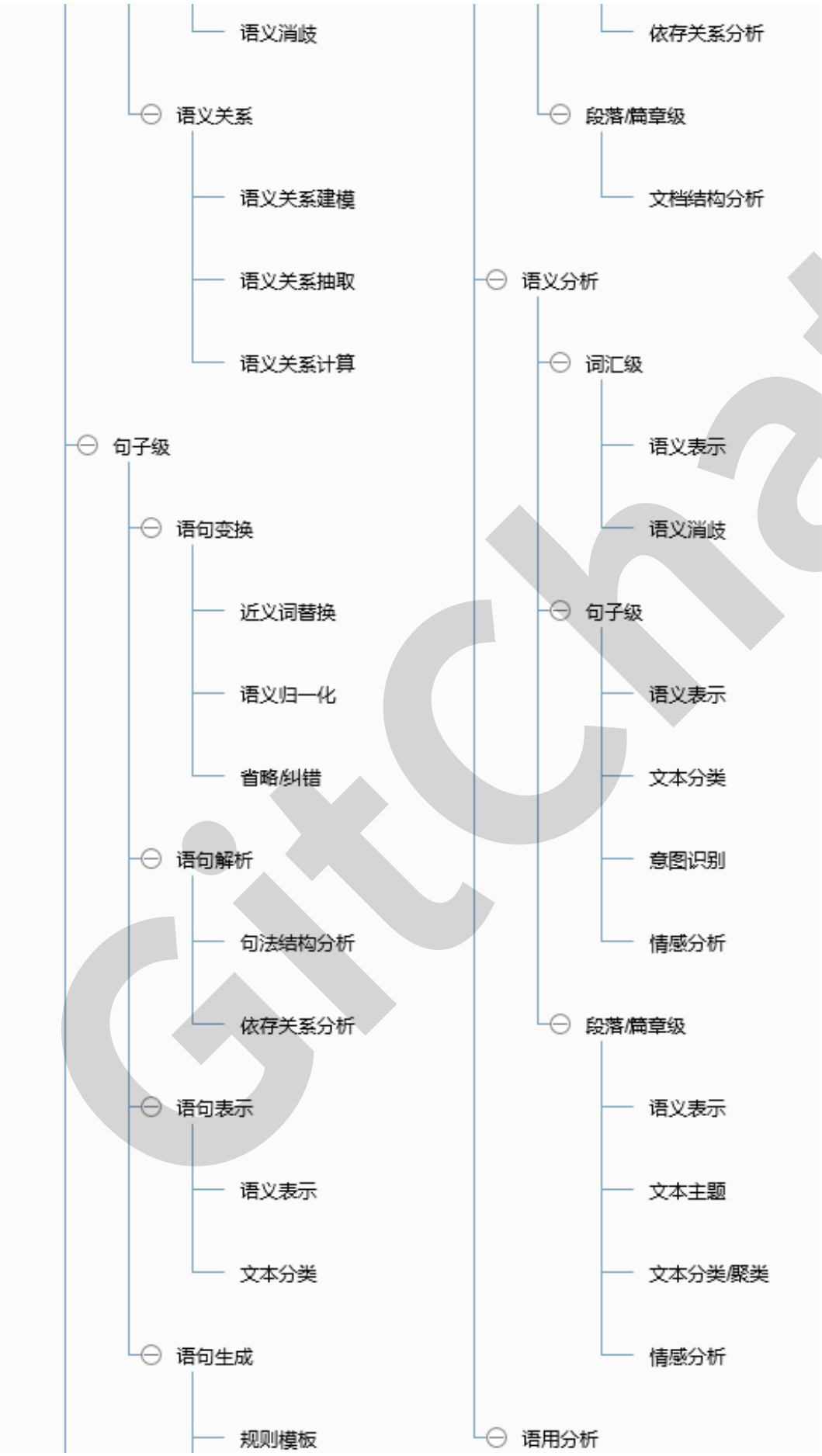
