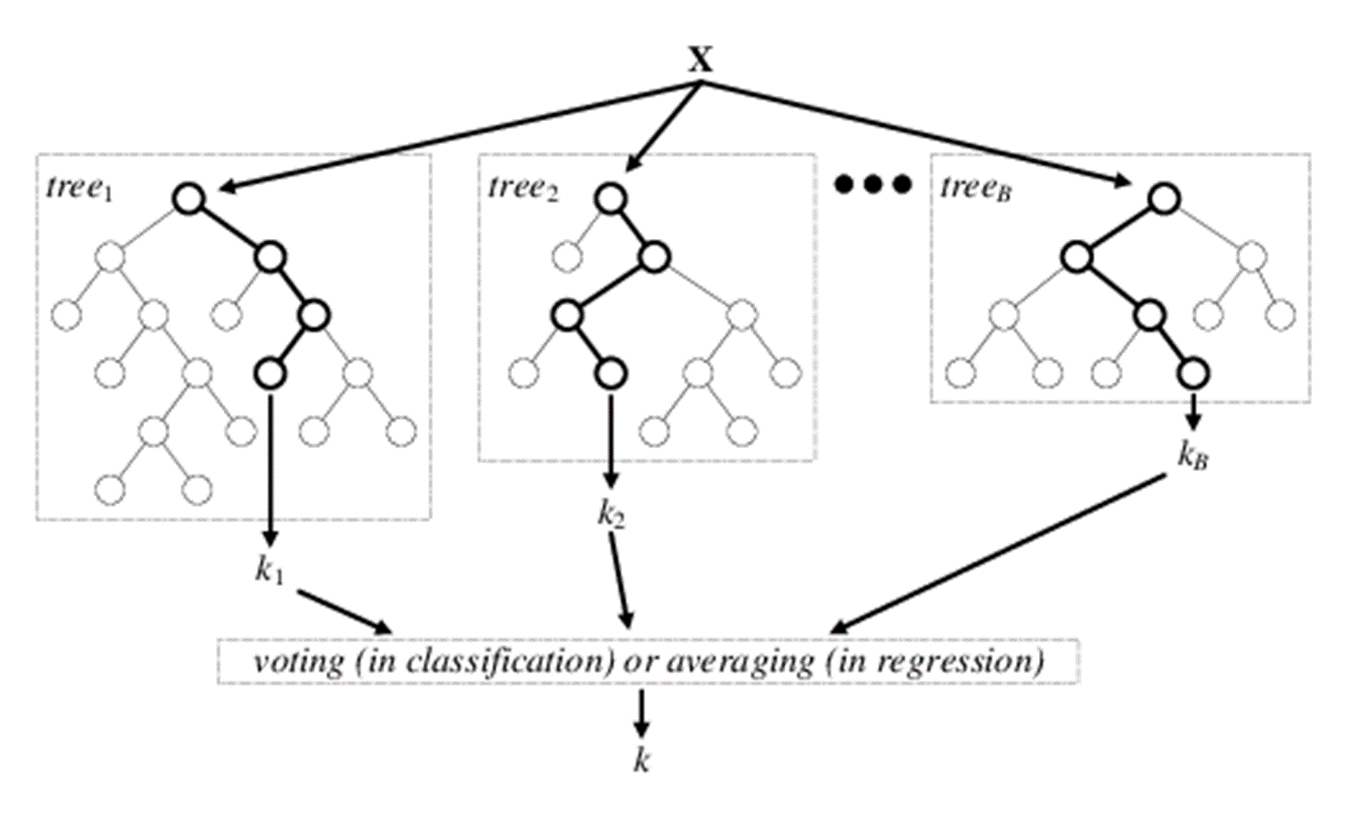
**相关模型说明文档**

1. **随机森林**

**(1). 模型介绍**

随机森林模型是由Breiman于2001年提出的一种集成学习模型。该模型使用决策树作为基学习器，它一方面像Bagging那样使用自助采样法训练各个决策树，另一方面创造性地在决策树训练过程中引入了随机属性选择，这就使得随机森林中基学习器地多样性不仅来自于样本扰动，还来自属性扰动，让最终集成的泛化性能可以通过个体学习器之间差异度的增加而进一步提升。随机森林可用于做分类和回归，大致流程如下图所示。



