ROB311 Quiz 1

Hanhee Lee

April 13, 2025

Contents

Learning Problems

1 Intro

Definition: Assume that there is some (unknown) relationship,

$$f : \mathcal{X} \to \mathcal{Y} \text{ s.t. } x \mapsto_f y$$

- $\mathcal{X}$: Input Space
- $\mathcal{Y}$: Output Space (i.e. information we desire about input)

Find $h : \mathcal{X} \to \mathcal{Y}$ (hypothesis) s.t. $h \approx f$, given some data about f:

$$\mathcal{D} = \left\{ \left(x^{(i)}, y_i \right), x^{(i)} \in \mathcal{X}, y_i = f\left(x^{(i)} \right) \in \mathcal{Y}, i = 1 \dots N \right\}$$

- $\text{in}(\mathcal{D}) = \{x \text{ s.t. } (x, y) \in \mathcal{D}\}$
- $\text{out}(\mathcal{D}) = \{y \text{ s.t. } (x, y) \in \mathcal{D}\}$

1.1 Classification vs. Regression Problems

Definition:
- **Classification Problems:** $\mathcal{X} \subseteq \mathbb{R}^M$ and $\mathcal{Y} \subseteq \mathbb{N}$
- **Regression Problems:** $\mathcal{X} \subseteq \mathbb{R}^M$ and $\mathcal{Y} \subseteq \mathbb{R}$

1.2 Feature Spaces

Definition: Easier to learn relationships from high-level features (instead of the raw input). Need mapping b/w input space and feature space:

$$\phi : \mathcal{X} \to \mathcal{F}$$

2 PAC Learning

2.1 Probably Approximately Correct (PAC) Estimations

Motivation: More than one fcn may be consistent w/ the data, how to find the best one?

2.1.1 Hoeffding's Inequality

Motivation: Bound $|\mu - \nu|$ w.r.t. N.

Definition: For any $\epsilon > 0$,

$$\mathbb{P}(|\nu - \mu| \geq \epsilon) \leq 2e^{-2\epsilon^2 N} \quad (1)$$

- μ: Probabillity of an event.
- ν: Relative frequency in a sample size N.
- ϵ: Tolerance (i.e. how close we want ν to be to μ).
  - $\epsilon \to 0$: $\nu = \mu$
- $\mu \overset{?}{\approx} \nu$: μ is probably approximately equal to ν. As $N \to \infty$: $\nu \to \mu$

Warning: Approx. the true dist. w/ high prob. by taking a large enough N (i.e. empirical dist. converges to true dist.).
- i.e. Probability of a sig. deviation shrinks exp. w/ N.

Notes: A smaller value for ϵ results in a tighter and less certain bound. If we make ϵ half as small, we need to make N 4 times larger to acheive the same bound.

2.2 PAC Learning

2.2.1 Error

Definition:
- **Out-Sample Error:**
$$E_{\text{out}} = \mathbb{P}[f \neq h]$$
- **In-Sample Error:**
$$E_{\text{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[f(x^{(i)}) \neq h(x^{(i)})]$$

2.2.2 Union Bound Theorem

Theorem: The prob. of at least one of the events $E_1, \ldots, E_M$ occurring is bounded by the sum of the prob. of each event occurring:

$$\mathbb{P}\left[E_1 \vee \cdots \vee E_M\right] \leq \sum_{i=1}^{M} \mathbb{P}[E_i]$$

Notes:
- If the events are mutually exclusive, then the union bound is tight (i.e. equality holds).
- If the events are highly correlated, then the union bound is loose (i.e. inequality holds)
  - Some events may be more likely to occur together.

2.2.3 Generalization of Hoeffding's Inequality

Definition: Assuming that h is chosen from a set of hypotheses $\mathcal{H}$, derive a (loose) upper-bound on $|E_{\text{out}} - E_{\text{in}}|$:

$$\begin{aligned}
\mathbb{P}\left[\bigvee_{h\in\mathcal{H}} (|E_{\text{out}} - E_{\text{in}}(h)| > \varepsilon)\right] &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left[|E_{\text{out}} - E_{\text{in}}(h)| > \varepsilon\right] \\
&\leq \sum_{h\in\mathcal{H}} 2e^{-2\varepsilon^2 N} \\
&= 2|\mathcal{H}|e^{-2\varepsilon^2 N}
\end{aligned}$$

- Endow $\mathcal{F}$ (i.e. fcn space) w/ prob. distribution, $P : \mathcal{X} \to [0, 1]$, then
  - E_{out} (i.e. true error of a hyp. over entire dist. of data) is analogous to μ
  - $E_{\text{in}}(h)$ (i.e. empirical error of hyp. on a finite sample) is analogous to ν.