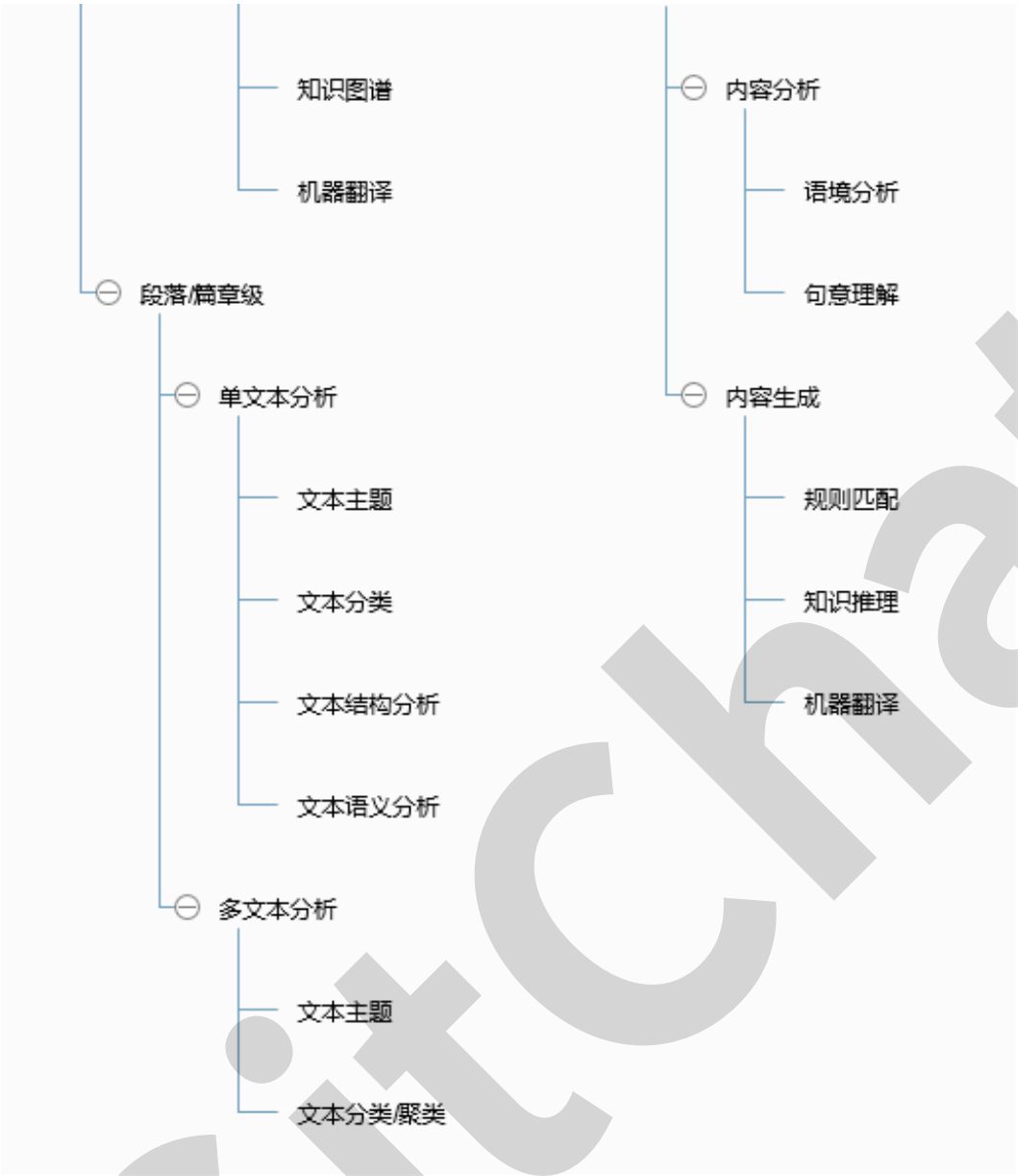
第01课：中文自然语言处理的完整机器处理流程

