**人工智能导论实验指导书**

**实验环境配置：**

* Anaconda3
* Keras/Tensorflow
* OpenCV（python版本）
* sklearn

**实验1 手写体识别实验**

**1.1 实验目标**

通过人工智能导论所学内容，应用多层感知机以及深度学习的方法完成图像分类的建模

**1.2 实验内容**

**1.2.1 卷积神经网络**

利用tensorflow或者keras等深度学习框架在jupyter notebook上完成：使用卷积神经网络在MNIST数据集上完成分类任务，并打印测试集准确率。

**1.2.2 多层感知器**

利用tensorflow或者keras深度学习框架在jupyter notebook上完成：使用MLP在MNIST数据集上完成分类任务，并打印测试集准确率。

多层感知器(Multi-Layer Perceptron，MLP)也叫人工神经网络(Artificial Neural Network，ANN)，除了输入输出层，它中间可以有多个隐层。最简单的MLP需要有一层隐层，即输入层、隐层和输出层才能称为一个简单的神经网络。

**1.3 实验数据**

模型使用的MNIST数据集，该数据集是目前最大的数字手写体数据集（0~9），总共包含60,000张训练图像和10,000张测试图像，每张图像的大小为28x28的灰度图。示例如下：

MNIST数据集来自美国国家标准与技术研究所，National Institute of Standards and Technology(NIST)。训练集 (training set) 由来自 250 个不同人手写的数字构成, 其中 50% 是高中学生, 50% 来自人口普查局 (the Census Bureau) 的工作人员.。测试集(test set) 也是同样比例的手写数字数据。训练数据集包含60,000 个样本, 测试数据集包含 10,000 样本. 在 MNIST 数据集中的每张图片由28 x 28 个像素点构成, 每个像素点用一个灰度值表示，手写数字的类标签(整数0-9)。图片数据集的下载以及导入可参考：

import tensorflow as tf  
from tensorflow.examples.tutorials.mnist import input\_data   
mnist = input\_data.read\_data\_sets("MNIST\_data/", one\_hot=True)

**1.4 实验验收**

保留jupyter notebook上的运行结果，现场验收。

**实验2 文本情感分析实验**

**2.1 实验目的**

通过人工智能导论所学内容，应用机器学习以及深度学习的方法完成文本情感分析的建模。实验目标：

* 初步掌握经典的机器学习方法SVM在文本分类的应用
* 初步掌握循环神经网络及卷积神经网络在文本分类的应用
* 比较机器学习以及深度学习方法的异同

**2.2 实验内容**

**2.2.1 机器学习模型**

利用sklearn中自带的SVM库函数在jupyter notebook上完成IMDB数据集分类任务

**2.2.2 深度学习模型**

利用tensorflow或者keras等深度学习框架在jupyter notebook上完成：

* 单独使用循环神经网络如GRU或者LSTM在IMDB数据集上完成分类任务，并且打印测试集准确率；
* 单独使用一维卷积神经网络在IMDB数据集上完成分类任务，并且打印测试集准确率；
* 组合使用一维卷积神经网络以及循环神经网络级联在IMDB数据集上完成分类任务，并且打印测试集准确率；

对于库函数或者各网络中的超参数，如句子长度、词嵌入大小、循环神经隐状态大小，卷积核尺寸以及数目等自行调整，最终取测试集上准确率最高时的超参数。

**2.3 实验数据集**

IMDB情绪分类数据集来自IMDB用户的50,000个电影评论组成，标签为二分类：positive（1）和negative（0）。请点击数据[下载](https://pan.baidu.com/s/1qy8YSkvMasTvmrBDP32Cmg)，提取码8td7。

本链接的评论数据都是预处理的，每一个评论样本都被编码为一个整数形式的单词索引序列，同时为了得到的向量数据不会太大，该链接里的数据集仅保留前10000个最常见出现的单词，低频单词已被舍弃。评论中的单词按照它们在数据集中的总体频率进行索引，即整数编码i为第i个最频繁的词。例如本链接训练集第一个样本为：

