在市场法中应用神经网络算法

分为以下步骤

1. 获取数据
2. 数据预处理
3. 数据归一化
4. 模型构建
5. 模型优化
6. 模型验证、偏差修正

获取数据

通过爬虫从数据购买网站上获得大量数据
期望这些数据的字段越丰富越好，比如数据量、数据使用范围、数据大小、数据来源方式，数据来源类型、流入数据数量、维护频率等等，可以参考这张图

特征维度	输入维度
颗粒度	数据属性数量、数据属性类型、数据属性精度、数据属性准确性、数据属性长度、数据属性完整性、数据属性合规性、维护频率、数据属性格式、编码方式、标准、命名规则
多维度	数据来源种类、数据来源数量、数据来源方式、数据来源类型、数据覆盖范围、数据重复率、数据一致性、数据采集方式
活性度	数据更新频率、数据访问频率、数据存在时间、数据更新差异大小、访问系统数量、常用属性数量、累积访问次数、累积更新次数
规模度	数据量、数据使用范围、数据大小、数据增长速度、数据获取难易程度、数据独占程度
关联度	流入数据数量、流出数据数量、流入数据频率、流出数据频率、流入数据大小、流出数据大小、流入数据关联强度、流出数据关联强度、数据依赖程度、数据独立程度

数据预处理

对数据进行清洗，使数据中不再存在不完整、无效、重复的数据；

修正错误、不一致数据，一些错误的数据可能会影响模型的效果；

结构化非结构化数据，比如数据来源是“京东万象”，没法拿中文去输入到模型里，那就先把各个来源的文字对应成数字再输入，”京东万象“对应0，”聚合数据“对应1

数据归一化

进行归一化，即将所有数据的取值映射到一个方便应用神经网络算法，符合激活函数值域的较小区间内。

因为有些数据的大小太大，可能不方便算法直接处理；有些数据可能会有0、负数，也不方便处理。

于是，把这些数据通过一个映射函数映射到比较好的一个范围内。

有这些"映射函数"

1. Min-Max标准化（线性标准化）

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

x'是结果，x是初始输入的值，x_{min}是同一输入中最小的值，x_{max}是同一输入中最大的值，取值范围为[0,1]

2. Z-Score标准化

$$x' = \frac{x - \mu}{\sigma}$$

x'是结果，x是初始输入的值，μ是样本数据的均值，σ是样本数据的标准差，没有确定取值范围，处理后的数据呈正态分布，对于含0较多且特征稀疏的数据样本不适用

3. 对数函数标准化

$$x' = \frac{log_{10} x}{log_{10} x_{max}}$$

x'是结果，x是初始输入的值，x_{max}是同一输入中最大的值，取值范围为[0,1]，需要输入数据大于1

4. 其他

   建议根据具体数据，自己设计一个归一化函数

模型构建

大体概括就是

特征维度	输入维度
颗粒度	数据属性数量、数据属性类型、数据属性精度、数据属性准确性、数据属性长度、数据属性完整性、数据属性合规性、维护频率、数据属性格式、编码方式、标准、命名规则
多维度	数据来源种类、数据来源数量、数据来源方式、数据来源类型、数据覆盖范围、数据重复率、数据一致性、数据采集方式
活性度	数据更新频率、数据访问频率、数据存在时间、数据更新差异大小、访问系统数量、常用属性数量、累积访问次数、累积更新次数
规模度	数据量、数据使用范围、数据大小、数据增长速度、数据获取难易程度、数据独占程度
关联度	流入数据数量、流出数据数量、流入数据频率、流出数据频率、流入数据大小、流出数据大小、流入数据关联强度、流出数据关联强度、数据依赖程度、数据独立程度

输入的东西：右边那一栏"输入维度"里的东西

输出的东西：左边那一栏的"特征维度"

比如，我输入数据属性数量、数据属性类型、数据属性精度、数据属性准确性、数据属性长度、数据属性完整性等等，输出的是这些数据的颗粒度

对于一批数据，把这5个特征维度都通过模型算出来，比如一个京东万象上象棋的数据集，颗粒度为9.8，多维度为2.6，活性度为5.6，规模度为4.5，关联度为2（瞎编的）

（后续会对这几个维度再进行处理，得到这个数据集的价值，可能会用到洛伦兹-pagerank那个论文的算法，现在就只需要通过RNN算出来数据的这5个维度）

模型优化

可选算法：贝叶斯优化、梯度下降、基于置信上界的乐观优化、随机搜索

应用市场法

(初步理解，主要参考论文《基于洛伦兹变换和PageRank算法的数据资产估值》)

根据上面的算法可以算出来每个数据集的颗粒度、多维度、活性度、规模度、关联度。把所有的已知的数据集（比如京东万象上收集的）都算出来这5个度，存到数据库里

现在来了一个新的数据集，想要上市，想要知道数据的价值

根据这篇论文，粗略地描述一下算法：找到和这个新的数据集相像的几个数据集，比较新的数据集和这几个数据集的颗粒度、多维度、活性度、规模度、关联度，然后根据比较的结果来调整价值
（乱举个例子：我的数据集的颗粒度更好，活性度更好，所以价格再加个500元，但是规模度差，价格减个200元；跟5个数据集比较后得到5个价格，取个平均值）

其中，“找到和这个新的数据集相像的几个数据集”会用到pagerank算法，表达几个数据集的关联度；“根据比较的结果来调整价值”要根据5个度的权值来修改，这个地方我觉得可以我们自己来定义，我有一个想法就是

$$\sum_{i=1}^{5}\left(\left(1+\left(\text{原}XX\text{度}i-\text{现}XX\text{度}i\right)\times 100\%\right)\times \text{权值}i\right)\times \text{数据集价格}$$

上面说的只是粗略的说法，还有各种修正系数之类的，保证算法的准确性