# Logistic Regression Questions

August 19, 2025

## Exercise 1 — Logistic Regression with One Variable

**Goal:** Implement binary logistic regression using gradient descent from scratch (no built-in logistic regression routines).

You are provided a dataset of exam scores and admission outcomes.

### Dataset

```
exam_score,admitted
35,0
42,0
50,0
60,0
67,1
75,1
80,1
90,1
95,1
100,1
110,1
120,1
130,1
140,1
150,1
160,1
175,1
190,1
210,1
230,1
```

### Tasks

1. **Data Exploration (EDA):**

- Load the CSV.
- Print the first 5 rows.
- Print dataset shape $(N, d)$.
- Print summary statistics: min, max, mean, std of `exam_score`.

2. **Model Definition:** Logistic hypothesis:

$$\hat{y} = \sigma(z), \quad z = \theta_0 + \theta_1 x, \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

Loss function:

$$J(\theta_0, \theta_1) = -\frac{1}{N} \sum_{i=1}^{N} \left( y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right)$$

3. **Gradient Descent:** Full-batch updates:

$$\theta_0 \leftarrow \theta_0 - \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} (\hat{y}^{(i)} - y^{(i)}), \quad \theta_1 \leftarrow \theta_1 - \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} (\hat{y}^{(i)} - y^{(i)}) x^{(i)}$$

Use: learning rate $\alpha = 0.01$, epochs $= 1000$.

4. **Training:** Print final values of $\theta_0$, $\theta_1$, and final loss.

5. **Evaluation & Prediction:**

- Compute training accuracy with threshold 0.5.
- Predict admission probability for exam score = 65.
- Predict admission probability for exam score = 155.

# Exercise 2 — Logistic Regression with Multiple Variables

**Goal:** Implement multivariate logistic regression using vectorized gradient descent from scratch (no built-in logistic regression routines).

You are provided a dataset of exam results and study hours.

## Dataset

```
exam1,exam2,hours_study,admitted
35,40,5,0
42,50,6,0
50,52,7,0
60,65,8,0
67,70,9,1
75,78,10,1
80,85,12,1
90,88,14,1
95,90,15,1
100,92,16,1
110,100,17,1
120,105,18,1
130,110,19,1
140,115,20,1
150,118,22,1
160,120,24,1
175,125,25,1
190,128,26,1
210,130,28,1
230,135,30,1
```

## Tasks

1. **Data Exploration (EDA):**
   - Load the CSV.
   - Print the first 5 rows.
   - Print dataset shape $(N, d)$.
   - Print column-wise summary statistics: min, max, mean, std.
   - Standardize all features (`exam1`, `exam2`, `hours_study`) to zero mean and unit variance.
2. **Model Definition:**

$$\hat{y} = \sigma(X'\theta), \quad \sigma(z) = \frac{1}{1 + e^{-z}}$$

Loss:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right)$$

3. **Gradient Descent (Vectorized):**

$$\theta \leftarrow \theta - \alpha \cdot \frac{1}{N}X'^{\top}(\hat{y} - y)$$

   Use: learning rate $\alpha = 0.01$, epochs $= 1500$, initialize $\theta = 0$.

4. **Training:** Print final parameter vector $\theta$ and final loss.

5. **Evaluation & Prediction:**

   - Compute training accuracy using threshold 0.5.
   - Predict admission probability for:
     (a) (exam1=72, exam2=80, hours_study=11)
     (b) (exam1=150, exam2=118, hours_study=20)

     Apply the same standardization as training data.