Introduction

In this assignment, you will use breast cancer data, which contains features computed from digitized images of fine needle aspirates (FNA) of breast masses. The dataset includes features such as radius, texture, perimeter, area, smoothness, and more. Your task is to build a classification model that can predict whether a tumor is malignant (cancerous) or benign (non-cancerous) for unseen data.

Dataset

The dataset contains the 30 features, derived from 10 values measured from digitized images:

- `radius`: Mean of distances from the center to points on the perimeter.
- `texture`: Standard deviation of gray-scale values.
- `perimeter`: Perimeter of the tumor.
- `area`: Area of the tumor.
- `smoothness`: Local variation in radius lengths.
- `compactness`: Perimeter^2 / area - 1.0.
- `concavity`: Severity of concave portions of the contour.
- `concave points`: Number of concave portions of the contour
- `symmetry`: Symmetry of the tumor.
- `fractal dimension`: "Coastline approximation" - 1.

The target variable is binary:

- `M`: Malignant (cancerous).
- `B`: Benign (non-cancerous).

More information can be found at `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic`

Goal

You will train a neural network model to classify tumors as malignant or benign. You will present results from both:

- A neural network built with `pyTorch`).
- Your hand-built neural network implementation using `jax`.

Deadlines

- **Deliverable 1: In-class presentation**: Slide presentation on data exploration results and planned approach.
- **Deliverable 2: Final submission**: Final paper and code on GitHub.

The Write-Up

Here's the overarching writing rule for this course: *you need to be sufficiently precise with your writing and include enough detail that a competent reader could reproduce your results.* You will submit a PDF. A paper from AAAI, a notable artificial intelligence conference, is provided for guidance. You may use any word processing platform you prefer for your report. AAAI provides a LaTeXtemplate for those interested. Your writeup must also include an ethics section in line with the NeurIPS (another notable AI conference) guidelines: `https://neurips.cc/public/EthicsGuidelines`, and a model card: `https://modelcards.withgoogle.com/about`.

Here are some additional things to address in your report, in no particular order. This is *not* meant to be an exhaustive list.

- What preprocessing did you perform on the data? Did you perform any exploratory data analysis? Generate any plots or charts? Describe these, along with any relevant findings, in your report.
- What classification models did you build? How do they compare in terms of performance? What was the best-performing model, and how did it do? Optionally, you can go beyond the methods we've seen in class and try other algorithms — if you go this route, you should describe how your chosen algorithm works (and don't forget to include citations!).

- What was your model-building and tuning regime? How did you address overfitting? How did you make hyperparameter choices?

Timetable

Here's how to budget your time over the next couple of weeks as you work on this project:

- **Week 1:** Explore the dataset, think about data scaling,, make model decisions. Outline potential biases and inequities that can result from this project. Make your data exploration presentation.
- **Week 1:** Build your first model, begin running thorough experiments (hyperparameter tuning, further feature engineering, etc.), analyze your results and iterate, search the literature for related work on the problem, write your *Introduction* and *Background* sections.
- **Week 2:** Complete experiments, take a step back and think about your report's narrative, write drafts of your *Experiments* and *Results* section.
- **Week 2:** Wrap-up any pending experiments, write the *Conclusions* section and the abstract, revise and proof-read the entire report and submit it for review.

Rubric

- Data exploration (presentation): /15
- Model choice/exploration: /10
- Code: /10
- Report:
  - Objective description: /5
  - Method description: /10
  - Analysis: /10
  - Results: /10
  - Ethics: /5
  - Writing: Clarity: /5