

CMPSCI 389 Homework 3

Spring 2022

Assigned: March 3, 2022; Due: March 10, 2022 @ 11:59pm Eastern

1 Introduction

In this assignment, you will train a machine learning model to classify breast tumors as benign or malignant, based on features computed from images of the tumors such as their radii, texture, and area. A fuller description of the dataset can be found at [https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)).

The main goal of this assignment is to give you hands-on experience using several popular Python libraries and tools to solve a machine learning problem like you might in the real world. These libraries and tools are

- Jupyter, a interactive “notebook” environment for editing and running Python code,
- scikit-learn, a library of implementations of popular machine learning models, data pre-processing algorithms, and more,
- NumPy, a library for fast numerical computing, and
- pandas, a library that makes loading, cleaning, and transforming data much easier.

2 Collaboration

This assignment should be completed independently. Though you may discuss the assignment at a high-level with your peers, you should write your code on your own.

3 Instructions

1. **Verify that pip is installed.** pip is a package manager for Python libraries and programs. It *should* have been installed when you installed Python, but sometimes that doesn’t happen. Open a terminal and run the command `pip --version`. You should see an output like “`pip 21.1.3 from /usr/share/lib/python3.9/site-packages/pip (python 3.9)`”. If you get an error telling you that the pip command was not found, try `pip3` instead. If *that* doesn’t work, you will need to figure out why pip is not installed and fix your system.
2. **Install the requirements.** cd to this project’s directory (the one containing `requirements.txt`) and run `pip install -r requirements.txt`. You should see some output indicating that pip is downloading a bunch of files from the internet, with a final line that says something like “Successfully installed ...”. If you see errors, fix them before proceeding. If the errors that say things like “permission denied” or similar, try running pip as an administrator (run the Command Prompt as an administrator on Windows, use sudo on Mac or Linux).
3. **Run Jupyter.** From a terminal, run the command `jupyter notebook`. Your web browser should open and show you the Jupyter interface. If it does not, open your browser yourself and visit `http://localhost:8888`. Use the file browser to navigate to `HWN.ipynb`, and open it.
4. **Do the assignment.** Follow the instructions in the notebook. Many of the prompts link you to external documentation, which you should visit.
5. **Submit to Moodle.** Please submit your modified `.ipynb` file. Do not export to some other format like PDF or extract a Python script from the notebook.

4 Checklist

It can be easy to miss questions tucked away in Jupyter notebook cells, so here is a checklist for your final submission. Have you ...

1. loaded `data.csv` into a pandas DataFrame and removed the 'id' column?
2. printed all of the required summary statistics for the columns of this DataFrame?
3. produced a 5x6 grid of histograms of the values for each column?
4. correctly created the `X_unprocessed` and `y_unprocessed` NumPy arrays and printed their shapes?
5. created the `X` and `y` arrays, with `y` being 0's and 1's matching `y_unprocessed`?
6. produced a correct train/test split?
7. explained F1 scores, and computed one for the results produced by the KNN code?
8. created a model that works better than KNN on unprocessed data and computed its F1 score?
9. explained why you chose the algorithm you did, your preprocessing steps, and how you tuned your hyperparameters?