ASSIGNMENT-5

K. Bhikshapathi
24489

1. Information Theory

(a) Properties of Entropy

We consider five statements about the entropy $H(P)$ of a distribution P:

1. **"is always defined in bits" — False.** The units depend on the base of the logarithm. For example, using base 2 yields bits, while natural logarithms give nats.
2. **"$H(P) > 0$" — False.** Although entropy is nonnegative, it can be zero (e.g., in a degenerate distribution).
3. **"is the expected value of the information content of a distribution" — True.** By definition,
$$H(P) = \mathbb{E}\big[-\log P(x)\big].$$
4. **"$H(P) \leq 1$" — False.** There is no universal upper bound of 1; for example, a uniform distribution over many outcomes can have entropy greater than 1 bit.
5. **"is defined for continuous distributions too" — True (with a caveat).** In the continuous case, differential entropy is defined, though it does not share all properties of the discrete case.

Thus, the correct statements are (iii) and (v).

(b) Domain of KL Divergence

The Kullback–Leibler divergence is defined as

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

For $D(P\|Q)$ to be well defined, every event x with $P(x) > 0$ must also have $Q(x) > 0$.

(c) Computing $D(P\|Q)$ and Comparing Cross-Entropy with Entropy

Given the distributions over X:

X	x_1	x_2
P	0.3	0.7
Q	0.5	0.5

The KL divergence is computed as

$$D(P\|Q) = 0.3 \log_2 \frac{0.3}{0.5} + 0.7 \log_2 \frac{0.7}{0.5}.$$

Calculations:

$$\log_2(0.3/0.5) = \log_2(0.6) \approx -0.737, \quad 0.3 \times (-0.737) \approx -0.221,$$
$$\log_2(0.7/0.5) = \log_2(1.4) \approx 0.485, \quad 0.7 \times 0.485 \approx 0.340.$$

Thus,

$$D(P\|Q) \approx -0.221 + 0.340 \approx 0.119 \text{ bits}.$$

The entropy of P is:

$$H(P) = -\Big(0.3 \log_2(0.3) + 0.7 \log_2(0.7)\Big).$$

With

$$0.3 \log_2(0.3) \approx 0.3 \times (-1.737) \approx -0.521,$$
$$0.7 \log_2(0.7) \approx 0.7 \times (-0.515) \approx -0.361,$$

we get

$$H(P) \approx 0.521 + 0.361 \approx 0.882 \text{ bits.}$$

The cross-entropy is given by:

$$H(P,Q) = H(P) + D(P\|Q) \approx 0.882 + 0.119 \approx 1.001 \text{ bits,}$$

verifying that $H(P,Q) > H(P)$.

2. Decision Trees

(a) Statements about Decision Trees

1. **"There are only binary trees" — False.** Decision trees can have multiway splits.
2. **"We cannot mix both categorical and numerical data while building trees" — False.** Many implementations can handle a mix of data types.
3. **"Trees can be imbalanced" — True.** The structure of a decision tree depends on the data, and imbalances often occur.
4. **"Number of leaves equals the number of rows in the training dataset" — False.** Generally, the number of leaves is determined by the splits and is much less than the number of training examples.

Thus, only statement (iii) is true.

(b) Maximum Number of Leaf Nodes

For a dataset with N datapoints and k binary attributes (where $N \gg k$), a fully grown decision tree can have at most

$$2^k \text{ leaves.}$$

(c) Average Number of Questions Asked

For classification, each non-leaf node represents a question (or test). Let the tree have leaves $i = 1, 2, \ldots, m$ with counts n_i (number of training examples) and depths d_i (number of questions). The average number of questions is given by:

$$\text{Average questions} = \frac{\sum_{i=1}^{m} d_i\, n_i}{\sum_{i=1}^{m} n_i}.$$

Method:

1. Identify all leaf nodes and record their counts n_i.
2. Determine the depth d_i (number of questions) for each leaf.
3. Compute the weighted average.

For example, if a tree has leaves at depths 2, 3, and 3 with counts 4, 3, and 5 respectively, then:

$$\text{Average questions} = \frac{4 \times 2 + 3 \times 3 + 5 \times 3}{4 + 3 + 5} = \frac{8 + 9 + 15}{12} \approx 2.67.$$