实验 A：基于 MindSpore 的鸢尾花二分类实验

实验介绍：

逻辑回归（Logistic Regression）是机器学习最经典的算法之一，与线性回归有很多不同，这两种回归都属于广义线性回归（Generalized Linear Regression）的范畴。逻辑回归具有如下特点：

1）逻辑回归对自变量分布没有要求；

2）因变量是离散型变量，即分类变量；

3）逻辑回归分析的是因变量取某个值的概率与自变量的关系。

本实验使用 MindSpore 在 2 分类数据集上进行逻辑回归实验，分析自变量和因变量（概率）之间的关系，即求得一个概率函数。

实验目的：

1）掌握逻辑回归的基本概念。

2）掌握机器学习的一般流程。

3）掌握如何使用 MindSpore 进行逻辑回归实验。

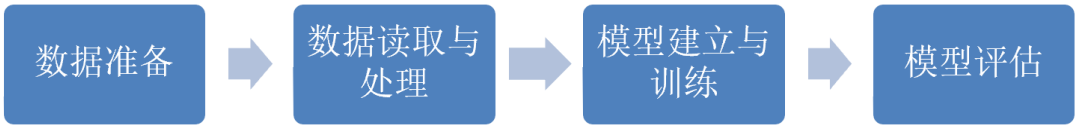
4） 掌握如何使用华为云 ModelArts Notebook 上传数据、执行 Python 代码。

实验环境要求：

华为云 ModelArts

MindSpore 1.2（MindSpore 版本会定期更新）

实验总体设计：



实验内容：

1. 数据准备

1)下载数据

Iris 数据集是模式识别最著名的数据集之一。数据集包含 3 类，每类 50 个实例，其中每个类都涉及一种鸢尾植物。 第一类与后两类可线性分离，后两类之间不能线性分离，所以本实验取前两类数据，做一个 2 分类数据集。

每个样本含有 4 个数值属性和一个类别属性：

sepal length in cm

sepal width in cm

petal length in cm

petal width in cm

class:

- Iris Setosa
- Iris Versicolour
- Iris Virginica

2)上传数据到 ModelArts

2. 数据读取与处理

1）导入 MindSpore 模块和辅助模块如图 1 所示。

```
import os
# os.environ['DEVICE_ID'] = '6'
import csv
import numpy as np
```

```
import mindspore as ms
from mindspore import nn
from mindspore import context
from mindspore import dataset
from mindspore.train.callback import LossMonitor
from mindspore.common.api import ms_function
from mindspore.ops import operations as P

context.set_context(mode=context.GRAPH_MODE, device_target="Ascend")
```

