**机器学习作业一**

计算机1602 1611640205 黎庭乐

1. **题目**

使用 Python 实现对数几率回归模型

1. **问题描述**

通过Python编程实现对数几率回归模型（Logistics Regression），利用批量梯度下降法训练模型实现对鸢尾花类型的分类

1. **数据集描述**

鸢尾花数据集是原则20世纪30年代的经典数据集。它是用统计进行分类的鼻祖。数据包括三种鸢尾花的四个特征，features分别是：花萼长度（cm）、花萼宽度（cm）、花瓣长度（cm）和花瓣宽度（cm），labels分别是：山鸢尾花(Iris Setosa)、变色鸢尾花(Iris Versicolor)和维吉尼亚鸢尾花(Iris Virginica)，这些features可用来被识别物种，因此我们可以通过这些数据训练模型，来实现对新数据的分类。在本次实验中，为作图和分类方便，我仅选取了前两个features、前两种labels进行训练和预测

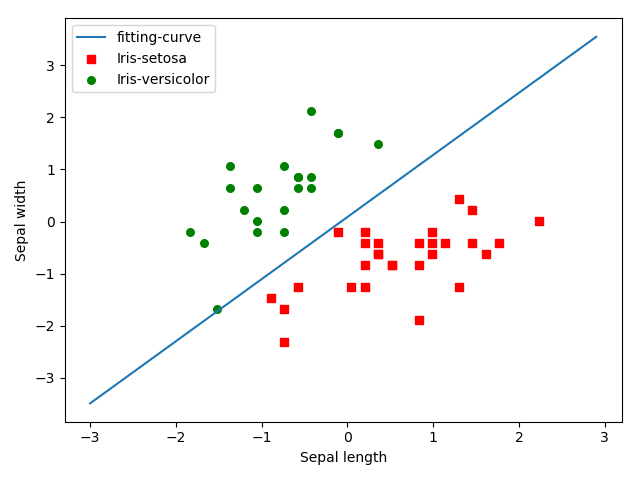
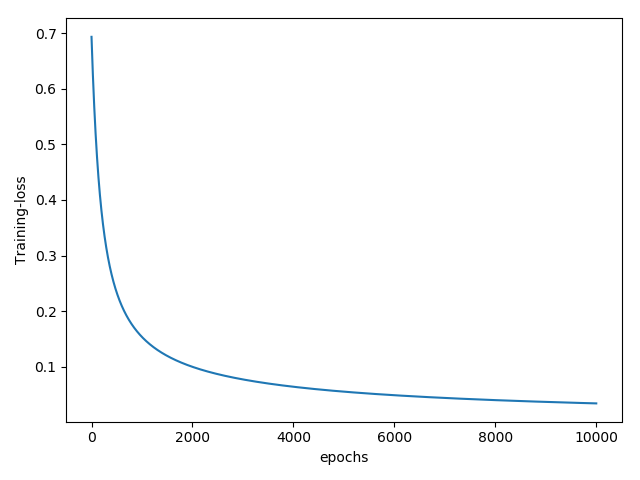
1. **实验结果图**

注：为保证训练模型的数据中label的值在train\_set和test\_set上分布均匀，数据在输入模型之前都会被打乱，因此每次运行后的accuracy会有所差异。另外，Accuracy在程序运行后的控制台上显示，没有在图片中显示，在报告中未截图

