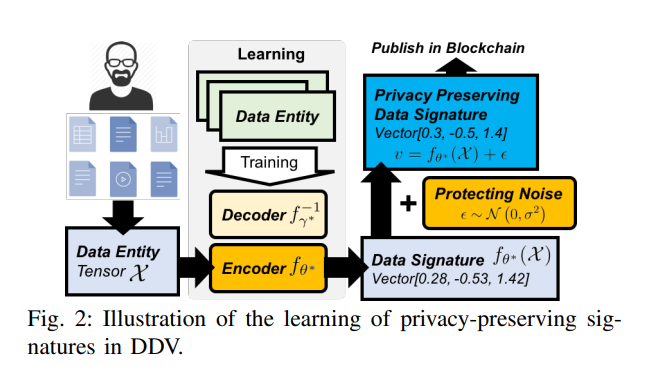
# 利用数据嵌入和距离度量技术实现数据检索

## 需求分析

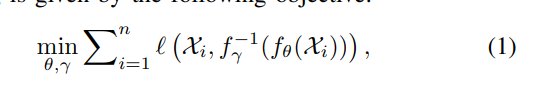
在我们的DDM(decentralized data marketplace)中，卖家需要将自己的数据的详细信息发布到链上以供买方检索，如何在保证数据隐私不被泄露的前提下，让买方能够检索到想要的数据是一件具有挑战性的事情。论文*ZHOU J,TANG F,ZHU H,et al．Distributed data vending on blockchain*提出一种结合数据嵌入和距离度量学习的技术，它首先学习一个隐私保护签名来有效地总结原始隐私数据，然后使用通过选择的相似性度量来计算相似性，最后从数据库中实现智能合约的检索，无需损害私有数据的完整性。

## 实现原理

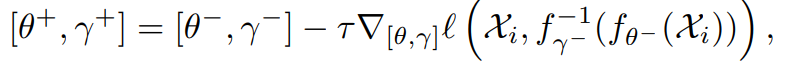


### 数据嵌入

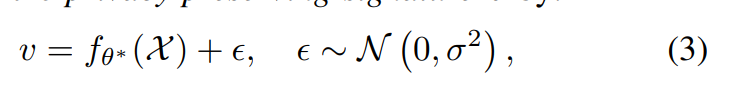
其中θ和γ分别表示加密函数f和其反函数的一系列参数的向量。

论文中给出了计算θ和γ的目标函数

在有一定的数据样本的情况下，可以通过机器学习算出其最优解，即获得了将高维数据映射成向量的函数f，这个向量可以在最大程度上保持原有数据特征，买家将这个一维向量发布到链上。

由于这个方法需要样本量的支撑，在运营初期可能没有那么多的样本供机器学习，所以论文提出了一种随样本量不断优化学习的方法：

另一方面，由于加密函数f是公开的，其反函数f-1也是公开的。在某个买家在购买了一定量的数据之后，可能能够推算出γ从而恢复原始数据，鉴于此，论文做出了一个trade-off，加入了一个高斯噪音来破坏这个可逆的推算过程（然而，这样会影响检索性能）：



对于我们的DDM，可以利用这项技术来实现数据的发布。

### 相似性检索

对于一个查询请求v`, 它和数据嵌入向量v之间的马氏距离决定了它们之间的相似程度：



其中Mt描述了该特定查询的参数矩阵。

## 计划

由于我之前没怎么弄过机器学习的项目，但是我在读研之前学习过机器学习相关的基础知识，所以我预计这个部分将花费1.5个月时间来实现。具体安排如下：

* 两周周时间攻坚机器学习，两周后给出详细设计文档
* 两周写代码搭建模型
* 两周测试实验数据

预计1月中旬完成。