TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Shannon's Source Coding Theorem

Huffman Coding

Perils of Continuous Information Theory

Source Coding Theorem

Consider a probability distribution Pop on a finite set S.

Consider a code C assigning a bit string code word $C(y_1, \ldots, y_B)$ to each possible batch of B elements with $y_i \sim \mathrm{Pop}$.

Source coding theorem: As $B \to \infty$ the optimal coding uses exactly $H(\mathrm{Pop})$ bits per batch element.

Prefix Free Codes

Let S be a finite set.

Let C be assignment of a bit string $C(y)$ to each $y \in S$.

C is called *prefix-free* if for $x \neq y$ we have that $C(x)$ is not a prefix of $C(y)$.

A concatenation of sequence of prefix-free code words can be uniquely segmented (parsed) back into a sequence of code words.

Prefix-Free Codes as Trees and as Probabilities

A prefix-free code defines a binary branching tree — branch on the first code bit, then the second, and so on.

The leaves of this tree are labeled with the elements of S.

The code defines a probability distribution on S by randomly selecting branches.

We have $P_C(y) = 2^{-|C(y)|}$.

The Source Coding Theorem

(1) There exists a prefix-free code C such that

$$|C(y)| <= (-\log_2 \mathrm{Pop}(y)) + 1$$

and hence

$$E_{y\sim\mathrm{Pop}}|C(y)| \leq H(\mathrm{Pop}) + 1$$

(2) For any prefix-free code C

$$E_{y\sim\mathrm{Pop}}\ |C(y)| = E_{y\sim\mathrm{Pop}} - \ln\ P_C(y) = H(\mathrm{Pop}, P_C) \geq H(\mathrm{Pop})$$

Code Construction

We construct a code by iterating over $y \in S$ in order of decreasing probability (most likely first).

For each y select a code word $C(y)$ (a tree leaf) with length (depth)

$$|C(y)| = \lceil -\log_2 \mathrm{Pop}(y) \rceil$$

and where $C(y)$ is not an extension of (under) any previously selected code word.

Code Existence Proof

At any point before coding all elements of S we have

$$\sum_{y \in \text{Defined}} 2^{-|C(y)|} \leq \sum_{y \in \text{Defined}} \text{Pop}(y) < 1$$

Therefore there exists an infinite descent into the tree that misses all previous code words.

Hence there exists a code word $C(x)$ not under any previous code word with $|C(x)| = \lceil -\log_2 \text{Pop}(y) \rceil$.

Furthermore $C(x)$ is at least as long as all previous code words and hence $C(x)$ is not a prefix of any previously selected code word.

Huffman Coding

Maintain a list of trees $T_1, \ldots, T_N$.

Inititally each tree is just one root node labeled with an element of S.

Each tree T_i has a weight equal to the sum of the probabilities of the nodes on the leaves of that tree.

Repeatedly merge the two trees of lowest weight into a single tree until all trees are merged.

Optimality of Huffman Coding

Theorem: The Huffman code T for Pop is optimal — for any other tree T' we have $d(T; \mathrm{Pop}) \leq d(T'; \mathrm{Pop})$.

Proof: The algorithm maintains the invariant that there exists an optimal tree including all the subtrees on the list.

To prove that a merge operation maintains this invariant we consider any tree containing the given subtrees.

Consider the two subtrees T_i and T_j of minimal weight. Without loss of generality we can assume that T_i is at least as deep as T_j.

Swapping the sibling of T_i for T_j brings T_i and T_j together and can only improve the average depth.

Optimality of Huffman Coding

Why the swap operation cannot increase entropy.
Let $d(n)$ be the depth of node n, let $P(n)$ be $2^{-}d(n)$, and let $W(n)$ be the weight of n.

For $P(n_1) \geq P(n_2)$ and $W_1 \geq W_2$ then

$$P(n_1)W_1 + P(n_2)W_2 \geq P(n_1)W_2 + P(n_2)W_1$$

Perils of Differential Entropy

Consider a continuous density $p(x)$. For example

$$p(x) = \frac{1}{\sqrt{2\pi}\ \sigma}\ e^{\frac{-x^2}{2\sigma^2}}$$

Differential entropy is often defined as

$$H(p) \doteq \int \left(\ln \frac{1}{p(x)} \right) p(x) dx$$

Finite Differential Entropy is Not Meaningful

$$
\begin{aligned}
H(\mathcal{N}(0,\sigma)) &= +\int \left(\ln(\sqrt{2\pi}\sigma) + \frac{x^2}{2\sigma^2} \right) p(x)dx \\
&= \ln(\sigma) + \ln(\sqrt{2\pi}) + \frac{1}{2}
\end{aligned}
$$

But if we take $y \doteq x/2$ we get $H(y) = H(x) - \ln 2$.

Also for $\sigma << 1$, we get $H(p) < 0$

Hence differential entropy then depends on the choice of units — a distributions on lengths will have a different entropy when measuring in inches than when measuring in feet.

Differential Entropy is Always Infinite

Consider quantizing the the real numbers into bins.

A continuous probability densisty p assigns a probability $p(B)$ to each bin.

As the bin size decreases toward zero the entropy of the bin distribution increases toward ∞.

A meaningful convention is that $H(p) = +\infty$ for any continuous density p.

Differential KL-divergence is Meaningful

$$KL(p,q) = \int \left(\ln \frac{p(x)}{q(x)} \right) p(x)dx$$

This integral can be computed by dividing the real numbers into bins and computing the KL divergence between the distributions on bins.

The KL divergence between the bin distribution typically approaches a finite limit as the bin size goes to zero.

KL-Divergence can also be Infinite

$$KL(p, q) = E_{x \sim p} \ \log \frac{p(x)}{q(x)}$$

In either the discrete or continuous case, if a set is assigned nonzero probability by p but zero probability by q then $KL(p, q) = +\infty$.

If every set assigned nonzero probability by p is also assigned nonzero probability by q then we say that p is absolutely continuous with respect to q.

Random Variables

We consider variables where a single draw form the population determines a value for each variable.

This is the formal definition of a "random variable".

Each random variable has a probability distribution defined by the distribution on the population.

We write $H(x)$ for the entropy of the distribution on x.

Mutual Information

For two random variables x and y there is a distribution on pairs (x, y) determined by the population distribution.

Mutual information concerns the relationship between the distribution on (x, y) and the marginal distributions on x and y.

For the discrete case we can write.

$$I(x, y) \doteq H(x) + H(y) - H(x, y)$$

This can be viewed as a quantity of non-independence — independent variables have zero mutual information.

Conditional Entropy

For the discrete case conditional entropy $H(y|x)$ is defined by

$$\begin{aligned} H(y|x) &\doteq \sum_x \mathrm{Pop}(x) \sum_y \mathrm{Pop}(y|x) - \log \mathrm{Pop}(y|x) \\ &= E_{x \sim \mathrm{Pop}}\; E_{y \sim \mathrm{Pop}|x}\; - \log \mathrm{Pop}(y|x) \\ &= E_{x \sim \mathrm{Pop}}\; H(\mathrm{Pop}(y|x)) \end{aligned}$$

More Identities

For the discrete case we have.

$$
\begin{aligned}
I(x,y) &= H(x) - H(x|y) \\
&= H(y) - H(y|x) \\
&= KL(\mathrm{Pop}(x,y), \mathrm{Pop}(x) \times \mathrm{Pop}(y))
\end{aligned}
$$

The last identity can be taken as a definition of $I(x,y)$ in the continuous case.

END