**线性回归：**

文本, 信件

描述已自动生成

**生成与判别：**

* 生成模型:生成模型从概率层面的角度考虑，基本思想为先计算联合概率分布，再得到后验概率，利用其进行分类。生成方法关心给定输出x产生输出y的生成关系。
* 判别模型：判别模型由数据直接学习决策函数或者条件概率分布。判别方法关心输入怎样的X，应该预测怎样的输出Y。
* 自回归模型（autoregressive model）的特点:

GPT是典型的自回归模型，相比于RNN算法，使用Transformer特征抽取器使得GPT能够捕捉到更长范围得信息，同时计算速度比循环神经网络要快，易于并行优化。自回归模型所需要得数据更少，可用自身变量数列来进行预测。

**知识大竞赛，没你我不来：**

自然语言处理：

* 文本分类的概念：

文本分类（Text Classification 或 Text Categorization，TC），又称自动文本分类（Automatic Text Categorization），是指计算机将载有信息的一篇文本映射到预先给定的某一类别或某几类别主题的过程，实现这一过程的算法模型叫做分类器。文本分类问题算是自然语言处理领域中一个非常经典的问题。

根据预定义的不同，文本分类分为两种：二分类和多分类，其中多分类可由二分类来实现。

根据文章的类别不同，文本分类又可分为但标签和多标签，因为一篇文章可以有多个标签

* 文本分类历史：

文本分类最初是通过专家规则（Pattern）进行分类，利用知识工程建立专家系统，这样做的好处是比较直观地解决了问题，但费时费力，覆盖的范围和准确率都有限。后来伴随着统计学习方法的发展，特别是 90 年代后互联网在线文本数量增长和机器学习学科的兴起，逐渐形成了一套解决大规模文本分类问题的经典做法，也即特征工程 + 浅层分类模型。又分为传统机器学习方法和深度学习文本分类方法。

* 文本分类的应用场景：

情感分析：sentiment analysis ( SA)

话题标记：topic labeling(TL)

新闻分类：news classification (NC)

问答系统：question answering(QA)

对话行为分类：dialog act classification (DAC)

自然语言推理：natural language inference (NLD),

关系分类：relation classification (RC)

事件预测：event prediction (EP)

* 文本分类的流程：

文本分类流程现在主要基于机器学习与深度学习的文本分类，专家系统现在已经基本淘汰。

目前文本分类器学习过程主要分为浅层学习和深度学习

* 对于浅层学习

不考虑训练集与模型评估部分，浅层学习中整个文本分类问题分为特征工程和分类器两个部分。

* 获取训练集：

训练集的获取可采用任何方法获取，但主要包括网络爬虫技术和页面处理两种方法。

* 文本特征工程（浅层学习）

1. 文本预处理：

获得的任何原始数据，都需要进行预处理才可以被计算机进行操作使用。学会做好原始数据处理这一步骤对模型的拟合程度以及模型的训练速度都有很大的帮助。

1. 对于英文文本的预处理：

对于英文文本的预处理主要包括分割单词以及删除空格以及标点符号。另外值得一提的是大小写转换，以及英文单词词根的还原方面都可以是英文文本预处理的考虑因素。

1. 中文文本的预处理：

中文与英文语法上有较大的差距，故文本预处理方式上也有较大的差异。

中文文本预处理包括文本分词和去停用词两个阶段。

其中中文分词算法可分为：

1. 基于词典的分词方法。
2. 基于理解的分词方法。
3. 基于统计的分词方法。

而去停用词主要是从停用词表中提出定义为停用词的常用词。

* 文本特征提取（特征选择）

npl任务非常重要的一步就是特征提取（对应机器学习中的特征工程步骤，也叫做降维），在向量空间模型中，文本的特征包括字、词组、短语等多种元素表示 。在文本数据集上一般含有数万甚至数十万个不同的词组，如此庞大的词组构成的向量规模惊人，计算机运算非常困难。

进行特征选择，对文本分类具有重要的意义。特征选择就是要想办法选出那些最能表征文本含义的词组元素 。特征选择不仅可以降低问题的规模，还有助于分类性能的改善，选取不同的特征对文本分类系统的性能有非常重要的影响。

* 文本表示

文本是一种非结构化的数据信息，是不可以直接被计算的。文本表示的作用就是将这些非结构化的信息转化为计算机可以理解的结构化的信息，这样就可以针对文本信息做计算，来完成后面的任务。

