Lecture 2: Information Theory and Lower bound on learning unstructured distribution

Daniel Kane
Scribe: Sankeerth Rao

April 7, 2017

Abstract

We setup the framework of Information theory and use it to prove lower bounds on the number of samples needed to learn an arbitrary distribution p on $[n]$

1 Information Theory

Consider a random variable p. How much information does p encode ?

Motivation: If told that that $p = x$ what is the "surprisingness" of this event. We expect it to be dependent on $\frac{1}{\Pr(p=x)}$, since the more likely it is the less surprising it is. Also if we are told about two independent events we expect the "surprisingness" to be additive rather than multiplicative thus we take the logarithm to get $\log \frac{1}{\Pr(p=x)}$. The mathematical notion that captures the average "surprisingness" is the Shannon Entropy.

Shannon Entropy:

$$H(p) = \sum_x \Pr(p = x) \log \frac{1}{\Pr(p = x)}$$

Lets look at few examples:

- p is uniform on $[n]$

$$H(p) = \sum_{i=1}^{n} \frac{1}{n} \log n = \log n$$

- p weighted coin. Probability (q, q')

$$H(p) = q \log \frac{1}{q} + q' \log \frac{1}{q'}$$

Note. *p is supported on $[n]$. Then*

$$0 \leq H(p) \leq \log n.$$

with equality iff p is uniform.

Intuition: The context in which entropy was defined is - Given n iid instances $X_1, X_2, \dots, X_n$ of X the number of bits that need to be communicated to convey these is equal to $nH(X)$. However we do not need this interpretation in the course.

Complicated random variables can be constructed from simple random variables. For instance we can look at a pair of random variables as a new random variable. Then the entropy would be

$$H(p,q) = \sum_{x,y} \Pr(p = x, q = y) \log \frac{1}{\Pr(p = x, q = y)}.$$

If p, q are independent,

$$H(p,q) = \sum_{x,y} p_x q_y \log \frac{1}{p_x q_y} = \sum_{x,y} p_x q_y \log \frac{1}{p_x} + \sum_{x,y} p_x q_y \log \frac{1}{q_y} = H(p) + H(q)$$

Relative Entropy $H(p|q)$ captures how much entropy does p have if one already knows q.

$$\begin{aligned} H(p|q) &= \mathbb{E}_{q=y}[H(p|q = y)] \\ &= \sum_y q_y \sum_x \frac{\Pr(p = x, q = y)}{q_y} \log \frac{q_y}{\Pr(p = x, q = y)} \\ &= \sum_{x,y} \Pr(p = x, q = y) \log \frac{1}{\Pr(p = x, q = y)} - \sum_{x,y} \Pr(p = x, q = y) \log \frac{1}{q_y} \\ &= H(p,q) - H(q) \geq 0 \end{aligned}$$

Its the total information learned from p and q minus the information learned from q.

Entropy satisfies the Chain rule. Given one bit at a time how much information do we learn.

$$H(p_1, p_2, \dots, p_n) = H(p_1) + H(p_2|p_1) + H(p_3|p_1, p_2) + \dots$$

These notions would be useful because we would like to determine how much information do the samples say about the underlying distribution. We define mutual information $I(p;q)$ How much information does q tell us about p.

$$I(p;q) = H(p) - H(p|q) = H(p) + H(q) - H(p,q)$$

Lets look at few examples:

- Random element of $\mathbb{F}_2^n$.
  p gives the coordinates in positions in S,
  q gives the coordinates in positions in T,

  $$\begin{aligned} H(p) = |S|, H(q) = |T|, H(p,q) = |S \cup T| \\ I(p;q) = |S| + |T| - |S \cup T| = |S \cap T| \end{aligned}$$

- p, q are individually fair but correlated coins, $p = q$ with probability x.

  $$\begin{aligned} H(q) = \log 2 = 1, H(q|p) = x \log(\frac{1}{x}) + (1-x) \log(\frac{1}{1-x}). \\ I(p;q) = 1 - x \log(\frac{1}{x}) + (1-x) \log(\frac{1}{1-x}) \end{aligned}$$

  Note that $I(p;q) \geq 0$ with equality iff $x = 1/2$. If $x = 0$ or $x = 1$ then $I(p;q) = 1$. If $x = \frac{1}{2} + \epsilon$ then Taylor expanding we have $I(p;q) = \Theta(\epsilon^2)$. This makes sense because think if we are given all the tosses of p and now we are given tosses of q one by one then this equivalent to learning the bias of a coin that is being tossed(heads corresponds to $p = q$) then we need $\frac{1}{\epsilon^2}$ samples to learn it.

Mutual information is always non negative.

Lemma 1.1.

$$I(p;q) \geq 0.$$

Proof.

$$\begin{aligned} I(p;q) &= H(p) - H(p|q) \\ &= H(\mathbb{E}[p|q]) - \mathbb{E}[H(p|q)] \geq 0(\text{ H is a concave function}) \end{aligned}$$

□

Relative shared information: This is the expected shared information between p and q when you know R and thus positive.

$$I(p;q|R) = H(p|R) + H(q|R) - H(p,q|R)$$

Now we could get a chain rule for mutual information. The information in p about q and R is equal to the information about p in R and the extra information in q about p knowing R.

$$I(p;q,R) = I(p;R) + I(p;q|R)$$

Information Processing Inequality:

Lemma 1.2. $I(p;q) \geq I(f(p);q)$

Proof.

$$\begin{aligned} I(p;q) &= I(p, f(p); q) \\ &= I(f(p);q) + \overbrace{I(p;q|f(p))}^{\geq 0} \\ &\geq I(f(p);q) \end{aligned}$$

□

2 Lower bounds:

Now we use this machinery to prove lower bounds. We use the adversarial method where in we see it as a game between Nature and the Algorithm. Nature comes up with a distribution $\mathcal{D}$ over all possible p, the algorithm gets N samples from a random $p \in \mathcal{D}$ and tries to come up with a distribution that is close to p.