Notes:
- $E_{\text{in}}(h) \overset{?}{\approx} E_{\text{out}}$ requires small $|\mathcal{H}|$ (generalization)
  - Look at inequality, small $|\mathcal{H}| \to$ small $E_{\text{out}} - E_{\text{in}}$ (i.e. prevents overfitting but leads to underfitting)
- $E_{\text{in}}(h) \approx 0$ requires large $|\mathcal{H}|$ (discrimination)
  - Need large $|\mathcal{H}|$ to capture the true dist. (i.e. prevents underfitting but leads to overfitting)

Example:

1. **Given:** An opaque box containing red and blue balls. Take N IID samples.
   - μ: Probability of drawing a blue balls (unknown).
   - ν: Relative frequency of blue balls in the sample (known).
2. **Problem 1:** What is ν in this case? 8 balls total, 5 are blue.
3. **Solution 1:** $\nu = \frac{5}{8}$
4. **Problem 2:** How to partition $\mathcal{F}$ into regions where $f = h$ and $f \neq h$?
5. **Solution 2:**

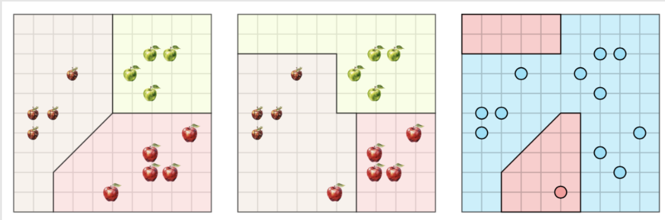

Figure 1: LS h, MS f

6. **Problem 3:** What is the out-sample error?
7. **Solution 3:** In words, the probability of the hypothesis being wrong.

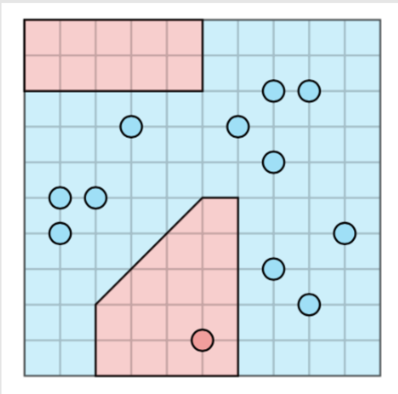

Figure 2

8. **Problem 4:** What is the in-sample error given this sample of 11 balls s.t. $f = h$, 1 ball s.t. $f \neq h$?
9. **Solution 4:** $E_{\text{in}} = \frac{1}{12}$

3 Decision Trees

3.1 Structure

Definition: Each vertex in a decision tree is either:
1. A **condition vertex**: a vertex that sorts points based on a question.
2. A **decision vertex**: a vertex that assigns all points a specific class.

Notes: We want to find the minimum # of condition vertices (or questions) needed to "sufficiently discriminate" (identify the class of every point in $\mathcal{D}$).
- More condition vertices improve discrimination.
- Less condition vertices improve generalization.

3.2 Building a Decision Tree

Definition: Consider determining the class of a randomly chosen target point.
- If we ask a K-ary question abt. the pts. in $\mathcal{D}$, we can form K subsets, $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$, using the answers s.t.
  - $|\mathcal{D}^{(k)}| \in \{0, \ldots, |\mathcal{D}|\}$
  - $|\mathcal{D}| = \sum_{k=1}^{K} |\mathcal{D}^{(k)}|$

3.3 Special Case

Notes: Suppose each pt. belongs to a unique class (i.e. the # of classes is $|\mathcal{D}|$).

