Assignment 3[1]

1.考虑如下数据产生过程

$$y = 2 + cos(2x) + 0.5\varepsilon, \tag{1}$$

其中，$x \in [-10, 10]$(1000等分)，$\varepsilon \sim N(0,1)$，样本量为2000。

(i)将样本随机分为两份，训练集样本量1000，测试集样本量1000。

(ii)基于训练集，绘制(x,y)散点图。

(iii)基于训练集，使用双隐藏层（第一个隐藏层20个神经元，第二个隐藏层10个神经元）的神经网络模型对数据进行拟合，训练300步，每30步绘制训练效果图。

(iv)基于测试集和均方误差损失函数，评估神经网络的样本外预测表现。

[1]Due date: 6:30pm 5/23/2025

2.修改*CBOW.ipynb*，使其包括两个隐藏层。两个隐藏层神经元的维度可以相同，也可以不同。输入层到第一个隐藏层的结构跟经典CBOW模型一致（即将上下文窗口中每个词的one-hot representation通过矩阵W转化为对应的隐状态并以所有词对应隐状态的平均作为第一个隐藏层的神经元），第一个隐藏层的输出经过ReLU激活函数后传给第二个隐藏层（加入偏置参数），第二个隐藏层的输出经过ReLU激活函数后传给输出层（加入偏置参数），输出层神经元的维度等于词典大小，最后使用softmax变换将输出转化为预测概率。将修改以后的模型应用于一个中文文本，并查看生成的词向量。

3.以倚天屠龙记.*txt*为文本库，使用gensim包对其进行词向量建模，寻找与下列词语最相近的关联词语：张无忌、武当派、冰火岛。

4.下载腾讯英文词向量（100维/200维均可），修改*影评情感分析.ipynb*，将一篇影评中的每个词替换为其对应的腾讯词向量（腾讯词向量中没有的词可忽略），将一篇影评中所有词对应的词向量平均作为该影评的向量表示，运用logistic regression预测该影评的sentiment，并以混淆矩阵评估模型的分类效果。