AdaBoost on Categorical Data

Dataset (imbalanced: more +1 than −1)

ID	Color	Shape	Label Y
1	Red	Circle	+1
2	Red	Square	+1
3	Blue	Circle	+1
4	Blue	Square	+1
5	Red	Circle	+1
6	Blue	Square	−1
7	Blue	Circle	−1

We have **5 positives (+1)** and **2 negatives (−1)**.

Step 1. Initialization

Number of samples = 7.

Initial weights (equal):

$$w_j^{(0)} = \frac{1}{7} \approx 0.142857 \quad \text{for each } j.$$

Step 2. Train a weak learner M_1

Choose stump (rule):

👉 **If Color = Red → predict +1, else predict −1.**

Predictions:

ID	Color	True Y	Predicted Y	Correct?
1	Red	+1	+1	✅
2	Red	+1	+1	✅
3	Blue	+1	−1	❌
4	Blue	+1	−1	❌
5	Red	+1	+1	✅
6	Blue	−1	−1	✅
7	Blue	−1	−1	✅

Misclassified: IDs **3, 4** (positives predicted as −1).

Step 3. Compute weighted error

$$\text{error}(M_1) = w_3^{(0)} + w_4^{(0)} = 0.142857 + 0.142857 = 0.285714.$$

Step 4. Classifier weight (vote strength)

$$\alpha_1 = \ln \frac{1 - \text{error}}{\text{error}} = \ln \frac{1 - 0.285714}{0.285714} = \ln \frac{0.714286}{0.285714} = \ln(2.5) \approx 0.9163.$$

So M_1 will have weight ≈ 0.916 in the ensemble.

Step 5. Update sample weights

Ratio for correct classification:

$$r = \frac{\text{error}}{1 - \text{error}} = \frac{0.285714}{0.714286} = 0.4.$$

- Misclassified samples (3,4): keep same weight = 0.142857.
- Correctly classified samples (1,2,5,6,7): multiply weight by 0.4 = 0.057143.

Unnormalized weights:

ID	Weight
1	0.057143
2	0.057143
3	0.142857
4	0.142857
5	0.057143
6	0.057143
7	0.057143

Sum = 0.571429.

Normalize (divide each by 0.571429):

ID	Normalized weight
1	0.10
2	0.10
3	0.25
4	0.25
5	0.10
6	0.10
7	0.10

Final Ensemble after 1 Iteration

So far we have only one weak learner:

- M_1: Rule = *If Color = Red → +1 else −1*
- Weight = **0.916**

Ensemble prediction is just M_1 (since k=1).
Future rounds would focus on **examples 3 and 4**, which now have the largest weights.

---

n just one iteration: AdaBoost shifted more importance (0.25 each) onto the **difficult positives (3 and 4)** so the next learner will try harder to classify them correctly.
