

CSCI 360 Project 4: Machine Learning

Released: 4/14/2021

Due: **11:59 PM 4/30/2021**

1 Introduction

This programming assignment will focus on something different from the previous programming assignments, machine learning methods and frameworks. We will setup the programming assignment on Google Colab. We will use the Anaconda distribution through Google Colab, which will save significant work in setting up an Anaconda distribution for Jupyter notebooks. Google Colab will be used for hosting the Jupyter notebooks. We will also use the following frameworks in addition to Python and the Anaconda distribution: numpy, matplotlib, sklearn, and Pytorch. Lastly, we will use the iris dataset, the breast cancer dataset, the CIFAR-10 dataset, and the AG news dataset.


The hope of this programming assignment is to give you a hands-on learning experience with current frameworks that AI researchers and data scientists use to analyze and visualize data, run experiments involving machine learning, and get a closer look at practical applications of machine learning, such as computer vision and natural language processing. This assignment shouldn't be difficult, but fun. The goal should be for you guys to take a look at how to run simple experiments with datasets against some written code that you write based around machine learning algorithms learned in class.

Download Proj4.ipynb and the four sets of data (iris dataset, breast cancer dataset, CIFAR-10 dataset, and AG news dataset). Create a folder (any name to store your files) in Google drive (with any gmail). Put the ipynb and these three sets of data inside that folder, and open Proj4.ipynb with Google Colab.


The Jupyter Notebook structure is designed in a way for AI researchers and data scientists to not only write Python code to solve some tasks, but also explain and analyze the results of the code. The file is structured with different code blocks and different HTML blocks, in which you as the data scientist can separate code blocks with HTML to visualize and analyze the data.

In order to optimize the training process, running the code in Google Colab with the setting of hardware accelerator to GPU will make training significantly faster. Clicking Edit, Notebook Settings allows you to change the hardware accelerator to GPU.

Folders






























































 datasets


Files

 Proj4.ipynb

**CSCI 360 Project 4:
Machine Learning**

**Date Due: 11:59 PM
4/30/2021**

 Proj4.ipynb

Folders

 AG_news

 breast-cancer-dataset

 cifar-10

 iris-dataset


Runtime Tools Help [Last edited on April 13](#)

Project 4: Machine Learning

PM 4/30/2021

ab and the Anaconda distribution. In this project, you will work with the AG News dataset and help teach machine learning algorithms. The project is broken into four parts to be done. The first part focuses on the K-nearest neighbors algorithm using the iris dataset from UCI's ML repository. The second part focuses on the Linear Perceptron algorithm. The third part focuses on ensemble methods using the SKLearn package, run against the breast cancer dataset. The fourth part focuses on Recurrent Neural Networks built from Pytorch: you will build a nn.Module class and run it against the AG News dataset for text classification. This part focuses on the concepts of neural networks and natural language processing.

Notebook settings

Hardware accelerator
GPU 

To get the most out of Colab, avoid using a GPU unless you need one. [Learn more](#)

☐ Omit code cell output when saving this notebook

CANCEL SAVE

To summarize the tasks assigned for this homework, you will need to write code without sklearn to build a k-Nearest Neighbors class to work with the iris dataset. You need to also build a set with sklearn, tested against the iris dataset as well. You will then use Pytorch to create a simple Convolutional Neural Network (CNN) for a simple image classification task. Lastly, you will use Pytorch to create a simple Recurrent Neural Network (RNN) for a simple text classification task.

A key part to working with the ipynb framework is that the majority of the information can be stored within the ipynb.

2 Question 1 K-Nearest Neighbors - 5 Points

Build a k-nearest neighbors classifier to run against the iris dataset. See the ipynb for pseudocode and guidelines on how to create the predict function in particular.

Each assert statement is worth a point.

3 Question 2 Ensemble Learning - 6 points

Build 3 different ensembles using sklearn's packages, ran against the breast cancer dataset. The limit here is that only sklearn and numpy packages can be used- no other package is allowed (e.g. don't use Pytorch or Tensorflow here).

Each assert statement is worth 2 points. There are 3 assert statements, each requiring each of the 3 ensembles to be above 62 percent accuracy.

4 Question 3 Convolutional Neural Network - 5 points

Build a convolutional neural network using nn.Sequential from Pytorch; run it against CIFAR-10.

If you score above 65 percent accuracy, you score all 5 points.

5 Question 4 Recurrent Neural Network - 5 points

Build a recurrent neural network using nn.Module from Pytorch; run it against the AG news dataset.

If you score above 88 percent accuracy, you score all 5 points.

6 Submission

Turn in just Proj4.ipynb- MAKE SURE TO RUN ALL CODE BLOCKS AND SHOW RESULTS. This project is completely autograded using nbgrader. If any issues arise in submission/grading, please contact Michael Yuen (I will monitor and grade the assignments).

7 Final thoughts

Sorry for the lateness of assigning the project. This project is a rather quick project, with minimal code necessary, but some reading and thinking required. This project is built mostly from scratch with some references below. This project is experimental, which means there might be some rough edges (especially in Question 4 in my opinion), therefore the grading of the project is lenient. I expect approximate time spent to complete the project is 1 day despite having 2 1/2 weeks to do the project. I hope everyone enjoys the project despite the hiccups in developing the project.

References:

Sklearn: <https://scikit-learn.org/stable/index.html>

Pytorch: <https://pytorch.org/docs/stable/index.html>

Iris dataset: <https://archive.ics.uci.edu/ml/datasets/iris>

Breast cancer dataset: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))

Cifar-10 dataset: <https://www.cs.toronto.edu/~kriz/cifar.html>

AG news dataset: <https://www.kaggle.com/amananandrai/ag-news-classification-dataset>