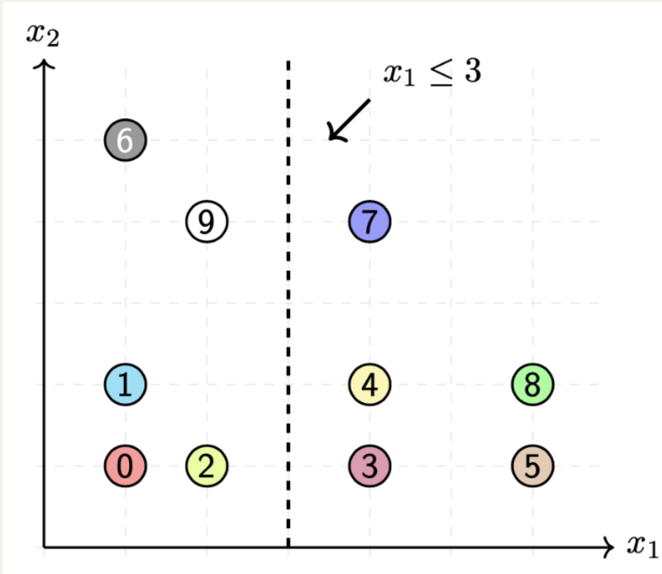

Figure 3

1. Before asking the question: $|\mathcal{D}|$ possible guesses for the target point's class.
2. After asking the question: Either
   - $|\mathcal{D}^{(1)}|, \dots, |\mathcal{D}^{(K-1)}|$ or
   - $|\mathcal{D}^{(K)}|$

   guesses, depending on the answer for the target point.
   - i.e. $|\mathcal{D}^{(K)}|$ if the target point belongs to class K (Yes)
   - i.e. $|\mathcal{D}^{(1)}|, \dots, |\mathcal{D}^{(K-1)}|$ if the target point belongs to class $1, \dots, K-1$ (No)
3. Goal: Minimize the # of guesses needed in the worst-case, which would be

$$\max\{|\mathcal{D}^{(1)}|, \dots, |\mathcal{D}^{(K)}|\}.$$

   - i.e. Target point falls into the largest subset after a question is asked.
4. Given the constraints on $|\mathcal{D}^{(1)}|, \dots, |\mathcal{D}^{(K)}|$, we can show that $\max\{|\mathcal{D}^{(1)}|, \dots, |\mathcal{D}^{(K)}|\}$ is minimized when

$$|\mathcal{D}^{(K)}| \in \left\{ \left\lfloor \frac{|\mathcal{D}|}{K} \right\rfloor, \left\lceil \frac{|\mathcal{D}|}{K} \right\rceil \right\}.$$

   Basically, the best question splits $\mathcal{D}$ into K sets of (roughly) the same size.

Warning: Roughly due to floor/ceil.

3.3.1 # of K-ary Questions Needed

Theorem: Given a classification data-set, $\mathcal{D}$, in which the class of each point is unique (i.e., $|\text{out}(\mathcal{D})| = |\mathcal{D}|$), the class of a randomly chosen target point can be determined within

$$\lceil \log_K(|\mathcal{D}|) \rceil$$

K-ary questions.

3.4 General Case

Motivation: Suppose points do not necessarily belong to a unique class.
- X is the class of a randomly chosen target point.
- Y is the answer to a K-ary question for X.

3.4.1 Expected # of Questions

Definition: Using the theorem above, since for each class, c, we can partition $\mathcal{D}$ into $\lceil 1/p_c \rceil$ subsets, with a subset containing all class c points
- p_c: Proportion of class c points.

If the target point's class is c, we can confirm it w/in $\lceil \log_K(\lceil 1/p_c \rceil) \rceil$ K-ary questions.

Thus, the expected # of Qs needed is

$$\sum_c p_c \lceil \log_2(\lceil 1/p_c \rceil) \rceil.$$

Notes: i.e. Reduces to special cases with each subset containing a unique class.