文本表示的方法有很多种，传统做法有词袋模型和向量空间模型。

* 分类器（算法模型）

传统机器学习算法（浅层学习）

主要分为4种模型：

1. 基于规则的模型

基于规则的模型使用许多条规则来进行分类判别，具有时间复杂度低，运算速度快等特点。主要代表算法为决策树算法。

1. 基于概率的模型

基于概率的模型通过计算联合概率分布再计算后验概率来判断分类结果，代表算法为朴素贝叶斯算法，基于概率的模型在分类中表现出了较好的结果与鲁棒性。

1. 基于几何学的模型

基于几何学的模型代表算法为SVM算法，通过引入高维映射与核函数，线性不可分空间转化为高维线性可分空间，在高维空间中构造线性决策函数来实现原线性不可分空间的决策函数。SVM具有较好的推广能力，并且准确度较高

1. 基于统计的模型

基于统计的模型的代表算法为KNN算法，KNN算法没有训练过程，但通过统计学中多数表决的思想，使得knn算法在文本分类算法中也有一席之地。

深度学习方法:

前馈神经网络和递归神经网络是用于文本分类任务的前两种深度学习方法，与浅层学习模型相比，它们可以提高性能。然后，将 CNN，RNN 和注意力机制用于文本分类。许多研究人员通过改进 CNN，RNN 和注意力，或模型融合和多任务方法，提高了针对不同任务的文本分类性能。可以生成上下文化词向量的 BERT 的出现，是文本分类和其他 NLP 技术发展的重要转折点，该模型在包括文本分类在内的多个 NLP 任务具有更好的性能。此外，一些研究人员研究了基于 GNN 的文本分类技术，以捕获文本中的结构信息，这是其他方法无法替代的。

根据结构对文本分类的 DNN 算法可分为几大类：

1. 基于递归神经网络的方法（ReNN）

递归神经网络(ReNN)可以自动递归学习文本的语义和语法树结构，无需特征设计。递归自动编码器（RAE）用来预测每个输入句子的情感标签分布，并学习多词短语的表示。为了学习每个输入文本的成分向量表示，矩阵向量递归神经网络(MV-RNN)引入了 ReNN 模型来学习短语和句子的表示。

1. 基于多层感知机的方法（MLP）

多层感知器(MLP, multilayer perceptron)，俗称”vanilla”神经网络，是一种用于自动捕获特征的简单神经网络结构。

1. 基于循环神经网络的方法（RNN、

循环神经网络（Recurrent Neural Network，RNN）是一种用于处理序列数据的神经网络。相比一般的神经网络来说，他能够处理序列变化的数据。比如某个单词的意思会因为上文提到的内容不同而有不同的含义，RNN 就能够很好地解决这类问题。

1. 基于卷积神经网络的方法（CNN）

卷积神经网络(Convolutional neural networks, CNNs)起初用于图像分类，卷积滤波器(convolving filters)可以提取图像的特征。与 RNN 不同，CNN 可以同时将不同核定义的卷积应用于序列的多个块。因此，CNN 用于许多自然语言处理任务，包括文本分类。对于文本分类，需要将文本表示为类似于图像表示的向量，可以从多个角度对文本特征进行过滤。

1. 基于注意力机制的方法

注意力模型(Attentional mechanism, AM)最初被用于机器翻译，现在已成为神经网络领域的一个重要概念。注意力机制借鉴了人类的注意力机制。深度学习中的注意力机制从本质上讲和人类的选择性视觉注意力机制类似，核心目标也是从众多信息中选择出对当前任务目标更关键的信息。同样，在涉及语言或视觉的问题中，输入的某些部分可能会比其他部分对决策更有帮助。例如，在翻译和总结任务中，输入序列中只有某些单词可能与预测下一个单词相关。

1. 基于Transformer的方法

Transformer 是一种预训练的语言模型，可以有效地学习全局语义表示，并显著提高包括文本分类在内的 NLP 任务。通常使用无监督的方法自动挖掘语义知识，然后构造预训练目标，使机器能够学习理解语义。Transformer 可以在不考虑连续信息(sequential information)的情况下并行计算，适用于大规模数据集，因此在 NLP 任务中很受欢迎。

* 文本分类技术挑战

数据层面：

1. 目前模型依赖大量的标记数据，对无标注的文本分类效果不佳。
2. 对于外部数据的收集以及添加仍然是一个问题。
3. 减少训练过程中层次语义的丢失仍然是一个问题
4. 特定领域的单词以及文本难以使用

模型层面：

如何在增加数据和计算资源，和预测性能之间权衡依然值得研究的。

性能评估层面：

1. 提高模型的稳定性即鲁棒性仍然是关键。
2. 深度学习是一个黑盒模型，输出可解释性比较差，提升模型的可解释性也依然值得研究