Q11 — Decision Trees (18 points)

Part 1 (3 points) — Entropy for Regression?

Answer: False.

House price y is a **continuous real number** — this is a **regression** task. Entropy and information gain are designed for **classification** (discrete labels). For regression, we use **variance reduction** (purity) as the splitting criterion:

$$\text{Gain} = |D| \cdot \text{Var}(D) - (|D_L| \cdot \text{Var}(D_L) + |D_R| \cdot \text{Var}(D_R))$$

Part 2 (3 points) — Best Training Accuracy (Self-Consistent Data)

Answer: 100% training accuracy.

With no constraints on the tree, we can split until every leaf contains a single example (or examples with identical x and hence identical y, by the self-consistency assumption). This perfectly memorizes the training set.

Problematic implication: This leads to **overfitting** — the tree captures noise in the training data and generalizes poorly to unseen examples.

Part 3 (9 points) — Information Gain of Three-Way Split

Dataset

$\mathbf{x^{(1)}}$	**y**
1.3	Yes
1.7	No
2.1	No
2.7	Yes
3.0	Yes
4.0	No

Split: a = 2.0, b = 2.5. Three branches:

Branch	Condition	Examples	Yes	No
Left	$x^{(1)} < 2.0$	{1.3, 1.7}	1	1
Middle	$2.0 \leqslant x^{(1)} < 2.5$	{2.1}	0	1
Right	$x^{(1)} \geqslant 2.5$	{2.7, 3.0, 4.0}	2	1

Step 1 — Root entropy

3 Yes, 3 No out of 6 total:

$$H(Y) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

Step 2 — Branch entropies

Left (1 Yes, 1 No): $H_L = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

Middle (0 Yes, 1 No): $H_M = 0$ (pure node)

Right (2 Yes, 1 No): $H_R = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} \approx 0.918$

Step 3 — Conditional entropy

$$H(Y|\text{split}) = \frac{2}{6}(1) + \frac{1}{6}(0) + \frac{3}{6}(0.918) = 0.333 + 0 + 0.459 = 0.792$$

Step 4 — Information gain

$$\boxed{IG = H(Y) - H(Y|\text{split}) = 1 - 0.792 \approx 0.21}$$

Part 4 (3 points) — Number of {a, b} Pairs

Sorted $x^{(1)}$ values: 1.3, 1.7, 2.1, 2.7, 3.0, 4.0

There are **5 intervals** between consecutive values where valid split points can be placed (ensuring each branch has $\geqslant 1$ example):

Interval	Between
1	$(1.3, 1.7)$
2	$(1.7, 2.1)$
3	$(2.1, 2.7)$
4	$(2.7, 3.0)$
5	$(3.0, 4.0)$

We choose 2 intervals out of 5 (the smaller becomes a, the larger becomes b):

$$\boxed{\binom{5}{2} = 10 \text{ pairs}}$$

This guarantees: at least one example in $x^{(1)} < a$, at least one in $a \leqslant x^{(1)} < b$, and at least one in $x^{(1)} \geqslant b$.