3.4.2 Entropy, Conditional Entropy, and Information Gain

Definition: The **entropy** of a random variable X (in K-its) is defined as

$$H(X) = -\sum_{\forall x \in X} p_X(x) \log_K(p_X(x)).$$

The **conditional entropy** of a random variable, X, given a random variable Y, is

$$H(X|Y) = -\sum_{\forall y \in Y} \sum_{\forall x \in X} p_{X|Y}(x|y) \log_K(p_{X|Y}(x|y)).$$

The **information gain** from Y is:

$$IG(X|Y) = H(X) - H(X|Y).$$

- Maximize $IG(X|Y)$ (i.e. choose the question to maximize the information gained).

Process:

1. Calculate $H(X)$ (i.e. entropy before the split).
2. Calculate $H(X|Y)$ (i.e. entropy after the split).
   (a) Calculate entropy for each subset of X based on the question, Y.
   (b) Calculate the weighted average of the entropies.
3. Calculate $IG(X|Y) = H(X) - H(X|Y)$.

Example: Consider a classification problem where $\mathcal{X} = \{0, \dots, 9\}^2$, $\mathcal{Y} = \{0, 1, 2\}$ and suppose we are given

$$\mathcal{D} = \left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, 0 \right), \left(\begin{bmatrix} 4 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 4 \\ 2 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 6 \\ 1 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 1 \\ 5 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 6 \\ 2 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 2 \\ 4 \end{bmatrix}, 2 \right) \right\}.$$

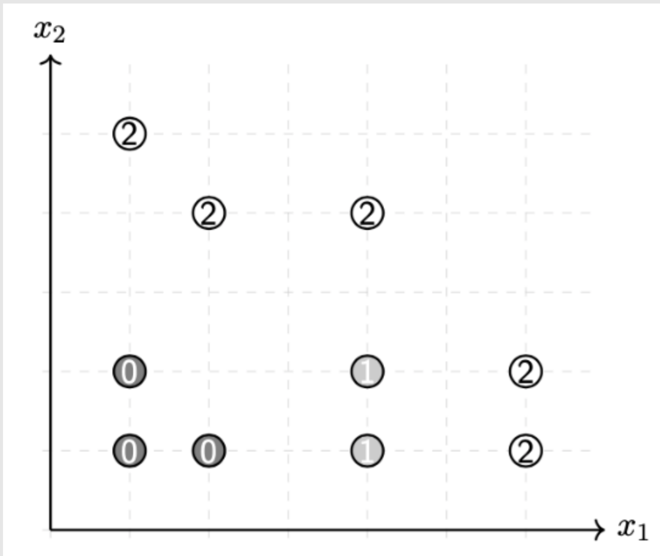

Figure 4

Example: **2-Ary Question**

1. **Given:** $X = \{0, 1, 2\}$, $Y = \begin{cases} 1, & \text{if } x_1 \leq 3 \quad \text{(Yes)} \\ 0, & \text{if } x_1 > 3 \quad \text{(No)} \end{cases}$,
2. **Problem:** $IG(X|Y) = ?$
3. **Solution:**
   (a) Entropy before the split: $H(X) = \frac{3}{10} \log_2 \left(\frac{10}{3}\right) + \frac{2}{10} \log_2 \left(\frac{10}{2}\right) + \frac{5}{10} \log_2 \left(\frac{10}{5}\right)$
   (b) Entropy after the split:
      i. $H(X \mid x_1 \leq 3) = \frac{3}{5} \log_2 \left(\frac{5}{3}\right) + \frac{2}{5} \log_2 \left(\frac{5}{2}\right)$
      ii. $H(X \mid x_1 > 3) = \frac{2}{5} \log_2 \left(\frac{5}{2}\right) + \frac{3}{5} \log_2 \left(\frac{5}{3}\right)$.
      iii. Weighted Avg. Entropy: $H(X|Y) = \frac{5}{10} H(X \mid x_1 \leq 3) + \frac{5}{10} H(X \mid x_1 > 3)$
   (c) $IG(X|Y) = H(X) - H(X|Y)$

