6 对比学习 和 互信息?

问题导向：

为什么最大互信息就是对比学习?

对比学习可以看成一个距离度量，度量准则就是互信息

参考资料：

两个损失函数https://blog.csdn.net/yyhaohaoxuexi/article/details/113824125 (https://blog.csdn.net/yyhaohaoxuexi/article/details/113824125)

互信息计算：https://zhuanlan.zhihu.com/p/149743192 (https://zhuanlan.zhihu.com/p/149743192)

交叉熵的理解：https://zhuanlan.zhihu.com/p/61944055 (https://zhuanlan.zhihu.com/p/61944055)

softmax函数：https://blog.csdn.net/lz_peter/article/details/84574716?utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1.no_search_link&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1.no_search_link (https://blog.csdn.net/lz_peter/article/details/84574716?utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1.no_search_link&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7Edefault-1.no_search_link)

6.1 pre-knowledge：

内积 存在两个向量 $a = [a_1, a_2, a_3, \dots] \; b = [b_1, b_2, b_3, \dots]$

那么a，b 内积为 $a \cdot b = a_1 b_1 + a_2 b_2 + a_3 b_3$

余弦相似度 a·b = |a||b|cos(a, b)

cos(a, b) = (a·b) / (|a||b|)

理解：余弦相似性可看作内积的归一化，缩小范围 余弦相似性只考虑向量夹角大小，而内积不仅考虑向量夹角大小，也考虑了向量的长度差

比如两个向量 A 和 B，A=(1,1,0) B=(0, 1, 1)，AB余弦相似度为 1/(sqrt(2) * sqrt(2)) = 1/2， 余弦相似度不考虑向量长度，(1,1,0) 和 (0, 3, 3) 的相似度 等于 AB的相似度

但是，如果入股向量的长度对相似性有真实影响，那么 A(1, 1) B(4, 4) C(5, 5) 三个向量，相似度相同，但 BC 内积 大于 AB 内积，故 BC 更相似

交叉熵可在神经网络(机器学习)中作为损失函数，p表示真实标记的分布，q则为训练后的模型的预测标记分布，交叉熵损失函数可以衡量p与q的相似性。交叉熵作为损失函数还有一个好处是使用sigmoid函数在梯度下降时能避免均方误差损失函数学习速率降低的问题，因为学习速率可以被输出的误差所控制。 在特征工程中，可以用来衡量两个随机变量之间的相似度。

互信息量的下界估计一般是用来做互信息量增大的工作，互信息量上界估计一般是用来做互信息量减小的工作。

6.2 对比学习理解

对比学习前面提到过相关的概念和例子，简单来说，拉近正样本，远离负样本，更能学习到编码器的特征。

但如果从一般范式出发：

$$S((f(x), f(x^+))) << S((f(x), f(x^+)))$$

这里的x^+就是和x类似的样本，x^-就是和x不相似的样本，S(也可写成score)，这是一个度量样本之间相似程度的函数

如何定义目标函数？最简单的一种就是上面提到的内积函数，另外一种 triplet 的形式就是 $l = max(0, \eta + s(x, x^+) - s(x, x^-))$，也就是希望正例和负例隔开$\eta$至少的距离，或者写成另外一种形式：$s(x, x^+) = ||max(0, f(x) - f(x^+))||$

也就是说我们期望找到一个好的相似度函数

6.3 为什么要用到互信息?

一般的原则是：

好特征的基本原则应当是“能够从整个数据集中辨别出该样本出来”，即重点在于不是如何提取每一个细节，而是要想办法提取最特别的、能够与其他样本有所区分的特征。而这个时候，便采用互信息来衡量这个特征是否独特。

而且互信息还有个优点在于：互信息可以捕捉到变量间非线性的统计相关性，更值得信赖

$$\begin{aligned} I(X; Z) &= H(X) - H(X|Z) \\ &= \int_{XZ} \log \frac{dP_{XZ}}{dP_X \otimes dP_Z} \, dP_{XZ} \end{aligned}$$

第一个式子认为，互信息就是当给定变量Z时，观察者对变量X的不确定度的减少量； 根据第二个公式可得到互信息和KL散度的联系,可用于后续互信息的估计 $I(X; Z) = D_{KL}(P_{XZ} \| P_X \otimes P_Z)$

因此重点在于：对比学习要想利用互信息，该加大什么互信息，才能更好的设计正负样本，从而利用对比学习来设计任务提高

6.4 从对比学习的损失函数出发

对比学习中一个比较典型的 score 函数就是就是向量内积，即优化下面这一期望：

$$\mathop{\mathbb{E}}_{x, x^+, x^-} \left[-\log \left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)} + e^{f(x)^T f(x^-)}} \right) \right]$$

如果对于一个x，我们有1个正例和 N-1个负例，那么这个 loss 就可以看做是一个 N 分类问题，实际上就是一个交叉熵，而这个函数在对比学习的文章中被称之为 InfoNCE。

查阅资料，一些经典论文的损失函数如下所示：

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$

$$\mathbb{E}_{x\sim p, x^+\sim p_x^+}\left[-\log\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+\frac{Q}{N}\sum_{i=1}^{N}e^{f(x)^T f(x_i^-)}}\right]$$

这两个式子形式上看起来不一样，但感觉其实本质上差不多一个道理，

都是进行点积（或者叫做得分？）分子体现为正样本得分，分母体现为总样本得分（正样本加负样本） 即mean(-log(正样本的得分/所有样本的得分))。

那么想学习到一个好的特征表示，那么就寻找这个loss的最小值，可以从一方面入手：加强式子中的分子成分，即加强f(x)与f(x+)的点积，换种角度来说，就是追求f(x)与f(x+)的相似度，因此可以采取互信息的角度来进行衡量，因此互信息和其对比学习就能联系起来，改向追求锚点数据与正样本的最大互信息。

所以最后总结为：从追求loss最小值转变为互信息最大值，加大互信息，学习是学习f函数

6.5 NCE到InforNCE的推导

从NLP领域推导：https://www.cnblogs.com/hansjorn/p/14314400.html (https://www.cnblogs.com/hansjorn/p/14314400.html) 看不太懂但大为震撼

6.6 互信息最大化的推导

苏神从Deep INFOMAX 的论文为背景开始分析，互信息的定义出发，开始推导如何互信息最大化

<https://www.jiqizhixin.com/articles/2018-10-12-11 (https://www.jiqizhixin.com/articles/2018-10-12-11)

或者参考这个：互信息评估：https://zhuanlan.zhihu.com/p/149743192 (https://zhuanlan.zhihu.com/p/149743192)