机器学习及大模型关键信息梳理

(NLP)

一、算法

【有监督学习】

经典算法：

Logistic 回归

SVM

Decision Tree 决策树

KNN K-Nearest-Neighbor

Naive Bayes

Random Forest 随机森林

【无监督学习】

传统机器学习算法：

1.聚类（Clustering）算法：

tf-idf

Kmeans

层次聚类（Hierarchical Clustering）

DBSCAN

图聚类

2.降维（Dimensionality Reduction）算法：

PCA（主成分分析）

SVD（奇异值分解）

LDA — 主题词挖掘

传统深度学习：

1.前馈网络（Feedforward Neural Network, FNN）算法

MLP

2.卷积神经网络（CNN 系列）算法：

LeNet

ResNet

DenseNet

3. 循环神经网络（RNN 系列）算法：

RNN

LSTM

自监督学习：

生成式学习　对比式学习　预测式学习

1.生成式（Generative）算法：

AutoEncoder

Mask Modeling

Transformer

2.对比式（Contrastive）算法：

SimCLR

MoCo

GAN

3.预测式（Predictive）算法：

GPT

CPC

Seq2Seq

4.衍生模型：多模态自监督算法（文本+图像）：

CLIP

5.衍生模型：强化学习（Reinforcement Learning）

二、模型训练流程

【有监督学习】

1.数据采集：

开源数据集

私有数据集：实验室数据、私有部门爬取数据、业务内部数据

2.数据清洗：

分词

停止词去除

3.数据标注：

4.模型训练；

5.模型实验及性能评估

【无监督学习】

1.数据采集：

开源数据集

私有数据集：实验室数据、私有部门爬取数据、业务内部数据

2. 数据清洗（非必要）

3. 数据增强（非必要）

4. 文本表示：

Tokenizer: word segmentation word pepresentation

5. 文本特征学习：生成 batch

WordEmbedding

BERT

Bow

AE

6. 模型训练及调优

Constrastive Learning

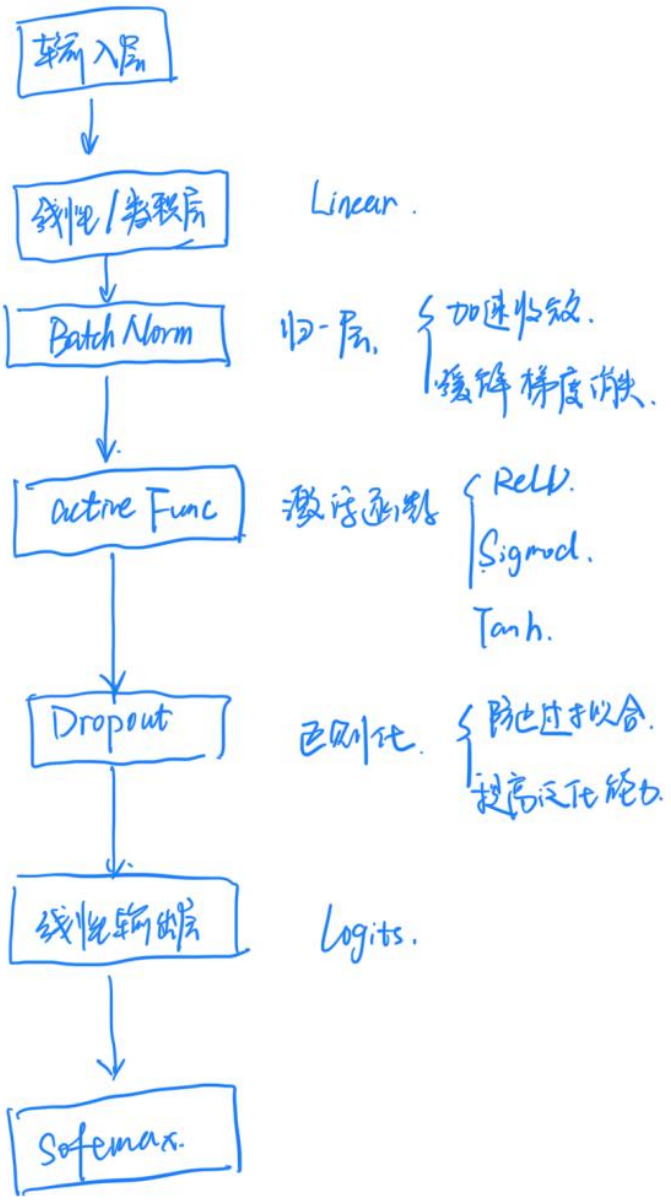
Mask Language Model

Pseudo label Generate

7. 模型实验及性能评估

AUC、F1 Score

三、深度学习技术模块（常见算法组件）



激活函数:

ReLU、LeakyReLU、GELU、Sigmoid、Tanh

正则化方法:

Dropout

Batch Normalization / LayerNorm

优化算法:

Adam、AdamW

SGD、Momentum

损失函数:

Cross Entropy

InfoNCE（自监督）

Focal Loss

MSE