Example: **2-Ary Question**

1. **Given:** $X = \{0, 1, 2\}$, $Y = \begin{cases} 1, & \text{if } x_2 \leq 3 \quad \text{(Yes)} \\ 0, & \text{if } x_2 > 3 \quad \text{(No)} \end{cases}$,
2. **Problem:** $IG(X|Y) = ?$
3. **Solution:**
   (a) Entropy before the split: $H(X) = \frac{3}{10} \log_2 \left(\frac{10}{3}\right) + \frac{2}{10} \log_2 \left(\frac{10}{2}\right) + \frac{5}{10} \log_2 \left(\frac{10}{5}\right)$
   (b) Entropy after the split:
      i. $H(X \mid x_2 > 3) = \frac{3}{3} \log_2 \left(\frac{3}{3}\right)$
      ii. $H(X \mid x_2 \leq 3) = \frac{3}{5} \log_2 \left(\frac{5}{3}\right) + \frac{2}{5} \log_2 \left(\frac{5}{2}\right) + \frac{2}{5} \log_2 \left(\frac{5}{2}\right)$.
      iii. Weighted Avg. Entropy: $H(X|Y) = \frac{3}{10} H(X \mid x_2 > 3) + \frac{7}{10} H(X \mid x_2 \leq 3)$
   (c) $IG(X|Y) = H(X) - H(X|Y)$

Example: **3-Ary Question**

1. **Given:** $X = \{0, 1, 2\}$, $Y = \begin{cases} 1, & \text{if } x_1 \leq 3 \text{ and } x_2 \leq 3 \\ 2, & \text{if } x_1 \leq 3 \text{ and } x_2 > 3 \\ 3, & \text{if } x_1 > 3 \end{cases}$
2. **Problem:** $IG(X|Y) = ?$
3. **Solution:**
   (a) Entropy before the split: $H(X) = \frac{3}{10} \log_2 \left(\frac{10}{3}\right) + \frac{2}{10} \log_2 \left(\frac{10}{2}\right) + \frac{5}{10} \log_2 \left(\frac{10}{5}\right)$
   (b) Entropy after the split:
      i. $H(X \mid x_1 \leq 3 \text{ and } x_2 \leq 3) = \frac{3}{3} \log_2 \left(\frac{3}{3}\right)$
      ii. $H(X \mid x_1 \leq 3 \text{ and } x_2 > 3) = \frac{2}{2} \log_2 \left(\frac{2}{2}\right)$
      iii. $H(X \mid x_1 > 3) = \frac{2}{5} \log_2 \left(\frac{5}{2}\right) + \frac{3}{5} \log_2 \left(\frac{5}{3}\right)$
      iv. $H(X|Y) = \frac{3}{10} H(X \mid x_1 \leq 3 \text{ and } x_2 \leq 3) + \frac{2}{10} H(X \mid x_1 \leq 3 \text{ and } x_2 > 3) + \frac{5}{10} H(X \mid x_1 > 3)$
   (c) $IG(X|Y) = H(X) - H(X|Y)$

Example: Decision Tree

1. **Given:** $X = \{0, 1, 2\}$
2. **Problem:** Draw a decision tree using binary conditions of the form, $x_i \leq k$, where $i \in \{1, 2\}$ and $k \in \mathbb{Z}$, that maximizes the information gained at each level.
3. **Solution (Level 1):**
   (a) Entropy before the split: $H(X) = \frac{3}{10}\log_2\left(\frac{10}{3}\right) + \frac{2}{10}\log_2\left(\frac{10}{2}\right) + \frac{5}{10}\log_2\left(\frac{10}{5}\right) = 1.485[\text{bits}]$
   (b) Entropy after the split and information gain (everything in base 2 since 2-ary).