随机森林简单、容易实现、计算开销小并且在许多现实任务中展现出强大的性能，被誉为“代表集成学习技术水平的方法”。所以本项目中我们采样随机森林进行预测。

**(2). 模型应用**

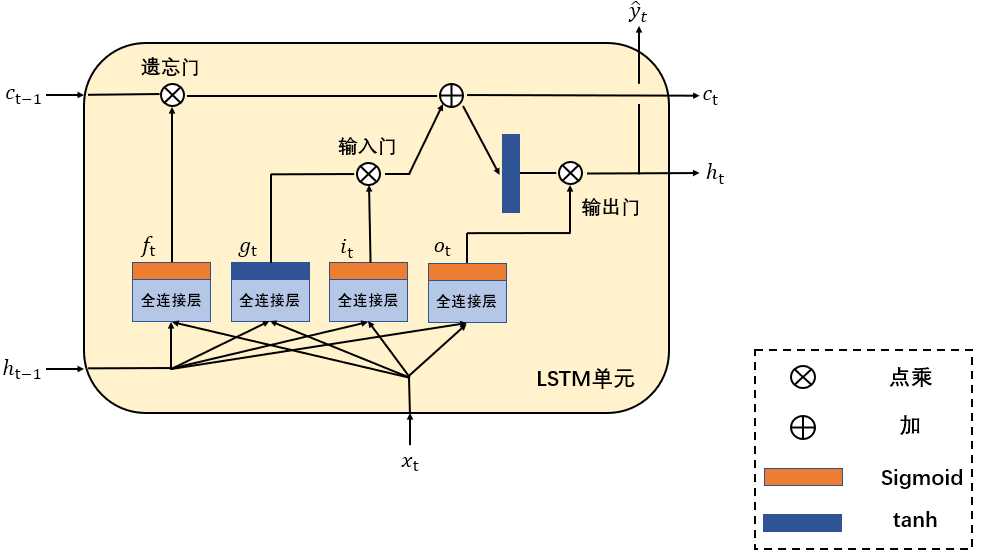
scikit-learn（简称sklearn）是一个重要的python机器学习库，支持包括分类、回归、降维和聚类四大机器学习算法，还包括了特征提取，数据处理和模型评估者三大模块。除此之外，sklearn拥有完善的文档，上手容易，具有丰富的API，在学术界颇受欢迎。因此在本项目中我们利用sklearn中的随机森林模型。

【具体应用在哪我不清楚】

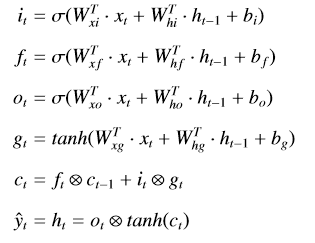
1. **LSTM**

**(1). 模型介绍**

长短时记忆模型（LSTM）现在已经被广泛地应用于序列数据处理问题中，比如自然语言处理、语音识别、DNA序列处理等。它是由Hochreiter 和 Schmidhuber 于 1997 年作为循环神经网络（RNN）的一个变种提出，与 RNN 不同的是，LSTM 通过内部独特的设计实现了对长期信息的记忆，其大致结构如下图所示：



LSTM中所涉及的运算有：



所有的W和b均为网络参数，在模型训练时更新。Sigmoid函数可以输出0到1之间的数值，这里被用来描述允许多少信息通过，0就表示不允许任何信息通过，1表示所有信息都可通过。LSTM内部使用三种记忆门(遗忘门，输入门和输出门)来控制记忆单元的状态，实现了对长期信息的记忆。

**(2). 模型应用**

由于LSTM模型在时间序列预测问题中具有广泛的应用，所以在本项目中我们利用LSTM模型进行多因子预测。我们首先利用连续若干天多因子数据训练LSTM模型，希望LSTM模型能学出多因子数据变化的潜在规律。训练结束后就可以利用该模型及当前掌握的最近若干天的多因子数据预测下一天的数据。

1. **CNN**

**(1). 模型介绍**

卷积神经网络（CNN）是指包含卷积运算并且具有较大深度的前馈神经网络。卷积神经网络的卷积运算可以使得网络可以提取输入数据的平移不变特征，网络的深度也可以让网络得以从原始输入开始一步步地提取出层次化的特征。卷积神经网络最引人注目的还是其表示学习（representation learning）能力，比如之前图像分类中人们想要手工提取合适的特征，而卷积神经网络可以自动从原始图片中提取合适特征，因此卷积神经网络现在被广泛应用于图像、语音的相关任务中。

