机器学习第五章实验内容

详细内容见第四周实验内的 jupyter notebook。

一、实验内容

1.1 使用 sklearn 的多层感知机

使用 sklearn 自带的手写数字数据集：

1. 学习标准化处理的方法
2. 使用 sklearn.neural_network.MLPClassifier 完成手写数字分类任务
3. 绘制学习率为 3，1，0.1，0.01 训练集损失函数的变化曲线

1.2 神经网络：线性回归

1. 学会梯度下降的基本思想
2. 学会使用梯度下降求解线性回归
3. 了解标准化处理的效果

1.3 神经网络：对数几率回归

1. 完成对数几率回归
2. 使用梯度下降求解模型参数
3. 绘制模型损失值的变化曲线
4. 调整学习率和迭代轮数，观察损失值曲线的变化
5. 按照给定的学习率和迭代轮数，初始化新的参数，绘制新模型在训练集和测试集上损失值的变化曲线，完成表格内精度的填写

1.4 神经网络：三层感知机

1. 实现一个三层感知机
2. 对手写数字数据集进行分类
3. 绘制损失值变化曲线
4. 完成 kaggle MNIST 手写数字分类任务，根据给定的超参数训练模型，完成表格的填写

1.5 实现 n 层感知机（选做）

实验内容：

1. 数据集不限
2. 激活函数不限
3. 损失函数不限

要求给出以下内容的总结：
1. 数据集描述
2. 预处理方法及步骤
3. 模型架构：层数，激活函数，损失函数
4. 神经网络超参数：学习率，迭代轮数
5. 训练集和测试集精度
6. 损失值变化曲线
7. 代码注释

1.6 设计一种改良的优化算法（选做）

实验内容： 请你设计一个改进算法，能通过动态调整学习率显著提升收敛速度
1. 数据集不限
2. 激活函数不限
3. 损失函数不限

要求给出以下内容的总结：
1. 数据集描述
2. 预处理方法及步骤
3. 模型架构：层数，激活函数，损失函数
4. 神经网络超参数：学习率，迭代轮数
5. 训练集和测试集精度
6. 损失值变化曲线
7. 代码注释

二、数据介绍

2.1 sklearn 自带的手写数字

使用 from sklearn.datasets import load_digits 加载。
load_digits()['images']包含了原始图像， 8×8 大小的灰度图像，一共 1797 个样本。
load_digits()['data']包含了处理后的图像，将 8×8 的图像展开成 1×64 的向量，一共 1797 个样本。
load_digits()['target']包含了图像对应的标记，一共 1797 个样本标记。

2.2 kaggle 房价预测数据集

文件名：
1. 原始数据：kaggle_hourse_price_train.csv
2. 字段说明：kaggle 房价预测字段说明.txt

数据来源：https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

2.3 kaggle MNIST 手写数字训练集

MNIST 是最有名的手写数字数据集之一，主页：http://yann.lecun.com/exdb/mnist/
MNIST 手写数字数据集有 60000 个样本组成的训练集，10000 个样本组成的测试集，是 NIST 的子集。数字的尺寸都是归一化后的，且都在图像的中央。可以从上方的主页下载。

我们使用的数据集是 kaggle 手写数字识别比赛中的训练集。数据集一共 42000 行，785 列，其中第 1 列是标记，第 2 列到第 785 列是图像从左上角到右下角的像素值。图像大小为 28×28 像素，单通道的灰度图像。

文件名：
1. 原始数据：mnist_train.csv
2. 字段说明：kaggle_mnist_字段说明.txt

数据来源：https://www.kaggle.com/c/digit-recognizer/data

读取方法：
```
import pandas as pd
data = pd.read_csv('data/kaggle_mnist/mnist_train.csv')
X = data.values[:, 1:].astype('float32')
Y = data.values[:, 0]
```

2.4 白葡萄酒质量数据集

文件名：
1. 原始数据：winequality-white.csv
2. 字段说明：葡萄酒字段说明.txt

数据来源：http://archive.ics.uci.edu/ml/datasets/Wine+Quality

数据无需预处理，读取即可使用
```
import pandas as pd
data = pd.read_csv('data/wine_quality/winequality-white.csv', delimiter=";")
```

2.5 spambase 垃圾邮件数据集

文件名：
1. 原始数据：spambase.data
2. 字段说明：spambase 数据说明.txt

数据来源：http://archive.ics.uci.edu/ml/datasets/spambase

数据无需预处理，读取即可使用

```
import numpy as np
data = np.loadtxt('data/spambase/spambase.data', delimiter = ",")
```

2.6 Dota2 Games Results Data Set Dota2 游戏结果数据集

文件名：
1. 原始数据：dota2Train.csv
2. 字段说明：dota2 比赛结果字段说明.txt

数据来源：http://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results

数据无需预处理，读取即可使用

```
import numpy as np
data = np.loadtxt('data/dota2Dataset/dota2Train.csv', delimiter=',')
```