Split	**Entropy**
$x_1 \leq 1$	$H(X\|Y) = \frac{3}{10}\left[\frac{2}{3}\log\left(\frac{3}{2}\right) + \frac{1}{3}\log\left(\frac{3}{1}\right)\right] + \frac{7}{10}\left[\frac{1}{7}\log\left(\frac{7}{1}\right) + \frac{2}{7}\log\left(\frac{7}{2}\right) + \frac{4}{7}\log\left(\frac{7}{4}\right)\right] = 1.241[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.241 = 0.244[\text{bits}]$
$x_1 \leq 2, 3$	$H(X\|Y) = \frac{5}{10}\left[\frac{3}{5}\log\left(\frac{5}{3}\right) + \frac{2}{5}\log\left(\frac{5}{2}\right)\right] + \frac{5}{10}\left[\frac{2}{5}\log\left(\frac{5}{2}\right) + \frac{3}{5}\log\left(\frac{5}{3}\right)\right] = 0.971[\text{bits}]$ • $\boxed{IG(X\|Y) = 1.485 - 0.971 = 0.514[\text{bits}]}$
$x_1 \leq 4, 5$	$H(X\|Y) = \frac{8}{10}\left[\frac{3}{8}\log\left(\frac{8}{3}\right) + \frac{2}{8}\log\left(\frac{8}{2}\right) + \frac{3}{8}\log\left(\frac{8}{3}\right)\right] + \frac{2}{10}\left[\frac{2}{2}\log\left(\frac{2}{2}\right)\right] = 1.249[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.249 = 0.236[\text{bits}]$
$x_1 \leq 6$	$H(X\|Y) = \frac{10}{10}\left[\frac{3}{10}\log\left(\frac{10}{3}\right) + \frac{2}{10}\log\left(\frac{10}{2}\right) + \frac{5}{10}\log\left(\frac{10}{5}\right)\right] = 1.485[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.485 = 0[\text{bits}]$
$x_2 \leq 1$	$H(X\|Y) = \frac{4}{10}\left[\frac{2}{4}\log\left(\frac{4}{2}\right) + \frac{1}{4}\log\left(\frac{4}{1}\right) + \frac{1}{4}\log\left(\frac{4}{1}\right)\right] + \frac{6}{10}\left[2\cdot\frac{1}{6}\log\left(\frac{6}{1}\right) + \frac{4}{6}\log\left(\frac{6}{4}\right)\right] = 1.351[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.351 = 0.134[\text{bits}]$
$x_2 \leq 2, 3$	$H(X\|Y) = \frac{7}{10}\left[\frac{3}{7}\log\left(\frac{7}{3}\right) + \frac{2}{7}\log\left(\frac{7}{2}\right) + \frac{2}{7}\log\left(\frac{7}{2}\right)\right] + \frac{3}{10}\left[\frac{3}{3}\log\left(\frac{3}{3}\right)\right] = 1.090[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.090 = 0.395[\text{bits}]$
$x_2 \leq 4$	$H(X\|Y) = \frac{9}{10}\left[\frac{3}{9}\log\left(\frac{9}{3}\right) + \frac{2}{9}\log\left(\frac{9}{2}\right) + \frac{4}{9}\log\left(\frac{9}{4}\right)\right] + \frac{1}{10}\left[\frac{1}{1}\log\left(\frac{1}{1}\right)\right] = 1.377[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.377 = 0.108[\text{bits}]$
$x_2 \leq 5$	$H(X\|Y) = \frac{10}{10}\left[\frac{3}{10}\log\left(\frac{10}{3}\right) + \frac{2}{10}\log\left(\frac{10}{2}\right) + \frac{5}{10}\log\left(\frac{10}{5}\right)\right] = 1.485[\text{bits}]$ • $IG(X\|Y) = 1.485 - 1.485 = 0[\text{bits}]$

Example: Decision Tree Continued:

4. **Solution (Level 2):** $x_1 \leq 2, 3$ has the highest information gain. For clarity, choose $x_1 \leq 3$ as the question.
   (a) Entropy before the split (treat as 2 indep. problems)
      i. $H(X_L) = \frac{3}{5}\log\left(\frac{5}{3}\right) + \frac{2}{5}\log\left(\frac{5}{2}\right) = 0.971$
      ii. $H(X_R) = \frac{2}{5}\log\left(\frac{5}{2}\right) + \frac{3}{5}\log\left(\frac{5}{3}\right) = 0.971$
   (b) Entropy after the split and information gain (everything in base 2 since 2-ary).