[1,14,22,16,43,530,973,1622,1385,65,458,4468,66,3941,4,173,36,256,5,25,100,43,838,112,50,670,2,9,35,480,284,5,150,4,172,112,167,2,336,385,39,4,172,4536,1111,17,546,38,13,447,4,192,50,16,6,147,2025,19,14,22,4,1920,4613,469,4,22,71,87,12,16,43,530,38,76,15,13,1247,4,22,17,515,17,12,16,626,18,2,5,62,386,12,8,316,8,106,5,4,2223,5244,16,480,66,3785,33,4,130,12,16,38,619,5,25,124,51,36,135,48,25,1415,33,6,22,12,215,28,77,52,5,14,407,16,82,2,8,4,107,117,5952,15,256,4,2,7,3766,5,723,36,71,43,530,476,26,400,317,46,7,4,2,1029,13,104,88,4,381,15,297,98,32,2071,56,26,141,6,194,7486,18,4,226,22,21,134,476,26,480,5,144,30,5535,18,51,36,28,224,92,25,104,4,226,65,16,38,1334,88,12,16,283,5,16,4472,113,103,32,15,16,5345,19,178,32]

该样本真实评论数据为：

"? this film was just brilliant casting location scenery story direction everyone’s really suited the part they played and you could just imagine being there robert ? is an amazing actor and now the same being director ? father came from the same scottish island as myself so i loved the fact there was a real connection with this film the witty remarks throughout the film were great it was just brilliant so much that i bought the film as soon as it was released for ? and would recommend it to everyone to watch and the fly fishing was amazing really cried at the end it was so sad and you know what they say if you cry at a film it must have been good and this definitely was also ? to the two little boy’s that played the ? of norman and paul they were just brilliant children are often left out of the ? list i think because the stars that play them all grown up are such a big profile for the whole film but these children are amazing and should be praised for what they have done don’t you think the whole story was so lovely because it was true and was someone’s life after all that was shared with us all"

在本数据集的50,000个评论中，25000个作为训练集，另25000个作为测试集，训练集和测试集的positive样本和negative样本数目均衡。加载方式是通过python的pickle模块加载链接的数据：

import pickle  
with open("./data.pkl","rb") as f:  
   train\_data, train\_labels, test\_data, test\_labels = pickle.load(f)

其中train\_data和train\_labels是训练数据的样本以及标签， test\_data和test\_labels是测试集的样本及标签。

**2.4 实验验收**

保留jupyter notebook上运行的结果，现场验收。

**实验3 风机结冰数据分析(待完善)**

**3.1 实验目的**

通过人工智能导论所学内容，应用机器学习以及深度学习的方法完成行业数据分析的建模。

**3.2 实验内容**

**3.2.1 机器学习模型**

利用经典的机器学习方法随机森林、XGBoost等在行业数据分析的应用

**3.2.2 深度学习模型**

利用循环神经网络/LSTM进行建模分析时序数据

**3.3 实验数据**

* 问题描述:叶片结冰是风电领域的一个全球范围难题。目前风机运行的实时数据主要由SCADA系统进行存储，对叶片结冰故障的监测手段主要是比较风机实际功率与理论功率之间的偏差，当偏差达到一定值后会触发风机的报警和停机。然而，触发报警时往往已经发生叶片大面积结冰现象，在这样的情况下运行会增加叶片折断损坏的风险。对结冰过程的预测准确度决定了除冰系统的效率、风机的效率损失和风机运行的风险。原问题链接：[http://www.industrial-bigdata.com/competition/competition/competitionAction!showDetail.action?competition.competitionId=1](https://www.kaggle.com/c/stayalert)
* 数据描述：
* SCADA系统的数据通常有上百个变量，经过筛选保留了其中28个连续数值型变量，涵盖了风机的工况参数、环境参数和状态参数等多个维度。
* 训练集包含两个风机的数据。同一个风机有3个csv数据文件，将其聚合为一张表，为样本打上标签。