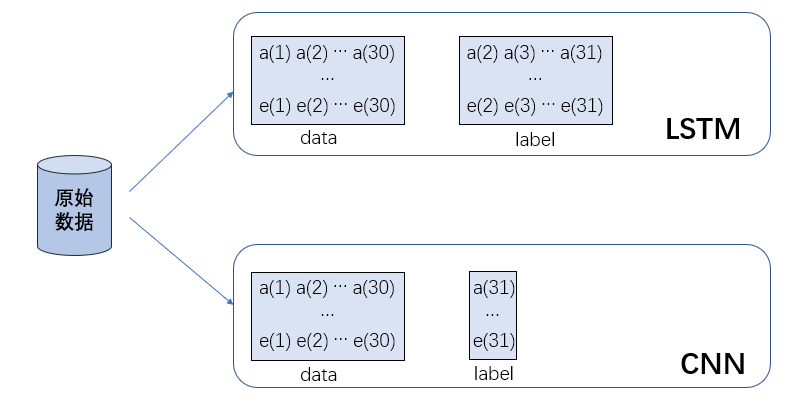
【可以插入一张CNN的示意图】

【可以介绍一下别人研究的CNN在金融方面的应用】

**(2). 模型应用**

这里我们利用CNN进行多因子数据的预测。由于CNN具有捕捉数据平移不变特征的能力以及表示学习能力，所以我们认为此处利用CNN模型是合理的。我们首先将多因子历史数据转化为训练集，训练集中每个样本的数据是5个因子连续30天的值，这里可以看成一张5\*30的单通道图片；样本的标签为5个因子在第31天的值。CNN模型训练结束后便可以用来由过去若干天的多因子值预测未来的多因子值。

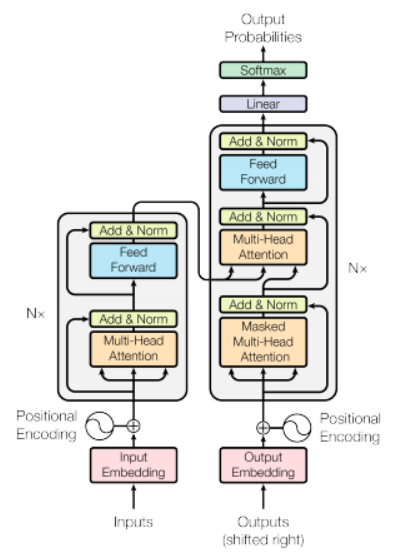
下面的图展示了LSTM和CNN训练样本的差别：



1. **Transformer**

**(1). 模型介绍**

Transformer是Google的Vaswani等人于2017年提出的一种新颖的序列到序列模型，该模型在翻译任务中取得了很好的结果。该模型的大致结构如下：



该模型左侧是编码器，右侧是解码器。编码器首先利用自注意力模块对输入信息进行编码并提取出有用的部分，然后利用前馈神经网络综合所得的信息传给解码器。解码器同样首先利用自注意力模块对已输出的前缀序列进行编码，该编码好的信息可以通过注意力机制帮助从编码器的信息中提取有用的部分，最后同样利用前馈神经网络综合所的信息，然后经过全连接层和softmax层得到最终输出。注意到之前的输出会在下个值的预测中作为输入，所以Transformer是一种典型的自回归模型（Autoregressive Model）。

**(2). 模型应用**

Transformer作为一种“序列到序列”的模型，我们认为也可以用在我们的项目中。Transformer对训练集结构的要求和LSTM一致，因此我们首先按照之前准备LSTM数据集的方法为Transformer准备训练集，之后训练Transformer。但我们在训练时发现损失函数在下降一段时间后便不再下降，而且预测精度不如LSTM和CNN，这可能是由于Transformer的embedding机制导致的，Transformer原本的embedding是为自然语言处理任务设计的，所以可能不适用于本任务。

1. **XGBoost**

**(1). 模型介绍**

XGBoost（eXtreme Gradient Boosting）是由陈天奇提出的一种著名boosting算法。XGBoost是一种提升树模型，它将许多树模型集成在一起，形成一个很强的分类/回归器。算法的核心思想是每轮添加一棵新树来拟合上一轮预测的残差，并通过最小化损失函数来不断优化模型参数。XGBoost被广泛地应用于各种数据科学竞赛和工业界中，得益于该算法的一些优点：可以有效防止过拟合、支持并行化、不用人为提取特征（即具有表示学习能力）。

**(2). 模型应用**

XGBoost和随机森林一样用于预测。。。，这两种方法的一个共同优点是几乎不需要人工特征工程。实验表明随机森林的效果要好于XGBoost。