2016年全球瞩目的围棋大战中，人类以失败告终，更是激起了各种“机器超越、控制人类”的讨论，然而机器真的懂人类吗？机器能感受到人类的情绪吗？机器能理解人类的语言吗？如果能，那它又是如何做到呢？带着这样好奇心，本文将带领大家熟悉和回顾一个完整的自然语言处理过程，后续所有章节所有示例开发都将遵从这个处理过程。

首先我们通过一张图（来源：网络）来了解 NLP 所包含的技术知识点，这张图从分析对象和分析内容两个不同的维度来进行表达，个人觉得内容只能作为参考，对于整个 AI 背景下的自然语言处理来说还不够完整。




语义消歧
语义关系
语义关系建模
语义关系抽取
语义关系计算
句子级
语句变换
近义词替换
语义归一化
省略/纠错
语句解析
句法结构分析
依存关系分析
语句表示
语义表示
文本分类
语句生成
规则模板
依存关系分析
段落/篇章级
文档结构分析
语义分析
词汇级
语义表示
语义消歧
句子级
语义表示
文本分类
意图识别
情感分析
段落/篇章级
语义表示
文本主题
文本分类/聚类
情感分析
语用分析



有机器学习相关经验的人都知道，中文自然语言处理的过程和机器学习过程大体一致，但又存在很多细节上的不同点，下面我们就来看看中文自然语言处理的基本过程有哪些呢？

获取语料

语料，即语言材料。语料是语言学研究的内容。语料是构成语料库的基本单元。所以，人们简单地用文本作为替代，并把文本中的上下文关系作为现实世界中语言的上下文关系的替代品。我们把一个文本集合称为语料库（Corpus），当有几个这样的文本集合的时候，我们称之为语料库集合(Corpora)。（定义来源：百度百科）按语料来源，我们将语料分为以下两种：

1.已有语料

很多业务部门、公司等组织随着业务发展都会积累有大量的纸质或者电子文本资料。那么，对于这些资料，在允许的条件下我们稍加整合，把纸质的文本全部电子化就可以作为我们的语料库。

2.网上下载、抓取语料

如果现在个人手里没有数据怎么办呢？这个时候，我们可以选择获取国内外标准开放数据集，比如国内的**中文汉语有搜狗语料**、**人民日报语料**。国外的因为大都是英文或者外文，这里暂时用不到。也可以选择通过爬虫自己去抓取一些数据，然后来进行后续内容。

语料预处理

这里重点介绍一下语料的预处理，在一个完整的中文自然语言处理工程应用中，语料预处理大概会占到整个50%-70%的工作量，所以开发人员大部分时间就在进行语料预处理。下面通过数据洗清、分词、词性标注、去停用词四个大的方面来完成语料的预处理工作。

1.语料清洗

数据清洗，顾名思义就是在语料中找到我们感兴趣的东西，把不感兴趣的、视为噪音的内容清洗删除，包括对于原始文本提取标题、摘要、正文等信息，对于爬取的网页内容，去除广告、标签、HTML、JS 等代码和注释等。常见的数据清洗方式有：人工去重、对齐、删除和标注等，或者规则提取内容、正则表达式匹配、根据词性和命名实体提取、编写脚本或者代码批处理等。

2.分词

中文语料数据为一批短文本或者长文本，比如：句子，文章摘要，段落或者整篇文章组成的一个集合。一般句子、段落之间的字、词语是连续的，有一定含义。而进行文本挖掘分析时，我们希望文本处理的最小单位粒度是词或者词语，所以这个时候就需要分词来将文本全部进行分词。

常见的分词算法有：基于字符串匹配的分词方法、基于理解的分词方法、基于统计的分词方法和基于规则的分词方法，每种方法下面对应许多具体的算法。

当前中文分词算法的主要难点有歧义识别和新词识别，比如：“羽毛球拍卖完了”，这个可以

切分成“羽毛 球拍 卖 完 了”，也可切分成“羽毛球 拍卖 完 了”，如果不依赖上下文其他的句子，恐怕很难知道如何去理解。

3.词性标注

词性标注，就是给每个词或者词语打词类标签，如形容词、动词、名词等。这样做可以让文本在后面的处理中融入更多有用的语言信息。词性标注是一个经典的序列标注问题，不过对于有些中文自然语言处理来说，词性标注不是非必需的。比如，常见的文本分类就不用关心词性问题，但是类似情感分析、知识推理却是需要的，下图是常见的中文词性整理。

词性编码	词性名称	注 解
Ag	形语素	形容词性语素。形容词代码为 a，语素代码 g 前面置以A。
a	形容词	取英语形容词 adjective 的第1个字母。
ad	副形词	直接作状语的形容词。形容词代码 a和副词代码d并在一起。
an	名形词	具有名词功能的形容词。形容词代码 a和名词代码n并在一起。
b	区别词	取汉字“别”的声母。
c	连词	取英语连词 conjunction 的第1个字母。
dg	副语素	副词性语素。副词代码为 d，语素代码 g 前面置以D。
d	副词	取 adverb 的第2个字母，因其第1个字母已用于形容词。
e	叹词	取英语叹词 exclamation 的第1个字母。
f	方位词	取汉字“方”
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。
h	前接成分	取英语 head 的第1个字母。
i	成语	取英语成语 idiom 的第1个字母。
j	简称略语	取汉字“简”的声母。
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。
m	数词	取英语 numeral 的第3个字母，n，u已有他用。
Ng	名语素	名词性语素。名词代码为 n，语素代码 g 前面置以N。
n	名词	取英语名词 noun 的第1个字母。
nr	人名	名词代码 n和“人(ren)”的声母并在一起。
ns	地名	名词代码 n和处所词代码s并在一起。
nt	机构团体	“团”的声母为 t，名词代码n和t并在一起。
nz	其他专名	“专”的声母的第 1个字母为z，名词代码n和z并在一起。
o	拟声词	取英语拟声词 onomatopoeia 的第1个字母。
p	介词	取英语介词 prepositional 的第1个字母。
q	量词	取英语 quantity 的第1个字母。
r	代词	取英语代词 pronoun 的第2个字母,因p已用于介词。
s	处所词	取英语 space 的第1个字母。
tg	时语素	时间词性语素。时间词代码为 t,在语素的代码g前面置以T。
t	时间词	取英语 time 的第1个字母。
u	助词	取英语助词 auxiliary
vg	动语素	动词性语素。动词代码为 v。在语素的代码g前面置以V。
v	动词	取英语动词 verb 的第一个字母。
vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。
w	标点符号	
x	非语素字	非语素字只是一个符号，字母 x通常用于代表未知数、符号。
y	语气词	取汉字“语”的声母。
z	状态词	取汉字“状”的声母的前一个字母。
un	未知词	不可识别词及用户自定义词组。取英文Unkonwn首两个字母。(非北大标准，CSW分词中定义)

