NLP方向总结-激活函数

- 原理
  - 本质：神经网络加入加入非线性
  - 激活函数设计需要考虑的因素：
    非线性
    连续可微性
    有界性
    单调性
    平滑性
    原点附近近似Identity
- Sign
- Sigmoid
- Tanh
  - 双SS函数
  - tanh(z) = f(z) = e^z - e^-z / e^z + e^-z
- Softsign
- ReLU
  - The Rectified Linear Unit, 修正线性单元relu
  - ReLu = f(x) = max(0,x)
  - 缺点:
    没有边界，可以使用变种ReLU: min(max(0,x), 6)
    比较脆弱，比较容易陷入出现"死神经元"的情况, 当为负数时，即导数为0了，w不更新了，就是死神经元了
    解决方案:较小的学习率
    有可能大的梯度设置的权重使ReLU 单元始终为 0。这些"无效"的单元将始终为 0，很多计算在训练中被浪费了。
    ReLU有一个主要的缺点，即ReLU死亡问题，在这种情况下，多达50%的神经元在网络训练期间死亡。
    为了克服ReLU的不足，近年来提出了大量的激活方法，其中Leaky ReLU、Parametric ReLU 、ELU、Softplus、随机化Leaky ReLU是其
    中的几种，它们在一定程度上改善了ReLU的性能。
    Swish是谷歌脑组提出的非线性激活函数，对ReLU有一定的改善；GELU是另一种常用的平滑激活函数。可以看出，Swish和GELU都是ReLU的光滑近似。近年来，人们提出了一些提高ReLU、Swish或GELU性能的非线性激活方法，其中一些是ReLU或Leaky ReLU的光滑逼近方法，还有TanhSoft、EIS、Padé激活单元、正交Padé激活单元、Mish、ErfAct等。
  - ReLU 激活函数是你可以使用的最简单非线性激活函数。当输入是整数时，导数是 1，所以没有 S 型函数的反向传播错误导致的消失效果。研究表明，对于大型神经网络来说，ReLU 的训练速度要快很多。TensorFlow 和 TFLearn 等大部分框架使你能够轻松地在隐藏
    层使用 ReLU，你不需要自己去实现这些 ReLU。
  - 相比于Sigmoid和Tanh，提升收敛速度
    梯度求解公式简单，不会产生梯度消失和梯度爆炸, 解决了sigmoid的梯度消失问题
  - 1. 单侧抑制;
    2. 相对宽阔的兴奋边界;
    3. 稀疏激活性;
    4. 更快的收敛速度;
- P-ReLU
  - 类似Leaky ReLU，也是解决死神经元问题
  - f(x) = max(x,ax)
- Leaky-ReLU
  - Leaky Rectified Linear Units
  - 在ReLU函数的基础上，对x≤0的部分进行修正;目的是为了解决ReLU激活函数中容易存在的"死神经元"情况的;不过实际场景中:
    效果不>
    是太好。
  - LeakyReLU = f(x) = x when x>0; =αx when x≤0
- ELU
  - 指数线性激活函数，同样属于对ReLU激活函数的x≤0部分的转换进行指数修正，而不是和Leaky ReLU中的线性修正
- Maxout
  - 可以看作是在深度学习网络中加入一层激活函数层，包含一个参数k，拟合能力特别强。特殊在于:增加了k个神经元进行激活，然后输出激活值最大的值
- gelu
  - bert用的激活函数
  - Gaussian Error Linerar Units
  - 0.5 * x * (1 + torch.tanh(math.sqrt(2 / math.pi) * (x + 0.044715 * torch.pow(x, 3))))
- SwiGLU
- Swish
  - https://arxiv.org/abs/2002.05202
- softmax
  - import numpy as np
    logits = [2.0, 1.0, 0.1]
    a = np.exp(logits)
    softmax = a/np.sum(a)
    print(softmax)
    print(sum(softmax))
  - S\left(y_{i}\right)=\frac{e^{y_{i}}}{\sum_{j} e^{y_{j}}}
  - softmax function是指数族函数，输出元素在(0,1)之间，总和为1，所以可以解释为概率，softmax输出的向量代表了输入列的可能的概率
    softmax是把差距和不同加大，把结果压缩推向1或者0, softmax最主要的目的是把数字转换称可能性