So we want to come up with a distribution $\mathcal{D}$ on the set of possible distributions on $[n]$.

Essentially we want to pick a collection of distributions p on $[n]$. We don't want all the elements of our collection to be close together because then the algorithm could just output one p that is close to all the elements in the collection no matter what the samples are. Also we don't want all the elements in the collection to be far apart because then the algorithm could easily differentiate the distributions.

Also in the upper bound we bounded $|p-q|_1$ by $|p-q|_2$ using Cauchy Schwarz. We sort of want $|p-q|_1$ to be as large as possible so for equality in Cauchy Schwarz we need $|p_i - q_i|$ to be equal for all i. Thus we choose the collection to be around the uniform distribution but at ϵ distance away so that we are not too clustered or too far away.

$\mathcal{D}$ Our collection is made up of distribution over $2n$ bins. The probabilities of ith bin in every distribution in our collection is $p_i = \frac{1}{2n} \pm \frac{5\epsilon}{2n}$, but we need all the p_i's to sum to 1 so we pick (p_{2i-1}, p_{2i}) is randomly $(\frac{1+5\epsilon}{2n}, \frac{1-5\epsilon}{2n})$ or $(\frac{1-5\epsilon}{2n}, \frac{1+5\epsilon}{2n})$, this gives us a collection of distributions that are 5ϵ away from the uniform distribution. Now Distribution $\mathcal{D}$ picks one element of this collection uniformly at random. Note that this also specifies a distribution on all possible p where we never pick a distribution outside our collection.

Now suppose a $p \sim \mathcal{D}$ is picked and the algorithm is given N independent samples $X_1, \ldots, X_N$ from p. The algorithm processes these samples and comes up with $f(X_1, \ldots, X_N) = q$ such that $d_{TV}(p,q) < \epsilon$ with probability at least 2/3.

Now define q' to be the closest distribution in $Supp(\mathcal{D})$ to q.

Lemma 2.1.

$$d_{TV}(p, q') \leq 2\epsilon$$

Proof. Consider the indices $(2i-1, 2i)$. Now if $(q'_{2i-1}, q'_{2i}) = (p_{2i-1}, p_{2i})$, then we are done. So let $(q'_{2i-1}, q'_{2i}) = (p_{2i}, p_{2i-1})$. But note that (q'_{2i-1}, q'_{2i}) is the closest rounding of (q_{2i-1}, q_{2i}) among the two possibilities (p_{2i-1}, p_{2i}) and (p_{2i}, p_{2i-1}), so it should be the case that $|q_{2i-1} - q'_{2i-1}| + |q_{2i} - q'_{2i}| < |q_{2i-1} - p_{2i-1}| + |q_{2i} - p_{2i}|$. Thus $|q'_{2i-1} - p_{2i-1}| + |q'_{2i} - p_{2i}| \leq |q'_{2i-1} -$

$q_{2i-1}| + |q'_{2i} - q_{2i}| + |q_{2i-1} - p_{2i-1}| + |q_{2i} - p_{2i}| < 2(|q_{2i-1} - p_{2i-1}| + |q_{2i} - p_{2i}|)$. This when summed over all $i \in [n]$ proves the lemma. □

But now that both p, q' both lie in the $Supp(\mathcal{D})$ we have

$$d_{TV}(p, q') = \frac{10\epsilon}{n}[\text{ no of pairs of bins q' is wrong on }].$$

but since $d_{TV}(p, q') < 2\epsilon$, q' will agree with p on all but at most $n/5$ pairs of bins. Note that q' is a deterministic function (say g) of the samples $X_1, \ldots, X_N$ (denote this by X_1^N). Now we use the IT inequalities to lower bound the amount of information that the samples contain about $\mathcal{D}$:

$$I(\mathcal{D}; X_1^N) \geq I(\mathcal{D}; g(X_1^N)) = H(\mathcal{D}) - H(\mathcal{D}|g(X_1^N)).$$

Let each pair bins correspond to a bit.$H(\mathcal{D}) = n$. We know that w.p atleast $2/3, \mathcal{D}$ agrees with g on $\frac{4n}{5}$ of the n bits. So

$$H(\mathcal{D}|g(X_1^N)) \leq \frac{2}{3}\log(\binom{n}{\leq n/5}) + \frac{1}{3}n \leq \frac{2}{3}\log(\frac{en}{n/5})^{n/5} + \frac{1}{3}n < n(2\frac{\log 5e}{15} + \frac{1}{3}) < 0.9n$$

Thus we have

$$I(\mathcal{D}; X_1^N) \geq \Omega(n).$$

In the next lecture we will show that each instance of this sample can contribute at most $O(\epsilon^2)$ information about $\mathcal{D}$ and that would prove that we need at least $\Omega(\frac{n}{\epsilon^2})$ samples to learn the arbitrary distribution p.