常见的词性标注方法可以分为基于规则和基于统计的方法。其中基于统计的方法，如基于最大熵的词性标注、基于统计最大概率输出词性和基于 HMM 的词性标注。

4.去停用词

停用词一般指对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等一些词。所以在一般性的文本处理中，分词之后，接下来一步就是去停用词。但是对于中文来说，去停用词操作不是一成不变的，停用词词典是根据具体场景来决定的，比如在情感分析中，语气词、感叹号是应该保留的，因为他们对表示语气程度、感情色彩有一定的贡献和意义。

特征工程

做完语料预处理之后，接下来需要考虑如何把分词之后的字和词语表示成计算机能够计算的类型。显然，如果要计算我们至少需要把中文分词的字符串转换成数字，确切的说应该是数学中的向量。有两种常用的表示模型分别是词袋模型和词向量。

词袋模型（Bag of Word, BOW)，即不考虑词语原本在句子中的顺序，直接将每一个词语或者符号统一放置在一个集合（如 list），然后按照计数的方式对出现的次数进行统计。统计词频这只是最基本的方式，TF-IDF 是词袋模型的一个经典用法。

词向量是将字、词语转换成向量矩阵的计算模型。目前为止最常用的词表示方法是 One-hot，这种方法把每个词表示为一个很长的向量。这个向量的维度是词表大小，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。还有 Google 团队的 Word2Vec，其主要包含两个模型：跳字模型（Skip-Gram）和连续词袋模型（Continuous Bag of Words，简称 CBOW），以及两种高效训练的方法：负采样（Negative Sampling）和层序 Softmax（Hierarchical Softmax）。值得一提的是，Word2Vec 词向量可以较好地表达不同词之间的相似和类比关系。除此之外，还有一些词向量的表示方式，如 Doc2Vec、WordRank 和 FastText 等。

特征选择

同数据挖掘一样，在文本挖掘相关问题中，特征工程也是必不可少的。在一个实际问题中，构造好的特征向量，是要选择合适的、表达能力强的特征。文本特征一般都是词语，具有语义信息，使用特征选择能够找出一个特征子集，其仍然可以保留语义信息；但通过特征提取找到的特征子空间，将会丢失部分语义信息。所以特征选择是一个很有挑战的过程，更多的依赖于经验和专业知识，并且有很多现成的算法来进行特征的选择。目前，常见的特征选择方法主要有 DF、 MI、 IG、 CHI、WLLR、WFO 六种。

模型训练

在特征向量选择好之后，接下来要做的事情当然就是训练模型，对于不同的应用需求，我们使用不同的模型，传统的有监督和无监督等机器学习模型， 如 KNN、SVM、Naive Bayes、决策树、GBDT、K-means 等模型；深度学习模型比如 CNN、RNN、LSTM、 Seq2Seq、FastText、TextCNN 等。这些模型在后续的分类、聚类、神经序列、情感分析等示例中都会用到，这里不再赘述。下面是在模型训练时需要注意的几个点。

1.注意过拟合、欠拟合问题，不断提高模型的泛化能力。

过拟合：模型学习能力太强，以至于把噪声数据的特征也学习到了，导致模型泛化能力下降，在训练集上表现很好，但是在测试集上表现很差。

常见的解决方法有：

- 增大数据的训练量；
- 增加正则化项，如 L1 正则和 L2 正则；
- 特征选取不合理，人工筛选特征和使用特征选择算法；
- 采用 Dropout 方法等。