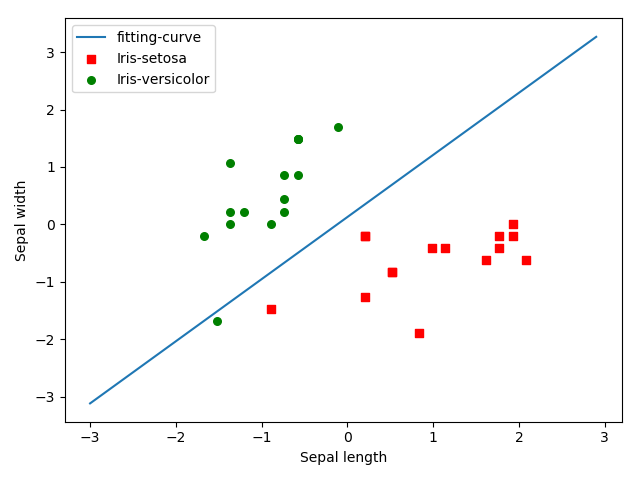
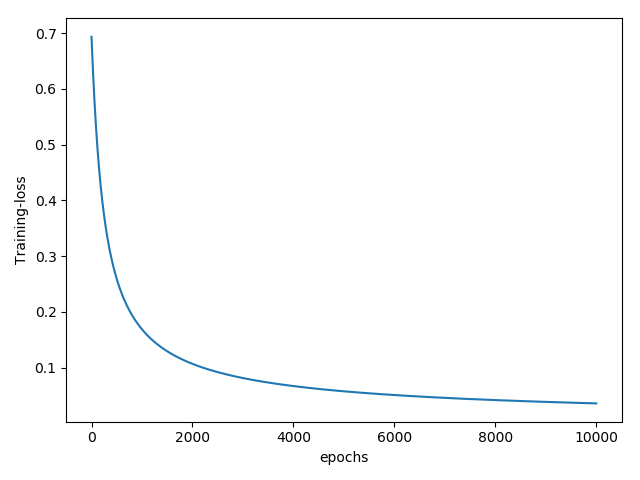
**参数解释：**

Epoch：预分得的批量数、train\_size：训练集占数据集的总比例、α：学习率

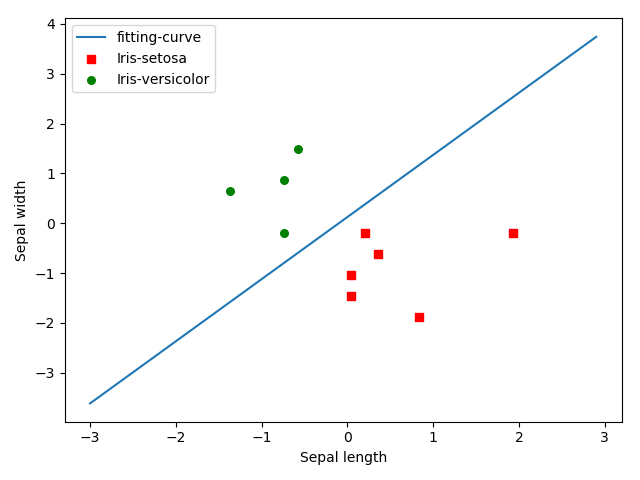
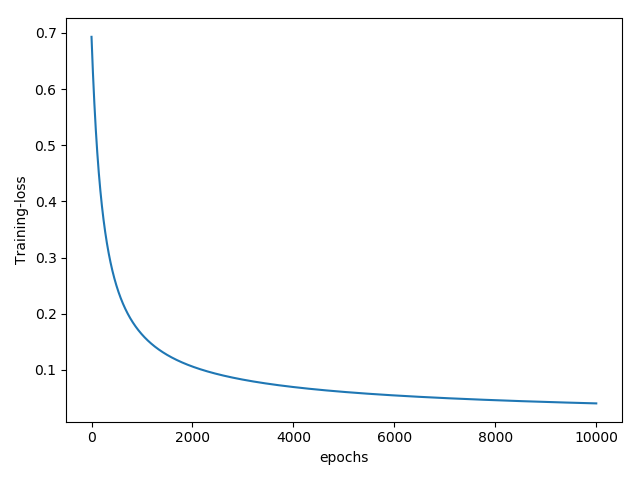
* 1. Epoch=10000，train\_size=0.5，α=0.01：

****

* 1. Epoch=10000，train\_size=0.7，α=0.01：

****

* 1. Epoch=10000，train\_size=0.9，α=0.01：

****

1. **实验结果分析**
   1. 由于数据规模有限，在使用预设超参数下，train\_size对accuracy的影响不大
   2. 由于α=0.01，epoch=10000，因此利用梯度下降法时，cost function收敛的较快，导致train\_size不同对cost function的影响不大
   3. 在将数据划分为train\_set和test\_set之前，为了训练效果更好，我对数据进行了标准化处理，从多次实验的标准化处理前后的图像中可以看出，处理后训练效果优于处理前
   4. 由于本数据集的规模较小，从cost function变化曲线图上可看出，正则化对本数据集的影响不大，因此本次作业没有加入L2正则化方法
   5. 从实验结果可以看出，accuracy稳定于0.95-1之间，说明使用此模型对鸢尾花种类的分类效果较好