Split	Entropy
Left Split	
$x_1 \leq 1$	$H(X_L\|Y) = \frac{3}{5}\left[\frac{2}{3}\log\left(\frac{3}{2}\right) + \frac{1}{3}\log\left(\frac{3}{1}\right)\right] + \frac{2}{5}\left[\frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right] = 0.151[\text{bits}]$
	• $IG(X\|Y) = 0.971 - 0.151 = 0.820[\text{bits}]$
$x_2 \leq 1$	$H(X_L\|Y) = \frac{2}{5}\left[\frac{2}{2}\log\left(\frac{2}{2}\right)\right] + \frac{3}{5}\left[\frac{1}{3}\log\left(\frac{3}{1}\right) + \frac{2}{3}\log\left(\frac{3}{2}\right)\right] = 0.551[\text{bits}]$
	• $IG(X\|Y) = 0.971 - 0.551 = 0.420[\text{bits}]$
$x_2 \leq 2, 3$	$H(X_L\|Y) = \frac{3}{5}\left[\frac{3}{3}\log\left(\frac{3}{3}\right)\right] + \frac{2}{5}\left[\frac{2}{2}\log\left(\frac{2}{2}\right)\right] = 0[\text{bits}]$
	• $\boxed{IG(X_L\|Y) = 0.971 - 0 = 0.971[\text{bits}]}$
Right Split	
$x_1 \leq 4, 5$	$H(X_R\|Y) = \frac{3}{5}\left[\frac{2}{3}\log\left(\frac{3}{2}\right) + \frac{1}{3}\log\left(\frac{3}{1}\right)\right] + \frac{2}{5}\left[\frac{2}{2}\log\left(\frac{2}{2}\right)\right] = 0.551[\text{bits}]$
	• $\boxed{IG(X_L\|Y) = 0.971 - 0.551 = 0.420[\text{bits}]}$
$x_2 \leq 1$	$H(X_R\|Y) = \frac{2}{5}\left[\frac{1}{2}\log\left(\frac{2}{1}\right) + \frac{1}{2}\log\left(\frac{2}{1}\right)\right] + \frac{3}{5}\left[\frac{2}{3}\log\left(\frac{3}{2}\right) + \frac{1}{3}\log\left(\frac{3}{1}\right)\right] = 0.951[\text{bits}]$
	• $IG(X_L\|Y) = 0.971 - 0.951 = 0.020[\text{bits}]$
$x_2 \leq 2, 3$	$H(X_R\|Y) = \frac{4}{5}\left[\frac{2}{4}\log\left(\frac{4}{2}\right) + \frac{2}{4}\log\left(\frac{4}{2}\right)\right] + \frac{1}{5}\left[\frac{1}{1}\log\left(\frac{1}{1}\right)\right] = 0.8[\text{bits}]$
	• $IG(X_L\|Y) = 0.971 - 0.8 = 0.171[\text{bits}]$

Example: Decision Tree Continued:

5. **Solution (Level 3):** $x_2 \leq 2, 3$ and $x_1 \leq 4, 5$ has the highest information gain. For clarity, choose $x_2 \leq 3$ as the question for the left split and choose $x_1 \leq 5$ as the question for the right split.
   (a) Since 3 are pure splits already, therefore, look at right-left side only.
   (b) Entropy before the split for the right-left side
      i. $H(X_{RL}) = \frac{2}{3} \log\left(\frac{3}{2}\right) + \frac{1}{3} \log\left(\frac{3}{1}\right) = 0.918[\text{bits}]$
   (c) Entropy after the split and information gain (everything in base 2 since 2-ary).

Split	**Entropy**
$x_2 \leq 1$	$H(X_{RL}\|Y) = \frac{1}{3}\left[\frac{1}{1}\log\left(\frac{1}{1}\right)\right] + \frac{2}{3}\left[\frac{1}{2}\log\left(\frac{2}{1}\right) + \frac{1}{2}\log\left(\frac{2}{1}\right)\right] = 0.667[\text{bits}]$ • $IG(X\|Y) = 0.971 - 0.667 = 0.304[\text{bits}]$
$x_2 \leq 2, 3$	$H(X_{RL}\|Y) = \frac{1}{3}\left[\frac{1}{1}\log\left(\frac{1}{1}\right)\right] + \frac{2}{3}\left[\frac{2}{2}\log\left(\frac{2}{2}\right)\right] = 0[\text{bits}]$ • $\boxed{IG(X\|Y) = 0.971 - 0 = 0.971[\text{bits}]}$

6. Now all regions in our graph contain a pure set (one class). Note this took more questions than needed, but IG is a heuristic so its not perfect.