欠拟合：就是模型不能够很好地拟合数据，表现在模型过于简单。

常见的解决方法有：

- 添加其他特征项；
- 增加模型复杂度，比如神经网络加更多的层、线性模型通过添加多项式使模型泛化能力更强；
- 减少正则化参数，正则化的目的是用来防止过拟合的，但是现在模型出现了欠拟合，则需要减少正则化参数。

2.对于神经网络，注意梯度消失和梯度爆炸问题。

评价指标

训练好的模型，上线之前要对模型进行必要的评估，目的让模型对语料具备较好的泛化能力。具体有以下这些指标可以参考。

1.错误率、精度、准确率、精确度、召回率、F1 衡量。

错误率：是分类错误的样本数占样本总数的比例。对样例集 D，分类错误率计算公式如下：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$$

精度：是分类正确的样本数占样本总数的比例。这里的分类正确的样本数指的不仅是正例分类正确的个数还有反例分类正确的个数。对样例集 D，精度计算公式如下：

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)=y_i)\\ &= 1-E(f;D)\ .\end{aligned}$$

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例（True Positive）、假正例（False Positive）、真反例（True Negative)、假反例（False Negative）四种情形，令 TP、FP、TN、FN 分别表示其对应的样例数，则显然有 TP+FP++TN+FN=样例总数。分类结果的“混淆矩阵”（Confusion Matrix）如下：

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

准确率，缩写表示用 P。准确率是针对我们预测结果而言的，它表示的是预测为正的样例中有多少是真正的正样例。定义公式如下：

$$\mathrm{P}=\frac{TP}{TP+FP}$$

精确度，缩写表示用 A。精确度则是分类正确的样本数占样本总数的比例。Accuracy 反应了分类器对整个样本的判定能力（即能将正的判定为正的，负的判定为负的）。定义公式如下：

$$A = \frac{TP+TN}{TP+FN+FP+TN}$$

召回率，缩写表示用 R。召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确。定义公式如下：

$$R = \frac{TP}{TP+FN}$$

F1 衡量，表达出对查准率/查全率的不同偏好。定义公式如下：

$$F1 = \frac{2 \times P \times R}{P+R} = \frac{2 \times TP}{样例总数 + TP - TN}$$

2.ROC 曲线、AUC 曲线。

ROC 全称是“受试者工作特征”（Receiver Operating Characteristic）曲线。我们根据模型的预测结果，把阈值从0变到最大，即刚开始是把每个样本作为正例进行预测，随着阈值的增大，学习器预测正样例数越来越少，直到最后没有一个样本是正样例。在这一过程中，每次计算出两个重要量的值，分别以它们为横、纵坐标作图，就得到了 ROC 曲线。

ROC 曲线的纵轴是“真正例率”（True Positive Rate, 简称 TPR)，横轴是“假正例率”（False Positive Rate,简称FPR），两者分别定义为：

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{TN+FP}$$

ROC 曲线的意义有以下几点：

1. ROC 曲线能很容易的查出任意阈值对模型的泛化性能影响；
2. 有助于选择最佳的阈值；
3. 可以对不同的模型比较性能，在同一坐标中，靠近左上角的 ROC 曲所代表的学习器准确性最高。

如果两条 ROC 曲线没有相交，我们可以根据哪条曲线最靠近左上角哪条曲线代表的学习器性能就最好。但是实际任务中，情况很复杂，若两个模型的 ROC 曲线发生交叉，则难以一般性的断言两者孰优孰劣。此时如果一定要进行比较，则比较合理的判断依据是比较 ROC 曲线下的面积，即AUC（Area Under ROC Curve）。

AUC 就是 ROC 曲线下的面积，衡量学习器优劣的一种性能指标。AUC 是衡量二分类模型优劣的一种评价指标，表示预测的正例排在负例前面的概率。

前面我们所讲的都是针对二分类问题，那么如果实际需要在多分类问题中用 ROC 曲线的话，一般性的转化为多个“一对多”的问题。即把其中一个当作正例，其余当作负例来看待，画出多个 ROC 曲线。

模型上线应用

模型线上应用，目前主流的应用方式就是提供服务或者将模型持久化。

第一就是线下训练模型，然后将模型做线上部署，发布成接口服务以供业务系统使用。

第二种就是在线训练，在线训练完成之后把模型 pickle 持久化，然后在线服务接口模板通过读取 pickle 而改变接口服务。

模型重构（非必须）

随着时间和变化，可能需要对模型做一定的重构，包括根据业务不同侧重点对上面提到的一至七步骤也进行调整，重新训练模型进行上线。

参考文献

1. 周志华《机器学习》
2. 李航《统计学习方法》

3. 伊恩·古德费洛《深度学习》