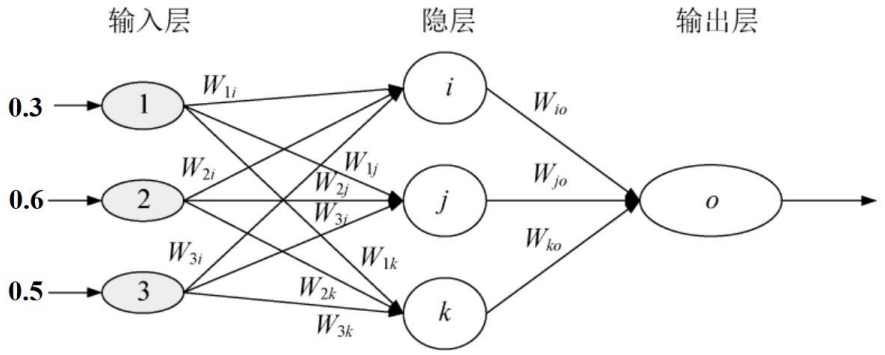
暨南大学本科作业专用纸

数据挖掘课程作业

学院 信息科学技术学院 系 计算机科学系 成绩评定

学生姓名 陈宇 学号 2020101642

一．对于输入实例［0.3，0.6，0.5］，如图所示的神经网络



其中：

w_{1i}	w_{2i}	w_{3i}	w_{1j}	w_{2j}	w_{3j}	w_{1k}	w_{2k}	w_{3k}	w_{io}	w_{jo}	w_{ko}
0.10	0.20	0.30	–0.20	–0.10	0.10	0.10	–0.10	0.20	0.3	0.5	0.4

计算：

答：假设在隐含层中选择使用 relu 激活函数、输出层使用 sigmoid 激活函数：

（1）计算节点 *i*、节点 *j* 和节点 *k* 的输入值和输出值。

对于 i 点输入=0.1*0.3+0.6*0.2+0.5*0.3=0.3

对于 i 点输出=relu(0.3)=0.3

对于 j 点输入=(-0.2)*0.3+(-0.1)*0.6+0.1*0.5=-0.07

对于 j 点输出=relu(-0.07)=0

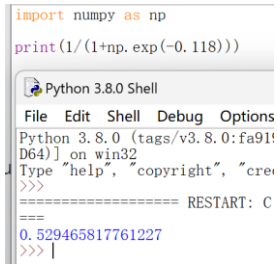
对于 k 点输入=0.1*0.3+(-0.1)*0.6+0.2*0.5=0.07

对于 k 点输出=relu(0.07)=0.07

（2）计算节点 *o* 的输入值和输出值。

对于 o 点输入=0.3*0.3+0*0.5+0.07*0.4=0.118

对于 o 点输出=sigmoid(0.118)= 0.529465817761227



二、数据挖掘十大经典算法是什么，请对每个算法进行简要概述（凝练语言，每个算法不要超过 200 字）
答：
1. Apriori 算法：用于频繁项集挖掘，通过迭代生成候选项集，再计算其支持度，剪枝得到频繁项集。基本原理是迭代生成候选项集，并计算每个候选项集的支持度。如果一个项集的支持度超过了预定的最小支持度阈值，则将其视为频繁项集。然后，通过组合频繁项集来得到更长的项集，并进行支持度计算和剪枝操作，直到没有更多的频繁项集可以生成为止。Apriori 算法的缺点是需要生成大量的候选项集，计算支持度的时间复杂度较高。

2. EM 算法：EM 算法是一种迭代算法，用于处理含有隐变量的概率模型。EM 算法通过交替进行两步操作来优化模型参数：E 步骤计算隐变量的后验概率，M 步骤更新模型参数。EM 算法的优点是能够处理缺失数据、噪声数据和不完整数据等情况，但需要选择合适的初始参数和停止条件，且在计算后验概率时可能会陷入局部最优解。