图 1 参考代码

2）读取 Iris 数据集，并查看部分数据

3）抽取样本

取前两类样本（共 100 条），将数据集的 4 个属性作为自变量 X。将数据集的 2 个类别映射为{0, 1}，作为因变量 Y。

4）样本可视化

取样本的前两个属性进行 2 维可视化，如图 2 所示，可以看到在前两个属性上两类样本是线性可分的。

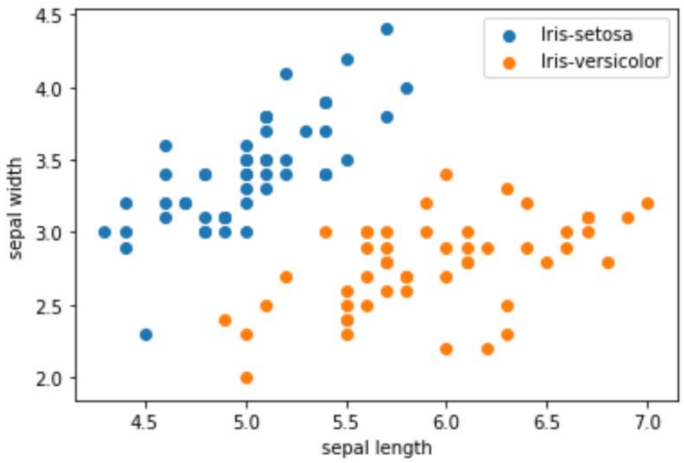


图 2 2 维可视化

5）分割数据集

将数据集按 8:2 划分为训练集和验证集。

6）数据类型转换

使用 MindSpore 的 GeneratorDataset 接口将 numpy.ndarray 类型的数据转换为 Dataset。

3. 模型建立与训练

1）可视化逻辑回归函数

逻辑回归常用的联系函数是 Sigmoid（S 形函数），Sigmoid 函数如下图 3 所示，可以将连续值映射到{0, 1}，同时也是单调可微的。

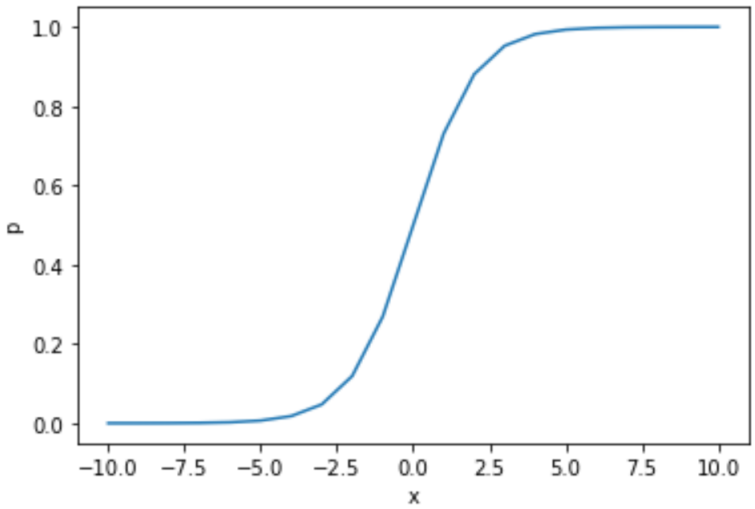


图 3 Sigmoid 函数

2）建模

使用 MindSpore 提供的 nn.Dense(4, 1)算子（https://www.mindspore.cn/api/zh-CN/0.2.0-alpha/api/python/mindspore/mindspore.nn.html#mindspore.nn.Dense）作为线性部分，其中(4, 1)表示每个样本的输入是含 4 个元素的向量，输出是含 1 个元素的向量，即 W 是 1x4 的矩阵。算子会随机初始化权重 W 和偏置 b。使用 SigmoidCrossEntropyWithLogits 算子（https://www.mindspore.cn/api/zh-CN/0.3.0-alpha/api/python/mindspore/mindspore.ops.operations.html?#mindspore.ops.operations.SigmoidCrossEntropyWithLogits)作为非线性部分：

对于每个样本 N_i，模型的计算方式如下：

$$z_i = wx_i + b$$

$$p_i = sigmoid(z_i) = \frac{1}{1 + e^{-z_i}}$$

$$loss = \frac{1}{n}\sum_{i=1}^{n}(y_i * \ln(p_i) + (1 - y_i) * \ln(1 - p_i))$$

其中，x_i是 1D Tensor（含 4 个元素），z_i是 1D Tensor（含 1 个元素），y_i是真实类别（2 个类别{0, 1}中的一个），p_i是 1D Tensor（含 1 个元素，表示属于类别 1 的概率，值域为[0, 1]），loss 是标量。

3）模型训练

使用 2 分类的 Iris 数据集对模型进行几代（Epoch）训练。

4. 模型评估

计算模型在测试集上精度，如果测试集上的精度达到了 1.0 左右，即逻辑回归模型学会了区分 2 类鸢尾花。

需要提交的文件：

1） 源代码.ipynb（里面要有注释,包括版本和功能注释，运行结果）

2） 实验报告（实验名称、作者、实验目的、实验内容（简述）、实验流程图，实验代码、实验结果（截图）、实验心得，请注意格式的美观）

3) 实验运行视频（带个人标识的视频，比如带电脑边框或者电脑背景图等）