3. K-Means 算法：一种聚类算法，将数据分为 k 个簇，每个簇的中心点为簇的质心，通过迭代优化质心位置实现聚类。首先，随机初始化 k 个质心，然后将每个数据点分配到最近的质心中，形成 k 个簇。接着，计算每个簇的质心，并将质心作为新的聚类中心。重复以上步骤直到簇不再发生变化或达到最大迭代次数。K-Means 算法的优点是简单易懂、计算速度快，但需要预先确定聚类簇数 k，且对初始质心的选择敏感。

4. 分类与回归树（CART）算法：一种决策树算法，通过对数据的分裂与剪枝来生成分类或回归树。CART 算法的基本思路是将数据集递归地划分为更小的子集，直到子集中的数据属于同一类别或达到预定的停止条件。在分类问题中，CART 算法通过计算基尼不纯度来选择最优的分裂点，生成分类树。在回归问题中，CART 算法通过计算平方误差和来选择最优的分裂点，生成回归树。CART 算法具有可解释性强、易于实现等优点，但容易过拟合，需要使用剪枝等方法来缓解。

5. 朴素贝叶斯算法：一种分类算法，基于贝叶斯定理，通过计算先验概率和条件概率来预测新数据的类别。朴素贝叶斯算法假设每个特征之间相互独立，因此可以将多维特征向量的联合概率分解为各个特征的条件概率的乘积。在分类前，需要从训练数据中计算各个类别的先验概率和每个特征在各个类别下的条件概率。对于新的数据，通过计算数据在各个类别下的后验概率，选择具有最大后验概率的类别作为预测结果。朴素贝叶斯算法适用于高维数据和大规模数据集，但对于特征之间存在依赖性的情况，表现可能不佳。

6. 支持向量机（SVM）算法：一种分类算法，通过在高维空间中找到最优的超平面来区分不同类别的数据。SVM 算法的基本思路是通过选择具有最大间隔的超平面来实现分类，间隔是指每个类别最靠近超平面的数据点到超平面的距离。SVM 算法可以通过核函数将低维特征空间映射到高维空间，从而处理非线性问题。在训练时，SVM 算法通过优化损失函数来寻找最优的超平面；在预测时，

将新数据映射到高维空间，根据所处位置来判断其所属类别。SVM 算法具有泛化能力强、对于小样本效果好等优点，但需要预先确定核函数和正则化参数等参数。

7. AdaBoost 算法：一种集成学习算法，通过迭代训练多个弱分类器，并根据分类器的表现来调整样本权重，最终组合成一个强分类器。在每一轮迭代中，AdaBoost 算法通过调整样本权重来聚焦于错误分类的样本，使得下一轮训练的分类器更加关注错误分类的样本。同时，AdaBoost 算法通过对每个分类器的表现进行加权，来产生最终的分类器。AdaBoost 算法具有较高的精度和泛化能力，但对于噪声数据和异常值等情况，容易过拟合。

8.PageRank 算法： PageRank 算法是一种链接分析算法，用于评估网页的重要性。PageRank 算法基于网页之间的链接关系，通过计算每个网页的入链数和入链质量来评估网页的重要性。PageRank 算法将网页之间的链接关系表示为一个图，通过迭代计算每个网页的 PageRank 值，最终得到每个网页的重要性排名。PageRank 算法的优点是能够处理大规模的网络，且可以通过调整阻尼因子来平衡网页的重要性和随机浏览的可能性。

9.C4.5 算法： C4.5 算法是一种决策树算法，用于分类和回归任务。C4.5 算法通过计算信息增益比来选择最优的分裂点，生成决策树。C4.5 算法具有高可解释性、易于实现等优点，能够处理连续型和离散型特征，但对于缺失数据的处理较为复杂。

10.KNN 算法：KNN 算法是一种基于实例的分类算法，通过计算新数据点与已知数据点之间的距离来确定其类别。KNN 算法将数据集存储在内存中，对于每个新数据点，计算其与所有已知数据点的距离，然后选取距离最近的 k 个数据点，并根据这 k 个数据点的类别来预测新数据点的类别。KNN 算法的优点是简单易懂、无需训练、对于异常值和噪声数据有较好的容错性，但需要选择合适的 k 值，且对于高维数据和大规模数据集效果较差。